SIAM J. IMAGING SCIENCES
Vol. 16, No. 2, pp. 671–705

On the Convergence of Stochastic Gradient Descent for Linear Inverse Problems in Banach Spaces*

Bangti Jin† and Željko Kereta‡

Abstract. In this work we consider stochastic gradient descent (SGD) for solving linear inverse problems in Banach spaces. SGD and its variants have been established as one of the most successful optimization methods in machine learning, imaging, and signal processing, to name a few. At each iteration SGD uses a single datum, or a small subset of data, resulting in highly scalable methods that are very attractive for large-scale inverse problems. Nonetheless, the theoretical analysis of SGD-based approaches for inverse problems has thus far been largely limited to Euclidean and Hilbert spaces. In this work we present a novel convergence analysis of SGD for linear inverse problems in general Banach spaces: we show the almost sure convergence of the iterates to the minimum norm solution and establish the regularizing property for suitable a priori stopping criteria. Numerical results are also presented to illustrate features of the approach.

Key words. stochastic gradient descent, Banach spaces, linear inverse problems, convergence rate, regularizing property, almost sure convergence

MSC codes. 60G48, 46N10, 49N45, 65J22

DOI. 10.1137/22M1518542

1. Introduction.

This work considers (stochastic) iterative solutions for linear operator equations of the form

$$\mathbf{A}\mathbf{x} = \mathbf{y}, \tag{1.1}$$

where $\mathbf{A} : \mathcal{X} \to \mathcal{Y}$ is a bounded linear operator between Banach spaces $\mathcal{X}$ and $\mathcal{Y}$ (equipped with the norms $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$, respectively), and $\mathbf{y} \in \mathrm{range}(\mathbf{A})$ is the exact data. In practice, we only have access to noisy data $\mathbf{y}^\delta = \mathbf{y} + \boldsymbol{\xi}$, where $\boldsymbol{\xi}$ denotes the measurement noise with a noise level $\delta \geq 0$ such that $\|\mathbf{y}^\delta - \mathbf{y}\|_{\mathcal{Y}} \leq \delta$. Linear inverse problems arise naturally in many applications in science and engineering, and also form the basis for studying nonlinear inverse problems. Hence, design and analysis of stable reconstruction methods for linear inverse problems have received much attention.

Iterative regularization is a powerful algorithmic paradigm that has been successfully employed for many inverse problems [14, Chapters 6 and 7], [30]. Classical iterative methods for

*Received by the editors August 29, 2022; accepted for publication (in revised form) December 22, 2022; published electronically May 15, 2023.
https://doi.org/10.1137/22M1518542
Funding: This work was supported by UK EPSRC grant EP/T000864/1. The work of the first author was also supported by UK EPSRC grant EP/V026259/1 and a start-up fund of The Chinese University of Hong Kong.

†Department of Mathematics, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong (bangti.jin@gmail.com, b.jin@cuhk.edu.hk).
‡Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK (z.kereta@ucl.ac.uk).

inverse problems include the (accelerated) Landweber method, conjugate gradient method, Levenberg–Marquardt method, and Gauss–Newton method, to name a few. The per-iteration computational bottleneck of many iterative methods lies in utilizing all the data at each iteration, which can be of a prohibitively large size. For example, this occurs while computing the derivative of an objective. One promising strategy to overcome this challenge is stochastic gradient descent (SGD), due to Robbins and Monro [40]. SGD decomposes the original problem into (finitely many) subproblems, and then at each iteration uses only a single datum, or a minibatch of data, typically selected uniformly at random. This greatly reduces the per iteration computational complexity and enjoys excellent scalability with respect to data size. In the standard, and best-studied, setting, $\mathcal{X}$ and $\mathcal{Y}$ are finite-dimensional Euclidean spaces and the corresponding data fitting objective is the (rescaled) least squares $\Psi(\mathbf{x}) = \frac{1}{2N}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_{\mathcal{Y}}^2 = \frac{1}{N}\sum_{i=1}^N \frac{1}{2}\|\mathbf{A}_i\mathbf{x} - \mathbf{y}_i\|_{\mathcal{Y}}^2$. In this setting SGD takes the form

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mu_{k+1}\mathbf{A}_{i_{k+1}}^*(\mathbf{A}_{i_{k+1}}\mathbf{x}_k - \mathbf{y}_{i_{k+1}}), \quad k = 0, 1, \ldots,$$

where $\mu_k >$ is the step-size, i_{k+1} is a randomly selected index, $\mathbf{A}_i$ is the ith row of a matrix $\mathbf{A}$, and $\mathbf{y}_i$ is the ith entry of $\mathbf{y}$. In the seminal work [40], Robbins and Monro presented SGD as a Markov chain, laying the groundwork for the field of stochastic approximation [32]. SGD has since had a major impact on statistical inference and machine learning, especially for the training of neural networks. SGD has been extensively studied in the Euclidean setting; see [2] for an overview of the convergence theory from the viewpoint of optimization.

SGD has also been a popular method for image reconstruction, especially in medical imaging. For example, the (randomized) Kaczmarz method is a reweighted version of SGD that has been extensively used in computed tomography [18, 37]. Other applications of SGD and its variants include optical tomography [6], phonon transmission coefficient recovery [15], positron emission tomography [31], as well as general sparse recovery [42, 43]. For linear inverse problems in Euclidean spaces, Jin and Lu [24] gave a first proof of convergence of SGD iterates towards the minimum norm solution, and analyzed the regularizing behavior in the presence of noise; see [22, 25, 26, 34, 39] for further convergence results, a posteriori stopping rules (discrepancy principle), nonlinear problems, general step-size schedules, etc.

Iterative methods in Euclidean and Hilbert spaces are effective for reconstructing smooth solutions but fail to capture special features of the solutions, such as sparsity and piecewise constancy. In practice, many imaging inverse problems are more adequately described in non-Hilbert settings, including sequence spaces $\ell^p(\mathbb{R})$ and Lebesgue spaces $\mathcal{L}^p(\Omega)$, with $p \in [1, \infty] \setminus \{2\}$, which requires changing either the solution, the data space, or both. For example, inverse problems with impulse noise are better modeled by setting the data space $\mathcal{Y}$ to a Lebesgue space $\mathcal{L}^p(\Omega)$ with $p \approx 1$ [11], whereas the recovery of sparse solutions is modeled by doing the same to the solution space $\mathcal{X}$ [4]. Thus, it is of great importance to develop and analyze algorithms for inverse problems in Banach spaces, and this has received much attention [41, 46]. For the Landweber method for linear inverse problems in Banach spaces, Schöpfer, Louis, and Schuster [44] were the first to prove strong convergence of the iterates under a suitable step-size schedule for a smooth and uniformly convex Banach space $\mathcal{X}$ and an arbitrary Banach space $\mathcal{Y}$. This has since been extended and refined in various aspects, e.g., regarding acceleration [45, 49, 17, 51], nonlinear forward models [12, 35], and Gauss–Newton methods [29].

In this work, we investigate SGD for inverse problems in Banach spaces, which has thus far lagged behind due to outstanding challenges in extending the analysis of standard Hilbert space approaches to the Banach space setting. The main challenges in analyzing SGD-like gradient-based methods in Banach spaces are twofold:

1. The use of duality maps results in nonlinear update rules, which greatly complicates the convergence analysis. For example, the (expected) difference between successive updates can no longer be identified as the (sub)gradient of the objective.
2. Due to geometric characteristics of Banach spaces, it is more common to use the Bregman distance for the convergence analysis, which results in the loss of useful algebraic tools, e.g., triangle inequality and bias-variance decomposition, that are typically needed for the analysis.

In this work, we develop an SGD approach for the numerical solution of linear inverse problems in Banach spaces, using the subgradient approach based on duality maps, and present a novel convergence analysis. We first consider the case of exact data and show that SGD iterates converge to a minimizing solution (first almost surely and then in expectation) under standard assumptions on summability of step-sizes, and geometric properties of the space $\mathcal{X}$; cf. Theorems 3.8 and 3.10. This solution is identified as the minimum norm solution if the initial guess $\mathbf{x}_0$ satisfies the range condition $\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_0) \in \overline{\text{range}(\mathbf{A}^*)}$. Further, we give a convergence rate in Theorem 3.14 when the forward operator $\mathbf{A}$ satisfies a conditional stability estimate. In the case of noisy observations, we show the regularizing property of SGD for properly chosen stopping indices; cf. Theorem 4.3. The analysis rests on a descent property in Lemma 3.6 and the Robbins–Siegmund theorem for almost supermartingales. In addition, we perform extensive numerical experiments on a model inverse problem (linear integral equation) and computed tomography (with parallel beam geometry) to illustrate distinct features of the proposed Banach space SGD, and we examine the influence of various factors, such as the choice of the spaces $\mathcal{X}$ and $\mathcal{Y}$, minibatch size, and noise characteristics (Gaussian or impulse).

When finalizing the paper, we became aware of the independent and simultaneous work [27] on a stochastic mirror descent method for linear inverse problems between a Banach space $\mathcal{X}$ and a Hilbert space $\mathcal{Y}$. The method is a randomized version of the well-known Landweber–Kaczmarz method. The authors prove convergence results under a priori stopping rules, and also establish an order-optimal convergence rate when the exact solution $\mathbf{x}^\dagger$ satisfies a benchmark source condition, by interpreting the method as a randomized block gradient method applied to the dual problem. Thus, the current work differs significantly from [27] in terms of problem setting, main results, and analysis techniques.

The rest of the paper is organized as follows. In section 2, we recall background materials on the geometry of Banach spaces, e.g., duality maps and Bregman distance. In section 3, we present the convergence of SGD for exact data, and in section 4, we discuss the regularizing property of SGD for noisy observations. Finally, in section 5, we provide some experimental results on a model inverse problem and computed tomography. In the appendix we collect several useful inequalities and auxiliary estimates.

Throughout, let $\mathcal{X}$ and $\mathcal{Y}$ be two real Banach spaces, with their norms denoted by $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$, respectively. $\mathcal{X}^*$ and $\mathcal{Y}^*$ are their respective dual spaces, with their norms denoted by $\|\cdot\|_{\mathcal{X}^*}$ and $\|\cdot\|_{\mathcal{Y}^*}$, respectively. For $\mathbf{x} \in \mathcal{X}$ and $\mathbf{x}^* \in \mathcal{X}^*$, we denote the corresponding duality

pairing by $\langle \mathbf{x}^*, \mathbf{x}\rangle = \langle \mathbf{x}^*, \mathbf{x}\rangle_{\mathcal{X}^*\times\mathcal{X}} = \mathbf{x}^*(\mathbf{x})$. For a continuous linear operator $\mathbf{A}: \mathcal{X} \to \mathcal{Y}$, we use $\|\mathbf{A}\|_{\mathcal{X}\to\mathcal{Y}}$ to denote the operator norm (often with the subscript omitted). The adjoint of $\mathbf{A}$ is denoted by $\mathbf{A}^*: \mathcal{Y}^* \to \mathcal{X}^*$, and it is a continuous linear operator, with $\|\mathbf{A}\|_{\mathcal{X}\to\mathcal{Y}} = \|\mathbf{A}^*\|_{\mathcal{Y}^*\to\mathcal{X}^*}$. The conjugate exponent of $p \in (1,\infty)$ is denoted by p^*, such that $1/p + 1/p^* = 1$ holds. The Cauchy–Schwarz inequality of the following form holds for any $\mathbf{x} \in \mathcal{X}$ and $\mathbf{x}^* \in \mathcal{X}^*$:

$$|\langle \mathbf{x}^*, \mathbf{x}\rangle| \le \|\mathbf{x}^*\|_{\mathcal{X}^*}\|\mathbf{x}\|_{\mathcal{X}}. \tag{1.2}$$

For reals a, b we write $a\wedge b = \min\{a,b\}$ and $a\vee b = \max\{a,b\}$. By $(\mathcal{F}_k)_{k\in\mathbb{N}}$, we denote the natural filtration, i.e., a growing sequence of σ-algebras such that $\mathcal{F}_k \subset \mathcal{F}_{k+1} \subset \mathcal{F}$ for all $k \in \mathbb{N}$ and a σ-algebra $\mathcal{F}$, and $\mathcal{F}_k$. In the context of SGD, $k \in \mathbb{N}$ is the iteration number and $\mathcal{F}_k$ denotes the iteration history generated by random indices i_j for $j \le k$, that is, it denotes information available at time k. For a given initialization $\mathbf{x}_0$, we can identify $\mathcal{F}_k = \sigma(\mathbf{x}_1, \dots, \mathbf{x}_k)$. For a filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$ we denote by $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathbf{x}_1, \dots, \mathbf{x}_k]$ the conditional expectation with respect to $\mathcal{F}_k$. A sequence of random variables $(x_k)_{k\in\mathbb{N}}$ (adapted to the filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$) is a called supermartingale if $\mathbb{E}_k[x_{k+1}] \le x_k$. Throughout, the notation *a.s.* denotes almost sure events.

2. Preliminaries on Banach spaces.

In this section we recall relevant concepts from Banach space theory and the geometry of Banach spaces.

2.1. Duality map.

In a Hilbert space $\mathcal{H}$, for every $\mathbf{x} \in \mathcal{H}$, there exists a unique $\mathbf{x}^* \in \mathcal{H}^*$ such that $\langle \mathbf{x}^*, \mathbf{x}\rangle = \|\mathbf{x}\|_{\mathcal{H}}\|\mathbf{x}^*\|_{\mathcal{H}^*}$ and $\|\mathbf{x}^*\|_{\mathcal{H}^*} = \|\mathbf{x}\|_{\mathcal{H}}$ by the Riesz representation theorem. For Banach spaces, however, such an $\mathbf{x}^*$ is not necessarily unique, motivating the notion of duality maps.

Definition 2.1 (duality map). *For any $p > 1$, a* duality map $\mathcal{J}_p^{\mathcal{X}}: \mathcal{X} \to 2^{\mathcal{X}^*}$ *is the subdifferential of the (convex) functional $\frac{1}{p}\|\mathbf{x}\|_{\mathcal{X}}^p$,*

$$\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}) = \left\{\mathbf{x}^* \in \mathcal{X}^* : \langle \mathbf{x}^*, \mathbf{x}\rangle = \|\mathbf{x}\|_{\mathcal{X}}\|\mathbf{x}^*\|_{\mathcal{X}^*}, \text{ and } \|\mathbf{x}\|_{\mathcal{X}}^{p-1} = \|\mathbf{x}^*\|_{\mathcal{X}^*}\right\}, \tag{2.1}$$

with gauge function $t \mapsto t^{p-1}$. A single-valued selection of $\mathcal{J}_p^{\mathcal{X}}$ is denoted by $\jmath_p^{\mathcal{X}}$.

In practice, the choice of the power parameter p depends on geometric properties of the space $\mathcal{X}$. For single-valued duality maps, we use $\mathcal{J}_p^{\mathcal{X}}$ and $\jmath_p^{\mathcal{X}}$ interchangeably. Next we recall standard notions of smoothness and convexity of Banach spaces. For an overview of Banach space geometry, we refer the interested reader to the monographs [9, 10, 46].

Definition 2.2. *Let $\mathcal{X}$ be a Banach space. $\mathcal{X}$ is said to be reflexive if the canonical map $\mathbf{x} \mapsto \widehat{\mathbf{x}}$ between $\mathcal{X}$ and the bidual $\mathcal{X}^{**}$, defined by $\widehat{\mathbf{x}}(\mathbf{x}^*) = \mathbf{x}^*(\mathbf{x})$, is surjective. $\mathcal{X}$ is smooth if for every $0 \ne \mathbf{x} \in \mathcal{X}$ there is a unique $\mathbf{x}^* \in \mathcal{X}^*$ such that $\langle \mathbf{x}^*, \mathbf{x}\rangle = \|\mathbf{x}\|_{\mathcal{X}}$ and $\|\mathbf{x}^*\|_{\mathcal{X}^*} = 1$. The function $\delta_{\mathcal{X}}: (0,2] \to \mathbb{R}$, defined as*

$$\delta_{\mathcal{X}}(\tau) = \inf\left\{1 - \tfrac{1}{2}\|\mathbf{z}+\mathbf{w}\|_{\mathcal{X}} : \|\mathbf{z}\|_{\mathcal{X}} = \|\mathbf{w}\|_{\mathcal{X}} = 1, \|\mathbf{z}-\mathbf{w}\|_{\mathcal{X}} \ge \tau\right\},$$

is the modulus of convexity *of $\mathcal{X}$. A space $\mathcal{X}$ is said to be* uniformly convex *if $\delta_{\mathcal{X}}(\tau) > 0$ for all $\tau \in (0,2]$, and* p-convex, *for $p > 1$, if $\delta_{\mathcal{X}}(\tau) \ge K_p\tau^p$ for some $K_p > 0$ and all $\tau \in (0,2]$. The function $\rho_{\mathcal{X}}: [0,\infty) \to [0,\infty)$, defined as*

$$\rho_{\mathcal{X}}(\tau) = \sup\left\{\frac{\|\mathbf{z}+\tau\mathbf{w}\|_{\mathcal{X}} + \|\mathbf{z}-\tau\mathbf{w}\|_{\mathcal{X}}}{2} - 1 : \|\mathbf{z}\|_{\mathcal{X}} = \|\mathbf{w}\|_{\mathcal{X}} = 1\right\},$$

is the modulus of smoothness *of $\mathcal{X}$, and is a convex and continuous function such that $\frac{\rho_{\mathcal{X}}(\tau)}{\tau}$ is a nondecreasing function with $\rho_{\mathcal{X}}(\tau) \le \tau$. $\mathcal{X}$ is said to be* uniformly smooth *if $\lim_{\tau \searrow 0} \frac{\rho_{\mathcal{X}}(\tau)}{\tau} = 0$, and* p-smooth, *for $p > 1$, if $\rho_{\mathcal{X}}(\tau) \le K_p \tau^p$ for some $K_p > 0$ and all $\tau \in (0, \infty)$.*

The following relationships between Banach spaces and duality maps will be used extensively.

Theorem 2.3 ([46, Theorems 2.52 and 2.53 and Lemma 5.16]).

(i) *For every $\mathbf{x} \in \mathcal{X}$, the set $\mathcal{J}_p^{\mathcal{X}}(\mathbf{x})$ is nonempty, convex, and weakly-$\star$ closed in $\mathcal{X}^*$.*
(ii) *$\mathcal{X}$ is p-smooth if and only if $\mathcal{X}^*$ is p^*-convex. $\mathcal{X}$ is p-convex if and only if $\mathcal{X}^*$ is p^*-smooth.*
(iii) *$\mathcal{X}$ is smooth if and only if $\mathcal{J}_p^{\mathcal{X}}$ is single valued. If $\mathcal{X}$ is convex of power type and smooth, then $\mathcal{J}_p^{\mathcal{X}}$ is invertible and $(\mathcal{J}_p^{\mathcal{X}})^{-1} = \mathcal{J}_{p^*}^{\mathcal{X}^*}$. If $\mathcal{X}$ is uniformly smooth and uniformly convex, then $\mathcal{J}_p^{\mathcal{X}}$ and $\mathcal{J}_{p^*}^{\mathcal{X}^*}$ are both uniformly continuous.*
(iv) *Let $\mathcal{X}$ be a uniformly smooth Banach space with duality map $\mathcal{J}_p^{\mathcal{X}}$ with $p \ge 2$. Then, for all $\mathbf{x}, \widetilde{\mathbf{x}} \in \mathcal{X}$, there holds*

$$\|\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}) - \mathcal{J}_p^{\mathcal{X}}(\widetilde{\mathbf{x}})\|_{\mathcal{X}^*}^{p^*} \le C \max\{1, \|\mathbf{x}\|_{\mathcal{X}}, \|\widetilde{\mathbf{x}}\|_{\mathcal{X}}\}^p \overline{\rho}_{\mathcal{X}}(\|\mathbf{x} - \widetilde{\mathbf{x}}\|_{\mathcal{X}})^{p^*},$$

where $\overline{\rho}_{\mathcal{X}}(\tau) = \rho_{\mathcal{X}}(\tau)/\tau$ is a modulus of smoothness function such that $\overline{\rho}(\tau) \le 1$.

Next we list some common Banach spaces, the corresponding duality maps, and convexity and smoothness properties.

Example 2.4.

(i) A Hilbert space $\mathcal{X}$ is 2-smooth and 2-convex, and $\mathcal{J}_2^{\mathcal{X}}$ is the identity.
(ii) If $\mathcal{X}$ is smooth, then $\mathcal{J}_p^{\mathcal{X}}$ is the Gateaux derivative of the functional $\mathbf{x} \mapsto \frac{1}{p}\|\mathbf{x}\|_{\mathcal{X}}^p$.
(iii) If $\mathcal{X} = \ell^r(\mathbb{R})$ with $1 < r < \infty$, then $\mathcal{J}_p^{\mathcal{X}}$ is single-valued, and the duality map is given by $\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}) = \|\mathbf{x}\|_r^{p-r} |\mathbf{x}|^{r-1} \operatorname{sign}(\mathbf{x})$. Moreover, $\mathcal{J}_p^{\mathcal{X}} = \nabla(\frac{1}{p}\|\cdot\|_{\mathcal{X}}^p)$ since $\mathcal{X}$ is smooth.
(iv) Lebesgue spaces $\mathcal{L}^p(\Omega)$, Sobolev spaces $W^{s,p}(\Omega)$, with $s > 0$ (for an open bounded domain Ω), and sequence spaces $\ell^p(\mathbb{R})$ are $p \wedge 2$-smooth and $p \vee 2$-convex for $1 < p < \infty$. For $p \in \{1, \infty\}$, they are neither smooth nor strictly convex.

2.2. Bregman distance. Due to the geometry of Banach spaces, it is often more convenient to use the Bregman distance than the standard Banach space norm $\|\cdot\|_{\mathcal{X}}$ in the convergence analysis.

Definition 2.5 (Bregman distance). *For a smooth Banach space $\mathcal{X}$, the functional*

$$\mathbf{B}_p(\mathbf{z}, \mathbf{w}) = \frac{1}{p^*}\|\mathbf{z}\|_{\mathcal{X}}^p + \frac{1}{p}\|\mathbf{w}\|_{\mathcal{X}}^p - \langle \mathcal{J}_p^{\mathcal{X}}(\mathbf{z}), \mathbf{w} \rangle$$

is called the Bregman distance, *where $1/p + 1/p^* = 1$.*

Note that the dependence of the Bregman distance $\mathbf{B}_p(\mathbf{z}, \mathbf{w})$ on the space $\mathcal{X}$ is omitted, which is often clear from the context. The Bregman distance does not satisfy the triangle inequality and is generally non-symmetric. Thus, it is not a norm. The next theorem lists useful properties of the Bregman distance which show the relationship between the geometry of the underlying Banach space and duality maps.

Theorem 2.6 ([46, Theorem 2.60, Lemmas 2.62 and 2.63]). *The following properties hold:*

(i) *If $\mathcal{X}$ is smooth and reflexive, then $\mathbf{B}_p(\mathbf{z},\mathbf{w}) = \mathbf{B}_{p^*}(\mathcal{J}_p^{\mathcal{X}}(\mathbf{w}), \mathcal{J}_p^{\mathcal{X}}(\mathbf{z}))$.*

(ii) *Bregman distance satisfies the three-point identity*

$$\mathbf{B}_p(\mathbf{z},\mathbf{w}) = \mathbf{B}_p(\mathbf{z},\mathbf{v}) + \mathbf{B}_p(\mathbf{v},\mathbf{w}) + \left\langle \mathcal{J}_p^{\mathcal{X}}(\mathbf{v}) - \mathcal{J}_p^{\mathcal{X}}(\mathbf{z}), \mathbf{w}-\mathbf{v}\right\rangle. \tag{2.2}$$

(iii) *If $\mathcal{X}$ is p-convex, then it is reflexive, $p \geq 2$, and there exists $C_p > 0$ such that*

$$\mathbf{B}_p(\mathbf{z},\mathbf{w}) \geq p^{-1} C_p \|\mathbf{w}-\mathbf{z}\|_{\mathcal{X}}^p. \tag{2.3}$$

(iv) *If $\mathcal{X}^*$ is p^*-smooth, then it is reflexive, $p^* \leq 2$, and there exists $G_{p^*} > 0$ such that*

$$\mathbf{B}_{p^*}(\mathbf{z}^*,\mathbf{w}^*) \leq (p^*)^{-1} G_{p^*} \|\mathbf{w}^*-\mathbf{z}^*\|_{\mathcal{X}^*}^{p^*}. \tag{2.4}$$

(v) *$\mathbf{B}_p(\mathbf{z},\mathbf{w}) \geq 0$, and if $\mathcal{X}$ is uniformly convex, we have $\mathbf{B}_p(\mathbf{z},\mathbf{w}) = 0$ if and only if $\mathbf{z} = \mathbf{w}$.*

(vi) *$\mathbf{B}_p(\mathbf{z},\mathbf{w})$ is continuous in the second argument. If $\mathcal{X}$ is smooth and uniformly convex, then $\mathcal{J}_p^{\mathcal{X}}$ is continuous on bounded subsets and $\mathbf{B}_p(\mathbf{z},\mathbf{w})$ is continuous in its first argument.*

3. Convergence analysis for exact data.

Now we develop an SGD-type approach for problem (1.1) and analyze its convergence. Throughout, we make the following assumption on the Banach spaces $\mathcal{X}$ and $\mathcal{Y}$, unless indicated otherwise.

Assumption 3.1. The Banach space $\mathcal{X}$ is p-convex and smooth, and $\mathcal{Y}$ is arbitrary.

To recover the solution $\mathbf{x}^\dagger$, we minimize a least-squares-type objective $\operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} \frac{1}{p}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_{\mathcal{Y}}^p$ for some $p > 1$. By $\mathcal{X}_{\min}$, we denote the (nonempty) set of minimizers over $\mathcal{X}$. Among the elements of $\mathcal{X}_{\min}$, the regularization theory focuses on the so-called minimum norm solution.

Definition 3.2. *An element $\mathbf{x}^\dagger \in \mathcal{X}$ is called a* minimum norm solution *(MNS) of (1.1) if*

$$\mathbf{A}\mathbf{x}^\dagger = \mathbf{y} \quad \text{and} \quad \|\mathbf{x}^\dagger\|_{\mathcal{X}} = \inf\{\|\mathbf{x}\|_{\mathcal{X}} : \mathbf{x}\in\mathcal{X},\, \mathbf{A}\mathbf{x} = \mathbf{y}\}.$$

The MNS $\mathbf{x}^\dagger$ is not unique for general Banach spaces. The following lemma states sufficient geometric assumptions on $\mathcal{X}$ for uniqueness.

Lemma 3.3 ([46, Lemma 3.3]). *Let Assumption 3.1 hold. Then there exists a unique MNS $\mathbf{x}^\dagger$. Furthermore, $\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}^\dagger) \in \overline{\operatorname{range}(\mathbf{A}^*)}$ for $1 < p < \infty$. If some $\widehat{\mathbf{x}} \in \mathcal{X}$ satisfies $\mathcal{J}_p^{\mathcal{X}}(\widehat{\mathbf{x}}) \in \overline{\operatorname{range}(\mathbf{A}^*)}$ and $\widehat{\mathbf{x}} - \mathbf{x}^\dagger \in \operatorname{null}(\mathbf{A})$, then $\widehat{\mathbf{x}} = \mathbf{x}^\dagger$.*

By Lemma 3.3, the MNS $\mathbf{x}^\dagger$ is unique modulo the null space of $\mathbf{A}$, under certain smoothness and convexity assumptions on $\mathcal{X}$. These conditions exclude, for example, Lebesgue space $\mathcal{L}^1(\Omega)$ and and sequence space $\ell^1(\mathbb{R})$; cf. Example 2.4(iv). The standard Landweber method [33, 44] constructs an approximation to the MNS $\mathbf{x}^\dagger$ by running the iterations

$$\mathbf{x}_{k+1} = \mathcal{J}_{p^*}^{\mathcal{X}^*}\left(\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k) - \mu_{k+1}\mathbf{A}^* \jmath_p^{\mathcal{Y}}(\mathbf{A}\mathbf{x}_k - \mathbf{y})\right), \quad k = 0,1,\ldots, \tag{3.1}$$

where $\mu_{k+1} > 0$ is the step-size. Asplund's theorem [46, Theorem 2.28] allows for characterizing the duality map as the subdifferential, $\mathcal{J}_p^{\mathcal{X}} = \partial(\frac{1}{p}\|\cdot\|_{\mathcal{X}}^p)$ for $p > 1$. This identifies

the descent direction $\mathbf{A}^* \jmath_p^{\mathcal{Y}}(\mathbf{A}\mathbf{x}_k - \mathbf{y})$ as the subgradient of the objective: $\mathbf{A}^* \jmath_p^{\mathcal{Y}}(\mathbf{A}\mathbf{x}_k - \mathbf{y}) = \partial(\frac{1}{p}\|\mathbf{A}\cdot - \mathbf{y}\|_{\mathcal{Y}})(\mathbf{x}_k)$. Note that $\mathcal{J}_p^{\mathcal{X}}$ is single-valued by Assumption 3.1 and Theorem 2.3, though $\mathcal{J}_p^{\mathcal{Y}}$ is not. For well-selected step-sizes, Landweber iterations (3.1) converge to an MNS of (1.1) [44, Theorem 3.3].

The evaluation of the subgradient $\mathbf{A}^* \jmath_p^{\mathcal{Y}}(\mathbf{A}\mathbf{x}_k - \mathbf{y})$ represents the main per-iteration cost of the iteration (3.1). In this work, we consider the following Kaczmarz-type setting:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_N \end{pmatrix} \quad \text{and} \quad \mathbf{A}\mathbf{x} = \begin{pmatrix} \mathbf{A}_1\mathbf{x} \\ \vdots \\ \mathbf{A}_N\mathbf{x} \end{pmatrix} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{pmatrix}, \tag{3.2}$$

where $\mathbf{A}_i : \mathcal{X} \to \mathcal{Y}_i$, $\mathbf{y}_i \in \mathcal{Y}_i$, for $i \in [N] = \{1, \dots, N\}$. Problem (3.2) is defined on the direct product $(\otimes_{i=1}^N \mathcal{Y}_i, \ell^r)$, equipped with the ℓ^r-norm, for $r \geq 1$,

$$\|\mathbf{y}\|_{\mathcal{Y}} := \|(\mathbf{y}_1, \dots, \mathbf{y}_N)\|_{\mathcal{Y}} = \|(\|\mathbf{y}_1\|_{\mathcal{Y}_1}, \dots, \|\mathbf{y}_N\|_{\mathcal{Y}_N})\|_{\ell^r} = \Big(\sum_{i=1}^N \|\mathbf{y}_i\|_{\mathcal{Y}_i}^r\Big)^{1/r}. \tag{3.3}$$

Below we identify $\mathcal{Y}_i = \mathcal{Y}$ for notational brevity and use $\|\cdot\|_{\mathcal{Y}}$ to denote both the norm of the direct product space and the component spaces, though all the relevant proofs and concepts easily extend to the general case. Then the objective $\Psi(\mathbf{x})$ is given by

$$\Psi(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^N \Psi_i(\mathbf{x}), \quad \text{with } \Psi_i(\mathbf{x}) = \frac{1}{p}\|\mathbf{A}_i\mathbf{x} - \mathbf{y}_i\|_{\mathcal{Y}}^p.$$

Note that for many common imaging problems we use $\mathcal{Y} = \ell^p(\mathbb{R})$, which then naturally gives $\Psi(\mathbf{x}) = \frac{1}{pN}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_{\mathcal{Y}}^p$. To reduce the computational cost per iteration, we exploit the finite-sum structure of the objective $\Psi(\mathbf{x})$ and adopt SGD iterations of the form

$$\mathbf{x}_{k+1} = \mathcal{J}_{p^*}^{\mathcal{X}^*}\left(\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k) - \mu_{k+1}\mathbf{g}_{k+1}\right), \tag{3.4}$$

where $\mathbf{g}_{k+1} = g(\mathbf{x}_k, \mathbf{y}, i_{k+1})$ is the stochastic update direction given by

$$g(\mathbf{x}, \mathbf{y}, i) = \mathbf{A}_i^* \jmath_p^{\mathcal{Y}}(\mathbf{A}_i\mathbf{x} - \mathbf{y}_i) = \partial\Big(\tfrac{1}{p}\|\mathbf{A}_i\cdot - \mathbf{y}_i\|_{\mathcal{Y}}^p\Big)(\mathbf{x}), \tag{3.5}$$

and the random index i_k is sampled uniformly over the index set $[N]$, independent of $\mathbf{x}_k$. Clearly, it is an unbiased estimator of the subgradient $\partial\Psi(\mathbf{x})$, i.e., $\mathbb{E}[g(\mathbf{x}, \mathbf{y}, i)] = \partial\Psi(\mathbf{x})$, and the per-iteration cost is reduced by a factor of N.

Remark 3.4. In the model (3.2), if $\mathcal{Y}$ admits a complemented sum $\mathcal{Y} = \sum_{i=1}^N \mathcal{Y}_i$, we can take the (internal) direct sum $(\oplus_{i=1}^N \mathcal{Y}_i, \ell^r)$, so that $\mathbf{y} = \mathbf{y}_1 + \cdots + \mathbf{y}_N$ and the corresponding norm $\|\mathbf{y}\| = \|(\|\mathrm{Proj}_{\mathcal{Y}_1}(\mathbf{y})\|_{\mathcal{Y}_1}, \dots, \|\mathrm{Proj}_{\mathcal{Y}_N}(\mathbf{y})\|_{\mathcal{Y}_N})\|_r$. With this identification the spaces $(\otimes_{i=1}^N \mathcal{Y}_i, \ell^r)$ and $(\oplus_{i=1}^N \mathcal{Y}_i, \ell^r)$ are isometrically isomorphic [48] and the norms are equivalent for all $r \geq 1$.

We now collect some useful properties about the objective Ψ and the Bregman divergence. Throughout, $L_{\max} = \max_{i\in[N]} \|\mathbf{A}_i\|$. Note that $c_N = 1/N$ if $\mathcal{Y} = \mathcal{L}^p(\Omega)$ or $\ell^p(\mathbb{R})$.

Lemma 3.5. *For all $i \in [N]$, $\mathbf{x} \in \mathcal{X}$, and any $\widehat{\mathbf{x}} \in \mathcal{X}_{\min}$ (such that $\mathbf{A}\widehat{\mathbf{x}} = \mathbf{y}$), we have*

$$\langle \partial \Psi_i(\mathbf{x}), \mathbf{x} - \widehat{\mathbf{x}} \rangle = p\Psi_i(\mathbf{x}) \quad \text{and} \quad \langle \partial \Psi(\mathbf{x}), \mathbf{x} - \widehat{\mathbf{x}} \rangle = p\Psi(\mathbf{x}). \tag{3.6}$$

Moreover, $\Psi_i(\mathbf{x}) \leq \frac{\|\mathbf{A}_i\|^p}{C_p} \mathbf{B}_p(\mathbf{x}, \widehat{\mathbf{x}})$, $\Psi(\mathbf{x}) \leq \frac{L_{\max}^p}{C_p} \mathbf{B}_p(\mathbf{x}, \widehat{\mathbf{x}})$, and for some $C_N > 0$ we have $\Psi(\mathbf{x}) \geq \frac{C_N}{p} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_{\mathcal{Y}}^p$.

Proof. It follows from the identity $\mathbf{A}\widehat{\mathbf{x}} = \mathbf{y}$ that

$$\langle \partial \Psi_i(\mathbf{x}), \mathbf{x} - \widehat{\mathbf{x}} \rangle = \langle \mathbf{A}_i^* j_p^{\mathcal{Y}}(\mathbf{A}_i \mathbf{x} - \mathbf{y}_i), \mathbf{x} - \widehat{\mathbf{x}} \rangle = \langle j_p^{\mathcal{Y}}(\mathbf{A}_i \mathbf{x} - \mathbf{y}_i), \mathbf{A}_i \mathbf{x} - \mathbf{y}_i \rangle = p\Psi_i(\mathbf{x}).$$

Since $\partial \Psi(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N \partial \Psi_i(\mathbf{x})$, the second identity in (3.6) follows from the linearity of the dual product. By the p-convexity of the space $\mathcal{X}$ and Theorem 2.6(iii), we get

$$\Psi_i(\mathbf{x}) = \frac{1}{p} \|\mathbf{A}_i \mathbf{x} - \mathbf{y}_i\|_{\mathcal{Y}}^p = \frac{1}{p} \|\mathbf{A}_i(\mathbf{x} - \widehat{\mathbf{x}})\|_{\mathcal{Y}}^p \leq \frac{\|\mathbf{A}_i\|^p}{p} \|\mathbf{x} - \widehat{\mathbf{x}}\|_{\mathcal{X}}^p \leq \frac{\|\mathbf{A}_i\|^p}{C_p} \mathbf{B}_p(\mathbf{x}, \widehat{\mathbf{x}}).$$

The second claim follows since $\Psi(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N \Psi_i(\mathbf{x})$. Lastly, by the norm equivalence (3.3) for $1 < r < \infty$, there exists $C_N > 0$ such that

$$\Psi(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N \frac{1}{p} \|\mathbf{A}_i \mathbf{x} - \mathbf{y}_i\|_{\mathcal{Y}}^p \geq \frac{C_N}{p} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_{\mathcal{Y}}^p. \qquad \blacksquare$$

We now focus on the convergence study of the iterations (3.4), without and with noise in the data, and discuss convergence rates under conditional stability.

3.1. Convergence for the Kaczmarz model. Below, the notation $\mathbb{E}[\cdot]$ denotes taking expectation with respect to the sampling of the random indices i_k, and $\mathbb{E}_k[\cdot]$ denotes taking conditional expectation with respect to $\mathcal{F}_k$. The remaining variables, e.g., $\mathbf{x}$ and $\mathbf{y}$, are measurable with respect to the underlying probability measure. To study the convergence of SGD (3.4), we first establish a descent property in terms of the Bregman distance.

Lemma 3.6. *Let Assumption 3.1 hold. For any $\widehat{\mathbf{x}} \in \mathcal{X}$, the iterates in (3.4) satisfy*

$$\mathbf{B}_p(\mathbf{x}_{k+1}, \widehat{\mathbf{x}}) \leq \mathbf{B}_p(\mathbf{x}_k, \widehat{\mathbf{x}}) - \mu_{k+1} \langle \mathbf{g}_{k+1}, \mathbf{x}_k - \widehat{\mathbf{x}} \rangle + \frac{G_{p^*}}{p^*} \mu_{k+1}^{p^*} \|\mathbf{g}_{k+1}\|_{\mathcal{X}^*}^{p^*}. \tag{3.7}$$

Proof. Let $\Delta_k := \mathbf{B}_p(\mathbf{x}_k, \widehat{\mathbf{x}})$. By Definition 2.5 and expression (3.4), we have

$$\begin{aligned} \Delta_{k+1} &= \frac{1}{p} \|\widehat{\mathbf{x}}\|_{\mathcal{X}}^p + \frac{1}{p^*} \|\mathbf{x}_{k+1}\|_{\mathcal{X}}^p - \langle \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_{k+1}), \widehat{\mathbf{x}} \rangle \\ &= \frac{1}{p} \|\widehat{\mathbf{x}}\|_{\mathcal{X}}^p + \frac{1}{p^*} \|\mathcal{J}_{p^*}^{\mathcal{X}^*}\left(\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k) - \mu_{k+1} \mathbf{g}_{k+1}\right)\|_{\mathcal{X}}^p - \langle \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_{k+1}), \widehat{\mathbf{x}} \rangle. \end{aligned}$$

Using Definition 2.1, the identity $p(p^* - 1) = p^*$, and Theorem 2.3(iii), we deduce

$$\begin{aligned} \Delta_{k+1} &= \frac{1}{p} \|\widehat{\mathbf{x}}\|_{\mathcal{X}}^p + \frac{1}{p^*} \|\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k) - \mu_{k+1} \mathbf{g}_{k+1}\|_{\mathcal{X}^*}^{p(p^*-1)} - \langle \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k) - \mu_{k+1} \mathbf{g}_{k+1}, \widehat{\mathbf{x}} \rangle \\ &= \frac{1}{p} \|\widehat{\mathbf{x}}\|_{\mathcal{X}}^p + \frac{1}{p^*} \|\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k) - \mu_{k+1} \mathbf{g}_{k+1}\|_{\mathcal{X}^*}^{p^*} - \langle \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k) - \mu_{k+1} \mathbf{g}_{k+1}, \widehat{\mathbf{x}} \rangle. \end{aligned}$$

Since $\mathcal{X}$ is p-convex, $\mathcal{X}^*$ is p^*-smooth; cf. Theorem 2.3(i). By [9, Corollary 5.8], this implies

$$\frac{1}{p^*}\|\mathbf{x}^* - \tilde{\mathbf{x}}^*\|_{\mathcal{X}^*}^{p^*} \leq \frac{1}{p^*}\|\mathbf{x}^*\|_{\mathcal{X}^*}^{p^*} + \frac{G_{p^*}}{p^*}\|\tilde{\mathbf{x}}^*\|_{\mathcal{X}^*}^{p^*} - \left\langle \mathcal{J}_{p^*}^{\mathcal{X}^*}(\mathbf{x}^*), \tilde{\mathbf{x}}^* \right\rangle \quad \forall \mathbf{x}^*, \tilde{\mathbf{x}}^* \in \mathcal{X}^*.$$

Using identities $p^*(p-1) = p$ and $(\mathcal{J}_p^{\mathcal{X}})^{-1} = \mathcal{J}_{p^*}^{\mathcal{X}^*}$ (cf. Theorem 2.3(iii)), we get

$$\begin{aligned}\frac{1}{p^*}\|\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k) - \mu_{k+1}\mathbf{g}_{k+1}\|_{\mathcal{X}^*}^{p^*} &\leq \frac{1}{p^*}\|\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k)\|_{\mathcal{X}^*}^{p^*} + \frac{G_{p^*}}{p^*}\|\mu_{k+1}\mathbf{g}_{k+1}\|_{\mathcal{X}^*}^{p^*} - \left\langle \mu_{k+1}\mathbf{g}_{k+1}, \mathbf{x}_k \right\rangle \\ &= \frac{1}{p^*}\|\mathbf{x}_k\|_{\mathcal{X}}^{p} + \frac{G_{p^*}}{p^*}\mu_{k+1}^{p^*}\|\mathbf{g}_{k+1}\|_{\mathcal{X}^*}^{p^*} - \mu_{k+1}\left\langle \mathbf{g}_{k+1}, \mathbf{x}_k \right\rangle.\end{aligned}$$

Combining the preceding estimates gives the desired assertion through

$$\begin{aligned}\Delta_{k+1} &\leq \frac{1}{p}\|\widehat{\mathbf{x}}\|_{\mathcal{X}}^{p} + \frac{1}{p^*}\|\mathbf{x}_k\|_{\mathcal{X}^*}^{p} - \left\langle \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k), \widehat{\mathbf{x}} \right\rangle + \frac{G_{p^*}}{p^*}\mu_{k+1}^{p^*}\|\mathbf{g}_{k+1}\|_{\mathcal{X}^*}^{p^*} - \mu_{k+1}\left\langle \mathbf{g}_{k+1}, \mathbf{x}_k - \widehat{\mathbf{x}} \right\rangle \\ &= \Delta_k - \mu_{k+1}\left\langle \mathbf{g}_{k+1}, \mathbf{x}_k - \widehat{\mathbf{x}} \right\rangle + \frac{G_{p^*}}{p^*}\mu_{k+1}^{p^*}\|\mathbf{g}_{k+1}\|_{\mathcal{X}^*}^{p^*}.\end{aligned}$$

■

Lemma 3.6 allows showing that the sequence of Bregman distances $(\mathbf{B}_p(\mathbf{x}_k, \widehat{\mathbf{x}}))_{k\in\mathbb{N}}$ forms an almost supermartingale (in the Robbins–Siegmund sense defined below) for $\widehat{\mathbf{x}} \in \mathcal{X}_{\min}$ and well-chosen step-sizes $(\mu_k)_{k\in\mathbb{N}}$. We will show almost sure convergence of the iterates using the Robbins–Siegmund theorem.

Theorem 3.7 (Robbins–Siegmund theorem on the convergence of almost supermartingales [38, Lemma 11]). *Consider a filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$ and four nonnegative, $(\mathcal{F}_k)_{k\in\mathbb{N}}$ adapted processes $(\alpha_k)_{k\in\mathbb{N}}$, $(\beta_k)_{k\in\mathbb{N}}$, $(\gamma_k)_{k\in\mathbb{N}}$, and $(\delta_k)_{k\in\mathbb{N}}$. Let $(\alpha_k)_{k\in\mathbb{N}}$ be an* almost supermartingale, *i.e., for all k we have $\mathbb{E}_k[\alpha_{k+1}] \leq (1+\beta_k)\alpha_k + \gamma_k - \delta_k$. Then the sequence $(\alpha_k)_{k\in\mathbb{N}}$ converges a.s. to a random variable α_∞, and $\sum_{k=1}^{\infty}\delta_k < \infty$ a.s. on the set $\{\sum_{k=1}^{\infty}\beta_k < \infty, \sum_{k=1}^{\infty}\gamma_k < \infty\}$.*

We can now show that under certain conditions on $\mathbf{x}_0$, the limit of the iterations (3.4) is the MNS $\mathbf{x}^\dagger$. Below, $\mathbb{E}_k$ denotes the conditional expectation with respect to the filtration $\mathcal{F}_k$.

Theorem 3.8. *Let $(\mu_k)_{k\in\mathbb{N}}$ satisfy $\sum_{k=1}^{\infty}\mu_k = \infty$ and $\sum_{k=1}^{\infty}\mu_k^{p^*} < \infty$, let Assumption 3.1 hold, and let $\mathbf{x}^\dagger$ be the MNS. Then the sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ converges a.s. to a solution of (1.1):*

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

Moreover, if $\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_0) \in \overline{\operatorname{range}(\mathbf{A}^*)}$, we have $\lim_{k\to\infty}\mathbf{B}_p(\mathbf{x}_k, \mathbf{x}^\dagger) = 0$ a.s.

Proof. By Lemma 3.5, we have $\langle \partial\Psi(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^\dagger\rangle = p\Psi(\mathbf{x}_k)$. Moreover,

$$\|g(\mathbf{x}, \mathbf{y}, i)\|_{\mathcal{X}^*} = \|\mathbf{A}_i^* j_p^{\mathcal{Y}}(\mathbf{A}_i\mathbf{x} - \mathbf{y}_i)\|_{\mathcal{X}^*} \leq \|\mathbf{A}_i\|\|j_p^{\mathcal{Y}}(\mathbf{A}_i\mathbf{x} - \mathbf{y}_i)\|_{\mathcal{Y}^*} \leq L_{\max}\|\mathbf{A}_i\mathbf{x} - \mathbf{y}_i\|_{\mathcal{Y}}^{p-1},$$

with $L_{\max} = \max_{i\in[N]}\|\mathbf{A}_i\|$. Thus, since $p^*(p-1) = p$, we have

$$\mathbb{E}[\|g(\mathbf{x}, \mathbf{y}, i)\|_{\mathcal{X}^*}^{p^*}] \leq pL_{\max}^{p^*}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{p}\|\mathbf{A}_i\mathbf{x} - \mathbf{y}_i\|_{\mathcal{Y}}^{p} = pL_{\max}^{p^*}\Psi(\mathbf{x}).$$

Upon taking the conditional expectation $\mathbb{E}_k[\cdot]$ of the descent property (3.7) (with $\widehat{\mathbf{x}} = \mathbf{x}^\dagger$), and using the measurability of $\mathbf{x}_k$ with respect to $\mathcal{F}_k$, we deduce

$$\mathbb{E}_k[\Delta_{k+1}] \leq \Delta_k - p\mu_{k+1}\Psi(\mathbf{x}_k) + pL_{\max}^{p^*}\frac{G_{p^*}}{p^*}\mu_{k+1}^{p^*}\Psi(\mathbf{x}_k).$$

Using Lemma 3.5 again we have $\Psi(\mathbf{x}_k) \leq \frac{L_{\max}^p}{C_p}\Delta_k$, which yields

$$\mathbb{E}_k[\Delta_{k+1}] \leq \left(1 + L_{\max}^{p^*+p}\frac{p}{C_p}\frac{G_{p^*}}{p^*}\mu_{k+1}^{p^*}\right)\Delta_k - p\mu_{k+1}\Psi(\mathbf{x}_k).$$

Since $\sum_{k=1}^\infty \mu_k^{p^*} < \infty$ by assumption, we can apply Theorem 3.7 and deduce that the sequence $(\Delta_k)_{k\in\mathbb{N}}$ converges a.s. to a random variable Δ_∞ and $\sum_{k=0}^\infty \mu_{k+1}\Psi(\mathbf{x}_k) < \infty$ a.s. Let Ω be the measurable set on which $(\Delta_k)_{k\in\mathbb{N}}$ converges, $\sum_{k=0}^\infty \mu_{k+1}\Psi(\mathbf{x}_k) < \infty$, and $\mathbb{P}(\Omega) = 1$. Next we show $\liminf_k \Psi(\mathbf{x}_k) = 0$ a.s. Consider an event ω on which this is not the case, i.e., where $\liminf_k \Psi(\mathbf{x}_k) > 0$. Then there exist $\epsilon > 0$ and $k_\epsilon \in \mathbb{N}$ such that for all $k \geq k_\epsilon$, $\Psi(\mathbf{x}_k) \geq \epsilon$, giving $\sum_{k\geq k_\epsilon}\mu_{k+1}\Psi(\mathbf{x}_k) \geq \epsilon\sum_{k\geq k_\epsilon}\mu_{k+1}$. If ω were in Ω Since for all events in Ω this would lead to a contradiction—the right-hand side diverges ($\sum_{k=1}^\infty \mu_k = \infty$ by assumption), whereas the left-hand side is the remainder of a convergent series. Thus, we conclude $\omega \notin \Omega$. Since $\mathbb{P}(\Omega^c) = 0$, we have $\liminf_k \Psi(\mathbf{x}_k) = 0$ a.s. For every event in the set where $\liminf_k \Psi(\mathbf{x}_k) = 0$ holds we can then find a subsequence $(\mathbf{x}_{n_k})_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty}\Psi(\mathbf{x}_{n_k}) = 0$. Define also $\widehat{\Psi}(\mathbf{x}) = \sum_{i=1}^N \widehat{\Psi}_i(\mathbf{x})$, with $\widehat{\Psi}_i(\mathbf{x}) = \|\mathbf{A}_i\mathbf{x} - \mathbf{y}_i\|_{\mathcal{Y}}$. We have $\liminf_k \widehat{\Psi}(\mathbf{x}_k) = 0$ and $\lim_{j\to\infty}\widehat{\Psi}(\mathbf{x}_{n_j}) = 0$ (on the same subsequence), since by Young's inequality,

$$\left(\sum_{i=1}^N \|\mathbf{A}_i\mathbf{x} - \mathbf{y}_i\|_{\mathcal{Y}}^p\right)^{1/p} \leq \sum_{i=1}^N \|\mathbf{A}_i\mathbf{x} - \mathbf{y}_i\|_{\mathcal{Y}} \leq N\left(\frac{1}{N}\sum_{i=1}^N \|\mathbf{A}_i\mathbf{x} - \mathbf{y}_i\|_{\mathcal{Y}}^p\right)^{1/p}.$$

Moreover, $\widehat{\Psi}(\mathbf{x})^p \leq pN^p\Psi(\mathbf{x})$. The following argument is understood pointwise on the a.s. set Ω where $(\Delta_k)_{k\in\mathbb{N}}$ converges, $\sum_{k=0}^\infty \mu_{k+1}\Psi(\mathbf{x}_k) < \infty$, and $\liminf_k \widehat{\Psi}(\mathbf{x}_k) = 0$. Since $(\Delta_k)_{k\in\mathbb{N}}$ converges it is bounded. By the coercivity of the Bregman distance (see Lemma A.3), so are $(\mathbf{x}_k)_{k\in\mathbb{N}}$ and $(\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k))_{k\in\mathbb{N}}$. By further passing to a subsequence, we can find a subsequence of $(\mathbf{x}_{n_k})_{k\in\mathbb{N}}$, which we denote the same, such that $(\|\mathbf{x}_{n_k}\|_{\mathcal{X}})_{k\in\mathbb{N}}$ is convergent, $(\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_{n_k}))_{k\in\mathbb{N}}$ is weakly convergent, and

$$\lim_{k\to\infty}\widehat{\Psi}(\mathbf{x}_{n_k}) = 0 \quad \text{and} \quad \widehat{\Psi}(\mathbf{x}_{n_k}) \leq \widehat{\Psi}(\mathbf{x}_n) \quad \forall n < n_k. \tag{3.8}$$

The latter can be obtained by setting $n_1 = 1$, and then recursively defining $n_{k+1} = \min\{k > n_k : \Psi(\mathbf{x}_k) \leq \Psi(\mathbf{x}_{n_k})/2\}$ for $k \in \mathbb{N}$. Any following subsequence satisfies the same property. Using Theorem 2.6(ii), we have, for $k > \ell$,

$$\mathbf{B}_p(\mathbf{x}_{n_\ell}, \mathbf{x}_{n_k}) = \frac{1}{p^*}\left(\|\mathbf{x}_{n_\ell}\|_{\mathcal{X}}^p - \|\mathbf{x}_{n_k}\|_{\mathcal{X}}^p\right) + \left\langle \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_{n_k}) - \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_{n_\ell}), \mathbf{x}^\dagger\right\rangle + \left\langle \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_{n_k}) - \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_{n_\ell}), \mathbf{x}_{n_k} - \mathbf{x}^\dagger\right\rangle.$$

Since the first two terms involve Cauchy sequences, it suffices to treat the last term, denoted by $\mathrm{I}_{k,\ell}$. Using telescopic sum and applying the iterate update rule, we have

$$\mathrm{I}_{k,\ell} = \sum_{n=n_\ell}^{n_k-1}\left\langle \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_{n+1}) - \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_n), \mathbf{x}_{n_k} - \mathbf{x}^\dagger\right\rangle = \sum_{n=n_\ell}^{n_k-1}\mu_{n+1}\left\langle \mathbf{A}_{i_{n+1}}^* J_p^{\mathcal{Y}}(\mathbf{A}_{i_{k+1}}\mathbf{x}_n - \mathbf{y}_{i_{n+1}}), \mathbf{x}_{n_k} - \mathbf{x}^\dagger\right\rangle$$

$$= \sum_{n=n_\ell}^{n_k-1} \mu_{n+1} \left\langle J_p^{\mathcal{Y}}(\mathbf{A}_{i_{n+1}}\mathbf{x}_n - \mathbf{y}_{i_{k+1}}), \mathbf{A}_{i_{n+1}}\mathbf{x}_{n_k} - \mathbf{y}_{i_{n+1}} \right\rangle.$$

By the Cauchy–Schwarz inequality and properties of the duality map, we get

$$|\mathrm{I}_{k,\ell}| \le \sum_{n=n_\ell}^{n_k-1} \mu_{n+1} \|\mathbf{A}_{i_{n+1}}\mathbf{x}_n - \mathbf{y}_{i_{n+1}}\|_{\mathcal{Y}}^{p-1} \|\mathbf{A}_{i_{n+1}}\mathbf{x}_{n_k} - \mathbf{y}_{i_{n+1}}\|_{\mathcal{Y}} \le \sum_{n=n_\ell}^{n_k-1} \mu_{n+1} \widehat{\Psi}_{i_{n+1}}(\mathbf{x}_n)^{p-1} \widehat{\Psi}_{i_{n+1}}(\mathbf{x}_{n_k}).$$

Since $\widehat{\Psi}_i(\mathbf{x}) \le \widehat{\Psi}(\mathbf{x})$ for all $i \in [N]$, we use (3.8) and get

$$|\mathrm{I}_{k,\ell}| \le \sum_{n=n_\ell}^{n_k-1} \mu_{n+1} \widehat{\Psi}(\mathbf{x}_n)^{p-1} \widehat{\Psi}(\mathbf{x}_{n_k}) \le \sum_{n=n_\ell}^{n_k-1} \mu_{n+1} \widehat{\Psi}(\mathbf{x}_n)^p.$$

Since $\widehat{\Psi}(\mathbf{x})^p \le pN^p \Psi(\mathbf{x})$, the right-hand side of the inequality converges to 0 as $n_\ell \to \infty$. Therefore, by [44, Theorem 2.12(e)], it follows that $(\mathbf{x}_{n_k})_{k\in\mathbb{N}}$ is a Cauchy sequence, and thus converges strongly to an $\widehat{\mathbf{x}}$ such that $\Psi(\widehat{\mathbf{x}}) = 0$.

The above argument showing the a.s. convergence of $(\Delta_k)_{k\in\mathbb{N}}$ can be applied pointwise to any solution. Namely, on the event where $(\mathbf{x}_{n_k})_{k\in\mathbb{N}}$ converges strongly to an $\widehat{\mathbf{x}} \in \mathcal{X}_{\min}$ (i.e., $\mathbf{A}\widehat{\mathbf{x}} = \mathbf{y}$), define $\widehat{\Delta}_k := \mathbf{B}_p(\mathbf{x}_k, \widehat{\mathbf{x}})$. By repeating the argument using Lemma 3.5, we deduce

$$\widehat{\Delta}_{k+1} \le \left(1 + L_{\max}^{p^*+p} \frac{p}{C_p} \frac{G_{p^*}}{p^*} \mu_{k+1}^{p^*}\right) \widehat{\Delta}_k - p\mu_{k+1} \Psi_{i_k}(\mathbf{x}_k).$$

Since $\sum_{k=1}^\infty \mu_k^{p^*} < \infty$, it follows that the (deterministic) sequence $(\widehat{\Delta}_k)_{k\in\mathbb{N}}$ converges to a $\widehat{\Delta}_\infty \ge 0$. The continuity of the Bregman distance in the first argument (Theorem 2.6(vi)) gives $\lim_{j\to\infty} \mathbf{B}_p(\mathbf{x}_{n_j}, \widehat{\mathbf{x}}) = \mathbf{B}_p(\widehat{\mathbf{x}}, \widehat{\mathbf{x}}) = 0$, and thus $\widehat{\Delta}_\infty = 0$. Moreover, by the p-convexity of $\mathcal{X}$ (Theorem 2.6(iii)), we have $0 \le \|\mathbf{x}_k - \widehat{\mathbf{x}}\|_{\mathcal{X}}^p \le \frac{p}{C_p}\widehat{\Delta}_k$. From the squeeze theorem it follows that $\lim_{k\to\infty} \|\mathbf{x}_k - \widehat{\mathbf{x}}\|_{\mathcal{X}} = 0$. Thus, for every event in an a.s. set Ω, the sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ strongly converges to some minimizing solution, that is,

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

Next assume $\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_0) \in \overline{\mathrm{range}(\mathbf{A}^*)}$. From (3.4), it follows that $\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k) \in \overline{\mathrm{range}(\mathbf{A}^*)}$ holds for all $k \ge 1$. By the continuity of $\mathcal{J}_p^{\mathcal{X}}$, we have $\mathcal{J}_p^{\mathcal{X}}(\widehat{\mathbf{x}}) \in \overline{\mathrm{range}(\mathbf{A}^*)}$. Thus, from $\mathbf{A}(\widehat{\mathbf{x}} - \mathbf{x}^\dagger) = 0$ and Lemma 3.3 it follows that $\widehat{\mathbf{x}} = \mathbf{x}^\dagger$. ∎

The assumptionsand conclusions of Theorem 3.8 can be broken down into two parts. The step-size conditions $\sum_{k=1}^\infty \mu_k = \infty$ and $\sum_{k=1}^\infty \mu_k^{p^*} < \infty$ are required to show the a.s. convergence of $(\mathbf{B}_p(\mathbf{x}_k, \widehat{\mathbf{x}}))_{k\in\mathbb{N}}$ to 0 for some nondeterministic $\widehat{\mathbf{x}} \in \mathcal{X}_{\min}$. The remaining assumption $\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_0) \in \mathrm{range}(\mathbf{A}^*)$ is needed to identify this limit as the MNS $\mathbf{x}^\dagger$, as the Landweber method [44, Remark 3.12]. If $\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_0) \notin \overline{\mathrm{range}(\mathbf{A}^*)}$, we can commonly establish convergence to an MNS relative to $\mathbf{x}_0$, i.e., a solution which minimizes $\|\mathbf{x} - \mathbf{x}_0\|_{\mathcal{X}}$, analogous to the Euclidean case [24].

Remark 3.9. The step-size conditions $\sum_{k=1}^\infty \mu_k = \infty$ and $\sum_{k=1}^\infty \mu_k^{p^*} < \infty$ are satisfied by a polynomially decaying step-size schedule $(\mu_k)_{k\in\mathbb{N}} = (\mu_0 k^{-\beta})_{k\in\mathbb{N}}$, with $\frac{1}{p^*} < \beta \le 1$.

Theorem 3.8 states sufficient conditions ensuring the a.s. convergence of $(\mathbf{B}_p(\mathbf{x}_k, \mathbf{x}^\dagger))_{k\in\mathbb{N}}$ to 0. To strengthen this to the convergence in expectation, we require an additional assumption to ensure that $(\mathbf{B}_p(\mathbf{x}_k, \mathbf{x}^\dagger))_{k\in\mathbb{N}}$ is a uniformly integrable supermartingale. Note that removing the assumptions of Theorem 3.8 from Theorem 3.10 would still result in convergence in expectation to some nonnegative random variable, but not necessarily to 0. Recall that a family $(X_t)_t$ of random variables is uniformly integrable provided $\lim_{k\to\infty} \sup_t \mathbb{E}[\|X_t\| \mid \mathbf{1}_{\|X_t\|\geq k}] = 0$, where $\mathbf{1}(\cdot)$ is the indicator function.

Theorem 3.10. *Let the conditions of Theorem 3.8 hold with* $\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_0) \in \overline{\mathrm{range}(\mathbf{A}^*)}$ *and let* $\mu_k^{p^*-1} \leq \frac{p^*}{G_{p^*} L_{\max}^{p^*}}$ *for all* $k \in \mathbb{N}$. *Then there holds* $\lim_{k\to\infty} \mathbb{E}[\mathbf{B}_p(\mathbf{x}_k, \mathbf{x}^\dagger)] = 0$. *Moreover, for* $1 \leq r \leq p$, *we have* $\lim_{k\to\infty} \mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^\dagger\|_{\mathcal{X}}^r] = 0$, *and if* $\mathcal{X}$ *is additionally uniformly smooth, then* $\lim_{k\to\infty} \mathbb{E}[\|\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k) - \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}^\dagger)\|_{\mathcal{X}^*}^{p^*}] = 0$.

Proof. The step-size conditions allow us to apply Lemma A.2, which yields $\mathbf{B}_p(\mathbf{x}_k, \mathbf{x}^\dagger) \leq \mathbf{B}_p(\mathbf{x}_0, \mathbf{x}^\dagger)$ for all k. It follows that $(\mathbf{B}_p(\mathbf{x}_k, \mathbf{x}^\dagger))_{k\in\mathbb{N}}$ is bounded, and is thus uniformly integrable, and by Theorem 3.8 it converges a.s. to 0. Then, by Vitali's convergence theorem [1, Theorem 4.5.4], we deduce that $(\Delta_k)_{k\in\mathbb{N}}$ converges to 0 in expectation as well. Using now the p-convexity of $\mathcal{X}$ and the monotonicity of expectation, we have

$$0 \leq \frac{C_p}{p} \lim_{k\to\infty} \mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^\dagger\|_{\mathcal{X}}^p] \leq \lim_{k\to\infty} \mathbb{E}[\mathbf{B}_p(\mathbf{x}_k, \mathbf{x}^\dagger)] = 0.$$

By the continuity of the power function and the Lyapunov inequality for $1 \leq r \leq p$, we have

$$0 \leq \lim_{k\to\infty} \mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^\dagger\|_{\mathcal{X}}^r] \leq \lim_{k\to\infty} (\mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^\dagger\|_{\mathcal{X}}^p])^{r/p} = 0.$$

To prove the last claim we use uniform smoothness of $\mathcal{X}$ and Theorem 2.3(iv), to deduce

$$\|\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k) - \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}^\dagger)\|_{\mathcal{X}^*}^{p^*} \leq C \max\{1, \|\mathbf{x}_k\|_{\mathcal{X}}, \|\mathbf{x}^\dagger\|_{\mathcal{X}}\}^p \overline{\rho}_{\mathcal{X}}(\|\mathbf{x}_k - \mathbf{x}^\dagger\|_{\mathcal{X}})^{p^*},$$

where $\overline{\rho}_{\mathcal{X}}(\tau) = \rho_{\mathcal{X}}(\tau)/\tau$ is a modulus of smoothness function such that $\overline{\rho}(\tau) \leq 1$ and $\lim_{\tau\to 0} \overline{\rho}(\tau) = 0$; cf. Definition 2.2. By Lemmas A.2 and A.3 $(\|\mathbf{x}_k\|_{\mathcal{X}}^p)_{k\in\mathbb{N}}$ is (uniformly) bounded, giving that the sequence $(\|\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k) - \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}^\dagger)\|_{\mathcal{X}^*}^{p^*})_{k\in\mathbb{N}}$ is bounded and thus uniformly integrable. Since $\lim_{k\to\infty} \mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^\dagger\|_{\mathcal{X}}] = 0$, it follows that $\|\mathbf{x}_k - \mathbf{x}^\dagger\|_{\mathcal{X}}$ converges to 0 in probability, and thus by the continuous mapping theorem $\overline{\rho}_{\mathcal{X}}(\|\mathbf{x}_k - \mathbf{x}^\dagger\|_{\mathcal{X}})^{p^*}$ also converges to 0 in probability. Applying Vitaly's theorem to the uniformly integrable sequence $(\|\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k) - \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}^\dagger)\|_{\mathcal{X}^*}^{p^*})_{k\in\mathbb{N}}$ yields that it converges to 0 in measure, and the claim follows. ∎

Remark 3.11. Note that the condition $\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_0) \in \overline{\mathrm{range}(\mathbf{A}^*)}$ on $\mathbf{x}_0$ is crucial for ensuring that all the limits are the same. Landweber iterations converge for uniformly convex and smooth $\mathcal{X}$, and any Banach space $\mathcal{Y}$ [44, Theorem 3.3]. In our analysis, we have assumed that $\mathcal{X}$ is p-convex to simplify the analysis. First, p-convexity is used in the proof of Lemma 3.6. If $\mathcal{X}$ were only uniformly convex (and $\mathcal{X}^*$ only uniformly smooth), then we may use the modulus of smoothness function $\rho_{\mathcal{X}}$ (cf. (2.2) and [46, Theorem 2.41]) to establish a suitable analogue of the descent property (3.7). Second, p-convexity is used in the proof of Theorem 3.8, allowing a more direct application of the Robbins–Siegmund theorem by relating the objective values to Bregman distances. Meanwhile, the Landweber method in [44] requires step-sizes that depend on the modulus of smoothness, the current iterate, and the objective value, which is more restrictive than the step-size regime used in this work.

3.2. Convergence analysis for the generalized Kaczmarz model. Schöpfer, Louis, and Schuster [44] studied general powers of the Banach space norm and subgradients of the form $\partial(\frac{1}{q}\|\mathbf{A}\cdot-\mathbf{y}\|_{\mathcal{Y}}^q)(\mathbf{x})$. Now we take an analogous perspective for the objective

$$\Psi(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\Psi_i(\mathbf{x}), \quad \text{with } \Psi_i(\mathbf{x}) := \frac{1}{q}\|\mathbf{A}_i\mathbf{x}-\mathbf{y}_i\|_{\mathcal{Y}}^q,$$

with $1 < q \le 2$. This model is herein called the generalized Kaczmarz model. (Note that this is different from the randomized extended Kaczmarz method [53].) We shall show the convergence of SGD with stochastic directions

$$g(\mathbf{x},\mathbf{y},i) = \mathbf{A}_i^* J_q^{\mathcal{Y}}(\mathbf{A}_i\mathbf{x}-\mathbf{y}_i) = \partial(\tfrac{1}{q}\|\mathbf{A}_i\cdot-\mathbf{y}_i\|_{\mathcal{Y}}^q)(\mathbf{x}). \tag{3.9}$$

The descent property (3.7) is unaffected, and a direct computation again yields

$$\mathbf{B}_p(\mathbf{x}_{k+1},\mathbf{x}^\dagger) \le \mathbf{B}_p(\mathbf{x}_k,\mathbf{x}^\dagger) - \mu_{k+1}\left\langle \mathbf{g}_{k+1}, \mathbf{x}_k-\mathbf{x}^\dagger\right\rangle + \frac{G_{p^*}}{p^*}\mu_{k+1}^{p^*}\|\mathbf{g}_{k+1}\|_{\mathcal{X}^*}^{p^*}. \tag{3.10}$$

However, the Robbins–Siegmund theorem cannot be applied directly. Instead, we pursue a different proof strategy by first establishing the uniform boundedness of iterates.

Lemma 3.12. *Let Assumption 3.1 hold. Consider SGD with descent directions (3.9) for $1<q\le 2$, and assume that $\mu_k^{p^*-1} < \frac{p^*}{G_{p^*}L_{\max}^{p^*}}$ holds for all $k\in\mathbb{N}$ and $\sum_{k=1}^{\infty}\mu_k^{p^*} =: \Gamma < \infty$. Then $(\mathbf{B}_p(\mathbf{x}_k,\mathbf{x}^\dagger))_{k\in\mathbb{N}}$ and $(\mathbf{x}_k)_{k\in\mathbb{N}}$ are uniformly bounded.*

Proof. Let $\overline{\Psi}_i(\mathbf{x}) = \|\mathbf{A}_i\mathbf{x}-\mathbf{y}_i\|_{\mathcal{Y}}^q$ and $\Delta_k = \mathbf{B}_p(\mathbf{x}_k,\mathbf{x}^\dagger)$. Then we have $\langle \mathbf{g}_{k+1}, \mathbf{x}_k-\mathbf{x}^\dagger\rangle = \overline{\Psi}_{i_{k+1}}(\mathbf{x}_k)$ and

$$\begin{aligned}\|\mathbf{g}_{k+1}\|_{\mathcal{X}^*}^{p^*} &= \|\mathbf{A}_{i_{k+1}}^* J_q^{\mathcal{Y}}(\mathbf{A}_{i_{k+1}}\mathbf{x}-\mathbf{y}_{i_{k+1}})\|_{\mathcal{X}^*}^{p^*} \le L_{\max}^{p^*}\|\mathbf{A}_{i_{k+1}}\mathbf{x}-\mathbf{y}_{i_{k+1}}\|_{\mathcal{Y}}^{p^*(q-1)}\\ &\le L_{\max}^{p^*}\overline{\Psi}_{i_{k+1}}(\mathbf{x}_k)^{p^*\frac{q-1}{q}} = L_{\max}^{p^*}\overline{\Psi}_{i_{k+1}}(\mathbf{x}_k)^{\frac{p^*}{q^*}},\end{aligned}$$

where $q^*\ge 2$ is the conjugate exponent of q. Plugging this into (3.10) gives

$$\Delta_{k+1} \le \Delta_k - \mu_{k+1}\overline{\Psi}_{i_{k+1}}(\mathbf{x}_k) + L_{\max}^{p^*}\frac{G_{p^*}}{p^*}\mu_{k+1}^{p^*}\overline{\Psi}_{i_{k+1}}(\mathbf{x}_k)^{\frac{p^*}{q^*}}. \tag{3.11}$$

Since $1<p^*\le 2$ by Theorem 2.6(iii), and $q^*\ge 2$, we have $\frac{p^*}{q^*}\le 1$. Now we define two sets of indices

$$\mathcal{I} = \{j\le k : \overline{\Psi}_{i_{j+1}}(\mathbf{x}_j)\ge 1\} \quad\text{and}\quad \mathcal{J} = \{j\le k : \overline{\Psi}_{i_{j+1}}(\mathbf{x}_j) < 1\},$$

so that $\mathcal{I}\cap\mathcal{J}=\emptyset$, and $\mathcal{I}\cup\mathcal{J}=[k]$. Note that $\mathcal{I}$ and $\mathcal{J}$ actually depend on the current iterate index k. Applying the inductive argument to (3.11) gives

$$\Delta_{k+1} \le \Delta_0 - \sum_{j=0}^{k}\mu_{j+1}\overline{\Psi}_{i_{j+1}}(\mathbf{x}_j) + L_{\max}^{p^*}\frac{G_{p^*}}{p^*}\sum_{j=0}^{k}\mu_{j+1}^{p^*}\overline{\Psi}_{i_{j+1}}(\mathbf{x}_j)^{\frac{p^*}{q^*}}$$

$$= \Delta_0 \underbrace{- \sum_{j\in\mathcal{I}} \mu_{j+1} \overline{\Psi}_{i_{j+1}}(\mathbf{x}_j) + L_{\max}^{p^*} \frac{G_{p^*}}{p^*} \sum_{j\in\mathcal{I}} \mu_{j+1}^{p^*} \overline{\Psi}_{i_{j+1}}(\mathbf{x}_j)^{\frac{p^*}{q^*}}}_{(\star)} \underbrace{- \sum_{j\in\mathcal{J}} \mu_{j+1} \overline{\Psi}_{i_{j+1}}(\mathbf{x}_j)}_{(\star\star)}$$
$$\underbrace{+ L_{\max}^{p^*} \frac{G_{p^*}}{p^*} \sum_{j\in\mathcal{J}} \mu_{j+1}^{p^*} \overline{\Psi}_{i_{j+1}}(\mathbf{x}_j)^{\frac{p^*}{q^*}}}_{(\star\star\star)}.$$

Next we analyze these three terms separately. First, for $j \in \mathcal{I}$, we have $\overline{\Psi}_{i_{j+1}}(\mathbf{x}_j) \geq 1$ and since $\frac{p^*}{q^*} < 1$, we have $\overline{\Psi}_{i_{j+1}}(\mathbf{x}_j)^{\frac{p^*}{q^*}} \leq \overline{\Psi}_{i_{j+1}}(\mathbf{x}_j)$, giving

$$(\star) \leq - \sum_{j\in\mathcal{I}} \mu_{j+1} \overline{\Psi}_{i_{j+1}}(\mathbf{x}_j) + L_{\max}^{p^*} \frac{G_{p^*}}{p^*} \sum_{j\in\mathcal{I}} \mu_{j+1}^{p^*} \overline{\Psi}_{i_{j+1}}(\mathbf{x}_j) = - \sum_{j\in\mathcal{I}} \Big(1 - L_{\max}^{p^*} \frac{G_{p^*}}{p^*} \mu_{j+1}^{p^*-1}\Big) \mu_{j+1} \overline{\Psi}_{i_{j+1}}(\mathbf{x}_j).$$

Since $\mu_{j+1}^{p^*-1} < \frac{p^*}{G_{p^*} L_{\max}^{p^*}}$ holds by assumption, the term $(\star)$ is nonpositive. Moreover, $(\star\star)$ is trivially nonpositive. Since $\overline{\Psi}_{i_{j+1}}(\mathbf{x}_j) < 1$ for $j \in \mathcal{J}$, the last term $(\star\star\star)$ can be bounded as

$$L_{\max}^{p^*} \frac{G_{p^*}}{p^*} \sum_{j\in\mathcal{J}} \mu_{j+1}^{p^*} \overline{\Psi}_{i_{j+1}}(\mathbf{x}_j)^{\frac{p^*}{q^*}} \leq L_{\max}^{p^*} \frac{G_{p^*}}{p^*} \sum_{j\in\mathcal{J}} \mu_{j+1}^{p^*} \leq L_{\max}^{p^*} \frac{G_{p^*}}{p^*} \sum_{j=1}^{\infty} \mu_j^{p^*} = L_{\max}^{p^*} \frac{G_{p^*}}{p^*} \Gamma.$$

By combining the last three bounds on $(\star)$, $(\star\star)$, and $(\star\star\star)$, we get

$$\Delta_{k+1} \leq \Delta_0 + L_{\max}^{p^*} \frac{G_{p^*}}{p^*} \Gamma \quad \forall k \geq 0.$$

Thus, $(\Delta_k)_{k\in\mathbb{N}}$ is uniformly bounded and by Lemma A.3, so is $(\mathbf{x}_k)_{k\in\mathbb{N}}$. ∎

The proof of Lemma 3.12 exposes the challenge in extending the convergence results to general stochastic directions. Namely, in the proof of Theorem 3.8, we showed the convergence by taking the conditional expectation of (3.7), recasting the resulting expression as an almost supermartingale, and then relating objective values to Bregman distances via $\Psi(\mathbf{x}_k) \leq C\Delta_k$ for some $C > 0$. Here, using $\frac{q}{q^*} = q - 1$ and $\frac{p^*}{p} = p^* - 1$, we instead have

$$\Psi(\mathbf{x}_k)^{\frac{p^*}{q^*}} \leq C \Delta_k^{(p^*-1)(q-1)}, \quad \text{with } C = q^{-\frac{p^*}{q^*}} L_{\max}^{p^*(q-1)} \Big(\frac{p}{C_p}\Big)^{(p^*-1)(q-1)},$$

which gives

$$\mathbb{E}_k[\Delta_{k+1}] \leq \Delta_k + C L_{\max}^{p^*} q^{\frac{p^*}{q^*}} \frac{G_{p^*}}{p^*} \mu_{k+1}^{p^*} \Delta_k^{(p^*-1)(q-1)} - q \mu_{k+1} \Psi(\mathbf{x}_k).$$

Here $0 < (p^* - 1)(q - 1) < 1$, provided $p^* \neq 2$ and $q \neq 2$. Therefore, the Robbins–Siegmund theorem cannot be applied directly. Nonetheless, we still have the following analogue of Theorem 3.10.

Theorem 3.13. *Consider iterations (3.4) with descent directions (3.9) for $1 < q \leq 2$, let Assumption 3.1 hold, and let $\mathbf{x}^\dagger$ be the MNS. Let the step-sizes $(\mu_k)_{k\in\mathbb{N}}$ satisfy $\sum_{k=1}^\infty \mu_k = \infty$, $\sum_{k=1}^\infty \mu_k^{p^*} < \infty$, and $\mu_k^{p^*-1} < \frac{p^*}{G_{p^*} L_{\max}^{p^*}}$ for all $k \in \mathbb{N}$. Then the sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ converges a.s. to a solution of (1.1):*

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

Moreover, if $\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_0) \in \overline{\operatorname{range}(\mathbf{A}^*)}$, we have

$$\lim_{k\to\infty} \mathbf{B}_p(\mathbf{x}_k, \mathbf{x}^\dagger) = 0 \text{ a.s.} \quad \text{and} \quad \lim_{k\to\infty} \mathbb{E}[\mathbf{B}_p(\mathbf{x}_k, \mathbf{x}^\dagger)] = 0.$$

Proof. To establish the a.s. convergence of iterates, we first take the conditional expectation of the descent property (3.10) and obtain

$$\mathbb{E}_k[\Delta_{k+1}] \leq \Delta_k - \mu_{k+1} \left\langle \mathbb{E}_k[\mathbf{g}_{k+1}], \mathbf{x}_k - \mathbf{x}^\dagger \right\rangle + \frac{G_{p^*}}{p^*} \mu_{k+1}^{p^*} \mathbb{E}_k[\|\mathbf{g}_{k+1}\|_{\mathcal{X}^*}^{p^*}]. \tag{3.12}$$

We now have $\langle \mathbb{E}_k[\mathbf{g}_{k+1}], \mathbf{x}_k - \mathbf{x}^\dagger\rangle = \langle \partial\Psi(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^\dagger\rangle = q\Psi(\mathbf{x}_k)$, and

$$\|g(\mathbf{x}, \mathbf{y}, i)\|_{\mathcal{X}^*} \leq L_{\max} \|\mathbf{A}_i\mathbf{x} - \mathbf{y}_i\|_{\mathcal{Y}}^{q-1}.$$

Then taking the conditional expectation of $\|g(\mathbf{x}, \mathbf{y}, i)\|_{\mathcal{X}^*}^{p^*}$ yields

$$\mathbb{E}[\|g(\mathbf{x}, \mathbf{y}, i)\|_{\mathcal{X}^*}^{p^*}] \leq L_{\max}^{p^*} \mathbb{E}\Big[\|\mathbf{A}_i\mathbf{x} - \mathbf{y}_i\|_{\mathcal{Y}}^{p^*(q-1)}\Big] = L_{\max}^{p^*} \mathbb{E}\Big[(\|\mathbf{A}_i\mathbf{x} - \mathbf{y}_i\|_{\mathcal{Y}}^q)^{\frac{p^*}{q^*}}\Big].$$

We have $0 < \frac{p^*}{q^*} \leq 1$, with the equality achieved only if $p^* = q^* = 2$. In the latter case, it trivially follows that $\mathbb{E}[\|g(\mathbf{x}, \mathbf{y}, i)\|_{\mathcal{X}^*}^{p^*}] \leq qL_{\max}^{p^*}\Psi(\mathbf{x})$. If $0 < \frac{p^*}{q^*} < 1$, by Jensen's inequality, we have

$$\mathbb{E}[\|g(\mathbf{x}, \mathbf{y}, i)\|_{\mathcal{X}^*}^{p^*}] \leq L_{\max}^{p^*} \mathbb{E}\Big[(\|\mathbf{A}_i\mathbf{x} - \mathbf{y}_i\|_{\mathcal{Y}}^q)^{\frac{p^*}{q^*}}\Big] \leq L_{\max}^{p^*} (\mathbb{E}[\|\mathbf{A}_i\mathbf{x} - \mathbf{y}_i\|_{\mathcal{Y}}^q])^{\frac{p^*}{q^*}} \leq L_{\max}^{p^*} q^{\frac{p^*}{q^*}} \Psi(\mathbf{x})^{\frac{p^*}{q^*}}.$$

Plugging this estimate into the conditional descent property (3.12) yields

$$\mathbb{E}_k[\Delta_{k+1}] \leq \Delta_k - q\mu_{k+1}\Psi(\mathbf{x}_k) + L_{\max}^{p^*} q^{\frac{p^*}{q^*}} \frac{G_{p^*}}{p^*} \mu_{k+1}^{p^*} \Psi(\mathbf{x}_k)^{\frac{p^*}{q^*}}.$$

Since the sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ is uniformly bounded by Lemma 3.12, so is $(\Psi(\mathbf{x}_k))_{k\in\mathbb{N}}$, and we thus have

$$\sum_{k=0}^\infty \mu_{k+1}^{p^*} \Psi(\mathbf{x}_k)^{\frac{p^*}{q^*}} \leq C \sum_{k=0}^\infty \mu_{k+1}^{p^*} < \infty.$$

Thus, we can apply the Robbins–Siegmund theorem for almost supermartingales, and deduce that $(\Delta_k)_{k\in\mathbb{N}}$ converges a.s. to a nonnegative random variable Δ_∞. Moreover, $\sum_{k=0}^\infty \mu_{k+1}\Psi(\mathbf{x}_k) < \infty$ holds a.s. By repeating the argument for Theorem 3.8, there exists a subsequence $(\mathbf{x}_{k_j})_{j\in\mathbb{N}}$ that a.s. converges to some $\widehat{\mathbf{x}} \in \mathcal{X}_{\min}$, and hence $\Delta_\infty = 0$, as desired.

Moreover, by Lemma 3.12, the sequence $(\Delta_k)_{k\in\mathbb{N}}$ is bounded, and thus uniformly integrable. Since it converges to 0 a.s., from Vitali's theorem it follows that $\lim_{k\to\infty}\mathbb{E}[\mathbf{B}_p(\mathbf{x}_k,\mathbf{x}^\dagger)]=0$. ∎

The results in Theorem 3.13 are similar to that of Theorem 3.10, but the generality of the latter is compensated for by an additional step-size assumption ensuring boundedness of iterates $(\mathbf{x}_k)_{k\in\mathbb{N}}$.

3.3. Convergence rates for conditionally stable operators. Theorem 3.10 states the conditions needed for the convergence of Bregman distances in expectation. However, it does not provide convergence rates. In order to obtain convergence rates, one needs additional conditions on the MNS $\mathbf{x}^\dagger$, which are collectively known as source conditions. One approach is via conditional stability: for a locally conditionally stable operator, we can extract convergence in expectation and quantify the convergence speed. Conditional stability is known for many inverse problems for PDEs and has been used extensively to investigate regularized solutions [8, 13]. It is useful for analyzing ill-posed problems that are locally well-posed, and in case of a (possibly) nonlinear forward operator F it is of the form

$$\|\mathbf{x}_1-\mathbf{x}_2\|_{\mathcal{X}}\le\Phi(\|F(\mathbf{x}_1)-F(\mathbf{x}_2)\|_{\mathcal{Y}})\quad\forall\mathbf{x}_1,\mathbf{x}_2\in\mathcal{M}\subset\mathcal{X}, \tag{3.13}$$

where $\Phi:[0,\infty)\to[0,\infty)$ with $\Phi(0)=0$ is a continuous, nondecreasing function, and $\mathcal{M}$ is typically a ball in the ambient norm [19] that is stronger than $\mathcal{X}$, and is thus conditional. In Banach space settings, the conditional stability needs to be adjusted by replacing the left-hand side of (3.13) with a nonnegative error measure [7]. Since the most relevant error measure for Banach space analysis is the Bregman distance $\mathbf{B}_p(\mathbf{x}_1,\mathbf{x}_2)$, a Hölder-type stability estimate then reads as follows: for some $\alpha\ge1$ and $C_\alpha>0$

$$\mathbf{B}_p(\mathbf{x},\mathbf{x}^\dagger)^\alpha\le C_\alpha^{-1}\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}^\dagger\|_{\mathcal{Y}}^p,\quad\forall\mathbf{x}\in\mathcal{X}. \tag{3.14}$$

Note that we relaxed the condition $\mathbf{x}\in\mathcal{M}\subset\mathcal{X}$ to $\mathbf{x}\in\mathcal{X}$, which makes the problem well-posed. Now we give a convergence rate under conditional stability bound (3.14). The constant C_N appears in Lemma 3.5 and denotes the norm equivalence constant.

Theorem 3.14. *Let the forward operator $\mathbf{A}$ satisfy the conditional stability bound (3.14) for some $\alpha\ge1$ and $C_\alpha>0$. Let $\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_0)\in\overline{\mathrm{range}(\mathbf{A}^*)}$, and for $C_k=C_NC_\alpha(1-L_{\max}^{p^*}\frac{G_{p^*}}{p^*}\mu_k^{p^*-1})>0$, the step-sizes satisfy $\sum_{k=1}^\infty\mu_kC_k=\infty$. Then there holds*

$$\lim_{k\to\infty}\mathbb{E}[\mathbf{B}_p(\mathbf{x}_k,\mathbf{x}^\dagger)]=0.$$

Moreover,

$$\mathbb{E}[\mathbf{B}_p(\mathbf{x}_k,\mathbf{x}^\dagger)]\le\begin{cases}\dfrac{\mathbf{B}_p(\mathbf{x}_0,\mathbf{x}^\dagger)}{\left(1+(\alpha-1)\mathbf{B}_p(\mathbf{x}_0,\mathbf{x}^\dagger)^{\alpha-1}\sum_{j=1}^k\mu_jC_j\right)^{\frac{1}{\alpha-1}}} & \text{if } \alpha>1,\\ \exp\Big(-\sum_{j=1}^k\mu_jC_j\Big)\mathbf{B}_p(\mathbf{x}_0,\mathbf{x}^\dagger) & \text{if } \alpha=1.\end{cases}$$

Proof. Let $\Delta_k:=\mathbf{B}_p(\mathbf{x}_k,\mathbf{x}^\dagger)$. The proof of Theorem 3.8 and the conditional stability bound (3.14) imply

$$\mathbb{E}_k[\Delta_{k+1}]\le\Delta_k-p\mu_{k+1}\left(1-L_{\max}^{p^*}\frac{G_{p^*}}{p^*}\mu_{k+1}^{p^*-1}\right)\Psi(\mathbf{x}_k) \tag{3.15}$$

$$\leq \Delta_k - p\mu_{k+1}\frac{C_N C_\alpha}{p}\left(1 - L_{\max}^{p^*}\frac{G_{p^*}}{p^*}\mu_{k+1}^{p^*-1}\right)\Delta_k^\alpha,$$

since by Lemma 3.5, there exists a $C_N > 0$ such that $\Psi(\mathbf{x}) \geq \frac{C_N}{p}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_{\mathcal{Y}}^p$. Taking the full expectation and using Jensen's inequality lead to

$$\mathbb{E}[\Delta_{k+1}] \leq \mathbb{E}[\Delta_k] - \mu_{k+1}C_{k+1}\mathbb{E}[\Delta_k]^\alpha.$$

Since $C_{k+1} > 0$ by assumption, $(\mathbb{E}[\Delta_k])_{k\in\mathbb{N}}$ is a monotonically decreasing sequence. By the convexity of the function $x \mapsto x^\alpha$ (for $\alpha \geq 1$), for any $\epsilon > 0$ and $x \geq \epsilon$, we have $\epsilon^\alpha \geq \frac{\epsilon}{x}x^\alpha$. We claim that for every $\epsilon > 0$, there exists a $k_\epsilon \in \mathbb{N}$ such that $\mathbb{E}[\Delta_k] \leq \epsilon$ for all $k \geq k_\epsilon$. Assuming the contrary, $\mathbb{E}[\Delta_k] \geq \epsilon$ for all k, gives

$$\mathbb{E}[\Delta_{k+1}] \leq \mathbb{E}[\Delta_k] - \mu_{k+1}C_{k+1}\mathbb{E}[\Delta_k]^\alpha \leq \mathbb{E}[\Delta_k] - \mu_{k+1}C_{k+1}\epsilon^\alpha \leq \Delta_0 - \epsilon^\alpha\sum_{j=1}^{k+1}\mu_j C_j \to -\infty,$$

since $\sum_{j=1}^\infty \mu_j C_j = \infty$ by assumption, which is a contradiction. Therefore, $\lim_{k\to\infty}\mathbb{E}[\Delta_k] = 0$. For $\alpha > 1$, by Polyak's inequality (cf. Lemma A.1), we have

$$\mathbb{E}[\Delta_{k+1}] \leq \frac{\Delta_0}{\left(1 + (\alpha-1)\Delta_0^{\alpha-1}\sum_{j=1}^{k+1}\mu_j C_j\right)^{\frac{1}{\alpha-1}}}.$$

Meanwhile, for $\alpha = 1$, using the inequality $1 - x \leq e^{-x}$ for $x \geq 0$, a direct computation yields

$$\mathbb{E}[\Delta_{k+1}] \leq (1 - \mu_{k+1}C_{k+1})\mathbb{E}[\Delta_k] \leq \prod_{j=1}^{k+1}(1 - \mu_j C_j)\Delta_0 \leq \exp\left(-\sum_{j=1}^{k+1}\mu_j C_j\right)\Delta_0,$$

completing the proof of the theorem. ∎

Remark 3.15. We have the following comments on Theorem 3.14:

(i) The estimates for $\alpha > 1$ and $\alpha = 1$ in Theorem 3.14 are consistent in the sense that

$$\lim_{\alpha\searrow 1}\frac{\mathbf{B}_p(\mathbf{x}_0, \mathbf{x}^\dagger)}{\left(1 + (\alpha-1)\mathbf{B}_p(\mathbf{x}_0, \mathbf{x}^\dagger)^{\alpha-1}\sum_{j=1}^k \mu_j C_j\right)^{\frac{1}{\alpha-1}}} = \exp\left(-\sum_{j=1}^k \mu_j C_j\right)\mathbf{B}_p\left(\mathbf{x}_0, \mathbf{x}^\dagger\right).$$

(ii) While it might seem counterintuitive, $\alpha = 1$ gives a better convergence rate than $\alpha > 1$, because of the following:

$$\mathbf{B}_p(\mathbf{x}, \mathbf{x}^\dagger)^\alpha \geq \mathbf{B}_p(\mathbf{x}, \mathbf{x}^\dagger)^{\tilde{\alpha}} \text{ if and only if } \alpha \log \mathbf{B}_p(\mathbf{x}, \mathbf{x}^\dagger) \geq \tilde{\alpha}\log \mathbf{B}_p(\mathbf{x}, \mathbf{x}^\dagger).$$

Hence, whenever $\mathbf{B}_p(\mathbf{x}, \mathbf{x}^\dagger) < 1$, we have $\mathbf{B}_p(\mathbf{x}, \mathbf{x}^\dagger) \geq \mathbf{B}_p(\mathbf{x}, \mathbf{x}^\dagger)^\alpha$ for $\alpha > 1$. Plugging this into the conditional stability bound (3.14) yields

$$\mathbf{B}_p(\mathbf{x}, \mathbf{x}^\dagger)^\alpha \leq \mathbf{B}_p(\mathbf{x}, \mathbf{x}^\dagger) \leq C_1^{-1}\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}^\dagger\|_{\mathcal{Y}}^p = C_1^{-1}pN\Psi(\mathbf{x}).$$

Meanwhile, the proof of Theorem 3.14 uses the conditional stability bound to establish a relationship between the objective value and the Bregman distance to the MNS $\mathbf{x}^\dagger$; cf. (3.15). Putting these together gives that $\alpha = 1$ provides a greater decrease of the expected Bregman distance, once we are close enough to the solution.

The conditional stability estimate (3.14) for a linear operator $\mathbf{A}$ implies its injectivity, and that the objective $\Psi(\mathbf{x})$ is strongly convex. Under condition (3.14), there can indeed be only one solution: if $\mathbf{A}\tilde{\mathbf{x}} = \mathbf{A}\mathbf{x}$, then $\mathbf{B}_p(\tilde{\mathbf{x}}, \mathbf{x}) = 0$ follows from (3.14). The step-size condition $\sum_{k=1}^{\infty} \mu_k C_k = \infty$ is weaker than that in Theorem 3.10. Namely, it follows from step-size conditions in Theorem 3.8, since

$$\sum_{k=1}^{\infty} \mu_k C_k = C_N C_\alpha \left(\sum_{k=1}^{\infty} \mu_k - L_{\max}^{p^*} \frac{G_{p^*}}{p^*} \sum_{k=1}^{\infty} \mu_k^{p^*} \right) = \infty$$

holds if $\sum_{k=1}^{\infty} \mu_k = \infty$ and $\sum_{k=1}^{\infty} \mu_k^{p^*} < \infty$. Further, if there exists a $C > 0$ such that $1 - L_{\max}^{p^*} \frac{G_{p^*}}{p^*} \mu_k^{p^*-1} > C$ holds for all $k \in \mathbb{N}$, e.g., if μ_k is a constant satisfying this condition, then $\sum_{k=1}^{\infty} \mu_k C_k = \infty$ is weaker than the conditions in Theorem 3.8, since the condition $\sum_{k=1}^{\infty} \mu_k^{p^*} < \infty$ is no longer needed for convergence, and $\sum_{k=1}^{\infty} \mu_k = \infty$ suffices. Moreover, we can choose constant step-sizes. Instead, setting $\mu_k = \mu_0$, with $1 - L_{\max}^{p^*} \frac{G_{p^*}}{p^*} \mu_0^{p^*-1} = \frac{1}{2}$ so that $C_k = \frac{C_N C_\alpha}{2}$,, we get an exponential convergence rate for $\alpha = 1$, since $C_k = \frac{C_N C_\alpha}{2}$, we have

$$\begin{aligned}\mathbb{E}[\Delta_{k+1}] &\le (1 - \mu_0 C_{k+1}) \mathbb{E}[\Delta_k] \le \left(1 - 2^{-1-1/(p^*-1)} L_{\max}^{-p^*/p^*-1} \left(\frac{p^*}{G_{p^*}} C_N C_\alpha\right)^{1/p^*-1}\right)^k \mathbb{E}[\Delta_0] \\ &\le \left(1 - 2^{-p} L_{\max}^{-p} \left(\frac{p^*}{G_{p^*}}\right)^{p^*/p} C_N C_\alpha\right)^k \Delta_0.\end{aligned}$$

Note that this convergence rate is largely comparable with that in the Hilbert case: the conditional stability bound implies the strict convexity of the quadratic objective $\Psi(\mathbf{x})$, and the SGD is known to converge exponentially fast (see, e.g., [16, Theorem 3.1]), with the rate determined by a variant of the condition number.

Remark 3.16. The conditional stability bound (3.14) is stated globally. However, such conditions are often valid only locally. A local definition could have been employed in (3.14), with minor modifications of the argument. Indeed, by the argument of Theorem 3.10, we appeal to Lemma A.2, showing that the Bregman distances of the iterates are nonincreasing. Thus, it suffices to assume that the initial point $\mathbf{x}_0$ is sufficiently close to the MNS $\mathbf{x}^\dagger$.

Remark 3.17. Conditional stability is intimately tied with classical source conditions. For example, as shown in [41], assuming $\alpha = 1$ in (3.14) allows us to show a variational inequality:

$$\left\langle \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}^\dagger), \mathbf{x} - \mathbf{x}^\dagger \right\rangle \le \|\mathbf{x}^\dagger\|_{\mathcal{X}}^{p-1} C_\alpha^{-1} (p C_p^{-1})^{1/p} \|\mathbf{A}(\mathbf{x} - \mathbf{x}^\dagger)\|_{\mathcal{Y}}.$$

Then the Hahn–Banach theorem and [41, Lemma 8.21] give the canonical range–type condition $\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}^\dagger) = \mathbf{A}^* \mathbf{w}$ for $\mathbf{w} \in \mathcal{X}$ such that $\|\mathbf{w}\|_{\mathcal{X}} \le 1$. Connections between source conditions and conditional stability estimates have been studied, e.g., for linear operators in Hilbert spaces [47] and in $\mathcal{L}^p$ spaces [5]. Moreover, variational source conditions often imply conditional stability estimates [20], and in case of bijective and continuous operators they are trivially inferred by a standard source condition (albeit, possibly only in a small neighborhood around the solution). See the book [50] about the connections between source conditions and conditional stability estimates, and [21] for inverse problems for differential equations.

4. Regularizing property. In practice, we often do not have access to the exact data $\mathbf{y}$ but only to noisy observations $\mathbf{y}^\delta$, such that $\|\mathbf{y}^\delta - \mathbf{y}\|_{\mathcal{Y}} \le \delta$. The convergence study in the presence of observational noise requires a different approach, since the sequence of objective values $(\|\mathbf{A}\mathbf{x}_k^\delta - \mathbf{y}^\delta\|_{\mathcal{Y}}^p)_{k\in\mathbb{N}}$ generally will not converge to 0. In this section we show that SGD has a regularizing effect, in the sense that the expected error $\mathbb{E}[\mathbf{B}_p(\mathbf{x}_{k(\delta)}^\delta, \mathbf{x}^\dagger)]$ converges to 0 as the noise level δ decays to 0 for properly selected stopping indices $k(\delta)$.

Let $(\mathbf{x}_k)_{k\in\mathbb{N}}$ and $(\mathbf{x}_k^\delta)_{k\in\mathbb{N}}$ be the noiseless and noisy iterates, defined respectively by

$$\mathbf{x}_{k+1} = \mathcal{J}_{p^*}^{\mathcal{X}^*}\left(\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k) - \mu_{k+1}\mathbf{g}_{k+1}\right), \quad \text{with } \mathbf{g}_{k+1} = g(\mathbf{x}_k, \mathbf{y}, i_{k+1}), \tag{4.1}$$

$$\mathbf{x}_{k+1}^\delta = \mathcal{J}_{p^*}^{\mathcal{X}^*}\left(\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k^\delta) - \mu_{k+1}\mathbf{g}_{k+1}^\delta\right), \quad \text{with } \mathbf{g}_{k+1}^\delta = g(\mathbf{x}_k^\delta, \mathbf{y}^\delta, i_{k+1}). \tag{4.2}$$

The key step in proving the regularizing property is to show the stability of SGD iterates with respect to noise. The noise enters into the iterations through the update directions $\mathbf{g}_{k+1}^\delta$, and thus the stability of the iterates requires that of update directions. This, however, requires imposing suitable assumptions on the observation space $\mathcal{Y}$ since in general the single-valued duality maps $\jmath_p^{\mathcal{Y}}$ are continuous only at 0. If $\mathcal{Y}$ is uniformly smooth, the corresponding duality maps are also smooth. This assumption is also needed for deterministic iterates; cf. [46, Proposition 6.17] or [35, Lemma 9]. Thus, we make the following assumption.

Assumption 4.1. The Banach space $\mathcal{X}$ is p-convex and uniformly smooth, and $\mathcal{Y}$ is uniformly smooth.

Using Assumption 4.1 we can show the following stability result on the iterates with respect to noise, whose elementary but lengthy proof is deferred to the appendix.

Lemma 4.2. *Let Assumption 4.1 hold. Consider the iterations (4.1) and (4.2) with the same initialization $\mathbf{x}_0^\delta = \mathbf{x}_0$, and following the same path (i.e., using same random indices i_k). Then, for any fixed $k \in \mathbb{N}$, we have*

$$\lim_{\delta\searrow 0} \mathbb{E}[\mathbf{B}_p(\mathbf{x}_k^\delta, \mathbf{x}_k)] = \lim_{\delta\searrow 0} \mathbb{E}[\|\mathbf{x}_k^\delta - \mathbf{x}_k\|_{\mathcal{X}}] = \lim_{\delta\searrow 0} \mathbb{E}[\|\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k^\delta) - \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k)\|_{\mathcal{X}^*}] = 0.$$

Now we show the regularizing property of SGD for suitable stopping indices $k(\delta)$.

Theorem 4.3. *Let Assumption 4.1 hold, and the step-sizes $(\mu_k)_{k\in\mathbb{N}}$ satisfy $\sum_{k=1}^\infty \mu_k = \infty$, $\sum_{k=1}^\infty \mu_k^{p^*} < \infty$, and $1 - L_{\max}^{p^*}\frac{G_{p^*}}{p^*}\mu_k^{p^*-1} > C > 0$. If $\lim_{\delta\searrow 0} k(\delta) = \infty$ and $\lim_{\delta\searrow 0} \delta^p \sum_{\ell=1}^{k(\delta)} \mu_\ell = 0$, then*

$$\lim_{\delta\searrow 0} \mathbb{E}[\mathbf{B}_p(\mathbf{x}_{k(\delta)}^\delta, \mathbf{x}^\dagger)] = 0.$$

Proof. Let $\Delta_k = \mathbf{B}_p(\mathbf{x}_k, \mathbf{x}^\dagger)$ and $\Delta_k^\delta = \mathbf{B}_p(\mathbf{x}_k^\delta, \mathbf{x}^\dagger)$. Take any $\delta > 0$ and $k \in \mathbb{N}$. By the three-point identity (2.2), we have

$$\begin{aligned}\Delta_k^\delta &= \mathbf{B}_p(\mathbf{x}_k^\delta, \mathbf{x}_k) + \Delta_k + \left\langle \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k) - \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k^\delta), \mathbf{x}_k - \mathbf{x}^\dagger \right\rangle \\ &\le \mathbf{B}_p(\mathbf{x}_k^\delta, \mathbf{x}_k) + \Delta_k + \|\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k) - \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k^\delta)\|_{\mathcal{X}^*}\|\mathbf{x}_k - \mathbf{x}^\dagger\|_{\mathcal{X}}. \end{aligned} \tag{4.3}$$

Consider a sequence $(\delta_j)_{j\in\mathbb{N}}$ decaying to zero. Taking any $\epsilon > 0$, it suffices to find a $j_\epsilon \in \mathbb{N}$ such that for all $j \ge j_\epsilon$ we have $\mathbb{E}[\Delta_{k(\delta_j)}^{\delta_j}] \le 4\epsilon$. By Theorem 3.10, there exists a $k_\epsilon \in \mathbb{N}$ such that for all $k \ge k_\epsilon$ we have

$$\mathbb{E}[\Delta_k] < \epsilon \quad \text{and} \quad \mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^\dagger\|_{\mathcal{X}}] < \epsilon^{1/2}. \tag{4.4}$$

Moreover, for any fixed k_ϵ, by Lemma 4.2, there exists $j_1 \in \mathbb{N}$ such that for all $j \geq j_1$ we have

$$\mathbb{E}[\mathbf{B}_p(\mathbf{x}_{k_\epsilon}^{\delta_j}, \mathbf{x}_{k_\epsilon})] < \epsilon \quad \text{and} \quad \mathbb{E}[\|\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_{k_\epsilon}) - \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_{k_\epsilon}^{\delta_j})\|_{\mathcal{X}^*}] < \epsilon^{1/2}. \tag{4.5}$$

Thus, plugging the estimates (4.4) and (4.5) into (4.3), we have $\mathbb{E}[\Delta_{k_\epsilon}^{\delta_j}] < 3\epsilon$ for all $j \geq j_1$. Note, however, that the same does not necessarily hold for all $k \geq k_\epsilon$, and it thus also does not hold for a monotonically increasing sequence of stopping indices $k(\delta_j)$, since $\mathbb{E}[\Delta_{k(\delta_j)}^{\delta_j}]$ are not necessarily monotonic. Instead, taking the expectation of the descent property (3.7) with respect to $\mathcal{F}_k$ yields

$$\mathbb{E}_k[\Delta_{k+1}^\delta] \leq \Delta_k^\delta - \mu_{k+1}\left\langle \mathbb{E}_k[\mathbf{g}_{k+1}^\delta], \mathbf{x}_k^\delta - \mathbf{x}^\dagger \right\rangle + pL_{\max}^{p^*}\frac{G_{p^*}}{p^*}\mu_{k+1}^{p^*}\Psi(\mathbf{x}_k^\delta).$$

Then we decompose the middle term into

$$\begin{aligned}
\left\langle \mathbb{E}_k[\mathbf{g}_{k+1}^\delta], \mathbf{x}^\dagger - \mathbf{x}_k^\delta \right\rangle &= \frac{1}{N}\sum_{i=1}^N \left\langle J_p^{\mathcal{Y}}(\mathbf{A}_i\mathbf{x}_k^\delta - \mathbf{y}_i^\delta), -(\mathbf{A}_i\mathbf{x}_k^\delta - \mathbf{y}_i^\delta) + \mathbf{y}_i - \mathbf{y}_i^\delta \right\rangle \\
&= -p\Psi(\mathbf{x}_k^\delta) + \frac{1}{N}\sum_{i=1}^N \left\langle J_p^{\mathcal{Y}}(\mathbf{A}_i\mathbf{x}_k^\delta - \mathbf{y}_i^\delta), \mathbf{y}_i - \mathbf{y}_i^\delta \right\rangle \\
&\leq -p\Psi(\mathbf{x}_k^\delta) + \frac{1}{N}\sum_{i=1}^N \|\mathbf{A}_i\mathbf{x}_k^\delta - \mathbf{y}_i^\delta\|_{\mathcal{Y}}^{p-1}\|\mathbf{y}_i - \mathbf{y}_i^\delta\|_{\mathcal{Y}} \\
&\leq -p\Psi(\mathbf{x}_k^\delta) + \delta\frac{1}{N}\sum_{i=1}^N \|\mathbf{A}_i\mathbf{x}_k^\delta - \mathbf{y}_i^\delta\|_{\mathcal{Y}}^{p-1},
\end{aligned}$$

where we used (2.1) and the Cauchy–Schwarz inequality. Taking the full expectation gives

$$\begin{aligned}
\mathbb{E}[\Delta_{k+1}^\delta] &\leq \mathbb{E}[\Delta_k^\delta] - p\mu_{k+1}\mathbb{E}[\Psi(\mathbf{x}_k^\delta)] + pL_{\max}^{p^*}\frac{G_{p^*}}{p^*}\mu_{k+1}^{p^*}\mathbb{E}[\Psi(\mathbf{x}_k^\delta)] + \delta\mu_{k+1}\frac{1}{N}\sum_{i=1}^N \mathbb{E}[\|\mathbf{A}_i\mathbf{x}_k^\delta - \mathbf{y}_i^\delta\|_{\mathcal{Y}}^{p-1}] \\
&= \mathbb{E}[\Delta_k^\delta] - p\mu_{k+1}C_{k+1}\mathbb{E}[\Psi(\mathbf{x}_k^\delta)] + \delta\mu_{k+1}\frac{1}{N}\sum_{i=1}^N \mathbb{E}[\|\mathbf{A}_i\mathbf{x}_k^\delta - \mathbf{y}_i^\delta\|_{\mathcal{Y}}^{p-1}],
\end{aligned}$$

where $C_k = 1 - L_{\max}^{p^*}\frac{G_{p^*}}{p^*}\mu_k^{p^*-1} > C > 0$. Now using the Lyapunov inequality

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}[\|\mathbf{A}_i\mathbf{x}_k^\delta - \mathbf{y}_i^\delta\|_{\mathcal{Y}}^{p-1}] \leq \frac{1}{N}\sum_{i=1}^N \left(\mathbb{E}[\|\mathbf{A}_i\mathbf{x}_k^\delta - \mathbf{y}_i^\delta\|_{\mathcal{Y}}^p]\right)^{(p-1)/p} = p^{1/p^*}\frac{1}{N}\sum_{i=1}^N \left(\mathbb{E}[\Psi_i(\mathbf{x}_k^\delta)]\right)^{1/p^*},$$

we deduce

$$\mathbb{E}[\Delta_{k+1}^\delta] \leq \mathbb{E}[\Delta_k^\delta] - p\mu_{k+1}C_{k+1}\mathbb{E}[\Psi(\mathbf{x}_k^\delta)] + \delta\mu_{k+1}p^{1/p^*}\frac{1}{N}\sum_{i=1}^N \left(\mathbb{E}[\Psi_i(\mathbf{x}_k^\delta)]\right)^{1/p^*}. \tag{4.6}$$

Next we remove the exponent in the last term. Using Young's inequality $ab \leq \frac{a^p}{p}\omega^{-p} + \frac{b^{p^*}}{p^*}\omega^{p^*}$, with $a = \delta$ and $b = \mathbb{E}[\Psi_i(\mathbf{x}_k^\delta)]^{1/p^*}$, we have

$$\frac{1}{N}\sum_{i=1}^{N}\delta\Big(\mathbb{E}[\Psi_i(\mathbf{x}_k^\delta)]\Big)^{1/p^*} \le \delta^p\frac{\omega^{-p}}{p} + \mathbb{E}\Big[\frac{1}{N}\sum_{i=1}^{N}\Psi_i(\mathbf{x}_k^\delta)\Big]\frac{\omega^{p^*}}{p^*} \le \delta^p\frac{\omega^{-p}}{p} + \mathbb{E}[\Psi(\mathbf{x}_k^\delta)]\frac{\omega^{p^*}}{p^*}.$$

Plugging this back into (4.6) gives

$$\mathbb{E}[\Delta_{k+1}^\delta] \le \mathbb{E}[\Delta_k^\delta] - p\mu_{k+1}C_{k+1}\mathbb{E}[\Psi(\mathbf{x}_k^\delta)] + p^{1/p^*}(p^*)^{-1}\omega^{p^*}\mu_{k+1}\mathbb{E}[\Psi(\mathbf{x}_k^\delta)] + p^{-1/p}\delta^p\omega^{-p}\mu_{k+1}.$$

Taking $\omega > 0$ small enough so that $\omega^{p^*} \le p^* p^{1/p} C_k$ (which can be made uniformly on k, thanks to the positive lower bound on C_k), replacing $k+1$ with $k(\delta)$, and using the inductive argument, we have

$$\mathbb{E}[\Delta_{k(\delta)}^\delta] \le \mathbb{E}[\Delta_{k(\delta)-1}^\delta] + p^{-1/p}\omega^{-p}\delta^p\mu_{k(\delta)} \le \mathbb{E}[\Delta_{k_\epsilon}^\delta] + p^{-1/p}\omega^{-p}\delta^p\sum_{\ell=1}^{k(\delta)}\mu_\ell.$$

Since $\lim_{\delta\searrow 0}\delta^p\sum_{\ell=1}^{k(\delta)}\mu_\ell = 0$ and $\lim_{\delta\searrow 0}k(\delta) = \infty$, there exists $j_2 \in \mathbb{N}$ such that for all $j \ge j_2$ we have $k(\delta_j) \ge k_\epsilon$ and $p^{-1/p}\omega^{-p}\delta_j^p\sum_{\ell=1}^{k(\delta_j)}\mu_\ell < \epsilon$. Taking $j_\epsilon = j_1 \vee j_2$ shows $\mathbb{E}[\Delta_{k(\delta_j)}^{\delta_j}] < 4\epsilon$ for all $j \ge j_\epsilon$, and hence the desired claim follows. ∎

Remark 4.4. In the constant step-size regime, such as in the case of conditionally stable operators, the correspondence between the noise level and the step-size regime takes a more standard form. Namely, the condition in Theorem 4.3 reduces to $\lim_{\delta\searrow 0}\delta^p k(\delta) = 0$. In other words, we have $k(\delta) = \mathcal{O}(\delta^{-p})$, mirroring the traditional conditions in Euclidean spaces. Note that the condition on $k(\delta)$ is fairly broad and does not give useful concrete stopping rules directly. Generally, the issue of a posterior stopping rules for stochastic iterative methods is completely open, even for the Hilbert setting [22]. For polynomially decaying step-sizes $(\mu_k)_{k\in\mathbb{N}} = (c_0 k^{-\beta})_{k\in\mathbb{N}}$, the conditions $\frac{1}{p^*} < \beta \le 1$ and $c_0 < (\frac{p^*}{L_{\max}^{p^*}G_{p^*}})^{\frac{1}{p^*-1}}$ give a valid step-size choice, and the stopping index $k(\delta)$ should satisfy $\lim_{\delta\searrow 0}k(\delta) = \infty$ and $\lim_{\delta\searrow 0}k(\delta)\delta^{\frac{p}{1-\beta}} = 0$.

Remark 4.5. It is of much interest to derive a convergence rate for noisy data under a conditional stability condition as in Theorem 3.14, as a natural extension of the regularizing property. However, this is still unavailable. Within the current analysis strategy, deriving the rate would require quantitative versions of stability estimates in Lemma 4.2 in terms of δ and k. Generally the convergence rate analysis for iterative regularization methods in Banach space remains a very challenging task, and much more work is still needed.

5. Numerical experiments.

We present numerical results on two sets of experiments to illustrate distinct features of the SGD (3.4). The first set of experiments deals with an integral operator and the reconstruction of a sparse signal in the presence of either Gaussian or impulse noise. In this model example, we investigate the impact of the number of batches and the choice of the spaces $\mathcal{X}$ and $\mathcal{Y}$ on the performance of the algorithm. To simplify the study we investigate spaces $\mathcal{X}$ and $\mathcal{Y}$ that are smooth and convex of power type, and thus the corresponding duality maps are singletons. To facilitate a direct comparison of the SGD with the Landweber method, we count the computational complexity with respect to the number of epochs, i.e., the partition size N_b defined below. Note, moreover, that our implementation of the Landweber method does not use the stepsizes described in [44, Method 3.1], since the

latter requires knowledge of quantities that are inconvenient to compute in practice. The second set of experiments is about tomographic reconstruction, with respect to different types of noise. All the shown reconstructions are obtained with a single stochastic run, as is often done in practice, and the stopping index is determined in a trial-and-error manner so that the corresponding reconstruction yields small errors.

5.1. Model linear inverse problem. We first consider the following model inverse problem studied in [28]. Let $\kappa : \overline{\Omega} \times \overline{\Omega} \to \mathbb{R}^+$, with $\Omega = (0,1)$, be a continuous function, and define an integral operator $\mathcal{T}_\kappa : \mathcal{L}^{r_\mathcal{X}}(\Omega) \to \mathcal{L}^{r_\mathcal{Y}}(\Omega)$, for $1 < r_\mathcal{X}, r_\mathcal{Y} < \infty$, by

$$(\mathcal{T}_\kappa x)(t) = \int_\Omega \kappa(t,s)x(s)ds. \tag{5.1}$$

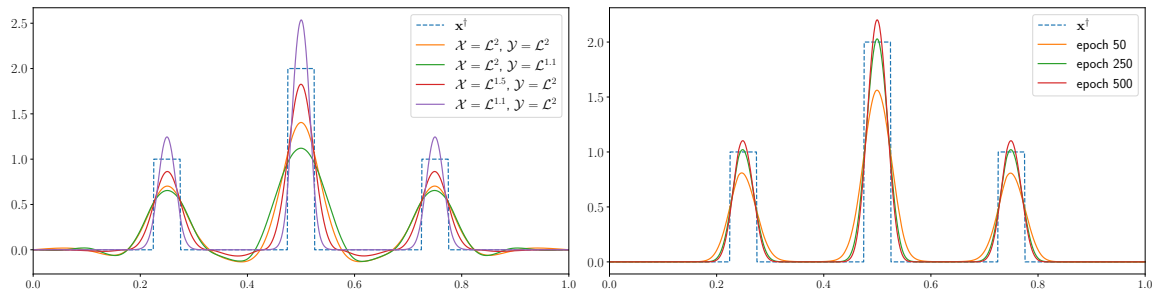
This is a compact linear operator between $\mathcal{L}^{r_\mathcal{X}}(\Omega)$ and $\mathcal{L}^{r_\mathcal{Y}}(\Omega)$, with the adjoint $\mathcal{T}_\kappa^* : \mathcal{L}^{r_\mathcal{Y}^*}(\Omega) \to \mathcal{L}^{r_\mathcal{X}^*}(\Omega)$ given by $(\mathcal{T}_\kappa^* y)(s) = \int_\Omega \kappa(t,s)y(t)dt$. To approximate the integrals, we subdivide the interval $\overline{\Omega}$ into $N = 1000$ subintervals $[\frac{k}{N}, \frac{k+1}{N}]$ for $k = 0, \ldots, N-1$, and then use quadrature, giving a finite-dimensional model $\mathbf{Ax} = \mathbf{y}$, with $\mathbf{A} = \frac{1}{N}\left(\kappa\left(\frac{j-1}{N}, \frac{2k-1}{N}\right)\right)_{j,k=1}^N$ and $\mathbf{x} = \left(x\left(\frac{2j-1}{2N}\right)\right)_{j=1}^N$. For SGD we use $N_b \leq N$ minibatches. To obtain equisized batches, we assume that N_b divides N. The minibatch matrices $\mathbf{A}_j$ are then constructed by taking every N_bth row of $\mathbf{A}$, shifted by j, resulting in well-balanced minibatches, in the sense that the norm $\|\mathbf{A}_j\|$ is (nearly) independent of j.

The kernel function $k(t,s)$ and the exact signal $x^\dagger$ are defined, respectively, by

$$\kappa(t,s) = \begin{cases} 40t(1-s) & \text{if } t \leq s, \\ 40s(1-t) & \text{otherwise} \end{cases} \quad \text{and} \quad x^\dagger(s) = \begin{cases} 1 & \text{if } s \in [\frac{9}{40}, \frac{11}{40}] \cup [\frac{29}{40}, \frac{31}{40}], \\ 2 & \text{if } s \in [\frac{19}{40}, \frac{21}{40}], \\ 0 & \text{otherwise.} \end{cases}$$

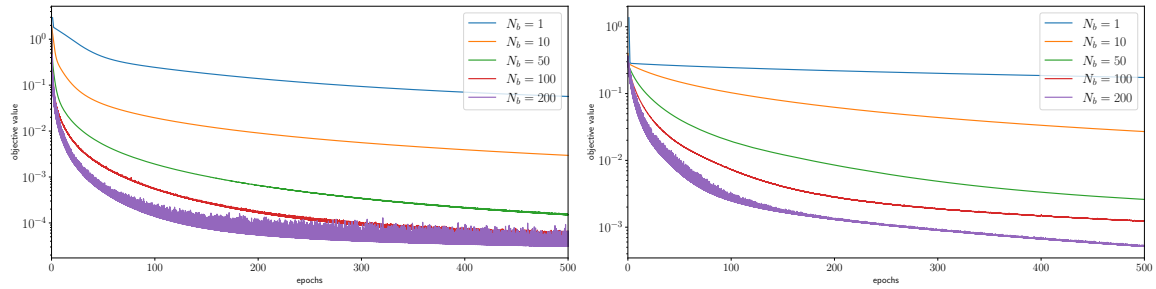
This is a sparse signal, and we expect sparsity promoting norms to perform well. To illustrate this, we compare the following four settings: (a) $\mathcal{X} = \mathcal{Y} = \mathcal{L}^2(\Omega)$; (b) $\mathcal{X} = \mathcal{L}^2(\Omega)$ and $\mathcal{Y} = \mathcal{L}^{1.1}(\Omega)$; (c) $\mathcal{X} = \mathcal{L}^{1.5}(\Omega)$ and $\mathcal{Y} = \mathcal{L}^2(\Omega)$; (d) $\mathcal{X} = \mathcal{L}^{1.1}(\Omega)$ and $\mathcal{Y} = \mathcal{L}^2(\Omega)$. Setting (a) is the standard Hilbert space setting, suitable for recovering smooth solutions from measurement data with independent and identically distributed Gaussian noise, whereas settings (b)–(d) use Banach spaces. Settings (c) and (d) both aim at sparse solutions, and we expect the latter to yield sparser solutions, since spaces $\mathcal{L}^r(\Omega)$ progressively enforce sparser solutions as the exponent r gets closer to 1. In the experiments, we employ the step-size schedule $\mu_k = \frac{L_{\max}}{1+0.05(k/N_b)^{1/p^*+0.01}}$, with $L_{\max} = \max_{j\in[N_b]} \|\mathbf{A}_j\|$. This satisfies the summability conditions $\sum_{k=1}^\infty \mu_k = \infty$ and $\sum_{k=1}^\infty \mu_k^{p^*} < \infty$ required by Theorem 3.8. The operator norm $\|\mathbf{A}_j\| = \|\mathbf{A}_j\|_{\mathcal{L}^{r_\mathcal{X}} \to \mathcal{L}^{r_\mathcal{Y}}} = \max_{\mathbf{x}\neq 0} \frac{\|\mathbf{A}_j\mathbf{x}\|_{\mathcal{L}^{r_\mathcal{Y}}}}{\|\mathbf{x}\|_{\mathcal{L}^{r_\mathcal{X}}}}$ is estimated using Boyd's power method [3]. All the reconstruction algorithms are initialized with a zero vector.

In Figure 1, we compare the reconstructions with settings (a)–(d) for exact data. We observe from Figure 1(a) that settings (a) and (b), with $\mathcal{X} = \mathcal{L}^2(\Omega)$, result in smooth solutions that fail to capture the sparsity structure of the true signal $\mathbf{x}^\dagger$. In contrast, the choice $\mathcal{X} = \mathcal{L}^{1.5}(\Omega)$ recovers a sparser solution, and the choice $\mathcal{X} = \mathcal{L}^{1.1}(\Omega)$ gives a truly sparse reconstruction, but with peaks that overshoot the magnitude of $\mathbf{x}^\dagger$. This might be related to



(a) Changing $\mathcal{X}$ and $\mathcal{Y}$ for $N_b = 100$ (b) Progression of iterates for $\mathcal{X} = \mathcal{L}^{1.1}(\Omega)$

Figure 1. *Comparison of reconstructed solutions after* 500 *epochs.*



(a) $\mathcal{X} = \mathcal{L}^{1.1}(\Omega)$ and $\mathcal{Y} = \mathcal{L}^2(\Omega)$ (b) $\mathcal{X} = \mathcal{L}^{1.5}(\Omega)$ and $\mathcal{Y} = \mathcal{L}^2(\Omega)$

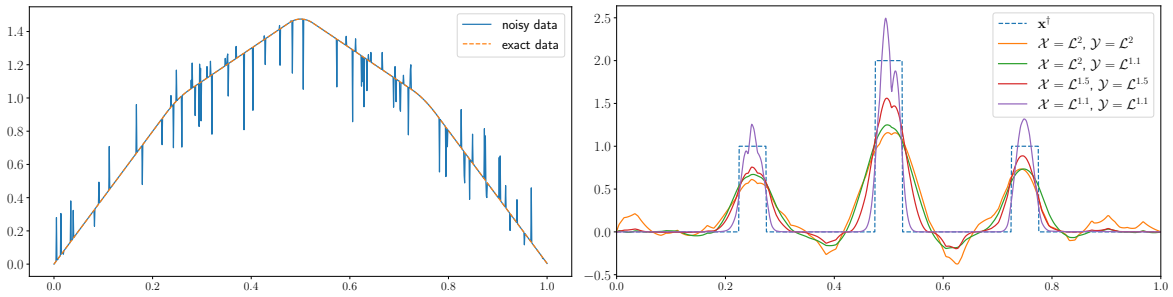
Figure 2. *The variation of* $\frac{1}{p}\sum_{i=1}^{N_b} \|\mathbf{A}_i\mathbf{x}_k - \mathbf{y}_i\|_{\mathcal{Y}}^p$ *with respect to the number of batches* N_b.

the fact that $\mathbf{x}^\dagger$ exhibits a cluster structure in addition to sparsity, which is not accounted for in the choice of the space $\mathcal{X} = \mathcal{L}^{1.1}(\Omega)$ [52, 23]. Figure 1(b) indicates that early stopping would result in lower peaks and significantly reduces the overshooting, but a more explicit form of regularization [52, 23] might allow faster convergence.

In Figure 2, we investigate the convergence of the objective value with respect to the number of batches N_b and the choice of the solution space $\mathcal{X}$. As expected, having a larger number of batches results in a faster initial convergence, but also in increased variance, as shown by the oscillations. Moreover, the variance is lower in the case of a smoother space $\mathcal{X}$ (promoting smoother solutions), where the variance existing in early epochs is dramatically reduced later on. This observation can be explained by the gradient expression $g(\mathbf{x}, \mathbf{y}, i) = \mathbf{A}_i^* J_p^{\mathcal{Y}}(\mathbf{A}_i\mathbf{x} - \mathbf{y}_i)$, which tends to zero as SGD iterates converge to the true solution $\mathbf{x}^\dagger$, as does its variance, and the larger the exponent p, the faster the convergence.

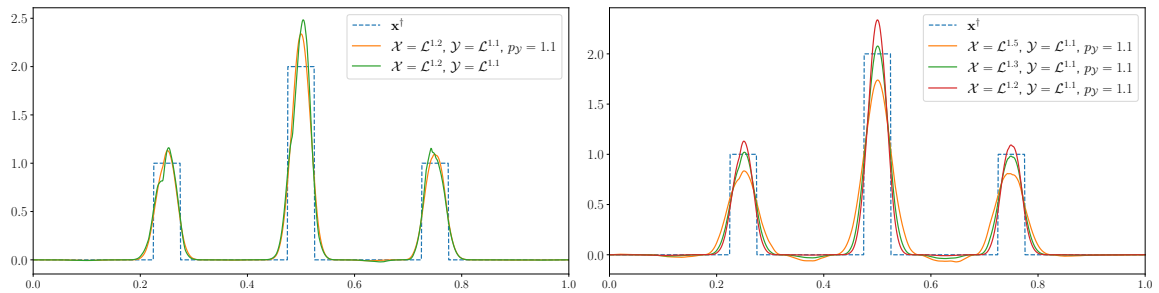
Next we examine the performance of the algorithm when the observational data $\mathbf{y}^\delta$ contains (random-valued) impulse noise (cf. Figure 3), which is generated by

$$y_i^\delta = \begin{cases} y_i^\dagger, & \text{with probability } 1-p, \\ (1-\xi)y_i^\dagger, & \text{with probability } p/2, \\ 1.4\xi + (1-\xi)y_i^\dagger, & \text{with probability } p/2, \end{cases}$$



(a) Data with impulse noise (b) Reconstructions with respect to $\mathcal{X}$ and $\mathcal{Y}$

Figure 3. *The reconstruction performance in case of impulse noise. The algorithms utilized* $N_b = 100$ *batches and were run for* 250 *epochs.*

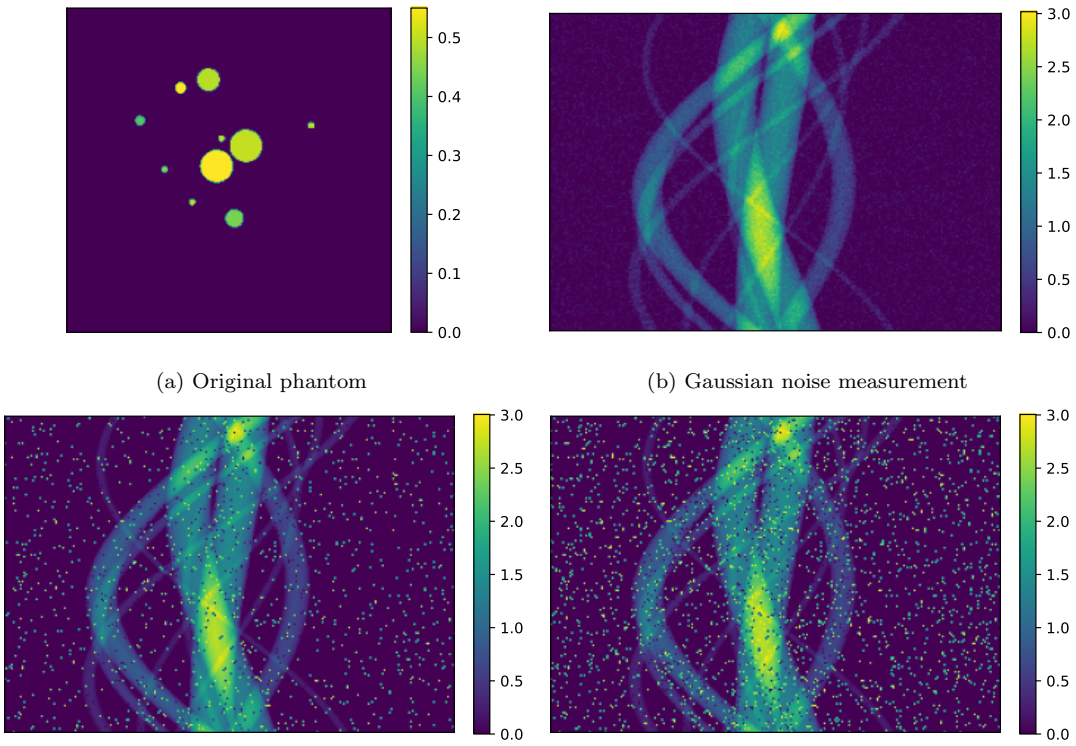


(a) Standard vs generalised Kaczmarz (b) Changing $\mathcal{X}$ in generalised Kaczmarz

Figure 4. *The dependence of the reconstructions in the case of impulse noise on the choice of* q *parameter in the generalized model* (3.9). *The results are obtained using* $N_b = 100$ *batches after* 250 *epochs.*

where $p \in (0,1)$ denotes the percentage of corruption (which is set to 0.05 in the experiment) and $\xi \sim \mathrm{Uni}(0.1, 0.4)$ follows a uniform distribution over the interval $(0.1, 0.4)$. It is known that $\mathcal{L}^r(\Omega)$ fittings with r close to 1 is suitable for impulsive noise. This allows us to investigate the role of not only the space $\mathcal{X}$ but also $\mathcal{Y}$. The results in Figure 3(b) show that the choice $\mathcal{Y} = \mathcal{L}^{r_\mathcal{Y}}(\Omega)$, with $r_\mathcal{Y}$ close to 1, performs significantly better. Indeed, the Hilbert setting $\mathcal{X} = \mathcal{Y} = \mathcal{L}^2(\Omega)$ produces overly smooth, nonsparse solutions with pronounced artifacts. In sharp contrast, setting $\mathcal{X} = \mathcal{Y} = \mathcal{L}^{1.1}(\Omega)$ yields solutions that can correctly identify the sparsity structure of the true solution, and have no artifacts. Similarly as before, the reconstruction in this setting overestimates the signal magnitude on its support, which is exacerbated as the exponent $r_\mathcal{Y}$ gets closer to 1.

Lastly, we investigate the convergence behavior of the method for the generalized model (3.9) in section 3.2, where stochastic directions $g(\mathbf{x}, \mathbf{y}, i)$ are defined as $g(\mathbf{x}, \mathbf{y}, i) = \mathbf{A}_i^* \jmath_q^\mathcal{Y}(\mathbf{A}_i \mathbf{x} - \mathbf{y}_i)$, with $q = r_\mathcal{Y}$ different from the convexity parameter p of the space $\mathcal{X}$. The results in Figure 4 show that this can indeed be beneficial for the performance of the method: the reconstructions are more accurate not only in terms of the solution support, but also in terms of the magnitudes of the nonzero entries. However, the precise mechanism of the excellent performance remains largely elusive.



(a) Original phantom

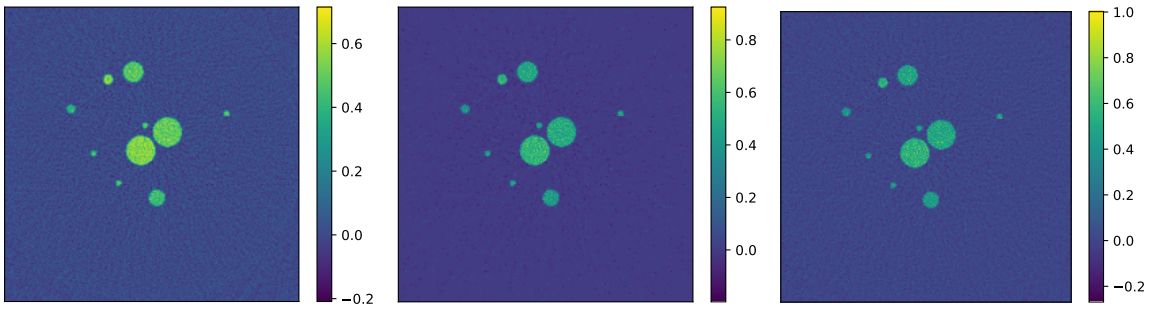
(b) Gaussian noise measurement

(c) Low noise salt-and-pepper measurement

(d) High noise salt-and-pepper measurement

Figure 5. *The plot in* (a) *shows the phantom to be recovered, and* (b)–(d) *show noisy measurements used in the recovery: in* (b)*, random Gaussian noise was added, and* (c)–(d) *are sinogram data degraded by salt-and-pepper noise in the low* (5%) *and high* (10%) *noise regimes.*

5.2. Computed tomography. Now we numerically investigate the behavior of SGD on computed tomography (CT), with respect to the model spaces $\mathcal{X}$ and $\mathcal{Y}$ and data noise. In CT reconstruction, we aim at determining the density of cross-sections of an object by measuring the attenuation of X-rays as they propagate through the object [36]. Mathematically, the forward map is given by the Radon transform. In the experiments, the discrete forward operator $\mathbf{A}$ is defined by a 2D parallel beam geometry, with 180 projection angles on a 1 angle separation, 256 detector elements, and pixel size of 0.1. The sought-for signal $\mathbf{x}^\dagger$ is a (sparse) phantom; cf. Figure 5(a). After applying the forward operator $\mathbf{A}$, either Gaussian (with mean zero and variance 0.01) or salt-and-pepper noise is added. In the latter setting we consider low (with 5% of values changed to either salt or pepper values) and high (10% of values changed) noise regimes. The resulting sinograms (i.e., measurement data) are shown in Figure 5(b)–(d).The experiments were conducted using the Core Imaging Library for the tomographic backend. Note that standard quality metrics in image assessment, such as peak signal-to-noise ratio or mean squared error, are computed using the distance between images in the ℓ^2-norm, which have an implicit bias towards Hilbert spaces and smooth signals, whereas using a metric that emphasizes sparsity is more pertinent to sparsity promoting spaces. To provide a balanced comparison, we report the following two metrics based on normalized ℓ^1-

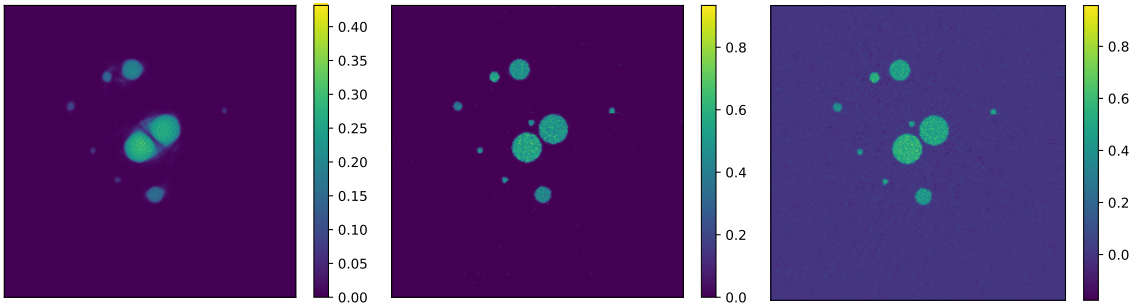


(a) $\mathcal{X} = \mathcal{Y} = \mathcal{L}^2$
$\delta_1(\mathbf{x})/\delta_2(\mathbf{x})$: 2.643/0.528

(b)$\mathcal{X} = \mathcal{L}^{1.1}$, $\mathcal{Y} = \mathcal{L}^2$
$\delta_1(\mathbf{x})/\delta_2(\mathbf{x})$: 0.711/0.341

(c) $\mathcal{X} = \mathcal{Y} = \mathcal{L}^{1.1}$
$\delta_1(\mathbf{x})/\delta_2(\mathbf{x})$: 2.195/0.620

Figure 6. *The reconstruction of the phantom from the observed sinograms degraded by Gaussian noise; cf. Figure 5*(b). *The algorithms use $N_b = 60$ batches and were run for* 200 *epochs.*



(a) 5 epochs
$\delta_1(\mathbf{x})/\delta_2(\mathbf{x})$: 0.702/0.627

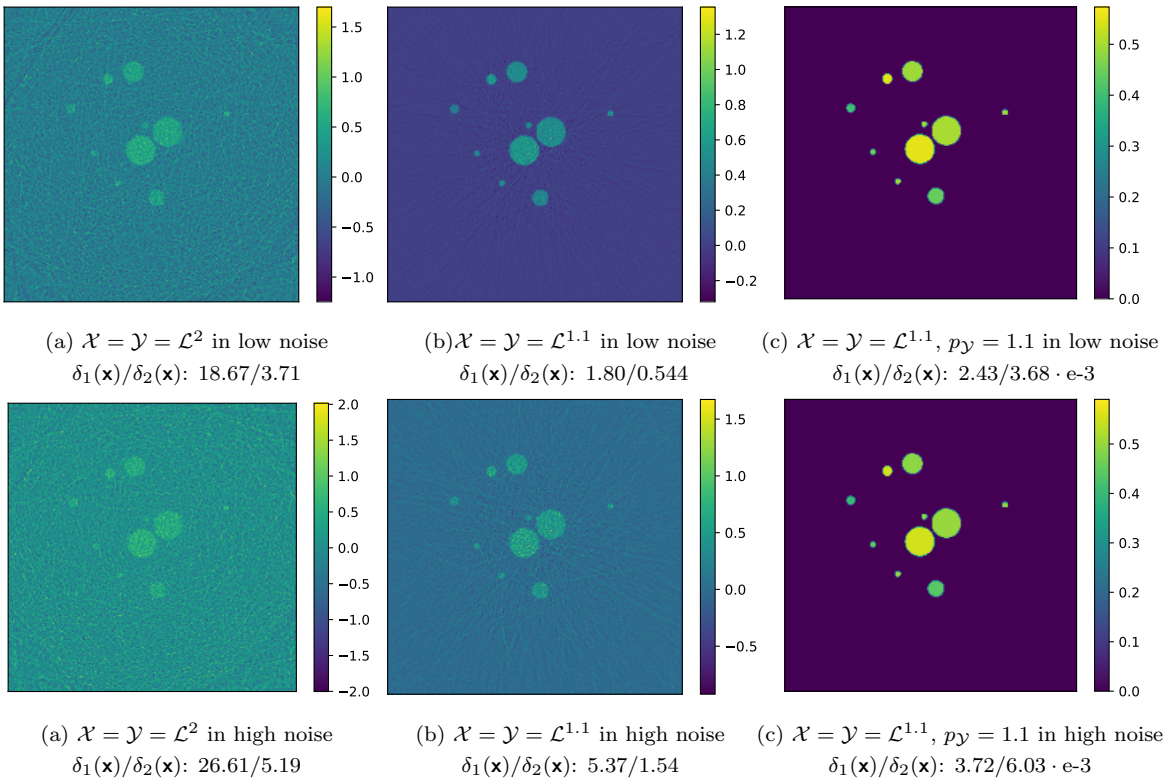
(b) 50 epochs
$\delta_1(\mathbf{x})/\delta_2(\mathbf{x})$: 0.263/0.235

(c) 200 epochs
$\delta_1(\mathbf{x})/\delta_2(\mathbf{x})$: 0.604/0.308

Figure 7. *The evolution of the quality of reconstruction from sinograms degraded by Gaussian noise with respect to the number of epochs. The algorithm uses $\mathcal{X} = \mathcal{Y} = \mathcal{L}^{1.1}$ and $p_{\mathcal{Y}} = 1.1$, with $N_b = 60$ batches.*

and ℓ^2-norms: $\delta_1(\mathbf{x}) = \|\mathbf{x}^\dagger - \mathbf{x}\|_{\ell^1}/\|\mathbf{x}^\dagger\|_{\ell^1}$ and $\delta_2(\mathbf{x}) = \|\mathbf{x}^\dagger - \mathbf{x}\|_{\ell^2}/\|\mathbf{x}^\dagger\|_{\ell^2}$.

First, we show the performance on Gaussian noise, where we compare the Hilbert setting ($\mathcal{X} = \mathcal{Y} = \mathcal{L}^2$) with two Banach settings ($\mathcal{X} = \mathcal{L}^{1.1}$, $\mathcal{Y} = \mathcal{L}^2$, and $\mathcal{X} = \mathcal{Y} = \mathcal{L}^{1.1}$). In the reconstruction, we employ step-sizes $\mu_k = \frac{L_{\max}/2}{1+0.05(k/N_b)^{1/p^*+0.01}}$, with $L_{\max} = \max_{j\in[N_b]} \|\mathbf{A}_j\|$. Figure 6 shows exemplary reconstructions. In all three settings much of the noise is retained in the reconstruction, and whereas the Hilbert setting is better at recovering the magnitude of nonzero entries, the Banach settings are better at recovering the support. Moreover, we observe that the Banach setting with a sparse signal space $\mathcal{X} = \mathcal{L}^{1.1}$ and a smooth observation space $\mathcal{Y} = \mathcal{L}^2$, has the best performance in terms of δ_1 and δ_2 metrics. The Hilbert model performs better than the fully sparse model $\mathcal{X} = \mathcal{Y} = \mathcal{L}^{1.1}$ in terms of the smooth metric δ_2, but worse in the sparsity promoting metric δ_1. We also consider the Banach setting for the generalized model (3.9), with $\mathcal{X} = \mathcal{Y} = \mathcal{L}^{1.1}$ and $p_{\mathcal{Y}} = 1.1$, where we study the effects of early stopping. Figure 7 shows that this setting recovers the support more accurately (and actually does so very early on) and recovers the magnitudes better, but that a form of regularization (through, e.g., early stopping) can be beneficial, since in the later epochs SGD iterates again



(a) $\mathcal{X} = \mathcal{Y} = \mathcal{L}^2$ in low noise
$\delta_1(\mathbf{x})/\delta_2(\mathbf{x})$: 18.67/3.71

(b) $\mathcal{X} = \mathcal{Y} = \mathcal{L}^{1.1}$ in low noise
$\delta_1(\mathbf{x})/\delta_2(\mathbf{x})$: 1.80/0.544

(c) $\mathcal{X} = \mathcal{Y} = \mathcal{L}^{1.1}$, $p_{\mathcal{Y}} = 1.1$ in low noise
$\delta_1(\mathbf{x})/\delta_2(\mathbf{x})$: 2.43/3.68 · e-3

(a) $\mathcal{X} = \mathcal{Y} = \mathcal{L}^2$ in high noise
$\delta_1(\mathbf{x})/\delta_2(\mathbf{x})$: 26.61/5.19

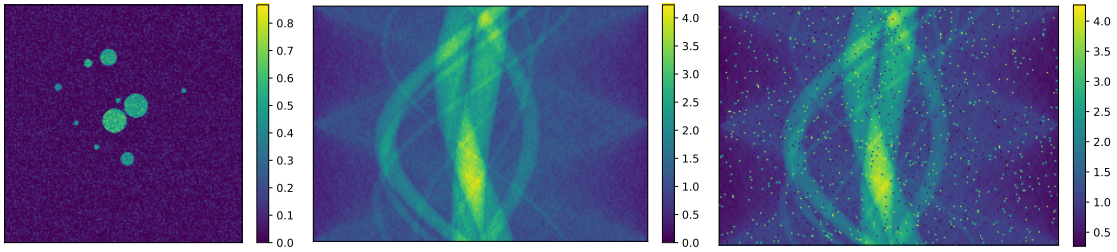
(b) $\mathcal{X} = \mathcal{Y} = \mathcal{L}^{1.1}$ in high noise
$\delta_1(\mathbf{x})/\delta_2(\mathbf{x})$: 5.37/1.54

(c) $\mathcal{X} = \mathcal{Y} = \mathcal{L}^{1.1}$, $p_{\mathcal{Y}} = 1.1$ in high noise
$\delta_1(\mathbf{x})/\delta_2(\mathbf{x})$: 3.72/6.03 · e-3

Figure 8. *The reconstruction of the phantom from the observed sinograms, degraded with low (top) and high (bottom) salt-and-pepper noise, respectively, obtained using the Hilbert space model ($\mathcal{X} = \mathcal{Y} = \mathcal{L}^2$) (left), the Banach model ($\mathcal{X} = \mathcal{Y} = \mathcal{L}^{1.1}$) (middle), and the Banach model ($\mathcal{X} = \mathcal{Y} = \mathcal{L}^{1.1}$) with the generalized Kaczmarz scheme ($p_{\mathcal{Y}} = 1.1$) (right). The algorithms use $N_b = 60$ batches and were run for 200 epochs.*

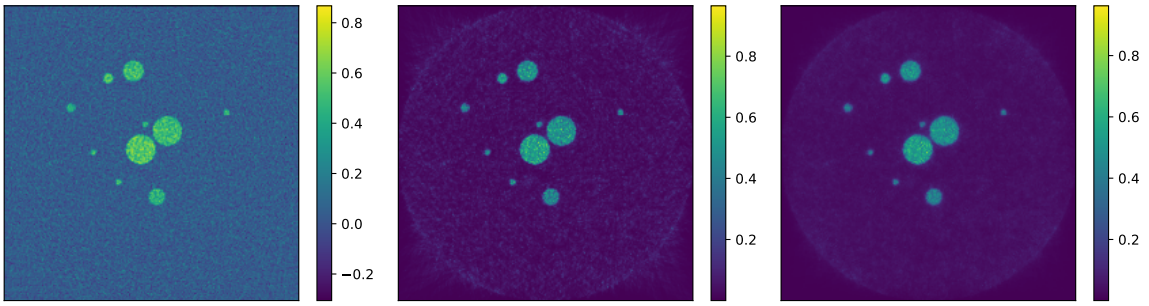
tend to overshoot on the support. Similar behavior can observed for other studied Banach space settings, but not for the Hilbert space setting, which does not recover the support.

We next investigate the performance for low and high salt-and-pepper noise. We compare the Hilbert setting with two Banach settings: the standard SGD with $\mathcal{X} = \mathcal{Y} = \mathcal{L}^{1.1}$ and the generalized model (3.9) with $\mathcal{X} = \mathcal{Y} = \mathcal{L}^{1.1}$ and $p_{\mathcal{Y}} = 1.1$. For the reconstruction, we employ step-sizes $\mu_k = \frac{0.5}{1+0.05(k/N_b)^{1/p^*+0.01}}$. The results in Figure 8 show the reconstructions after 200 epochs with $N_b = 60$ batches. In the low noise regime, the Hilbert setting can reconstruct the general shape of the phantom, but retains a lot of the noise and exhibits streaking artifacts in the background. The reconstruction in the high noise regime is of much poorer quality. The standard Banach SGD shows good behavior in the low noise setting, reconstructing well both the sparsity structure and the magnitudes, but its performance degrades in the high noise setting. In sharp contrast, the model (3.9) shows a nearly perfect reconstruction performance—the phantom is well recovered, with intensities on the correct scale, for both low and high noise regimes. Similarly as before, we observe that Banach methods tend to slightly overestimate the overall intensities, though the recovered values are comparable to the true solution. Overall, the Hilbert setting shows a qualitatively and quantitatively worst performance, in both the ℓ^1- and ℓ^2-norm sense, and the model (3.9) shows the best performance.



(a) Noisy Phantom (b) Gaussian measurement noise (c) Salt-and-pepper measurement noise

Figure 9. *The phantoms and sinograms for the forward model with both pre- and postmeasurement noise. The phantom on the left is degraded by Gaussian noise. After applying the forward operator, either Gaussian (middle) or salt-and-pepper noise (right) is added to the sinogram.*



(a) $\mathcal{X} = \mathcal{Y} = \mathcal{L}^2$
$\delta_1(\mathbf{x})/\delta_2(\mathbf{x})$: 5.65/1.12

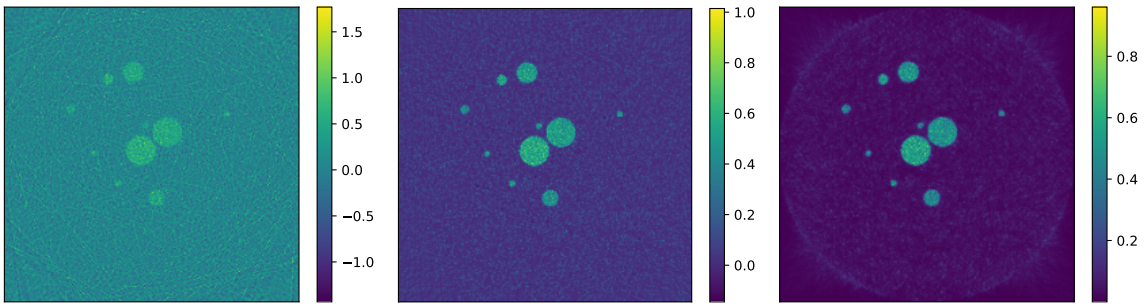
(b) $\mathcal{X} = \mathcal{Y} = \mathcal{L}^{1.1}$, $p_{\mathcal{Y}} = 1.1$
$\delta_1(\mathbf{x})/\delta_2(\mathbf{x})$: 3.16/0.632

(c) $\mathcal{X} = \mathcal{L}^{1.1}$, $\mathcal{Y} = \mathcal{L}^{1.9}$, $p_{\mathcal{Y}} = 1.9$
$\delta_1(\mathbf{x})/\delta_2(\mathbf{x})$: 2.99/0.561

Figure 10. *The reconstruction of the phantom from the observed sinograms with pre- and postmeasurement Gaussian noise. The algorithms use $N_b = 60$ batches and were run for 200 epochs.*

Lastly, we investigate a more challenging setting with noise affecting not only the sinograms, but also the original phantoms. Then the ground-truth image is only approximately sparse. The phantom is degraded with Gaussian noise (zero mean and variance 0.01) after which we apply the forward operator to the resulting noisy phantom. We then add either Gaussian (zero mean and variance 0.01) or salt-and-pepper noise (affecting 3% of measurements); see Figure 9 for representative images. The reconstruction algorithms use SGD with a decaying step-size schedule, $\mu_k = \frac{0.2}{1+0.05(k/N_b)^{1/p^*+0.01}}$.

The reconstructions for data with Gaussian noise in both the phantom and the sinogram are shown in Figure 10. As before, reconstructions in the Hilbert setting are comparable, but slightly worse than those computed with the Banach settings. Banach methods are better at recovering the sparsity structure of the solution and have better reconstruction quality metrics, though they do not completely remove the noise. In the second setting, with the Gaussian noise affecting the phantom and salt-and-pepper noise affecting the sinogram, the difference in reconstruction quality in the Hilbert space and Banach space settings is significantly more pronounced; cf. Figure 11. In both settings, the choice of spaces $\mathcal{X}$ and $\mathcal{Y}$ can have a big impact on the reconstruction quality, especially on the amount of noise retained in the



(a) $\mathcal{X} = \mathcal{Y} = \mathcal{L}^2$
$\delta_1(\mathbf{x})/\delta_2(\mathbf{x})$: 17.48/3.52

(b) $\mathcal{X} = \mathcal{Y} = \mathcal{L}^{1.3}$, $p_{\mathcal{Y}} = 1.3$
$\delta_1(\mathbf{x})/\delta_2(\mathbf{x})$: 3.46/0.86

(c) $\mathcal{X} = \mathcal{Y} = \mathcal{L}^{1.1}$, $p_{\mathcal{Y}} = 1.1$
$\delta_1(\mathbf{x})/\delta_2(\mathbf{x})$: 3.14/0.62

Figure 11. *The reconstructed phantom from the sinograms with a Gaussian premeasurement and a salt-and-pepper (post)measurement noise. The algorithms use $N_b = 60$ batches and were run for 400 epochs.*

background. Moreover, further improvements can be achieved by explicitly penalizing the objective function.

Appendix A. Technical results and proofs.

Lemma A.1 ([38, Lemma 6]). *Let $(\delta_n)_n$ be a sequence of nonnegative scalars, let $(\mu_n)_n$ be a sequence of positive scalars, and let $\alpha > 0$. If*

$$\delta_{n+1} \leq \delta_n - \mu_{n+1}\delta_n^{1+\alpha}, \quad \forall n = 0, \dots, N,$$

then

$$\delta_N \leq \delta_0 \left(1 + \alpha\delta_0^{\alpha} \sum_{n=1}^{N} \mu_n\right)^{-1/\alpha}.$$

A.1. Two elementary estimates.

In this section, we present two elementary estimates on the SGD iterates for exact data that are useful in establishing the regularizing property.

Lemma A.2. *Let the sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ be generated by iterations (3.4), let the step-sizes $(\mu_k)_{k\in\mathbb{N}}$ satisfy $\mu_k^{p^*-1} \leq \frac{p^*}{G_{p^*} L_{\max}^{p^*}}$ for all $k \in \mathbb{N}$, and let the stochastic update directions $\mathbf{g}_k$ be of the form (3.5). Then for any $\widehat{\mathbf{x}} \in \mathcal{X}_{\min}$, the sequence $(\mathbf{B}_p(\mathbf{x}_k, \widehat{\mathbf{x}}))_{k\in\mathbb{N}}$ is nonincreasing. In particular, if $\mathbf{B}_p(\mathbf{x}_0, \widehat{\mathbf{x}}) \leq \rho$, then $\mathbf{B}_p(\mathbf{x}_k, \widehat{\mathbf{x}}) \leq \rho$ for all k.*

Proof. Let $\Delta_k = \mathbf{B}_p(\mathbf{x}_k, \widehat{\mathbf{x}})$. By Lemma 3.6, we have

$$\Delta_{k+1} \leq \Delta_k - \mu_{k+1} \langle \mathbf{g}_{k+1}, \mathbf{x}_k - \widehat{\mathbf{x}} \rangle + \frac{G_{p^*}}{p^*} \mu_{k+1}^{p^*} \|\mathbf{g}_{k+1}\|_{\mathcal{X}^*}^{p^*}.$$

By the definition of duality map and the choice of the update directions $\mathbf{g}_k$, we have

$$\begin{aligned} \langle \mathbf{g}_{k+1}, \mathbf{x}_k - \widehat{\mathbf{x}} \rangle &= \langle j_p^{\mathcal{Y}}(\mathbf{A}_{i_{k+1}}\mathbf{x}_k - \mathbf{y}_{i_{k+1}}), \mathbf{A}_{i_{k+1}}\mathbf{x}_k - \mathbf{y}_{i_{k+1}} \rangle = \|\mathbf{A}_{i_{k+1}}\mathbf{x}_k - \mathbf{y}_{i_{k+1}}\|_{\mathcal{Y}}^p, \\ \|\mathbf{g}_{k+1}\|_{\mathcal{X}^*}^{p^*} &= \|\mathbf{A}_{i_{k+1}}^* j_p^{\mathcal{Y}}(\mathbf{A}_{i_{k+1}}\mathbf{x}_k - \mathbf{y}_{i_{k+1}})\|_{\mathcal{X}^*}^{p^*} \leq \|\mathbf{A}_{i_{k+1}}^*\|^{p^*} \|j_p^{\mathcal{Y}}(\mathbf{A}_{i_{k+1}}\mathbf{x}_k - \mathbf{y}_{i_{k+1}})\|_{\mathcal{Y}^*}^{p^*} \\ &\leq L_{\max}^{p^*} \|\mathbf{A}_{i_{k+1}}\mathbf{x}_k - \mathbf{y}_{i_{k+1}}\|_{\mathcal{Y}}^{(p-1)p^*} = L_{\max}^{p^*} \|\mathbf{A}_{i_{k+1}}\mathbf{x}_k - \mathbf{y}_{i_{k+1}}\|_{\mathcal{Y}}^p. \end{aligned}$$

Consequently,

$$\begin{aligned}\Delta_{k+1} &\le \Delta_k - \mu_{k+1}\langle \mathbf{g}_{k+1}, \mathbf{x}_k - \widehat{\mathbf{x}}\rangle + \frac{G_{p^*}}{p^*}\mu_{k+1}^{p^*}\|\mathbf{g}_{k+1}\|_{\mathcal{X}^*}^{p^*} \\ &\le \Delta_k - \Big(1 - L_{\max}^{p^*}\frac{G_{p^*}}{p^*}\mu_{k+1}^{p^*-1}\Big)\mu_{k+1}\|\mathbf{A}_{i_{k+1}}\mathbf{x}_k - \mathbf{y}_{i_{k+1}}\|_{\mathcal{Y}}^p.\end{aligned}$$

Since $\mu_k^{p^*-1} \le \frac{p^*}{G_{p^*}L_{\max}^{p^*}}$ by assumption, $\Delta_{k+1} \le \Delta_k \le \Delta_0$, completing the proof. ∎

Lemma A.3 (coercivity of the Bregman distance). *If $\Delta_k = \mathbf{B}_p(\mathbf{x}_k, \mathbf{x}^\dagger) \le C < \infty$ for all k, then $\|\mathbf{x}_k\|_{\mathcal{X}}^p \le (2p^*)^p(\|\mathbf{x}^\dagger\|_{\mathcal{X}}^p \vee C)$ for all $k \in \mathbb{N}$.*

Proof. By the definition of Δ_k and the Cauchy–Schwarz inequality, we have

$$\Delta_k \ge \frac{1}{p^*}\|\mathbf{x}_k\|_{\mathcal{X}}^p + \frac{1}{p}\|\mathbf{x}^\dagger\|_{\mathcal{X}}^p - \|\mathbf{x}^\dagger\|_{\mathcal{X}}\|\mathbf{x}_k\|_{\mathcal{X}}^{p-1}.$$

Then we have $\|\mathbf{x}_k\|_{\mathcal{X}}^{p-1}(\frac{1}{p^*}\|\mathbf{x}_k\|_{\mathcal{X}} - \|\mathbf{x}^\dagger\|_{\mathcal{X}}) \le \Delta_k$. If now $\frac{1}{p^*}\|\mathbf{x}_k\|_{\mathcal{X}} - \|\mathbf{x}^\dagger\|_{\mathcal{X}} \le \frac{1}{2p^*}\|\mathbf{x}_k\|_{\mathcal{X}}$, it follows that $\|\mathbf{x}_k\|_{\mathcal{X}}^p \le (2p^*)^p\|\mathbf{x}^\dagger\|_{\mathcal{X}}^p$. Otherwise, if $\frac{1}{p^*}\|\mathbf{x}_k\|_{\mathcal{X}} - \|\mathbf{x}^\dagger\|_{\mathcal{X}} \ge \frac{1}{2p^*}\|\mathbf{x}_k\|_{\mathcal{X}}$, we have

$$\frac{1}{2p^*}\|\mathbf{x}_k\|_{\mathcal{X}}^p \le \|\mathbf{x}_k\|_{\mathcal{X}}^{p-1}\left(\frac{1}{p^*}\|\mathbf{x}_k\|_{\mathcal{X}} - \|\mathbf{x}^\dagger\|_{\mathcal{X}}\right) \le \Delta_k.$$

Combining these two bounds gives $\|\mathbf{x}_k\|_{\mathcal{X}}^p \le (2p^*)^p(\|\mathbf{x}^\dagger\|_{\mathcal{X}}^p \vee \Delta_k)$. ∎

A.2. Proof of Lemma 4.2. To prove Lemma 4.2, we need the following simple fact.

Lemma A.4. *For any fixed $k \in \mathbb{N}$, the clean iterates $\mathbf{x}_k$ generated by (4.1) are uniformly bounded, i.e., there exists $C_k > 0$ such that $\sup_{\omega\in\mathcal{F}_k}\|\mathbf{x}_k\|_{\mathcal{X}} \le C_k < \infty$.*

Proof. If step-sizes μ_k satisfy the conditions of Lemma A.2, the statement is direct from Lemma A.3, and moreover C_k can be chosen to be independent of k. Otherwise we proceed by induction. The induction basis is trivial. Indeed, by the triangle inequality and the definition of duality maps, we have

$$\begin{aligned}\|\mathbf{x}_{k+1}\|_{\mathcal{X}}^{p-1} &= \|\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k) - \mu_{k+1}\mathbf{g}_{k+1}\|_{\mathcal{X}^*} \\ &\le \|\mathbf{x}_k\|_{\mathcal{X}}^{p-1} + L_{\max}\mu_{k+1}\|\jmath_p^{\mathcal{Y}}(\mathbf{A}_{i_{k+1}}\mathbf{x}_k - \mathbf{y}_{i_{k+1}})\|_{\mathcal{Y}^*} \\ &\le \|\mathbf{x}_k\|_{\mathcal{X}}^{p-1} + L_{\max}\mu_{k+1}\|\mathbf{A}_{i_{k+1}}\mathbf{x}_k - \mathbf{y}_{i_{k+1}}\|_{\mathcal{Y}}^{p-1} \\ &\le \|\mathbf{x}_k\|_{\mathcal{X}}^{p-1} + L_{\max}^p\mu_{k+1}\|\mathbf{x}_k - \mathbf{x}^\dagger\|_{\mathcal{X}}^{p-1}.\end{aligned}$$

Now under the inductive hypothesis $\sup_{\omega\in\mathcal{F}_k}\|\mathbf{x}_k\|_{\mathcal{X}} \le C_k < \infty$, we have

$$\begin{aligned}\|\mathbf{x}_{k+1}\|_{\mathcal{X}}^{p-1} &\le \|\mathbf{x}_k\|_{\mathcal{X}}^{p-1} + L_{\max}^p\mu_{k+1}(\|\mathbf{x}_k\|_{\mathcal{X}}^{p-1} + \|\mathbf{x}^\dagger\|_{\mathcal{X}}^{p-1}) \\ &\le C_k^{p-1}(1 + L_{\max}^p\mu_{k+1}) + L_{\max}^p\mu_{k+1}\|\mathbf{x}^\dagger\|_{\mathcal{X}}^{p-1}.\end{aligned}$$

This directly proves the statement of the lemma. ∎

Now we can present the proof of Lemma 4.2.

Proof of Lemma 4.2. For any sequence $(\delta_j)_{j\in\mathbb{N}}$, with $\lim_{j\to\infty}\delta_j = 0$, we consider a sequence of random vectors $(\mathbf{x}_k^{\delta_j}, \mathbf{x}_k)_{j\in\mathbb{N}}$. We will show by induction that (for any fixed $k\in\mathbb{N}$) the sequence $(\mathbf{B}_p(\mathbf{x}_k^{\delta_j}, \mathbf{x}_k))_j$ is uniformly bounded, i.e., $\sup_{\omega\in\mathcal{F}_k}\mathbf{B}_p(\mathbf{x}_k^{\delta_j}, \mathbf{x}_k) < \infty$, converges to 0 pointwise, and that $\mathbf{x}_k^{\delta_j}$ is uniformly bounded. The remaining two claims regarding the convergence of $\|\mathbf{x}_k^\delta - \mathbf{x}_k\|_{\mathcal{X}}$ and $\|\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k^\delta) - \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k)\|_{\mathcal{X}^*}$ then follow directly. For notational brevity, we also suppress the sequence notation δ_j, and only use δ. For the induction base, by Theorem 2.6(i) and (iv), we have

$$\begin{aligned}
&\mathbf{B}_p(\mathbf{x}_1^\delta, \mathbf{x}_1) = \mathbf{B}_{p^*}\Big(\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_1), \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_1^\delta)\Big)\\
&\le \frac{G_{p^*}}{p^*}\|\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_0) - \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_0) - \mu_1(\mathbf{g}_1^\delta - \mathbf{g}_1)\|_{\mathcal{X}^*}^{p^*} = \frac{G_{p^*}}{p^*}\mu_1^{p^*}\|\mathbf{g}_1^\delta - \mathbf{g}_1\|_{\mathcal{X}^*}^{p^*},
\end{aligned}$$

where $\mathbf{g}_1^\delta = g(\mathbf{x}_0, \mathbf{y}^\delta, i_1)$ and $\mathbf{g}_1 = g(\mathbf{x}_0, \mathbf{y}, i_1)$. Specifically, in the case (3.5), we have

$$\begin{aligned}
\|\mathbf{g}_1^\delta - \mathbf{g}_1\|_{\mathcal{X}^*}^{p^*} &= \|\mathbf{A}_{i_1}^*(\jmath_p^{\mathcal{Y}}(\mathbf{A}_{i_1}\mathbf{x}_0 - \mathbf{y}_{i_1}^\delta) - \jmath_p^{\mathcal{Y}}(\mathbf{A}_{i_1}\mathbf{x}_0 - \mathbf{y}_{i_1}))\|_{\mathcal{X}^*}^{p^*}\\
&\le L_{\max}^{p^*}\|\jmath_p^{\mathcal{Y}}(\mathbf{A}_{i_1}\mathbf{x}_0 - \mathbf{y}_{i_1}^\delta) - \jmath_p^{\mathcal{Y}}(\mathbf{A}_{i_1}\mathbf{x}_0 - \mathbf{y}_{i_1})\|_{\mathcal{Y}^*}^{p^*}.
\end{aligned}$$

Since $\mathcal{Y}$ is by assumption uniformly smooth, by Theorem 2.3(iv), we have

$$\begin{aligned}
&\|\jmath_p^{\mathcal{Y}}(\mathbf{A}_{i_1}\mathbf{x}_0 - \mathbf{y}_{i_1}^\delta) - \jmath_p^{\mathcal{Y}}(\mathbf{A}_{i_1}\mathbf{x}_0 - \mathbf{y}_{i_1})\|_{\mathcal{Y}^*}\\
&\le C\max\{1, \|\mathbf{A}_{i_1}\mathbf{x}_0 - \mathbf{y}_{i_1}^\delta\|_{\mathcal{Y}}, \|\mathbf{A}_{i_1}\mathbf{x}_0 - \mathbf{y}_{i_1}\|_{\mathcal{Y}}\}^{p-1}\bar{\rho}_{\mathcal{Y}}(\|\mathbf{y}_{i_1} - \mathbf{y}_{i_1}^\delta\|_{\mathcal{Y}}).
\end{aligned}$$

Upon maximizing over $\mathcal{F}_1$, the term in the maximum is uniformly bounded. Since $\bar{\rho}_{\mathcal{Y}} := \rho_{\mathcal{Y}}(\tau)/\tau \le 1$, $\mathbf{B}_p(\mathbf{x}_1^\delta, \mathbf{x}_1)$ is uniformly bounded. Since $\lim_{\tau\to 0}\bar{\rho}_{\mathcal{Y}}(\tau) = 0$, it follows that $\lim_{\delta\searrow 0}\mathbf{B}_p(\mathbf{x}_1^\delta, \mathbf{x}_1) = 0$ pointwise. By the p-convexity of $\mathcal{X}$, we have

$$0 \le \frac{C_p}{p}\|\mathbf{x}_1^\delta - \mathbf{x}_1\|_{\mathcal{X}}^p \le \mathbf{B}_p(\mathbf{x}_1^\delta, \mathbf{x}_1).$$

Thus, $\|\mathbf{x}_1^\delta - \mathbf{x}_1\|_{\mathcal{X}}$ is uniformly bounded and $\lim_{\delta\searrow 0}\|\mathbf{x}_1^\delta - \mathbf{x}_1\|_{\mathcal{X}} = 0$ point-wise. By the uniform boundedness of $\|\mathbf{x}_1^\delta - \mathbf{x}_1\|_{\mathcal{X}}$ and Lemma A.4, the sequence $\mathbf{x}_1^\delta$ is also uniformly bounded:

$$\|\mathbf{x}_1^\delta\|_{\mathcal{X}} \le \|\mathbf{x}_1^\delta - \mathbf{x}_1\|_{\mathcal{X}} + \|\mathbf{x}_1\|_{\mathcal{X}}. \tag{A.1}$$

For some $k > 0$, assume that $\mathbf{B}_p(\mathbf{x}_k^\delta, \mathbf{x}_k)$ is uniformly bounded and converges to 0 pointwise as $\delta\to 0^+$. Using the p-convexity of $\mathcal{X}$, it follows that $\|\mathbf{x}_k^\delta - \mathbf{x}_k\|_{\mathcal{X}}$ is uniformly bounded and converges to 0 pointwise, and, using again Lemma A.4, it follows that $\mathbf{x}_k^\delta$ is also uniformly bounded. Then by Theorem 2.6(i) and (iv), we have

$$\begin{aligned}
&\mathbf{B}_p(\mathbf{x}_{k+1}^\delta, \mathbf{x}_{k+1}) = \mathbf{B}_{p^*}\Big(\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_{k+1}), \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_{k+1}^\delta)\Big)\\
&\le \frac{G_{p^*}}{p^*}\|\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k^\delta) - \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k) - \mu_{k+1}(\mathbf{g}_{k+1}^\delta - \mathbf{g}_{k+1})\|_{\mathcal{X}^*}^{p^*}\\
&\le \frac{G_{p^*}}{p^*}(\|\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k^\delta) - \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k)\|_{\mathcal{X}^*} + \mu_{k+1}\|\mathbf{g}_{k+1}^\delta - \mathbf{g}_{k+1}\|_{\mathcal{X}^*})^{p^*}.
\end{aligned}$$

Now we separately analyze the two terms in parentheses. First, using the uniform smoothness of $\mathcal{X}$ (and Theorem 2.3(iv) with $\bar{\rho}_{\mathcal{X}^*}(\tau) < C\tau^{p^*-1}$; cf. Definition 2.2), we have

$$\|\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k^\delta) - \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k)\|_{\mathcal{X}^*} \le C\max\{1, \|\mathbf{x}_k^\delta\|_{\mathcal{X}}, \|\mathbf{x}_k\|_{\mathcal{X}}\}^{p-1}\bar{\rho}_{\mathcal{X}}(\|\mathbf{x}_k^\delta - \mathbf{x}_k\|_{\mathcal{X}}). \tag{A.2}$$

Since the right-hand side is uniformly bounded and converges to 0 pointwise by the induction hypothesis, the same holds for the left-hand side. Next we decompose the second term into a sum of two perturbation terms,

$$\begin{aligned}\|g(\mathbf{x}_k^\delta, \mathbf{y}^\delta, i_{k+1}) - g(\mathbf{x}_k, \mathbf{y}, i_{k+1})\|_{\mathcal{X}^*} \le{}& \|g(\mathbf{x}_k, \mathbf{y}^\delta, i_{k+1}) - g(\mathbf{x}_k, \mathbf{y}, i_{k+1})\|_{\mathcal{X}^*} \\ &+ \|g(\mathbf{x}_k^\delta, \mathbf{y}^\delta, i_{k+1}) - g(\mathbf{x}_k, \mathbf{y}^\delta, i_{k+1})\|_{\mathcal{X}^*} := \mathrm{I} + \mathrm{II}.\end{aligned}$$

First, by the assumption $\mathcal{Y}$ being uniformly smooth and Theorem 2.3(iv), we have

$$\begin{aligned}\mathrm{I} &= \|\mathbf{A}_{i_{k+1}}^*(\jmath_p^{\mathcal{Y}}(\mathbf{A}_{i_{k+1}}\mathbf{x}_k - \mathbf{y}_{i_{k+1}}^\delta) - \jmath_p^{\mathcal{Y}}(\mathbf{A}_{i_{k+1}}\mathbf{x}_k - \mathbf{y}_{i_{k+1}}))\|_{\mathcal{X}^*} \\ &\le L_{\max}\|\jmath_p^{\mathcal{Y}}(\mathbf{A}_{i_{k+1}}\mathbf{x}_k - \mathbf{y}_{i_{k+1}}^\delta) - \jmath_p^{\mathcal{Y}}(\mathbf{A}_{i_{k+1}}\mathbf{x}_k - \mathbf{y}_{i_{k+1}})\|_{\mathcal{Y}^*} \\ &\le CL_{\max}\max\{1, \|\mathbf{A}_{i_{k+1}}\mathbf{x}_k - \mathbf{y}_{i_{k+1}}^\delta\|_{\mathcal{Y}}, \|\mathbf{A}_{i_{k+1}}\mathbf{x}_k - \mathbf{y}_{i_{k+1}}\|_{\mathcal{Y}}\}^{p-1}\bar{\rho}_{\mathcal{Y}}(\|\mathbf{y}_{i_{k+1}} - \mathbf{y}_{i_{k+1}}^\delta\|_{\mathcal{Y}}).\end{aligned}$$

By the induction hypothesis and repeating the arguments from the base of induction, the right-hand side is uniformly bounded and converges to 0 pointwise. Second, similarly, we have

$$\begin{aligned}\mathrm{II} &= \|\mathbf{A}_{i_{k+1}}^*(\jmath_p^{\mathcal{Y}}(\mathbf{A}_{i_{k+1}}\mathbf{x}_k^\delta - \mathbf{y}_{i_{k+1}}^\delta) - \jmath_p^{\mathcal{Y}}(\mathbf{A}_{i_{k+1}}\mathbf{x}_k - \mathbf{y}_{i_{k+1}}^\delta))\|_{\mathcal{X}^*} \\ &\le L_{\max}\|\jmath_p^{\mathcal{Y}}(\mathbf{A}_{i_{k+1}}\mathbf{x}_k^\delta - \mathbf{y}_{i_{k+1}}^\delta) - \jmath_p^{\mathcal{Y}}(\mathbf{A}_{i_{k+1}}\mathbf{x}_k^\delta - \mathbf{y}_{i_{k+1}}^\delta)\|_{\mathcal{Y}^*} \\ &\le CL_{\max}\max\{1, \|\mathbf{A}_{i_{k+1}}\mathbf{x}_k^\delta - \mathbf{y}_{i_{k+1}}^\delta\|_{\mathcal{Y}}, \|\mathbf{A}_{i_{k+1}}\mathbf{x}_k - \mathbf{y}_{i_{k+1}}^\delta\|_{\mathcal{Y}}\}^{p-1}\bar{\rho}_{\mathcal{Y}}(\|\mathbf{A}_{i_{k+1}}(\mathbf{x}_k^\delta - \mathbf{x}_k)\|_{\mathcal{Y}}).\end{aligned}$$

By the same arguments, the right-hand side is uniformly bounded. Moreover, $\|\mathbf{A}_{i_{k+1}}(\mathbf{x}_k^\delta - \mathbf{x}_k)\|_{\mathcal{Y}} \le L_{\max}\|\mathbf{x}_k^\delta - \mathbf{x}_k\|_{\mathcal{X}}$, which by the induction hypothesis converges pointwise to 0. Putting all these bounds together yields that $\mathbf{B}_p(\mathbf{x}_{k+1}^\delta, \mathbf{x}_{k+1})$ is uniformly bounded and converges pointwise to 0. Using Vitaly's theorem, the desired statement follows directly. Since $\mathbf{B}_p(\mathbf{x}_k^\delta, \mathbf{x}_k)$ is uniformly bounded and converges pointwise to 0 for any k, then so does $\|\mathbf{x}_k^\delta - \mathbf{x}_k\|_{\mathcal{X}}$, and consequently, by inequality (A.2) (and (A.1)), so does $\|\mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k^\delta) - \mathcal{J}_p^{\mathcal{X}}(\mathbf{x}_k)\|_{\mathcal{X}^*}$. The second part of the claim thus follows. This completes the proof of the induction step, and hence also the lemma. ∎

Acknowledgments. We are very grateful to three anonymous referees for their constructive comments which have led to a significant improvement in the quality of the paper.

REFERENCES

[1] V. I. Bogachev, *Measure Theory* , Vol. I, Springer-Verlag, Berlin, 2007, https://doi.org/10.1007/978-3-540-34514-5.

[2] L. Bottou, F. E. Curtis, and J. Nocedal, *Optimization methods for large-scale machine learning*, SIAM Rev., 60 (2018), pp. 223–311, https://doi.org/10.1137/16M1080173.

[3] D. W. Boyd, *The power method for ℓ^p norms*, Linear Algebra Appl., 9 (1974), pp. 95–101, https://doi.org/10.1016/0024-3795(74)90029-9.

[4] E. J. Candès, J. K. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure Appl. Math., 59 (2006), pp. 1207–1223, https://doi.org/10.1002/cpa.20124.

[5] D.-H. Chen and I. Yousept, *Variational source conditions in L^p-spaces*, SIAM J. Math. Anal., 53 (2021), pp. 2863–2889, https://doi.org/10.1137/20M1334462.
[6] K. Chen, Q. Li, and J.-G. Liu, *Online learning in optical tomography: A stochastic approach*, Inverse Problems, 34 (2018), 075010, https://doi.org/10.1088/1361-6420/aac220.
[7] J. Cheng, B. Hofmann, and S. Lu, *The index function and Tikhonov regularization for ill-posed problems*, J. Comput. Appl. Math., 265 (2014), pp. 110–119, https://doi.org/10.1016/j.cam.2013.09.035.
[8] J. Cheng and M. Yamamoto, *One new strategy for a priori choice of regularizing parameters in Tikhonov's regularization*, Inverse Problems, 16 (2000), pp. L31–L38, https://doi.org/10.1088/0266-5611/16/4/101.
[9] C. Chidume, *Geometric Properties of Banach Spaces and Nonlinear Iterations* , Lecture Notes in Math. 1965, Springer-Verlag, London, 2009.
[10] I. Cioranescu, *Geometry of Banach Spaces, Duality mappings and Nonlinear Problems* , Math. Appl. 62, Kluwer Academic, Dordrecht, 1990, https://doi.org/10.1007/978-94-009-2121-4.
[11] C. Clason, B. Jin, and K. Kunisch, *A semismooth Newton method for L^1 data fitting with automatic choice of regularization parameters and noise calibration*, SIAM J. Imaging Sci., 3 (2010), pp. 199–231, https://doi.org/10.1137/090758003.
[12] M. V. de Hoop, L. Qiu, and O. Scherzer, *Local analysis of inverse problems: Hölder stability and iterative reconstruction*, Inverse Problems, 28 (2012), 045001, https://doi.org/10.1088/0266-5611/28/4/045001.
[13] H. Egger and B. Hofmann, *Tikhonov regularization in Hilbert scales under conditional stability assumptions*, Inverse Problems, 34 (2018), 115015, https://doi.org/10.1088/1361-6420/aadef4.
[14] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems* , Kluwer Academic, Dordrecht, 1996.
[15] I. M. Gamba, Q. Li, and A. Nair, *Reconstructing the thermal phonon transmission coefficient at solid interfaces in the phonon transport equation*, SIAM J. Appl. Math., 82 (2022), pp. 194–220, https://doi.org/10.1137/20M1381666.
[16] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik, *SGD: General analysis and improved rates* , in Proceedings of the 36th International Conference on Machine Learning, PMLR, 2019, pp. 5200–5209.
[17] R. Gu, B. Han, and Y. Chen, *Fast subspace optimization method for nonlinear inverse problems in Banach spaces with uniformly convex penalty terms*, Inverse Problems, 35 (2019), 125011, https://doi.org/10.1088/1361-6420/ab2a2b.
[18] G. T. Herman and L. B. Meyer, *Algebraic reconstruction techniques can be made computationally efficient*, IEEE Trans. Med. Imag., 12 (1993), pp. 600–609.
[19] B. Hofmann, *On the degree of ill-posedness for nonlinear problems*, J. Inverse Ill-Posed Probl., 2 (1994), pp. 61–76, https://doi.org/10.1515/jiip.1994.2.1.61.
[20] T. Hohage and F. Weidling, *Variational source conditions and stability estimates for inverse electromagnetic medium scattering problems*, Inverse Probl. Imaging, 11 (2017), pp. 203–220, https://doi.org/10.3934/ipi.2017010.
[21] V. Isakov, *Inverse Problems for Partial Differential Equations* , 3rd ed., Springer, Cham, 2017, https://doi.org/10.1007/978-3-319-51658-5.
[22] T. Jahn and B. Jin, *On the discrepancy principle for stochastic gradient descent*, Inverse Problems, 36 (2020), 095009, https://doi.org/10.1088/1361-6420/abaa58.
[23] B. Jin, D. A. Lorenz, and S. Schiffler, *Elastic-net regularization: Error estimates and active set methods*, Inverse Problems, 25 (2009), 115022, https://doi.org/10.1088/0266-5611/25/11/115022.
[24] B. Jin and X. Lu, *On the regularizing property of stochastic gradient descent*, Inverse Problems, 35 (2019), 015004, https://doi.org/10.1088/1361-6420/aaea2a.
[25] B. Jin, Z. Zhou, and J. Zou, *On the convergence of stochastic gradient descent for nonlinear ill-posed problems*, SIAM J. Optim., 30 (2020), pp. 1421–1450, https://doi.org/10.1137/19M1271798.
[26] B. Jin, Z. Zhou, and J. Zou, *On the saturation phenomenon of stochastic gradient descent for linear inverse problems*, SIAM/ASA J. Uncertain. Quantif., 9 (2021), pp. 1553–1588, https://doi.org/10.1137/20M1374456.
[27] Q. Jin, X. Lu, and L. Zhang, *Stochastic Mirror Descent Method for Linear Ill-Posed Problems in Banach Spaces*, preprint, https://arxiv.org/abs/2207.06584, 2022.

[28] Q. Jin and L. Stals, *Nonstationary iterated Tikhonov regularization for ill-posed problems in Banach spaces*, Inverse Problems, 28 (2012), 104011.
[29] B. Kaltenbacher and B. Hofmann, *Convergence rates for the iteratively regularized Gauss-Newton method in Banach spaces*, Inverse Problems, 26 (2010), 035007.
[30] B. Kaltenbacher, A. Neubauer, and O. Scherzer, *Iterative Regularization Methods for Nonlinear Ill-posed Problems* , Walter de Gruyter, Berlin, 2008, https://doi.org/10.1515/9783110208276.
[31] Ž. Kereta, R. Twyman, S. Arridge, K. Thielemans, and B. Jin, *Stochastic EM methods with variance reduction for penalised PET reconstructions*, Inverse Problems, 37 (2021), 115006.
[32] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications* , 2nd ed., Springer-Verlag, New York, 2003.
[33] L. Landweber, *An iteration formula for Fredholm integral equations of the first kind*, Amer. J. Math., 73 (1951), pp. 615–624, https://doi.org/10.2307/2372313.
[34] S. Lu and P. Mathé, *Stochastic gradient descent for linear inverse problems in Hilbert spaces*, Math. Comp., 91 (2022), pp. 1763–1788, https://doi.org/10.1090/mcom/3714.
[35] F. Margotti, *Inexact Newton regularization combined with gradient methods in Banach spaces*, Inverse Problems, 34 (2018), 075007, https://doi.org/10.1088/1361-6420/aac21f.
[36] F. Natterer, *The Mathematics of Computerized Tomography* , SIAM, Philadelphia, PA, 2001, https://doi.org/10.1137/1.9780898719284.
[37] D. Needell, R. Zhao, and A. Zouzias, *Randomized block Kaczmarz method with projection for solving least squares*, Linear Algebra Appl., 484 (2015), pp. 322–343, https://doi.org/10.1016/j.laa.2015.06.027.
[38] B. T. Polyak, *Introduction to Optimization* , Optimization Software, Inc., Publications Division, New York, 1987.
[39] J. C. Rabelo, Y. F. Saporito, and A. Leitão, *On stochastic Kaczmarz type methods for solving large scale systems of ill-posed equations*, Inverse Problems, 38 (2022), 025003, https://doi.org/10.1088/1361-6420/ac3f80.
[40] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Statistics, 22 (1951), pp. 400–407, https://doi.org/10.1214/aoms/1177729586.
[41] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen, *Variational Methods in Imaging* , Springer, New York, 2009.
[42] F. Schöpfer and D. A. Lorenz, *Linear convergence of the randomized sparse Kaczmarz method*, Math. Program., 173 (2019), pp. 509–536, https://doi.org/10.1007/s10107-017-1229-1.
[43] F. Schöpfer, D. A. Lorenz, L. Tondji, and M. Winkler, *Extended randomized Kaczmarz method for sparse least squares and impulsive noise problems*, Linear Algebra Appl., 652 (2022), pp. 132–154, https://doi.org/10.1016/j.laa.2022.07.003.
[44] F. Schöpfer, A. K. Louis, and T. Schuster, *Nonlinear iterative methods for linear ill-posed problems in Banach spaces*, Inverse Problems, 22 (2006), pp. 311–329, https://doi.org/10.1088/0266-5611/22/1/017.
[45] F. Schöpfer and T. Schuster, *Fast regularizing sequential subspace optimization in Banach spaces*, Inverse Problems, 25 (2009), 015013.
[46] T. Schuster, B. Kaltenbacher, B. Hofmann, and K. S. Kazimierski, *Regularization Methods in Banach Spaces* , Walter de Gruyter, Berlin, 2012, https://doi.org/10.1515/9783110255720.
[47] U. Tautenhahn, U. Hämarik, B. Hofmann, and Y. Shao, *Conditional stability estimates for ill-posed PDE problems by using interpolation*, Numer. Funct. Anal. Optim., 34 (2013), pp. 1370–1417, https://doi.org/10.1080/01630563.2013.819515.
[48] M. Unser and S. Aziznejad, *Convex optimization in sums of Banach spaces*, Appl. Comput. Harmonic Anal., 56 (2022), pp. 1–25.
[49] A. Wald, *A fast subspace optimization method for nonlinear inverse problems in Banach spaces with an application in parameter identification*, Inverse Problems, 34 (2018), 085008, https://doi.org/10.1088/1361-6420/aac8f3.
[50] F. Weidling, *Variational Source Conditions and Conditional Stability Estimates for Inverse Problems in PDEs*, Ph.D. thesis. University of Göttingen, 2019.
[51] M. Zhong, W. Wang, and Q. Jin, *Regularization of inverse problems by two-point gradient methods in Banach spaces*, Numer. Math., 143 (2019), pp. 713–747, https://doi.org/10.1007/s00211-019-01068-0.

[52] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. R. Stat. Soc. Ser. B Stat. Methodol., 67 (2005), pp. 301–320, https://doi.org/10.1111/j.1467-9868.2005.00503.x.
[53] A. Zouzias and N. M. Freris, *Randomized extended Kaczmarz for solving least squares*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 773–793, https://doi.org/10.1137/120889897.