Stochastic gradient descent for linear inverse problems in variable exponent Lebesgue spaces

Marta Lazzaretti[1,3], Zeljko Kereta[2], Luca Calatroni[3], and Claudio Estatico[1]

[1] Dip. di Matematica, Università di Genova, Via Dodecaneso 35, 16146, Italy
lazzaretti@dima.unige.it, estatico@dima.unige.it
[2] Dept. of Computer Science, University College London, UK
z.kereta@ucl.ac.uk
[3] CNRS, UCA, Inria, Laboratoire I3S, Sophia-Antipolis, 06903, France
calatroni@i3s.unice.fr

Abstract. We consider a stochastic gradient descent (SGD) algorithm for solving linear inverse problems (e.g., CT image reconstruction) in the Banach space framework of variable exponent Lebesgue spaces $\ell^{(p_n)}(\mathbb{R})$. Such non-standard spaces have been recently proved to be the appropriate functional framework to enforce pixel-adaptive regularisation in signal and image processing applications. Compared to its use in Hilbert settings, however, the application of SGD in the Banach setting of $\ell^{(p_n)}(\mathbb{R})$ is not straightforward, due, in particular to the lack of a closed-form expression and the non-separability property of the underlying norm. In this manuscript, we show that SGD iterations can effectively be performed using the associated modular function. Numerical validation on both simulated and real CT data show significant improvements in comparison to SGD solutions both in Hilbert and other Banach settings, in particular when non-Gaussian or mixed noise is observed in the data.

Keywords: Iterative regularisation · Stochastic gradient descent · Inverse problems in Banach spaces · Computed Tomography.

1 Introduction

The literature on iterative regularisation methods for solving ill-posed linear inverse problems in finite/infinite-dimensional Hilbert or Banach settings is very vast, see, e.g., [7,21] for surveys. Given two normed vector spaces $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ and $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$, we are interested in the inverse problem

$$\text{find} \quad x \in \mathcal{X} \quad \text{s.t.} \quad \mathcal{Y} \ni y = Ax + \eta, \tag{1}$$

where $A \in \mathcal{L}(\mathcal{X}; \mathcal{Y})$ is a bounded linear operator, and $\eta \in \mathcal{Y}$ denotes the (additive) noise perturbation of magnitude $\|\eta\|_{\mathcal{Y}} \leqslant \delta$, $\delta > 0$, corrupting the measurements. Due to the ill-posedness, the standard strategy for solving (1) consists in computing $x^\star \in \operatorname{argmin}_{x \in \mathcal{X}} \ \Psi(x)$, where the functional $\Psi : \mathcal{X} \to \mathbb{R}_+ \cup \{+\infty\}$ quantifies

the fidelity of a candidate reconstruction to the measurements, possibly combined with a penalty or regularisation term enforcing prior assumptions on the sought quantity $x \in \mathcal{X}$. A popular strategy for promoting implicit regularisation through algorithmic optimisation consists in designing iterative schemes solving instances of the minimisation problem $\operatorname{argmin}_{x\in\mathcal{X}} \ \|Ax-y\|_{\mathcal{Y}}$ or, more generally

$$\operatorname*{argmin}_{x\in\mathcal{X}} \ f(x) \qquad \text{with} \quad f(x) = \tilde{f}(Ax-y), \tag{P}$$

where, for $y \in \mathcal{Y}$, the function $f(\cdot) = \tilde{f}(A\cdot -y) : \mathcal{X} \to \mathbb{R}_{\geqslant 0}$ measures the discrepancy between the model observation Ax and y. The iterative scheme has to be endowed with a robust criterion for its early stopping in order to avoid that the computed reconstruction overfits the noise [16]. In this context, the role of the parameter tuning the amount of regularisation is thus played by nothing but the number of performed iterations. One-step gradient descent algorithms, such as the (accelerated) Landweber or the Conjugate Gradient, represent the main class of optimisation methods for the resolution of (P), see e.g. [6, 18, 19].

The most well-studied cases consider $\mathcal{X}$ and $\mathcal{Y}$ to be Hilbert spaces, e.g., $\mathcal{X} = \mathcal{Y} = \ell^2(\mathbb{R})$. In this setting, problem (P) takes the form $\operatorname{argmin}_{x\in\ell^2(\mathbb{R})} \ \frac{1}{2}\|Ax - y\|^2_{\ell^2(\mathbb{R})}$ and can be solved by a standard Landweber iterative scheme

$$x^{k+1} = x^k - \mu_{k+1}A^*(Ax^k - y), \tag{2}$$

for $k \geqslant 0$, where $\mu_{k+1} > 0$ denotes the algorithmic step-sizes. However, many inverse problems require a more complex setting to retrieve solutions with specific features, such as sharp edges, piecewise constancy, sparsity patterns and/or to model non-standard (e.g., mixed) noise in the data. Either $\mathcal{X}$ or $\mathcal{Y}$, or both, can thus be modelled as more general Banach spaces. Notable examples are standard Lebesgue spaces $L^p(\Omega)$ and, in discrete settings, sequence spaces $\ell^p(\mathbb{R})$ with $p \in [1, +\infty]\setminus\{2\}$. While the solution space $\mathcal{X}$ affects the choice of the specific iterative scheme to be used, the measurement (or data) space $\mathcal{Y}$ is naturally connected to the norm appearing in (P). For example, for Hilbert $\mathcal{X} = \ell^2(\mathbb{R})$ and Banach $\mathcal{Y} = \ell^p(\mathbb{R})$, an instance of (P) reads as

$$\operatorname*{argmin}_{x\in\ell^2(\mathbb{R})} \ \frac{1}{q}\|Ax - y\|^q_{\ell^p}, \quad \text{with } \ q > 1,$$

for which a gradient descent-type scheme can still be used in the form $x^{k+1} = x^k - A^*\mathbf{J}^q_{\ell^p}(Ax^k - y)$, where $\mathbf{J}^q_{\ell^p} : \ell^p(\mathbb{R}) \to \ell^{p^*}(\mathbb{R})$ is the so-called q-duality map of $\ell^p(\mathbb{R})$, defined as $\mathbf{J}^q_{\ell^p}(\cdot) = \partial\left(\frac{1}{q}\|\cdot\|^q_{\ell^p(\mathbb{R})}\right)$. When both $\mathcal{X}$ and $\mathcal{Y}$ are Banach spaces, a popular algorithm for solving

$$\operatorname*{argmin}_{x\in\mathcal{X}} \ \frac{1}{q}\|Ax - y\|^q_{\mathcal{Y}}, \quad \text{with } \ q > 1$$

is the dual Landweber method [22]

$$x^{k+1} = \mathbf{J}^{p^*}_{\mathcal{X}^*}\left(\mathbf{J}^p_{\mathcal{X}}(x^k) - \mu_{k+1}A^*\mathbf{J}^q_{\mathcal{Y}}(Ax^k - y)\right) \tag{3}$$

where $\mathbf{J}^p_{\mathcal{X}} : \mathcal{X} \to \mathcal{X}^*$, is the p-duality map of $\mathcal{X}$, $\mathbf{J}^{p^*}_{\mathcal{X}^*} : \mathcal{X}^* \to \mathcal{X}$ is its inverse with p^* denoting the conjugate exponent of p, i.e. $1/p + 1/p^* = 1$. For other references of gradient-descent-type solvers in Banach settings, see, e.g. [11, 21, 22].

A non-standard Banach framework for solving linear inverse problems is the one of variable exponent Lebesgue spaces $L^{p(\cdot)}(\Omega)$ and $\ell^{(p_n)}(\mathbb{R})$ [5]. These Banach spaces are defined in terms of a Lebesgue measurable function $p(\cdot) : \Omega \to [1, +\infty]$, or a real sequence $(p_n)_n$, respectively, that assigns coordinate-wise exponents to all points in the domain. Variable exponent Lebesgue spaces have proven useful in the design of adaptive regularisation, suited to model heterogeneous data and complex noise settings. Iterative regularisation procedures in this setting have been recently studied [2] and also extended to composite optimisation problems involving non-smooth penalty terms [14].

While benefiting from several convergence properties, the use of such (deterministic) iterative algorithms may be prohibitively expensive in large-size applications as they require the use of all data at each iteration. In this work, we follow the strategy performed by the seminal work of Robbins and Monro [20] and adapt a stochastic gradient descent (SGD) strategy to the non-standard setting of variable exponent Lebesgue space, in order to reduce the per-iteration complexity costs. Roughly speaking, this is done by defining a suitable decomposition of the original problem and implementing an iterative scheme where only a batch of data, typically one, is used to compute the current update. Note that the use of SGD schemes has recently attracted the attention of the mathematical imaging community [10, 13] due to its applicability in large-scale applications such as medical imaging [9, 17, 23]. However, its extension to variable exponent Lebesgue setting is not trivial due to some structural difficulties (e.g., non-separability of the norm), making the adaptation a challenging task.

Contribution. We consider an SGD-based iterative regularisation strategy for solving linear inverse problems in the non-standard Banach setting of variable exponent Lebesgue space $\ell^{(p_n)}(\mathbb{R})$. To overcome the non-separability of the norm in such space, we consider updates defined in terms of a separable function, the modular function. Numerical investigation of the methodology on CT image reconstruction are reported to show the advantages of considering such non-standard Banach setting in comparison to standard Hilbert scenarios. Comparisons between the modular-based deterministic and stochastic algorithms confirm improvements of the latter w.r.t. CPU times.

2 Optimisation in Banach spaces

In this section we revise the main definitions and tools useful for solving a general instance of (P) in the general context of Banach spaces $\mathcal{X}$ and $\mathcal{Y}$. For a real Banach space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$, we denote by $(\mathcal{X}^*, \|\cdot\|_{\mathcal{X}^*})$ its dual space and, for any $x \in \mathcal{X}$ and $x^* \in \mathcal{X}^*$, by $\langle x^*, x\rangle = x^*(x) \in \mathbb{R}$ its duality pairing.

The following definition is crucial for the development of algorithms solving (P) in Banach spaces. We recall that in Hilbert settings $\mathcal{H} \cong \mathcal{H}^*$ holds by the Riesz representation theorem, with $\cong$ denoting an isometric isomorphism.

Hence, for $x \in \mathcal{H}$, the element $\nabla f(x) \in \mathcal{H}^*$ can be implicitly identified with a unique element in $\mathcal{H}$ itself, up to the canonical isometric isomorphism, so that the design of gradient-type schemes is significantly simplified, as in (2). Since the same identification does not hold, in general, for a Banach space $\mathcal{X}$, we recall the notion of duality maps, which properly associate an element of $\mathcal{X}$ with an element (or a subset) of $\mathcal{X}^*$ [3].

Definition 1. *Let $\mathcal{X}$ be a Banach space and $p > 1$. The duality map $\mathbf{J}^p_{\mathcal{X}}$ with gauge function $t \mapsto t^{p-1}$ is the operator $\mathbf{J}^p_{\mathcal{X}} : \mathcal{X} \to 2^{\mathcal{X}^*}$ such that, for all $x \in \mathcal{X}$,*

$$\mathbf{J}^p_{\mathcal{X}}(x) = \left\{x^* \in \mathcal{X}^* : \langle x^*, x\rangle = \|x\|_{\mathcal{X}}\|x^*\|_{\mathcal{X}^*},\ \|x^*\|_{\mathcal{X}^*} = \|x\|^{p-1}_{\mathcal{X}}\right\}.$$

Under suitable smoothness assumptions on $\mathcal{X}$ [21], $\mathbf{J}^p_{\mathcal{X}}(x)$ is single valued at all $x \in \mathcal{X}$. For instance, for $\mathcal{X} = \ell^p(\mathbb{R})$, with $p > 1$, all duality maps are single-valued. The following Theorem (see [3]) provides an operative definition and a more intuitive interpretation of the duality maps.

Theorem 1 (Asplund's Theorem). *The duality map $\mathbf{J}^p_{\mathcal{X}}$ is the subdifferential of the convex functional $h : x \mapsto \frac{1}{p}\|x\|^p_{\mathcal{X}}$, that is, $\mathbf{J}^p_{\mathcal{X}}(\cdot) = \partial(\frac{1}{p}\|\cdot\|^p_{\mathcal{X}})$.*

The following result is needed for the invertibility of the duality map.

Proposition 1. *[21] Under suitable smoothness and convexity conditions on $\mathcal{X}$ and for $p > 1$, for all $x \in \mathcal{X}$ and all $x^* \in \mathcal{X}^*$, there holds*

$$\mathbf{J}^{p^*}_{\mathcal{X}^*}(\mathbf{J}^p_{\mathcal{X}}(x)) = x\,, \qquad \mathbf{J}^p_{\mathcal{X}}(\mathbf{J}^{p^*}_{\mathcal{X}^*}(x^*)) = x^*.$$

We notice that, if the gradient term $A^*\mathbf{J}^q_{\mathcal{Y}}(Ax^k - y)$ vanishes in iteration (3), then $x^{k+1} = \mathbf{J}^{p^*}_{\mathcal{X}^*}(\mathbf{J}^p_{\mathcal{X}}(x^k)) = x^k$ by Proposition 1.

For any $p, r > 1$ and for any $x, h \in \ell^p(\mathbb{R})$, the explicit formula for $\mathbf{J}^r_{\ell^p}$ is

$$\langle \mathbf{J}^r_{\ell^p}(x), h\rangle = \|x\|^{r-p}_p \sum_{n\in\mathbb{N}} \operatorname{sign}(x_n)|x_n|^{p-1}h_n. \tag{4}$$

Moreover, since $(\ell^p(\mathbb{R}))^* \cong \ell^{p^*}(\mathbb{R})$, then the inverse of the r-duality map $\mathbf{J}^r_{\ell^p}$ is nothing but $(\mathbf{J}^r_{\ell^p})^{-1} = \mathbf{J}^{r^*}_{(\ell^p)^*} = \mathbf{J}^{r^*}_{\ell^{p^*}}$. Hence, the explicit analytical expression of its inverse $(\mathbf{J}^r_{\ell^p})^{-1} = \mathbf{J}^{r^*}_{\ell^{p^*}}$ is also known [3].

2.1 Variable exponent Lebesgue spaces $\boldsymbol{\ell^{(p_n)}}(\mathbb{R})$

In the following, we will introduce the main concepts and definitions on the variable exponent Lebesgue spaces in the discrete setting of $\ell^{(p_n)}(\mathbb{R})$. For surveys, we refer the reader to [4, 5]. We define a family $\mathcal{P}$ of variable exponents as

$$\mathcal{P} := \left\{(p_n)_{n\in\mathbb{N}} \subset \mathbb{R} : 1 < p_- := \inf_{n\in\mathbb{N}} p_n \leqslant p_+ := \sup_{n\in\mathbb{N}} p_n < +\infty\right\}.$$

Definition 2. *For $(p_n)_{n\in\mathbb{N}} \in \mathcal{P}$ and any real sequence $x = (x_n)_{n\in\mathbb{N}}$,*

$$\rho_{(p_n)}(x) := \sum_{n\in\mathbb{N}} |x_n|^{p_n} \quad \text{and} \quad \bar{\rho}_{(p_n)}(x) := \sum_{n\in\mathbb{N}} \frac{1}{p_n}|x_n|^{p_n} \tag{5}$$

are called modular functions associated with the exponent map $(p_n)_{n\in\mathbb{N}}$.

Definition 3. *The Banach space $\ell^{(p_n)}(\mathbb{R})$ is the set of real sequences $x = (x_n)_{n\in\mathbb{N}}$ such that $\rho_{(p_n)}\left(\frac{x}{\lambda}\right) < 1$ for some $\lambda > 0$. For any $x = (x_n)_{n\in\mathbb{N}} \in \ell^{(p_n)}(\mathbb{R})$, the (Luxemburg) norm on $\ell^{(p_n)}(\mathbb{R})$ is defined as*

$$\|x\|_{\ell^{(p_n)}} := \inf\left\{\lambda > 0 :\ \rho_{(p_n)}\left(\frac{x}{\lambda}\right) \leqslant 1\right\}. \tag{6}$$

We now report a result from [2] where a characterisation of the duality map $\mathbf{J}^r_{\ell^{(p_n)}}$ is given, in relation with (4).

Theorem 2. *Given $(p_n)_{n\in\mathbb{N}} \in \mathcal{P}$, then for each $x = (x_n)_{n\in\mathbb{N}} \in \ell^{(p_n)}(\mathbb{R})$ and for any $r > 1$, the duality map $\mathbf{J}^r_{\ell^{(p_n)}}(x) : \ell^{(p_n)}(\mathbb{R}) \to (\ell^{(p_n)})^*(\mathbb{R})$ is the linear operator defined, for all $h = (h_n)_{n\in\mathbb{N}} \in \ell^{(p_n)}(\mathbb{R})$ by:*

$$\langle \mathbf{J}^r_{\ell^{(p_n)}}(x), h\rangle = \frac{1}{\sum_{n\in\mathbb{N}} \frac{p_n|x_n|^{p_n}}{\|x\|^{p_n}_{\ell^{(p_n)}}}} \sum_{n\in\mathbb{N}} \frac{p_n \operatorname{sign}(x_n)|x_n|^{p_n-1}}{\|x\|^{p_n-r}_{\ell^{(p_n)}}} h_n. \tag{7}$$

By (6), we note that $\|\cdot\|_{\ell^{(p_n)}}$ is not separable as its computation requires the solution of a minimisation problem involving all elements x_n and p_n at the same time. As a consequence, the expression (7) is not suited to be used in a computational optimisation framework. The following result from [14] provides more flexible expressions associated to the modular functions (5).

Proposition 2. *The functions $\rho_{(p_n)}$ and $\bar{\rho}_{(p_n)}$ in (5) are Gateaux differentiable at any $x = (x_n)_{n\in\mathbb{N}} \in \ell^{(p_n)}(\mathbb{R})$. For $h = (h_n)_{n\in\mathbb{N}} \in \ell^{(p_n)}(\mathbb{R})$ their derivatives read*

$$\langle \mathbf{J}_{\rho_{(p_n)}}(x), h\rangle = \sum_{n\in\mathbb{N}} p_n \operatorname{sign}(x_n)|x_n|^{p_n-1}h_n, \quad \langle \mathbf{J}_{\bar{\rho}_{(p_n)}}(x), h\rangle = \sum_{n\in\mathbb{N}} \operatorname{sign}(x_n)|x_n|^{p_n-1}h_n. \tag{8}$$

Notice that, although $\mathbf{J}_{\rho_{(p_n)}}$ and $\mathbf{J}_{\bar{\rho}_{(p_n)}}$ are formally not duality maps, we adopt the same notation for the sake of consistency with Asplund Theorem 1.

3 Modular-based gradient descent in $\boldsymbol{\ell^{(p_n)}}(\mathbb{R})$

Given $(p_n)_{n\in\mathbb{N}}, (q_n)_{n\in\mathbb{N}} \in \mathcal{P}$, we now discuss how to implement a deterministic gradient-descent (GD) type algorithm for solving an instance of (P) with $\mathcal{X} = \ell^{(p_n)}(\mathbb{R})$ and $\mathcal{Y} = \ell^{(q_n)}(\mathbb{R})$. Recalling (3), GD iterations in this setting require knowing the duality map $\mathbf{J}^r_{\ell^{(p_n)}}$ and its inverse. However, as shown in [5, Corollary 3.2.14], such an inverse does not directly relate to the point-wise conjugate exponents of $(p_n)_{n\in\mathbb{N}}$ as the isomorphism between $(\ell^{(p_n)})^*(\mathbb{R})$ and $\ell^{(p_n^*)}(\mathbb{R})$ -differing from the standard ℓ^p constant case- is not isometric. As discussed

Algorithm 1: Modular-based Gradient Descent in $\ell^{(p_n)}(\mathbb{R})$

Parameters: $\{\mu_k\}_k$ s.t. $0 < \bar{\mu} \leqslant \mu_k \leqslant \frac{\mathtt{p}c(1-\delta)}{K}$ with $0 < \delta < 1$, for all $k \geqslant 0$.
Initialisation: $x^0 \in \ell^{(p_n)}(\mathbb{R})$.
repeat

$$x^{k+1} = |\mathbf{J}_{\bar{\rho}_{(p_n)}}(x^k) - \mu_k \nabla f(x^k)|^{\frac{1}{p_n - 1}} \operatorname{sign}\left(\mathbf{J}_{\bar{\rho}_{(p_n)}}(x^k) - \mu_k \nabla f(x^k)\right) \tag{9}$$

until convergence

in [2], the approximation $\left(\mathbf{J}^r_{\ell^{(p_n)}}\right)^{-1} = \mathbf{J}^{r*}_{(\ell^{(p_n)})*} \approx \mathbf{J}^{r*}_{\ell^{(p_n^*)}}$ can be used as an inexact (but explicit) formula for computing the duality map of $(\ell^{(p_n)})^*(\mathbb{R})$. Under this assumption, the dual Landweber method can thus be used to solve the minimisation problem $\operatorname{argmin}_{x \in \ell^{(p_n)}(\mathbb{R})} \frac{1}{q}\|Ax - y\|^q_{\ell^{(q_n)}}, \quad q > 1$. Note, however, that the computation of the duality map $\mathbf{J}^p_{\ell^{(p_n)}}$ requires the computation of $\|x\|_{\ell^{(p_n)}}$ which, as previously discussed, makes the iterative scheme rather inefficient in terms of computational time. We thus follow [14] and define in Algorithm 1 a more efficient modular-based gradient descent iteration for the resolution of (P) in the general setting of variable exponent Lebesgue spaces. The following set of assumptions needs to hold:

A.1 $\nabla f : \ell^{(p_n)}(\mathbb{R}) \to (\ell^{(p_n)})^*(\mathbb{R})$ is $(\mathtt{p}-1)-$Hölder-continuous with exponent $1 < \mathtt{p} \leqslant 2$ and constant $K > 0$.

A.2 There exists $c > 0$ such that, for all $u, v \in \ell^{(p_n)}(\mathbb{R})$,

$$\langle \mathbf{J}_{\bar{\rho}_{(p_n)}}(u) - \mathbf{J}_{\bar{\rho}_{(p_n)}}(v), u - v \rangle \geqslant c \max\left\{ \|u - v\|^{\mathtt{p}}_{\ell^{(p_n)}}, \|\mathbf{J}_{\bar{\rho}_{(p_n)}}(u) - \mathbf{J}_{\bar{\rho}_{(p_n)}}(v)\|^{\mathtt{p}^*}_{(\ell^{(p_n)})*} \right\}.$$

The latter bound was previously used in [8,14]. It is a compatibility condition between the ambient space $\ell^{(p_n)}(\mathbb{R})$ and the Hölder smoothness properties of the residual function to minimise to achieve algorithmic convergence.

The minimisation of the specific function f of (P) is achieved solving at each iteration (9) the following minimisation problem:

$$x^{k+1} = \operatorname*{argmin}_{u \in \ell^{(p_n)}(\mathbb{R})} \bar{\rho}_{(p_n)}(u) - \langle \mathbf{J}_{\bar{\rho}_{(p_n)}}(x^k), u \rangle + \mu_k \langle \nabla f(x^k), u \rangle.$$

The following proof shows that the functional $\mathbf{J}_{\bar{\rho}_{(p_n)}}$ defined by (8) is invertible and gives a point-wise characterisation of its inverse.

Proposition 3. *The functional $\mathbf{J}_{\bar{\rho}_{(p_n)}}$ in (8) is invertible. For all $v \in (\ell^{(p_n)})^*(\mathbb{R})$,*

$$(\mathbf{J}_{\bar{\rho}_{(p_n)}})^{-1}(v) = \left(|v_n|^{\frac{1}{p_n - 1}} \operatorname{sign}(v_n) \right)_{n \in \mathbb{N}} \in \ell^{(p_n)}(\mathbb{R}).$$

Proof. By straightforward componentwise computation, we have

$$\begin{aligned} &|\mathbf{J}_{\bar{\rho}_{(p_n)}}(v_n)|^{\frac{1}{p_n - 1}} \operatorname{sign}(\mathbf{J}_{\bar{\rho}_{(p_n)}}(v_n)) = |\mathbf{J}_{\bar{\rho}_{(p_n)}}(v_n)|^{\frac{1}{p_n - 1} - 1} \mathbf{J}_{\bar{\rho}_{(p_n)}}(v_n) \\ &= |\mathbf{J}_{\bar{\rho}_{(p_n)}}(v_n)|^{\frac{2 - p_n}{p_n - 1}} \mathbf{J}_{\bar{\rho}_{(p_n)}}(v_n) = \big| |v_n|^{p_n - 1} \operatorname{sign}(v_n) \big|^{\frac{2 - p_n}{p_n - 1}} |v_n|^{p_n - 1} \operatorname{sign}(v_n) = v_n \,. \end{aligned}$$

By the Proposition above, the update rule (9) of Algorithm 1, can be rewritten as

$$x^{k+1} = (\mathbf{J}_{\bar{\rho}_{(p_n)}})^{-1}\Big(\mathbf{J}_{\bar{\rho}_{(p_n)}}(x^k) - \mu_k \nabla f(x^k)\Big).$$

As a consequence, whenever $\nabla f(x_k) = 0$ at some $k \geqslant 0$, a stationary point $x^{k+1} = (\mathbf{J}_{\bar{\rho}_{(p_n)}})^{-1}\Big(\mathbf{J}_{\bar{\rho}_{(p_n)}}(x^k)\Big) = x^k$ is found, as expected.

The following convergence result is a special case of [14, Proposition 3.4] providing an explicit convergence rate for the iterates of Algorithm 1.

Proposition 4. *Let $x^* \in \ell^{(p_n)}(\mathbb{R})$ be a minimiser of f and let $(x^k)_k$ be the sequence generated by Algorithm 1. If (x^k) is bounded, then:*

$$f(x^k) - f(x^*) \leqslant \frac{\eta}{k^{\boldsymbol{p}-1}},$$

*where $\boldsymbol{p} > 1$ is defined in assumption **A.1** and $\eta = \eta(\bar{\mu}, \delta, p_-, x^0, x^*)$.*

Note that when the measurement space $\mathcal{Y}$ is a variable exponent Lebesgue space $\ell^{(q_n)}(\mathbb{R})$, a more effective and consistent choice for the objective function is the modular of the discrepancy between the model observation and the data, i.e. $f(x) = \bar{\rho}_{(q_n)}(Ax - y)$. In this way, the heavy computations of the $\|\cdot\|_{\ell^{(q_n)}}$ norm and of its gradient are not required, making the iteration scheme faster.

4 Stochastic modular-based gradient-descent in $\boldsymbol{\ell^{(p_n)}}(\mathbb{R})$

The key challenge for the viability of many deterministic iterative methods for real-world image reconstruction problems is their scalability to data-size. For example, the highest per-iteration cost in emission tomography lies in the application of the entire forward operator at each iteration, whereas each image domain datum in computed tomography often requires several gigabytes of storage space. The same could thus be a bottleneck in the application of Algorithm 1. The stochastic gradient descent (SGD) paradigm addresses this issue [20].

We partition the forward operator A, and the forward model into a finite number of block-type operators $A_1, \ldots, A_{N_s}$, where $N_s \in \mathbb{N}$ is the number of subsets of data. The same partition is applied to the observations. Classical examples of this methodology include Kaczmarz methods in CT [9, 17]. The SGD version of the iteration (3) in Banach spaces takes the form

$$x^{k+1} = \mathbf{J}^{p^*}_{\mathcal{X}^*}\left(\mathbf{J}^{p}_{\mathcal{X}}(x^k) - \mu_{k+1} A^*_{i_k} \mathbf{J}^{q}_{\mathcal{Y}}(A_{i_k} x^k - y)\right), \tag{10}$$

where the indices $i_k \in \{1, \ldots, N_s\}$ are sampled uniformly at random. Sampling reduces the per-iteration computational cost in $\mathcal{Y}$ by a factor of N_s. In [13] convergence of the iterates to a minimum norm solution is shown.

Theorem 3. *Let $\sum_{k=1}^{\infty} \mu_k = +\infty$ and $\sum_{k=1}^{\infty} \mu_k^{p^*} < +\infty$. Then*

$$\mathbb{P}\Big(\lim_{k\to\infty} \inf_{\widetilde{x}\in\mathcal{X}_{\min}} \|x^{k+1} - \widetilde{x}\|_{\mathcal{X}} = 0\Big) = 1.$$

Algorithm 2: Stochastic Modular-based Gradient Descent in $\ell^{(p_n)}(\mathbb{R})$

Parameters: μ_0 s.t. $0 < \bar{\mu} \leqslant \mu_0 \leqslant \frac{p c(1-\delta)}{K}$, $0 < \delta < 1$, $N_s \geqslant 1$, $\gamma > 0$, $\eta > 0$.
Initialisation: $x^0 \in \ell^{(p_n)}(\mathbb{R})$.
repeat
Select uniformly at random $i_k \in \{1, \cdots, N_s\}$.
Set $\mu_k = \frac{\mu_0}{1+\eta(k/N_s)^\gamma}$
Compute

$$x^{k+1} = |\mathbf{J}_{\bar{\rho}_{(p_n)}}(x^k) - \mu_k \nabla f_{i_k}(x^k)|^{\frac{1}{p_n - 1}} \operatorname{sign}\left(\mathbf{J}_{\bar{\rho}_{(p_n)}}(x^k) - \mu_k \nabla f_{i_k}(x^k)\right)$$

until convergence

Let $\mathbf{J}^p_{\mathcal{X}}(x_0) \in \overline{\operatorname{range}(A^*)}$ *and let* $\mu_k^{p^*-1} \leqslant \frac{C}{L^{p^*}_{\max}}$ *for all* $k \geqslant 0$ *and some constant* $C > 0$, *where* $L_{\max} = \max_i \|A_i\|$. *Then* $\lim_{k\to\infty} \mathbb{E}[\|x^{k+1} - x^\dagger\|_{\mathcal{X}}] = 0$ $\lim_{k\to\infty} \mathbb{E}[\|\mathbf{J}^p_{\mathcal{X}}(x^{k+1}) - \mathbf{J}^p_{\mathcal{X}}(x^\dagger)\|^{p^*}] = 0$.

For noisy measurements, the regularising property of SGD should be established by defining suitable stopping criteria. However, robust stopping strategies are hard to use in practice and having methods that are less sensitive to data overfit is crucial for their practical use. Note that (10) is the standard form of SGD for separable objectives. Namely, for $f(x) = \|Ax - y\|_q^q$, we can choose $f_i(x; A, y) = \|A_i x - y_i\|_q^q$, so that $f(x) = \sum_{i=1}^{N_s} f_i(x)$. By Theorem 1, this decomposition shows that each step of (10) can thus be computed by simply taking a sub-differential of a single sum-function f_i.

To define a suitable SGD in variable exponent Lebesgue spaces, we take as objective function $f(x) = \bar{\rho}_{(q_n)}(Ax - y)$ and split it into $N_s \geqslant 1$ sub-objectives $f_i(x) := \bar{\rho}_{(q_n^i)}(A_i x - y_i)$, so that $\nabla f_i(x) = A_i^* \mathbf{J}_{\bar{\rho}_{(q_n^i)}}(A_i x - y_i)$. Exponents $(q_n^i)_n$ are obtained through the same partition of the exponents $(q_n)_n$ as the one used to split up the data. Then, at iteration k and a randomly sampled index $1 \leqslant i_k \leqslant N_s$, the corresponding stochastic iterates are given by

$$x^{k+1} = \operatorname*{argmin}_{u \in \ell^{(p_n)}(\mathbb{R})} \bar{\rho}_{(p_n)}(u) - \langle \mathbf{J}_{\bar{\rho}_{(p_n)}}(x^k), u\rangle + \mu_k \langle \nabla f_{i_k}(x^k), u\rangle.$$

The pseudocode of the resulting stochastic modular-based gradient descent in $\ell^{(p_n)}(\mathbb{R})$ is reported in Algorithm 2. We expect that through minimal modifications an analogous convegence result as Theorem 3 can be proved in this setting too. A detailed convergence proof, however, is left for future research.

5 Numerical results

We now present experimental results of the proposed Algorithm 2 on two exemplar problems in computed tomography (CT). The first set of experiments consider a simulated setting for quantitatively comparing the performance of Algorithm 2 with the corresponding Hilbert and Banach space versions (10). In

the second set of experiments we consider the dataset of real-world CT scans of a walnut taken from doi:10.5281/zenodo.4279549, with a fan beam geometry. For these data, we utilise the insights from the first set of experiments and apply Algorithm 2 in a setting with different noise modalities across the sinogram space. The experiments were conducted in `python`, using the open source package [12] for the tomographic backend.

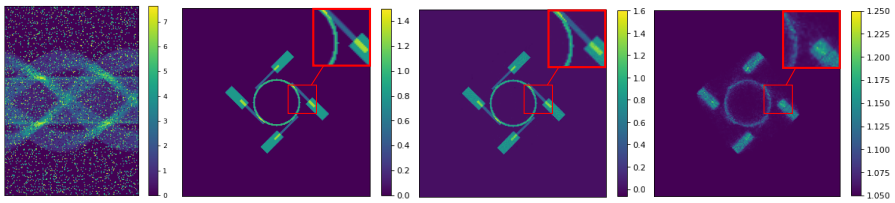
Hyper-parameter selection. In the following experiments, we employ a decaying stepsize regime such that it satisfies the conditions of Theorem 3 for the convergence of Banach space SGD, cf. [13]. A need for a decaying stepsize regime is common for stochastic gradient descent to mitigate the effects of inter-iterate variance. Specifically, we use $\mu_k = \frac{\mu_0}{1+c(k/N_s)^\gamma}$, where $\mu_0 > 0$ is the initial stepsize, and $\gamma > 0$ and $c > 0$ control the decay speed. For the Hilbert space setting, $\mathbf{SGD}_2$, initial stepsize μ_0 is given by the Lipschitz constant of the gradient of the objective function, namely $\mu_0 = 0.95/\max_i \|A_i\|^2$. For $\mathbf{SGD_p}$ and $\mathbf{SGD_{p_n,q_n}}$ the estimation of the respective Hölder continuity constant is more delicate and μ_0 has to be tuned to guarantee convergence. However, its tuning is rather easy and the employ of a decaying strategy makes the choice of μ_0 less critical.

As far as variable exponents are concerned, it is difficult (and somehow undesirable) to have a unified configuration as their selection is strictly problem-related. Parameters $(q_n)_n$ are related to the regularity of the measured sinograms as well as the different noise distributions considered. For instance, when impulsive noise is considered, values of q_- and q_+ closer to 1 are preferred while and for Gaussian noise values closer to 2 are more effective. Solution space parameters p_- and p_+ relate to the regularity of the solution to retrieve. As a consequence, their choice is intrinsically harder. We refer the reader to [2], where a comparison between different choices for p_- and p_+ and different interpolation strategies is carried out for image deblurring with gradient descent (3) in $\ell^{(p_n)}$.

Simulated data. We considered (1) with A given by the discrete Radon transform. For its definition we use a 2D parallel beam geometry, with 180 projection angles on a 1 angle separation, 256 detector elements, and pixel size of 0.1. The synthetic phantom was provided by the CIL library, see Figure 1(b). After applying the forward operator, a high level (15%) of salt-and-pepper noise is applied to the sinogram. The noisy sinogram is shown in Figure 1(a).

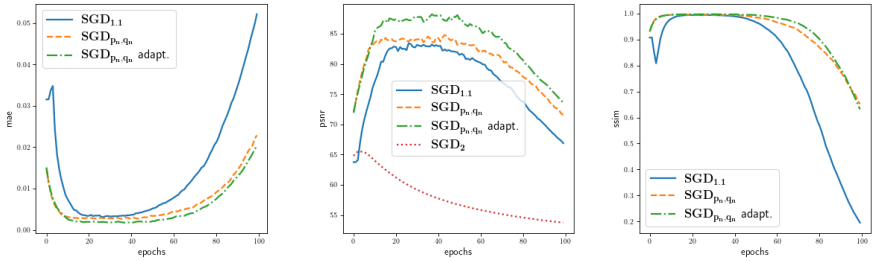
To compute subset data A_i and y_i, the forward operator and the sinogram are pre-binned according to equally spaced views (w.r.t. the number of subsets) of the scanner geometry. Subsequent subset data are offset from one another by the subset index i. We consider $N_s = 30$ batches. We compare results obtained by solving (P) by:

$\mathbf{SGD_2}$: $\mathcal{X} = \mathcal{Y} = \ell^2(\mathbb{R})$, $\mathcal{Y} = \ell^2(\mathbb{R})$, $f(x) = \frac{1}{2}\|Ax - y\|_2^2$ by SGD;
$\mathbf{SGD_p}$: $\mathcal{X} = \mathcal{Y} = \ell^p(\mathbb{R})$, $p = 1.1$, $f(x) = \frac{1}{p}\|Ax - y\|_p^p$ by Banach SGD (10);
$\mathbf{SGD_{p_n,q_n}}$: $\mathcal{X} = \ell^{(p_n)}(\mathbb{R})$, $\mathcal{Y} = \ell^{(q_n)}(\mathbb{R})$ for appropriately chosen exponent maps, $f(x) = \bar{\rho}_{(q_n)}(Ax - y)$ with modular-based SGD Algorithm 2.



(a) Sinogram (b) GT (c) $\mathbf{SGD_{p_n,q_n}}$ (d) (p_n) map

Fig. 1: In (c) reconstruction of noisy sinogram (a) by $\mathbf{SGD_{p_n,q_n}}$, where $1.05 = p_- \leqslant (p_n) \leqslant p_+ = 1.25$ is shown in (d) and $1.05 = q_- \leqslant (q_n) \leqslant q_+ = 1.25$ is based on the model observation corresponding to (p_n).



(a) MAE (b) PSNR (c) SSIM

Fig. 2: Quality metrics along the first 100 epochs of $\mathrm{SGD}_{\mathbf{2}}$; $\mathrm{SGD}_{\mathbf{1.1}}$; $\mathrm{SGD}_{\mathbf{p_n,q_n}}$ with and without adapting the exponent maps (p_n). $\mathrm{SGD}_{\mathbf{2}}$ is omitted from MAE and SSIM to improve the readability of the plots, due to its poor performance.

We considered step-sizes $\mu_k = \frac{\mu_0}{1+0.1(k/N_s)^\gamma}$, with μ_0 and γ which depend on the algorithm.[4] Spaces $\ell^{(p_n)}(\mathbb{R})$ allow for variable exponent maps sensitive to local assumptions on both the solution and the measured data. A possible strategy for informed pixel-wise variable exponents consists in basing them on observed data (for (q_n)) and an approximation of the reconstruction (for (p_n)), as done in [1, 2, 14]. To this end, we first compute an approximate reconstruction $\tilde{x} \in \ell^{(p_n)}(\mathbb{R})$ by running $\mathbf{SGD_p}$ in $\ell^{1.1}(\mathbb{R})$ for 5 epochs with a constant stepsize regime. The map (p_n) is then computed via a linear interpolation of $\tilde{x}$ between $p_- = 1.05$ and $p_+ = 1.25$. The map (q_n) is chosen as the linear interpolation between $q_- = 1.05$ and $q_+ = 1.25$ of $A(p_n)$. The bounds p_-, p_+ and q_-, q_+ are chosen by prior assumptions on y (sparse phantom) and on the noise observed (impulsive). We also tested an adaptive strategy by updating (p_n) based on the current solution estimate once every β_{updates} epochs to adapt the exponents along the iterations.

In Figure 2, we report the mean absolute error (MAE), peak signal to noise ratio (PSNR) and structural similarity index (SSIM) of the iterates x^k w.r.t. the

[4] For $\mathbf{SGD_2}$ μ_0 is set as $0.95/\max_i \|A_i\|^2$ and $\gamma = 0.51$. For $\mathbf{SGD_p}$ and $\mathbf{SGD_{p_n,q_n}}$, we use $\mu_0 = 0.015$ with $\gamma = (p-1)/p + 0.01$ and $\gamma = (p_- - 1)/p_- + 0.01$ respectively.

	Deterministic		Stochastic ($\cdot = \mathbf{S}$)					
Algorithm	It.	Tot.	It.	Epoch	Tot.	MAE	PSNR	SSIM
$\cdot\mathbf{GD_2}$	0.44s	1324s	0.02s	0.74s	74.4 s	2.582e-1	57.89	0.0304
$\cdot\mathbf{GD_{1.1}}$	0.43s	1297s	0.03s	0.81s	81.3s	3.671e-3	82.64	0.9897
$\cdot\mathbf{GD_{p_n,q_n}}$	0.47s	1403s	0.03s	0.96s	96.5s	2.887e-3	84.05	0.9927
$\cdot\mathbf{GD_{p_n,q_n}}$ adapt.	0.44s	1317s	0.03s	0.91s	91.2s	1.777e-3	88.10	0.9965
Compute (p_n), (q_n)	0.45s	16s	0.03s	0.8s	4.0s	-	-	-

Table 1: Comparison of per iteration cost and total CPU times after 3000 iterations for determistic algorithms and after 100 epochs for stochastic algorithm with $N_s = 30$. MAE, PSNR and SSIM values for stochastic algorithms are computed after 40 epochs (before noise overfitting).

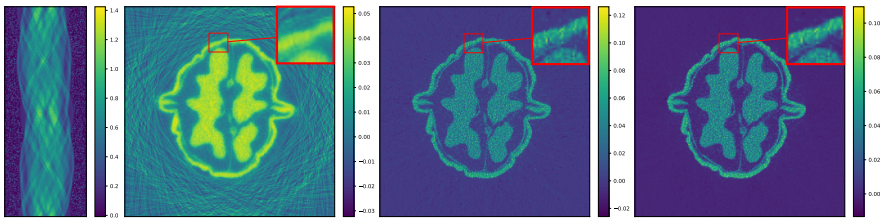
known ground-truth phantom along the first 100 epochs. Since PSNR favours smoothness, it is thus beneficial for $\mathbf{SGD_2}$, whereas MAE promotes sparsity hence is beneficial for both $\mathbf{SGD_p}$ and $\mathbf{SGD_{p_n,q_n}}$. Figure 2b shows that Banach space algorithms provide better performance than $\mathbf{SGD_2}$ in all three quality metrics. Note that all the results show the well-known semi-convergence behaviour with respect to the metrics considered. To avoid such behaviour an explicit regulariser or a sound early stopping criterion would be beneficial for reconstruction performance. We observe that the use variable exponents does not only improve all quality metrics, but also makes the algorithm more stable: the quality of the reconstructed solutions is significantly less sensitive to the number of epochs, making possible early stopping strategies more robust.

In Table 1, the CPU times for deterministic ($\mathbf{GD_2}$, $\mathbf{GD_p}$ and $\mathbf{GD_{p_n,q_n}}$) approaches and stochastic ones ($\mathbf{SGD_2}$, $\mathbf{SGD_p}$ and $\mathbf{SGD_{p_n,q_n}}$) are compared.

Real CT datasets: walnut. We consider a cone beam CT dataset of a walnut [15], from which we take a 2D fan beam sinograms from the centre plane of the cone. The cone beam data uses 0.5 angle separation over the range $[0, 360]$. The used sinogram is obtained by pre-binning the raw data by a factor of 8, resulting in 280 effective detector pixels. The measurements have been post-processed for dark current and flat-field compensation. As stepsize we used $\mu_k = \frac{\mu_0}{1+0.001(k/N_s)^\gamma}$, with $N_s = 10$ subsets, and suitable μ_0 and γ. [5] Initial images are computed by 5 epochs of $\mathbf{SGD_{1.4}}$ with a constant stepsize.

We consider a more delicate noise setting that requires exponential maps which vary in the acquisition domain. Here, we assume that noise has a different effect on the background (zero entries) and the foreground (non-zero entries) of the clean sinogram. Namely, we apply 10% salt and pepper noise to the background, and speckle noise with mean 0 and variance 0.01 to the foreground, cf. Fig. 3(a) for the resulting noisy sinogram. Notably, since this noise model has a non-uniform effect across the measurement data, Banach space methods favouring the adjustment of the Lebesgue exponents are expected to perform better than those making use of a constant value. Taking as a reference the result ob-

[5] For $\mathbf{SGD}_2$, $\mu_0 = 0.95/\max_i \|A_i\|^2$, $\gamma = 0.51$. For $\mathbf{SGD}_{p_n,q_n}$ we $\mu_0 = 0.001$, $\gamma = 0.58$.



(a) Sinogram (b) SGD (c) Constant exponents (d) Variable exponents

Fig. 3: (a) Noisy sinogram with 10% salt & pepper (background) and speckle noise with 0 mean and variance 0.01 (foreground). (b) $\mathbf{SGD_2}$ result. (c) $\mathbf{SGD_{p_n,1.1}}$ result (d) $\mathbf{SGD_{p_n,q_n}}$ result. $p_- = 1.2$, $p_+ = 1.3$, $q_- = 1.1$ and $q_+ = 1.9$.

tained by $\mathbf{SGD_2}$ (Fig. 3(b)), we compare here the effect of allowing variable exponents in the solution space only with the effect of allowing both maps (p_n) and (q_n) to be chosen. By choosing (p_n) based on the initial image and interpolating it between $p_- = 1.2$ and $p_+ = 1.3$ we then compare $\mathbf{SGD_{p_n,1.1}}$ (i.e., fixed exponent $q = 1.1$ in the measurement space), cf. Fig. 3(c), with $\mathbf{SGD_{p_n,q_n}}$ where (p_n) is as before while (q_n) is chosen from the sinogram by interpolating between $q_- = 1.1$ and $q_+ = 1.9$, cf. Fig. 3(d). The results show that a flexible framework where both maps (p_n) and (q_n) adapt to local contents are more suited for dealing with this challenging scenario.

6 Conclusions

We proposed a stochastic gradient descent algorithm for solving linear inverse problems in $\ell^{(p_n)}(\mathbb{R})$. After recalling its deterministic counterpart and the difficulties encountered due to the non-separability of the underlying norm, a modular-based stochastic algorithm enjoying fast scalability properties is proposed. Numerical results show improved performance in comparison to standard $\ell^2(\mathbb{R})$ and $\ell^p(\mathbb{R})$-based algorithms and significant computational gains. Future work should adapt the convergence result (Theorem 2) to this setting and consider proximal extensions for incorporating non-smooth regularisation terms.

7 Acknowledgements

CE and ML acknowledge the support of the Italian INdAM group on scientific calculus GNCS. LC acknowledges the support received by the ANR projects TASK-ABILE (ANR-22-CE48-0010) and MICROBLIND (ANR-21-CE48-0008), the H2020 RISE projects NoMADS (GA. 777826) and the GdR ISIS project SPLIN. ZK acknowledges support from EPSRC grants EP/T000864/1 and EP/X010740/1.

References

1. M. Alparone, F. Nunziata, C. Estatico, F. Lenti, and M. Migliaccio. An adaptive l^p -penalization method to enhance the spatial resolution of microwave radiometer measurements. *IEEE Trans. Geosci. Remote Sens.*, 57(9):6782–6791, 2019.

2. B. Bonino, C. Estatico, and M. Lazzaretti. Dual descent regularization algorithms in variable exponent Lebesgue spaces for imaging. *Numer. Algorithms*, 92(6), 2023.
3. I. Cioranescu. Geometry of Banach spaces, duality mappings and nonlinear problems. *Springer*, 1990.
4. D. V. Cruz-Uribe and A. Fiorenza. Variable Lebesgue spaces. *Springer Birkhäuser Basel*, 2013.
5. L. Diening, P. Harjulehto, P. Hästö, and M. Ruzicka. *Lebesgue and Sobolev Spaces with Variable Exponents*. Lecture Notes in Math. Springer-Verlag, Germany, 2011.
6. B. Eicke. Iteration methods for convexly constrained ill-posed problems in hilbert space. *Numer Funct Anal Optim*, 13(5-6):413–429, 1992.
7. H. W. Engl and A. Hanke, M. Neubauer. *Regularization of Inverse Problems*. Mathematics and Its Applications. Springer, 2000.
8. W.-B. Guan and W. Song. The Generalized Forward-Backward Splitting Method for the Minimization of the Sum of Two Functions in Banach Spaces. *Numer. Funct. Anal. Optim.*, 36(7):867–886, 2015.
9. G.T. Herman and L.B. Meyer. Algebraic reconstruction techniques can be made computationally efficient (positron emission tomography application). *IEEE Trans. Med. Imaging*, 12(3):600–609, 1993.
10. Q. Jin, X. Lu, and L. Zhang. Stochastic mirror descent method for linear ill-posed problems in Banach spaces, 2022. arXiv preprint: https://arxiv.org/abs/2207.06584.
11. Q. Jin and L. Stals. Nonstationary iterated Tikhonov regularization for ill-posed problems in Banach spaces. *Inverse Probl.*, 28(10):104011, oct 2012.
12. J. S. Jørgensen and et al. Core Imaging Library - Part I: a versatile Python framework for tomographic imaging. *Phil. Trans. R. Soc. A*, 2021.
13. Z. Kereta and B. Jin. On the convergence of stochastic gradient descent for linear inverse problems in Banach spaces. *SIAM J. Imaging Sci. (in press)*, 2023. arXiv preprint: https://arxiv.org/abs/2302.05197.
14. M. Lazzaretti, L. Calatroni, and C. Estatico. Modular-proximal gradient algorithms in variable exponent Lebesgue spaces. *SIAM J. Sci. Compu.*, 44(6), 2022.
15. A. Meaney. X-ray dataset of walnut (2020-11-11), November 2020.
16. F. Natterer. The mathematics of computerized tomography. *John Wiley*, 1986.
17. D. Needell, R. Zhao, and A. Zouzias. Randomized block Kaczmarz method with projection for solving least squares. *Linear Algebra Appl.*, 484:322–343, 2015.
18. A Neubauer. Tikhonov-regularization of ill-posed linear operator equations on closed convex sets. *J. Approx. Theory*, 53(3):304–320, 1988.
19. M. Piana and M. Bertero. Projected Landweber method and preconditioning. *Inverse Probl.*, 13(2):441–463, apr 1997.
20. H. Robbins and S. Monro. A Stochastic Approximation Method. *Ann. Math. Stat.*, 22(3):400 – 407, 1951.
21. T. Schuster, B. Kaltenbacher, B. Hofmann, and K. S. Kazimierski. Regularization methods in Banach spaces. *De Gruyter*, 2012.
22. F. Schöpfer, A. K. Louis, and T. Schuster. Nonlinear iterative methods for linear ill-posed problems in Banach spaces. *Inverse Probl.*, 22(1):311–329, 2006.
23. R. Twyman, S. Arridge, and et al. An investigation of stochastic variance reduction algorithms for relative difference penalized 3D PET image reconstruction. *IEEE Trans. Med. Imaging*, 42(1):29–41, 2023.