Nr. 96
2. Jun 2023

universität innsbruck
Institut für Mathematik

Preprint-Series: Department of Mathematics - Applied Mathematics

Convergence analysis of equilibrium methods for inverse problems

D. Obmann, M. Haltmeier

AppliedMathematics

Technikerstraße 13 - 6020 Innsbruck - Austria
Tel.: +43 512 507 53803 Fax: +43 512 507 53898
https://applied-math.uibk.ac.at

Convergence analysis of equilibrium methods for inverse problems

Daniel Obmann

Department of Mathematics, University of Innsbruck
Technikerstrasse 13, 6020 Innsbruck, Austria
E-mail: daniel.obmann@uibk.ac.at

Markus Haltmeier

Department of Mathematics, University of Innsbruck
Technikerstrasse 13, 6020 Innsbruck, Austria
E-mail: markus.haltmeier@uibk.ac.at

June 02, 2023

Abstract

Recently, the use of deep equilibrium methods has emerged as a new approach for solving imaging and other ill-posed inverse problems. While learned components may be a key factor in the good performance of these methods in practice, a theoretical justification from a regularization point of view is still lacking. In this paper, we address this issue by providing stability and convergence results for the class of equilibrium methods. In addition, we derive convergence rates and stability estimates in the symmetric Bregman distance. We strengthen our results for regularization operators with contractive residuals. Furthermore, we use the presented analysis to gain insight into the practical behavior of these methods, including a lower bound on the performance of the regularized solutions. In addition, we show that the convergence analysis leads to the design of a new type of loss function which has several advantages over previous ones. Numerical simulations are used to support our findings.

Keywords: inverse problems, regularization, equilibrium points, stability guarantees, stability estimates, convergence, convergence rates, learned reconstruction, neural networks

1 Introduction

In various imaging applications it is often not possible to measure the image of interest directly which is therefore measured indirectly. Assuming a linear measurement model,

recovering the sought for image $x \in \mathbb{X}$ requires solving the inverse problem

$$y^\delta = \mathbf{A}x + z_\delta \,, \tag{1.1}$$

where $\mathbf{A}\colon \mathbb{X} \to \mathbb{Y}$ is a linear operator between Hilbert spaces modeling the forward problem, z_δ is the data perturbation and $y^\delta \in \mathbb{Y}$ is the noisy data. In many cases problems of the form (1.1) are ill-posed, meaning that the operator $\mathbf{A}$ cannot be uniquely and stably inverted. To get reasonable approximate solutions one has to use regularization methods which essentially approximate (1.1) with a family of neighboring well-posed problems. The prime example is variational regularization [21] where approximate solutions are constructed as minimizers of $\|\mathbf{A}(\cdot) - y^\delta\|^2/2 + \alpha\mathcal{R}$, where $\mathcal{R}$ is the regularizer and $\alpha > 0$ is a tuning parameter that acts as a tradeoff between data-fitting and stability.

1.1 Equilibrium methods

For quite some time now, there has been a trend towards solving imaging problems by learned components [1, 2, 5, 9, 6, 7, 10, 11, 16, 18, 19, 24]. While these methods often achieve superior results compared to classical reconstruction methods, in many cases the theory is still rather underdeveloped. In this paper we are interested in a relatively recent development in this area, where the neighboring problem is defined by certain equilibrium equation; see [7] for a recent review. Specifically, in this paper we consider approximate solutions x_α^δ as solutions of the equilibrium equation

$$\mathbf{T}_\alpha(x, y^\delta) \coloneqq \mathbf{A}^*(\mathbf{A}x - y^\delta) + \alpha\mathbf{G}(x) = 0 \,, \tag{1.2}$$

where $\alpha > 0$ is a tuning parameter and $\mathbf{G}\colon \mathbb{X} \to \mathbb{X}$ is a potentially learned regularization operator. If $\mathbf{G} = \nabla\mathcal{R}$ is of gradient form, then equation (1.2) characterizes critical points of $\left\|\mathbf{A}(\cdot) - y^\delta\right\|^2/2 + \alpha\mathcal{R}$. However, in this paper we are interested in the more general case where $\mathbf{G}$ is not necessarily of gradient form. Equation (1.2) can equivalently be written in either of the fixed point forms

$$x = x - \beta(\mathbf{A}^*(\mathbf{A}x - y^\delta) + \alpha\mathbf{G}(x)) \tag{1.3}$$

$$x = (\mathrm{Id} + \beta\alpha\mathbf{G})^{-1}(x - \beta\mathbf{A}^*(\mathbf{A}x - y^\delta)) \tag{1.4}$$

where the characterization (1.4) requires $(\mathrm{Id} + \beta\alpha\mathbf{G})^{-1}$ to be invertible. While (1.3), (1.4) have been used to find appropriate numerical schemes to solve equation (1.2), in this paper we analyze (1.2) rather than specific algorithms for solving it.

To the best of our knowledge, theoretical considerations for methods of the form (1.2), (1.3), (1.4) mostly consider the convergence of iterative methods for its solution; see [19, 5, 24]. Opposed to this, theoretical questions such convergence of solutions of (1.2) as $\delta \to 0$ are still open. We mention here the paper [3] which considers an implicit form similar to (1.2)

motivated by the use of learned denoisers [12, 18, 23].

1.2 Main contributions

In this paper we analyze (1.2) and provide stability and convergence under relatively weak assumptions on $\mathbf{G}$. Besides this, we give some additional results about the behavior of these solutions including stability estimates and convergence rates. To be more specific, our main contributions are as follows.

- We show that for a fixed $\alpha > 0$ solutions of (1.2) depend stably on the data y^δ. Moreover we show convergence in the sense that for $y^\delta \to y \in \operatorname{ran}(\mathbf{A})$ as $\delta \to 0$ and a suitable parameter choice $\alpha = \alpha(\delta)$, solutions of (1.2) converge to solutions of the limiting problem

$$\mathbf{A}x = y \quad \text{and} \quad -\mathbf{G}(x) \in \ker(\mathbf{A})^\perp . \tag{1.5}$$

  In the special case that $\mathbf{G} = \nabla\mathcal{R}$ is of gradient form, this is the first order optimality condition of the constrained optimization problem $\arg\min_x\{\mathcal{R}(x) \mid \mathbf{A}x = y\}$ defining $\mathcal{R}$-minimizing solutions from variational regularization [21].

- Under the source-condition $\mathbf{G}(x_+) \in \operatorname{ran}(\mathbf{A}^*)$ we derive convergence rates in the absolute symmetric Bregman-distance $|\langle \mathbf{G}(x) - \mathbf{G}(x_+), x - x_+\rangle|$ and $\left\|\mathbf{A}x - y^\delta\right\|$.

- We strengthen our results for the case that $\mathbf{G} = \mathrm{Id} - \mathbf{N}$ is a residual operator with contractive residual part $\mathbf{N}$; compare [19, 5]. In particular, we derive uniqueness of the regularized problem (1.2) and the limiting problem (1.5). Moreover, we show that in this case convergence in the symmetric Bregman-distance is equivalent to convergence in norm. For this specific form we also study the case of inexact solution of (1.2).

- In the context of learning $\mathbf{G} = \mathrm{Id} - \mathbf{N}$, we provide a-priori lower bounds on $\operatorname{Lip}(\mathbf{N})$ necessary to guarantee that a set of desirable solutions can be recovered with (1.2) in the limit. Additionally, we provide lower bounds on the approximation error $\|x - x_\alpha^\delta\|$.

- Based on (1.5), we propose a novel loss-function for constructing the operator $\mathbf{G}$ that is independent of iterative methods for approximating solutions of (1.2).

1.3 Overview

We begin the paper with the general theoretical analysis in Section 2. In Section 3 we consider contractive residual parts showing that (1.2) indeed yields a regularization method, derive lower bounds on its performance and propose a new type of loss-function. Section 4 provides some numerical simulations to support the theoretical findings. The paper finishes with a brief summary and outlook in Section 5.

2 Convergence analysis

Let $\mathbf{A}\colon \mathbb{X} \to \mathbb{Y}$ be a linear and continuous mapping between real Hilbert-spaces $\mathbb{X}$ and $\mathbb{Y}$ and $\mathbf{G}\colon \mathbb{X} \to \mathbb{X}$ be an operator used for regularization. In this section we derive stability, convergence and quantitative estimates for solutions of (1.2) under relatively weak assumptions on the operator $\mathbf{G}$ given below. Recall that $g\colon \mathbb{X} \to \mathbb{R}$ is coercive if $g(x) \to \infty$ as $\|x\| \to \infty$.

Assumption 2.1 (Conditions for convergence)**.**

(A1) $\forall z \in \mathbb{X}\colon x \mapsto \langle \mathbf{G}(x), x - z\rangle$ is coercive.

(A2) $\mathbf{G}$ is weak-to-weak continuous.

If $\mathbf{G}(x) \in \partial\mathcal{R}(x)$ is a selection of subgradient of a convex and sub-differentiable function $\mathcal{R}$, then $\langle \mathbf{G}(x), x - z\rangle \geq \mathcal{R}(x) - \mathcal{R}(z)$. Thus Condition 2.1 is satisfied whenever $\mathcal{R}$ is coercive and $\mathbf{G}$ is weak-to-weak continuous. This shows that the conditions are in fact satisfied by subgradients of convex regularizers which are of particular interest in variational regularization [4, 21]. However, as we show later, Condition 2.1 is also satisfied for a large class of operators $\mathbf{G}$ that cannot be written as gradient. As an illustrative example we mention here the class of linear operators which are not self-adjoint.

In the following we make frequent use of the convexity of the data fidelity $\|\mathbf{A}(\cdot) - y^\delta\|^2/2$ which gives the inequality $2\langle \mathbf{A}^*(\mathbf{A}x - y^\delta), z - x\rangle \leq \|\mathbf{A}z - y^\delta\|^2 - \|\mathbf{A}x - y^\delta\|^2$.

2.1 Stability and convergence

We begin our analysis by deriving stability and convergence for solutions of equation (1.2). Conditions for their existence will be discussed later.

Theorem 2.2 (Stability)**.** *Let $\alpha > 0$ and $(y_k)_{k\in\mathbb{N}} \in \mathbb{Y}^{\mathbb{N}}$ with $y_k \to y^\delta$. Then, any sequence $(x_k)_{k\in\mathbb{N}}$ satisfying $\mathbf{T}_\alpha(x_k, y_k) = 0$ has a weakly convergent subsequence. Moreover, any weak cluster point of $(x_k)_{k\in\mathbb{N}}$ is a solution of $\mathbf{T}_\alpha(x, y^\delta) = 0$.*

Proof. We begin by showing that the sequence $(x_k)_k$ is bounded. By definition of x_k and the convexity of the data-fidelity term, for any $z \in \mathbb{X}$,

$$0 = \langle \mathbf{A}^*(\mathbf{A}x_k - y_k) + \alpha\mathbf{G}(x_k), z - x_k\rangle \leq \frac{1}{2}\|\mathbf{A}z - y_k\|^2 - \frac{1}{2}\|\mathbf{A}x_k - y_k\|^2 + \alpha\langle \mathbf{G}(x_k), z - x_k\rangle\,.$$

Hence, $\alpha\langle \mathbf{G}(x_k), x_k - z\rangle \leq \|\mathbf{A}z - y_k\|^2/2$ where $\|\mathbf{A}z - y_k\|$ is bounded. By Condition 2.1, $(x_k)_{k\in\mathbb{N}}$ is bounded and by reflexivity of $\mathbb{X}$ it has a weakly convergent subsequence. If x_+ is a weak cluster point of $(x_k)_{k\in\mathbb{N}}$, then taking the limit in equation (1.2) and using the weak continuity of $\mathbf{T}_\alpha$ shows $\mathbf{T}_\alpha(x_+, y^\delta) = 0$. □

The next goal is to show convergence of the regularized solutions for $\delta \to 0$.

Theorem 2.3 (Convergence)**.** *Let $y \in \operatorname{ran}(\mathbf{A})$, $(y_k)_{k\in\mathbb{N}} \in \mathbb{Y}^{\mathbb{N}}$ satisfy $\|y_k - y\| \leq \delta_k$ for $\delta_k \to 0$ and let $\alpha_k = \alpha(\delta_k)$ with $\lim_k \alpha_k = \lim_k \delta_k^2/\alpha_k = 0$. Any sequence $(x_k)_{k\in\mathbb{N}}$ with $\mathbf{T}_{\alpha_k}(x_k, y_k) = 0$ has at least one weak cluster point. Any such cluster point x_+ is a solution of (1.5), that is $\mathbf{A}x_+ = y$ and $\mathbf{G}(x_+) \in \ker(\mathbf{A})^\perp$. If the solution of (1.5) is unique, then $(x_k)_{k\in\mathbb{N}}$ weakly converges to x_+.*

Proof. Let x^* be any solution of $\mathbf{A}x = y$. By definition of x_k and the convexity of the data-fidelity term we have

$$\begin{aligned} 0 &= \langle \mathbf{A}^*(\mathbf{A}x_k - y_k) + \alpha_k \mathbf{G}(x_k), x^* - x_k \rangle \\ &\leq \frac{1}{2}\|\mathbf{A}x^* - y_k\|^2 - \frac{1}{2}\|\mathbf{A}x_k - y_k\|^2 + \alpha_k \langle \mathbf{G}(x_k), x^* - x_k \rangle \\ &\leq \delta_k^2/2 + \alpha_k \langle \mathbf{G}(x_k), x^* - x_k \rangle . \end{aligned}$$

Hence $2\langle \mathbf{G}(x_k), x_k - x^* \rangle \leq \delta_k^2/\alpha_k$. The choice of α_k and Condition 2.1 show that $(x_k)_k$ is bounded and hence has a weakly convergent subsequence.

Let x_+ be the limit of any weakly convergent subsequence denoted again by $(x_k)_k$. By the weak continuity of $\mathbf{G}$ we have that $(\mathbf{G}(x_k))_k$ is bounded and thus $0 = \lim_k \mathbf{T}_{\alpha_k}(x_k, y_k) = \lim_k \mathbf{A}^*(\mathbf{A}x_k - y_k) + \alpha_k \mathbf{G}(x_k) = \mathbf{A}^*(\mathbf{A}x_+ - y)$. Because $y \in \operatorname{ran}(\mathbf{A})$ this shows that x_+ is a solution of $\mathbf{A}x = y$. Moreover, for any $z_0 \in \ker(\mathbf{A})$ we have

$$\langle -\mathbf{G}(x_+), z_0 \rangle = \lim_k \langle -\mathbf{G}(x_k), z_0 \rangle = \lim_k \langle \mathbf{A}^*(\mathbf{A}x_k - y_k), z_0 \rangle / \alpha_k = 0$$

which is gives $\mathbf{G}(x_+) \in \ker(\mathbf{A})^\perp$. Finally, if the solution of (1.5) is unique, then any subsequence of $(x_k)_k$ has subsequence weakly converging to x_+, which implies that the full sequence weakly converges to x_+. □

Theorems 2.2 and 2.3 show that solutions to the equilibrium equation (1.2) are indeed stable and convergent whenever $\mathbf{G}$ satisfies Condition 2.1.

2.2 Quantitative estimates

The next goal is to derive quantitative results in the form of convergence rates and stability estimates. This will be done in the absolute symmetric Bregman-distance $D_{\mathbf{G}}$, defined by

$$D_{\mathbf{G}}(x, z) = |\langle \mathbf{G}(x) - \mathbf{G}(z), x - z \rangle| ,$$

for $x, z \in \mathbb{X}$.

Recall that $\mathbf{G}$ is called monotone if $\langle \mathbf{G}(z) - \mathbf{G}(x), z - x \rangle \geq 0$ for all $x, z \in \mathbb{X}$; see [17, 20] and [15] for recent developments in the context of inverse problems. If $\mathbf{G}$ is monotone, we call $D_{\mathbf{G}}(x, z) = |\langle \mathbf{G}(x) - \mathbf{G}(z), x - z \rangle| = \langle \mathbf{G}(x) - \mathbf{G}(z), x - z \rangle$ symmetric Bregman-distance. As the name suggests, the absolute symmetric Bregman-distance is indeed symmetric. If

$\mathbf{G} = \nabla\mathcal{R}$ for some convex functional $\mathcal{R}$, then the symmetric Bregman-distance is bounded from below by the classical Bregman-distance. As we will see, the absolute symmetric Bregman-distance is quite natural to consider for (1.2) and allows making use of the defining property efficiently. Before that, let us briefly discuss a possible interpretation of $D_{\mathbf{G}}$.

Remark 2.4 (Interpretation of symmetric Bregman-distance)**.** Assume that $\mathbf{G} = \mathbf{B}$ is a linear, bounded positive semidefinite not necessarily self-adjoint operator. Then, the symmetric Bregman-distance is non-negative and $0 \leq \langle \mathbf{B}x, x\rangle = \langle (\mathbf{B} + \mathbf{B}^*)x, x\rangle/2$ where $(\mathbf{B} + \mathbf{B}^*)/2$ is the projection of $\mathbf{B}$ onto the set of self-adjoint operators. Since $\mathbf{B} + \mathbf{B}^*$ is self-adjoint and positive semidefinite, there exists $\mathbf{C}$ with $\mathbf{C}^*\mathbf{C} = (\mathbf{B} + \mathbf{B}^*)$ and therefore $\langle \mathbf{B}x, x\rangle = \langle \mathbf{C}x, \mathbf{C}x\rangle/2 =: \|x\|^2_{\mathbf{C}}/2$. Thus, the symmetric Bregman-distance of a linear operator $\mathbf{G}$ can be interpreted as the square of a weighted norm $\|\cdot\|^2_{\mathbf{C}}$. In the case that $\mathbf{G}$ is non-linear but smooth, a similar interpretation can be given around a given point. Locally around $z \in \mathbb{X}$, it holds $\langle \mathbf{G}(z+h) - \mathbf{G}(z), z + h - z\rangle \simeq \langle D\mathbf{G}(z)h, h\rangle = \|h\|^2_{D\mathbf{G}(z)}/2$ where the weight $D\mathbf{G}(z)$ depends on z and on the local behavior of $\mathbf{G}$.

We write $\alpha \asymp \delta$ for $\alpha = \alpha(\delta)$ if there exist $C_1, C_2 > 0$ such that $C_1\delta \leq \alpha \leq C_2\delta$ as $\delta \to 0$.

Theorem 2.5 (Convergence rates)**.** *Let $y \in \operatorname{ran}(\mathbf{A})$ and $(y_k)_{k\in\mathbb{N}} \in \mathbb{Y}^{\mathbb{N}}$ be a sequence of data satisfying $\|y_k - y\| \leq \delta_k$ with $\delta_k \to 0$ and $\alpha_k \asymp \delta_k$. Let $(x_k)_k$ satisfy $\mathbf{T}_{\alpha_k}(x_k, y_k) = 0$ and denote by $x_\star$ the weak limit of $(x_k)_k$, possibly after restriction to a subsequence (see Theorem 2.3). Assume there exist $c, \varepsilon > 0$ such that*

$$\forall z \in \mathcal{M}_\varepsilon(x_\star)\colon \quad \langle \mathbf{G}(z), x_\star - z\rangle \leq c\,\|\mathbf{A}z - \mathbf{A}x_\star\| \tag{2.1}$$

$$\mathcal{M}_\varepsilon(x_\star) \coloneqq \{z \in \mathbb{X}\colon \mathbf{G}(z) \in \operatorname{ran}(\mathbf{A}^*) \wedge \langle \mathbf{G}(z), z - x_\star\rangle < \varepsilon\}\,. \tag{2.2}$$

Then, the source condition $\mathbf{G}(x_\star) \in \operatorname{ran}(\mathbf{A}^*)$ is sufficient and necessary for

(1) $\|\mathbf{A}x_k - y_k\| = \mathcal{O}(\delta_k)$ for $k \to \infty$

(2) $D_{\mathbf{G}}(x_k, x_\star) = \mathcal{O}(\delta_k)$ for $k \to \infty$.

Proof. We adapt the proof presented in [13]. Let us first assume that the source condition $\mathbf{G}(x_+) \in \operatorname{ran}(\mathbf{A}^*)$ holds. From the proof of Theorem 2.3 we see $2\alpha_k\langle \mathbf{G}(x_k), x_k - x_+\rangle \leq \delta_k^2$. By assumption on α_k we have $\limsup_k \langle \mathbf{G}(x_k), x_k - x_+\rangle \leq 0$ and hence $x_k \in \mathcal{M}_\varepsilon(x_+)$ for k sufficiently large. For the rest of the proof suppose $x_k \in \mathcal{M}_\varepsilon(x_+)$. By definition we have $D_{\mathbf{G}}(z, x_+) = \langle \mathbf{G}(z) - \mathbf{G}(x_+), z - x_+\rangle + \eta\langle \mathbf{G}(x_+) - \mathbf{G}(z), z - x_+\rangle$, where $\eta = 0$ if $\langle \mathbf{G}(z) - \mathbf{G}(x_+), z - x_+\rangle \geq 0$ and $\eta = 2$ otherwise. By assumption, $\mathbf{G}(x_+) \in \operatorname{ran}(\mathbf{A}^*)$ and $\langle \mathbf{G}(z), x_+ - z\rangle \leq c\,\|\mathbf{A}z - \mathbf{A}x_+\|$ for any $z \in \mathcal{M}_\varepsilon(x_+)$. Hence, $\langle \mathbf{G}(x_+) - \mathbf{G}(z), z - x_+\rangle \leq c\,\|\mathbf{A}(z - x_+)\|$ for some $c \geq 0$ and $D_{\mathbf{G}}(z, x_+) \leq \langle \mathbf{G}(z) - \mathbf{G}(x_+), z - x_+\rangle + c\,\|\mathbf{A}(z - x_+)\|$. In particular, for $z = x_k$,

$$D_{\mathbf{G}}(x_k, x_+) \leq \langle \mathbf{G}(x_k) - \mathbf{G}(x_+), x_k - x_+\rangle + c\,\|\mathbf{A}(x_k - x_+)\| \tag{2.3}$$

By construction of x_k, the convexity of the data-fidelity term, the equality $\mathbf{A}x_+ = y$ and the assumption $\|y - y_k\| \leq \delta_k$ we have

$$\begin{aligned}\frac{1}{2}\|\mathbf{A}x_k - y_k\|^2 + \alpha_k\langle \mathbf{G}(x_k), x_k - x_+\rangle \\ = \frac{1}{2}\|\mathbf{A}x_k - y_k\|^2 + \langle \mathbf{A}^*(\mathbf{A}x_k - y_k), x_+ - x_k\rangle \leq \frac{1}{2}\|\mathbf{A}x_+ - y_k\|^2 \leq \frac{1}{2}\delta_k^2 . \quad (2.4)\end{aligned}$$

By the source condition $\mathbf{G}(x_+) \in \operatorname{ran}(\mathbf{A}^*)$ it holds

$$\langle -\mathbf{G}(x_+), x_k - x_+\rangle = \langle \mathbf{A}^* w, x_+ - x_k\rangle \leq C\|\mathbf{A}x_k - \mathbf{A}x\| \leq C\left(\delta_k + \|\mathbf{A}x_k - y_k\|\right). \quad (2.5)$$

Finally, from (2.3)-(2.5) and Young's product-inequality we obtain

$$\begin{aligned}\frac{1}{2}\|\mathbf{A}x_k - y_k\|^2 + \alpha_k D_{\mathbf{G}}(x_k, x_+) &\leq \frac{1}{2}\delta_k^2 + C_1\alpha_k\delta_k + C_2\alpha_k\|\mathbf{A}x_k - y_k\| \\ &\leq \frac{1}{2}\delta_k^2 + C_1\alpha_k\delta_k + C_3\alpha_k^2 + \frac{1}{4}\|\mathbf{A}x_k - y_k\|^2 ,\end{aligned}$$

for some constants $C_1, C_2 > 0$. The rates (1), (2) then follow with $\alpha_k \asymp \delta_k$.

Assume now conversely that (1), (2) hold and define $w_k := (\mathbf{A}x_k - y_k)/\alpha_k$. Then $(w_k)_{k\in\mathbb{N}}$ is bounded and thus has a weakly convergent subsequence $(w_{k'})_{k'\in\mathbb{N}}$ with weak limit w^+. Along this subsequence we have $-\mathbf{G}(x_+) = \lim_k -\mathbf{G}(x_{k'}) = \lim_k \mathbf{A}^* w_{k'} = \mathbf{A}^* w^+$ and thus $-\mathbf{G}(x_+) \in \operatorname{ran}(\mathbf{A}^*)$. □

Remark 2.6 (Low order rates)**.** For $\mathbf{G}(x_+) \notin \operatorname{ran}(\mathbf{A}^*)$, Theorem 2.5 implies that $D_{\mathbf{G}}(x_k, x_+)$ cannot converge at rate δ_k. However, this does not mean that no convergence rate holds. Instead, if rates hold, these rates have to be slower than δ_k. For example, it is easy to construct examples where the convergence rate is exactly δ_k^ε for some $\varepsilon > 0$.

Remark 2.7 (Convergence rates in the stable case)**.** Note that in general, the condition $\mathbf{G}(x_+) \in \operatorname{ran}(\mathbf{A}^*)$ is stronger than the condition $\mathbf{G}(x_+) \in \ker(\mathbf{A})^\perp$. These conditions do, however, coincide if $\operatorname{ran}(\mathbf{A}^*)$ is closed, in which case $\mathbf{A}^+$ is bounded. Moreover, in this setting, Condition (2.1) holds trivially. As a consequence, in such a case one obtains convergence rates without any further assumptions on x_+, because Theorem 2.3 guarantees $-\mathbf{G}(x_+) \in \operatorname{ran}(\mathbf{A}^*)$. In particular, $\operatorname{ran}(\mathbf{A}^*)$ is closed whenever $\mathbb{X}$ is finite dimensional and hence in this case convergence rates hold.

For any monotone operator, Condition 2.1 holds whenever $-\mathbf{G}(x_+) \in \operatorname{ran}(\mathbf{A}^*)$. Indeed, in this case $\langle \mathbf{G}(z), x_+ - z\rangle = \langle \mathbf{G}(x_+), x_+ - z\rangle - \langle \mathbf{G}(x_+) - \mathbf{G}(z), x_+ - z\rangle \leq \langle \mathbf{A}^* w, x_+ - z\rangle \leq c\|\mathbf{A}(x_+ - z)\|$. In particular the convergence rates (1), (2) hold.

We next derive stability estimates for the class of monotone operators.

Theorem 2.8 (Stability estimates)**.** *Assume* $\mathbf{G}$ *is monotone, let* $\alpha > 0$, $y_1, y_2 \in \mathbb{Y}$ *and* $x_1, x_2 \in \mathbb{X}$ *with* $\mathbf{T}_\alpha(x_1, y_1) = \mathbf{T}_\alpha(x_2, y_2) = 0$. *Then*

(1) $\|\mathbf{A}(x_1 - x_2)\| \leq \|y_1 - y_2\|$,

(2) $D_{\mathbf{G}}(x_1, x_2) \leq 1/(2\alpha) \|y_1 - y_2\|^2$.

Proof. By construction of x_1, x_2 and Young's product inequality,

$$\begin{aligned}
\alpha D_{\mathbf{G}}(x_1, x_2) &= \alpha \langle \mathbf{G}(x_1) - \mathbf{G}(x_2), x_1 - x_2 \rangle \\
&= -\langle \mathbf{A}^*(\mathbf{A}x_1 - y_1) - \mathbf{A}^*(\mathbf{A}x_2 - y_2), x_1 - x_2 \rangle \\
&= -\|\mathbf{A}(x_1 - x_2)\|^2 + \langle y_1 - y_2, \mathbf{A}(x_1 - x_2) \rangle \\
&\leq -\|\mathbf{A}(x_1 - x_2)\|^2 + \|y_1 - y_2\| \, \|\mathbf{A}(x_1 - x_2)\| \\
&\leq -\|\mathbf{A}(x_1 - x_2)\|^2 / 2 + \|y_1 - y_2\|^2 / 2 \,.
\end{aligned}$$

Thus $\|\mathbf{A}(x_1 - x_2)\|^2 / 2 + \alpha D_{\mathbf{G}}(x_1, x_2) \leq \|y_1 - y_2\|^2 / 2$ which yields (1), (2). □

The theoretical analysis provided above suggests that monotone operators $\mathbf{G}$ are reasonable choices for regularizing inverse problems using (1.2).

3 Contractive residual operators

We now turn to a particular case where the regularization operator $\mathbf{G}$ satisfies

$$\forall x, z \in \mathbb{X}\colon \ \|(\mathbf{G} - \mathrm{Id})(x) - (\mathbf{G} - \mathrm{Id})(z)\| \leq L \|x - z\| \,, \tag{3.1}$$

where Id is the identity operator and $L < 1$. In this case, $\mathbf{G} = \mathrm{Id} - \mathbf{N}$ for some contractive residual part $\mathbf{N}$; compare [19, 5]. In what follows we will show that any $\mathbf{G}$ with (3.1) is monotone and satisfies Condition 2.1. Thus, the theory in Section 2 is applicable. We will show, however, that in this case even stronger results hold. We will consider the case of exact regularization where (1.2) is solved exactly (see Section 3.1) as well as the inexact case where where (1.2) is only solved up to a certain accuracy (see Section 3.2) .

Before we continue with our main results, we want to briefly discuss that, in general, variational regularization theory is not applicable. This is because $\mathbf{G}$ might not be of gradient form and hence the equilibrium points are not necessarily critical points of a Tikhonov functional. Indeed, even for $\mathbb{X} = \mathbb{R}^d$ there are simple vector fields $\mathbf{G}$ satisfying equation (3.1) which are not of gradient form. Concrete instances are operators of the form $\mathbf{G} = \mathrm{Id} - \mathbf{N}$ for some non-symmetric $\mathbf{N}$. Consider, for example, a single layer neural network function of the form $\mathbf{G}(x) = x - \mathbf{L}_2 \sigma(\mathbf{L}_1 x + b)$. In order be of gradient form this requires $\mathbf{L}_2 = \mathbf{L}_1^*$ and specific choices for σ, see for example [16]. Thus, application of variational regularization theory requires severe restrictions on the regularization operator. Instead, in this paper we focus on the more general case where $\mathbf{G}$ may not be of gradient form.

Lemma 3.1 (Equivalence of $D_{\mathbf{G}}$ and squared norm)**.** *Let* $\mathbf{G}$ *satisfy* (3.1) *with* $L < 1$*. Then,*

$$\forall x, z \in \mathbb{X}: \quad (1+L)\,\|x-z\|^2 \geq \langle \mathbf{G}(x) - \mathbf{G}(z), x - z\rangle \geq (1-L)\,\|x-z\|^2\,. \tag{3.2}$$

In particular, $\mathbf{G}$ *is monotone, one-to-one and* $\forall x, z \in \mathbb{X}$: $\|\mathbf{G}(x) - \mathbf{G}(z)\| \geq (1-L)\,\|x-z\|$.

Proof. Let $x, z \in \mathbb{X}$. From (3.1) it follows $\|\mathbf{G}(x) - \mathbf{G}(z)\| \leq (1+L)\,\|x-z\|$. With the Cauchy-Schwartz-inequality, we get $\langle \mathbf{G}(x) - \mathbf{G}(z), x - z\rangle \leq \|x-z\|\,\|\mathbf{G}(x) - \mathbf{G}(z)\| \leq (1+L)\,\|x-z\|^2$ which is the left inequality in (3.2). By (3.1) we further have

$$\begin{aligned}
\langle \mathbf{G}(x) - \mathbf{G}(z), x - z\rangle &= \langle (\mathbf{G}(x) - x) - (\mathbf{G}(z) - z), x - z\rangle + \langle x - z, x - z\rangle \\
&\geq -\|x-z\|\,\|(\mathbf{G}(x) - x) - (\mathbf{G}(z) - z)\| + \|x-z\|^2 \\
&\geq -L\,\|x-z\|^2 + \|x-z\|^2\,,
\end{aligned}$$

which is the right inequality in (3.2). This right inequality implies that $\mathbf{G}$ is monotone and one-to-one and with the Cauchy-Schwartz inequality $\|\mathbf{G}(x) - \mathbf{G}(z)\| \geq (1-L)\,\|x-z\|$. □

In the situation of Lemma 3.1 one can even show that $\mathbf{G}$ is bijective with Lipschitz-continuous inverse.

3.1 Exact regularization

We start with the case that (1.2) is solved exactly. Lemma 3.1 shows $\mathbf{G}$ is uniformly monotone and suggests that (1.2) behaves similar to a variational regularization with a uniformly convex regularizer. Indeed, the next theorem shows that we get similar results.

Theorem 3.2 (Exact equilibrium regularization for contractive residuals)**.** *Let* $\mathbf{G}$ *be weakly continuous and satisfy* (3.1) *with* $L < 1$*. Then, the following hold.*

(1) $\forall y^\delta \in \mathbb{Y}\ \forall \alpha > 0$: $\mathbf{T}_\alpha(x, y^\delta) = 0$ *has a unique solution.*

(2) Let $(y_k)_k \in \mathbb{Y}^{\mathbb{N}}$ *converge to* y^δ *and let* $(x_k)_k$ *satisfy* $\mathbf{T}_\alpha(x_k, y_k) = 0$*. Then,* $(x_k)_k$ *norm-converges to the unique solution of* $\mathbf{T}_\alpha(x, y^\delta) = 0$*.*

(3) Let $y_1, y_2 \in \mathbb{Y}$*,* $\alpha > 0$ *and* x_1, x_2 *satisfy* $\mathbf{T}_\alpha(x_1, y_1) = \mathbf{T}_\alpha(x_2, y_2) = 0$*. Then*

(a) $\|\mathbf{A}(x_1 - x_2)\| \leq \|y_1 - y_2\|$,

(b) $\langle \mathbf{G}(x_1) - \mathbf{G}(x_2), x_1 - x_2\rangle \leq 1/(2\alpha)\,\|y_1 - y_2\|^2$,

(c) $\|x_1 - x_2\| \leq \sqrt{1/(2\alpha(1-L))}\,\|y_1 - y_2\|$.

(4) Let $(y_k)_k \in \mathbb{Y}^{\mathbb{N}}$ *be a sequence converging to* $y \in \operatorname{ran}(\mathbf{A})$ *with* $\|y - y_k\| \leq \delta_k$*. Assume* $\alpha = \alpha(\delta)$ *is such that for* $\alpha_k = \alpha(\delta_k)$ *we have* $\lim_k \alpha_k = \lim_k \delta_k^2/\alpha_k = 0$*. For* $k \in \mathbb{N}$*,*

let x_k be the unique solution of $\mathbf{T}_{\alpha_k}(x, y_k) = 0$. Then, the sequence $(x_k)_k$ converges in norm to the unique solution x_+ of $\mathbf{A}x = y$ with $\mathbf{G}(x) \in \ker(\mathbf{A})^\perp$.

(5) In the setting of (4) with $\alpha_k \asymp \delta_k$, the source condition $-\mathbf{G}(x_+) \in \operatorname{ran}(\mathbf{A}^*)$ is necessary and sufficient for the rates

(a) $\|\mathbf{A}x_k - y_k\| = \mathcal{O}(\delta_k)$,

(b) $\langle \mathbf{G}(x_k) - \mathbf{G}(x_+), x_k - x_+ \rangle = \mathcal{O}(\delta_k)$,

(c) $\|x_k - x_+\| = \mathcal{O}(\sqrt{\delta_k})$.

Proof. The proof extends results derived above. To show (1), let $\alpha, \beta > 0$, $y^\delta \in \mathbb{Y}$ and define $\mathbf{F}(x) \coloneqq x - \beta \left(\mathbf{A}^*(\mathbf{A}x - y^\delta) + \alpha \mathbf{G}(x)\right)$. For $x, z \in \mathbb{X}$ we have

$$\begin{aligned}\|\mathbf{F}(x) - \mathbf{F}(z)\| &\leq \|((1-\beta\alpha)\operatorname{Id} - \beta\mathbf{A}^*\mathbf{A})(x-z)\| + \alpha\beta \|(\mathbf{G} - \operatorname{Id})(x) - (\mathbf{G} - \operatorname{Id})(z)\| \\ &\leq \|(1-\beta\alpha)\operatorname{Id} - \beta\mathbf{A}^*\mathbf{A}\| \|x - z\| + \alpha\beta L \|x - z\| \ .\end{aligned}$$

With $\beta \leq 1/(\|\mathbf{A}\|^2 + \alpha)$ we have $\|(1-\beta\alpha)\operatorname{Id} - \beta\mathbf{A}^*\mathbf{A}\| \leq (1 - \beta\alpha)$ and hence $\mathbf{F}$ is Lipschitz-continuous with constant $\gamma \leq (1-\beta\alpha) + L\alpha\beta = 1 - \alpha\beta(1-L) < 1$. Uniqueness and existence of a fixed-point of $\mathbf{F}$ follows from Banach's fixed point theorem which is (1). Lemma 3.1 shows $\langle \mathbf{G}(x), x - z \rangle \geq (1-L)\|x - z\|^2 + \langle \mathbf{G}(z), x - z \rangle$ and thus Condition 2.1 is satisfied whenever $L < 1$. Theorem 2.2 thus gives weak stability. The stability estimates in (3) are a consequence of Theorem 2.8 and Lemma 3.1 and in particular, norm-stability stated in (2) holds.

For proving (4), Theorem 2.3 gives the existence of a weakly convergent subsequence. We next show that the solution of the limiting problem (1.5) is unique. To that end let $\tilde{x}$ satisfy $\mathbf{A}\tilde{x} = y$ and $-\mathbf{G}(\tilde{x}) \in \ker(\mathbf{A})^\perp$. Then $x_+ - \tilde{x} \in \ker(\mathbf{A})$, $\mathbf{G}(x_+) - \mathbf{G}(\tilde{x}) \in \ker(\mathbf{A})^\perp$ and thus $\langle \mathbf{G}(x_+) - \mathbf{G}(\tilde{x}), x_+ - \tilde{x} \rangle = 0$. By Lemma 3.1, $0 = \langle \mathbf{G}(x_+) - \mathbf{G}(\tilde{x}), x_+ - \tilde{x} \rangle \geq (1-L)\|x_+ - \tilde{x}\|^2$ and thus $\tilde{x} = x_+$. Following the proof of Theorem 2.3 we find $2\alpha_k \langle \mathbf{G}(x_k), x_k - x_+ \rangle \leq \delta_k^2$. By Lemma 3.1, $(1-L)\|x_k - x_+\|^2 \leq \langle \mathbf{G}(x_k) - \mathbf{G}(x_+), x_k - x_+ \rangle \leq \delta_k^2/(2\alpha_k) + \langle \mathbf{G}(x_+), x_+ - x_k \rangle$. Because $(x_k)_k$ converges weakly to x_+ we get the norm convergence which completes the proof of (4). Finally, (5) follows with Theorem 2.5 and Lemma 3.1. □

The stability estimates in Theorem 3.2 demonstrate that (1.2) is stable independent of the noise-distribution, as empirically observed in [5]. More precisely, for $\alpha > 0$ fixed, the reconstruction operator defined by (3.1) is injective and Lipschitz-continuous. Further note that we derived stability with respect the norm and the symmetric Bregman-distance. Opposed to the norm estimate, the estimate in the symmetric Bregman-distance is independent of the constant L. Hence, for L close to 1 the stability estimate in the norm has a potentially large constant, whereas the stability estimate for the symmetric Bregman-distance does not depend on this factor. Additionally, it suggests that for $L = 1$ the derivation of stability estimates requires further assumptions.

Remark 3.3 (Interpretation of the limiting problem)**.** As we noted above, any $\mathbf{G}$ satisfying (3.1) is bijective. Hence the limiting problem (1.5) can be rewritten as

$$\text{Find } x \text{ with} \quad x \in \mathbf{G}^{-1}(-\ker(\mathbf{A})^{\perp}) \cap \mathbf{A}^{-1}(y)\,. \tag{3.3}$$

In particular, $\mathbf{G}^{-1}\colon \ker(\mathbf{A})^{\perp} \to \mathbb{X}$ is a parametrization of a manifold on which the solutions must lie. Taking the intersection of this manifold with the set of solution will then give the desired solution. Figure 3.1 provides an illustration of this.

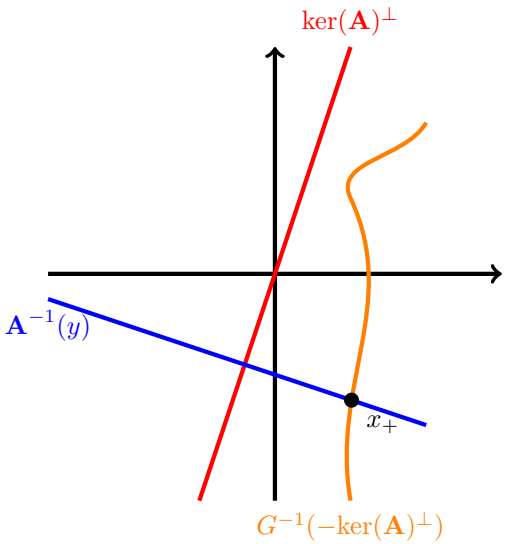

Figure 3.1: Interpretation of the limiting problem. According to (3.3) the limiting problem chooses the solution of $\mathbf{A}x = y$ with $x \in \mathbf{G}^{-1}(-\ker(\mathbf{A})^{\perp})$. Hence $\mathbf{G}^{-1}\colon \ker(\mathbf{A})^{\perp} \to \mathbb{X}$ might be interpreted as parametrization of the manifold of desired solutions.

Remark 3.4 (Implicit nullspace networks)**.** Consider the form $\mathbf{G} = \mathrm{Id} - \mathbf{P}_{\ker(\mathbf{A})}\mathbf{N}$ where $\mathbf{P}_{\ker(\mathbf{A})}$ is the projection on $\ker(\mathbf{A})$ and $\mathbf{N}\colon \mathbb{X} \to \mathbb{X}$ is Lipschitz. Then (1.2) is satisfied if and only if $\mathbf{A}^*(\mathbf{A}x - y^\delta) + \alpha \mathbf{P}_{\ker(\mathbf{A})^\perp} x = 0$ and $\mathbf{P}_{\ker(\mathbf{A})}(x - \mathbf{N}(x)) = 0$. With the decomposition $x = x_1 + x_0$ with $x_1 \in \ker(\mathbf{A})^{\perp}$ and $x_0 \in \ker(\mathbf{A})$ we find that $x_1 = (\mathbf{A}^*\mathbf{A} + \alpha\,\mathrm{Id})^{-1}\mathbf{A}^* y^\delta$ is defined by classical Tikhonov regularization (variational regularization with the regularizer $\mathcal{R}(x) = \|x\|^2/2$). The component x_0 is then implicitly defined by the equation $0 = x_0 - \mathbf{P}_{\ker(\mathbf{A})}\mathbf{N}(x_1 + x_0) = (\mathrm{Id} - \mathbf{P}_{\ker(\mathbf{A})}\mathbf{N})x_0 - \mathbf{P}_{\ker(\mathbf{A})}\mathbf{N}x_1$ which has solution $x_0 = \mathbf{G}^{-1}(x_1) - x_1$. Hence, the solution operators defined by the equilibrium equation (1.2) takes to form

$$\mathbf{R}_\alpha = \mathbf{G}^{-1} \circ \mathbf{B}_\alpha = \mathbf{G}^{-1} \circ ((\mathbf{A}^*\mathbf{A} + \alpha\,\mathrm{Id})^{-1}\mathbf{A}^*)\,. \tag{3.4}$$

This is an instance of the nullspace networks proposed in [22]. In particular, the null space network is defined implicitly by solving (1.2).

3.2 Inexact regularization

We next study the case where the equilibrium equation (1.2) is only solved up to a certain accuracy. We analyze the specific case that the source condition is satisfied.

Lemma 3.5 (Finite accuracy)**.** *Let* $\mathbf{G}$ *be weakly continuous and satisfy* (3.1) *with* $L < 1$*. Further, let* $(y_k)_k \in \mathbb{Y}^{\mathbb{N}}$ *be converge to* $y = \mathbf{A}x_\dagger$ *with* $\|y - y_k\| \leq \delta_k \to 0$*. Assume that* $(x_k)_k$ *satisfies* $\|\mathbf{T}_{\alpha_k}(x_k, y_k)\| \leq \varepsilon_k$ *for some* $\varepsilon_k > 0$ *and consider the parameter choice* $\alpha_k \asymp \delta_k$*. Then, if the source condition* $-\mathbf{G}(x_\dagger) \in \operatorname{ran}(\mathbf{A}^*)$ *holds, for constants* $C_1, C_2 > 0$ *we have*

$$\|x_k - x_\dagger\| \leq C_1\sqrt{\delta_k} + C_2\frac{\varepsilon_k}{\delta_k(1-L)}\,. \tag{3.5}$$

Proof. Denote $\mathbf{T}_k = \mathbf{T}_{\alpha_k}(\cdot, y_k)$ and let x_k, x_k^* satisfy $\|\mathbf{T}_k(x_k)\| \leq \varepsilon_k$ and $\mathbf{T}_k(x_k^*) = 0$. By Lemma 3.1, $\alpha_k(1-L)\|x_k - x_k^*\| \leq \|\mathbf{T}_k(x_1) - \mathbf{T}_k(x_k^*)\| \leq \varepsilon_k$. With the convergence rates in Theorem 3.2 this shows (3.5). □

Remark 3.6 (Semi-convergence to solution)**.** Suppose (1.2) is solved up to accuracy $\varepsilon > 0$ independent of the noise-level δ. Then, Lemma 3.5 then gives the bound $\|x_k - x_\dagger\| \leq C_1\sqrt{\delta_k} + C_2\varepsilon/((1-L)\delta_k)$. Hence, for fixed accuracy ε one cannot expect that $(x_k)_k$ converges to $x_\dagger$. Instead, one can expect some form of semi-convergence behavior. Note that the estimate includes the term $\varepsilon/(1-L)$ which means that for L close to 1 the approximation quality might be significantly reduced.

Theorem 3.7 (Early stopping)**.** *Consider the setting of Lemma 3.5 where* $\|\mathbf{T}_{\alpha_k}(x_k, y_k)\| \leq \delta_k\eta_k$ *with* $(\eta_k)_k \to 0$*. Then, as* $k \to \infty$*,*

1. $\|x_k - x_\dagger\| = \mathcal{O}(\sqrt{\delta_k} + \eta_k)$
2. $\langle \mathbf{G}(x_k) - \mathbf{G}(x_\dagger), x_k - x_\dagger\rangle = \mathcal{O}(\delta_k + \sqrt{\delta_k}\eta_k + \eta_k^2)$
3. $\|\mathbf{A}x_k - y_k\| = \mathcal{O}(\delta_k + \eta_k\sqrt{\delta_k})$.

Moreover, if $\eta_k = \sqrt{\delta_k}$ *then* x_k *can be constructed* $\mathcal{O}(\log(\delta_k)/\left((L-1)\delta_k\right))$ *iterations using a fixed point iteration applied to* $\mathbf{F}_k(x) = x - \beta(\mathbf{A}^*(\mathbf{A}x - y_k) + \alpha\mathbf{G}(x))$.

Proof. The rates for $\|x_k - x_\dagger\|$ and for the symmetric Bregman-distance follow from Lemmas and 3.5. Further, with $\mathbf{T}_k = \mathbf{T}_{\alpha_k}(\cdot, y_k)$ we have $\langle \mathbf{T}_k(x_k) - \mathbf{T}_k(x_k^*), x_k - x_k^*\rangle \leq C\delta_k\eta_k^2$. By Lemma 3.1, $\|\mathbf{A}(x_k - x_k^*)\|^2 \leq C\delta_k\eta_k^2$ and hence $\|\mathbf{A}x_k - y_k\| \leq \tilde{C}(\delta_k + \eta_k\sqrt{\delta_k})$ as desired. To show the last claim we assume without loss of generality that $\|\mathbf{A}\| = 1$. According to the proof of Theorem 3.2 the mapping $\mathbf{F}_k$ has contraction constant $\gamma_k = (1 + \alpha_k L)\,/\,(1 + \alpha_k)$. By Banach's fixed point theorem, the iterates $x_k^n := \mathbf{F}_k(x_k^{n-1})$ satisfy $\|x_k^* - x_k^n\| \leq C\gamma_k^n$. It is thus sufficient to have $C\gamma_k^n \leq \sqrt{\eta_k}\delta_k$. Rearranging this inequality we find that we need on the order of $\log(\delta_k)/\log(\gamma_k)$ iterations. Since $\log(\gamma_k) = \log(1 + \alpha_k L) - \log(1 + \alpha_k) \asymp (L-1)\alpha_k \asymp (L-1)\delta_k$, we thus get $\mathcal{O}(\log(\delta_k)/\left((L-1)\delta_k\right))$ iterations. □

Note that in the context of Theorem 3.7 stability is clear, since in this case the reconstruction is given by a finite number of fixed point iterations for a Lipschitz-continuous mapping. However, using a finite number of iterations, the regularized solution depends on the initial value and thus we loose uniqueness. Further note that the same proof can be given for a sequence $(\eta_k)_k$ with $\eta_k \geq \eta_* > 0$. In this case one might not obtain the solution x_+ as characterized by Theorem 3.2 but some element close to this solution where the closeness depends on η_*. Additionally, the rates in the data-fidelity term are only of order $\sqrt{\delta}$.

3.3 Limitations

For the rest of this section we assume that $\mathbb{S} \subseteq \mathbb{X}$ is a set of signals of interest. We assume further that $\mathbf{G}$ is weakly continuous and satisfies (3.1). The set $\mathbb{S}$ could, for example, be a set of natural images, such as a set of natural landscapes in a deblurring task. The solution operator for (1.2) is denoted by $\mathbf{R}_\alpha \colon \mathbb{Y} \to \mathbb{X} \colon y \mapsto \mathbf{R}_\alpha(y)$ and defined by $\mathbf{A}^*(\mathbf{A}(\mathbf{R}_\alpha(y)) - y) + \alpha \mathbf{G}(\mathbf{R}_\alpha(y)) = 0$. For $x \in \mathbb{S}$ we refer to $\mathbf{A}x$ as noise-free data and $y^\delta \in \mathbb{Y}$ with $\|y^\delta - \mathbf{A}x\| \leq \delta$ as noisy data. Further, for $y \in \operatorname{ran}(\mathbf{A})$ we denote by $x_+ = \mathbf{R}_0(y)$ the unique solution of the limiting problem (1.5) (see Theorem 3.2). For any closed subspace $\mathbb{V} \subseteq \mathbb{X}$ we denote by $\mathbf{P}_\mathbb{V}$ the orthogonal projection onto $\mathbb{V}$. The results so far positively answer the questions of stability and the convergence as $\alpha = \alpha(\delta) \to 0$ of solutions of the equilibrium equation (1.2). In this subsection we aim to address limitations on the reconstruction error for fixed α. The derived lower bounds highlight the need for our asymptotic analysis and motivate constructing the regularization operator $\mathbf{G}$ based on the limiting problem (1.5) as will be done below.

In the context of equilibrium methods learning $\mathbf{G}$ aims at fitting regularized solutions $\mathbf{R}_\alpha(y^\delta)$ as close as possible to the underlying signal $x \in \mathbb{S}$ that generated the data y^δ. Ideally, if exact data y is available, the regularized solution should coincide with x. Recall that according to Theorem 3.2 for any $x \in \mathbb{X}$ the following holds: If $\|y - y_k\| \leq \delta_k$ and $\alpha_k, \delta_k, \delta_k^2/\alpha_k \to 0$, then $\|\mathbf{R}_{\alpha_k}(x_k) - \mathbf{R}_0(\mathbf{A}x)\| \to 0$. This justifies the following definition.

Definition 3.8 (**G**-Recoverability)**.** An element $x \in \mathbb{X}$ is called **G**-recoverable if $\mathbf{R}_0\mathbf{A}x = x$. The signal-class $\mathbb{S} \subseteq \mathbb{X}$ is called **G**-recoverable if all elements in $\mathbb{S}$ are **G**-recoverable.

The following theorem gives a characterization of **G**-recoverability and provides a necessary condition on the interplay of $\mathbb{S}$ and $\ker(\mathbf{A})$ for this to be possible.

Theorem 3.9 (**G**-Recoverability)**.** *For any $\mathbb{S} \subseteq \mathbb{X}$ the following hold.*

*(1) $\mathbb{S}$ is **G**-recoverable if and only if $\mathbf{G}(\mathbb{S}) \subseteq \ker(\mathbf{A})^\perp$.*

(2) $\mathbf{G}(\mathbb{S}) \subseteq \ker(\mathbf{A})^\perp \Rightarrow L_* \coloneqq \sup_{x_1, x_2 \in \mathbb{S}} \|\mathbf{P}_{\ker(\mathbf{A})}(x_1 - x_2)\| / \|x_1 - x_2\| < 1$.

*(3) If $\mathbb{V}$ is a closed linear subspace with $\mathbb{S} \subseteq \mathbb{V}$ and $\|\mathbf{P}_{\ker(\mathbf{A})} P_\mathbb{V}\| < 1$, then there is a regularization operator **G** with (3.1) and $\mathbf{G}(\mathbb{S}) \subseteq \ker(\mathbf{A})^\perp$.*

(4) $\mathbf{G}(x) \notin \ker(\mathbf{A})^\perp \Rightarrow \|\mathbf{R}_0(\mathbf{A}x) - x\| \geq \|\mathbf{P}_{\ker(\mathbf{A})}\mathbf{G}(x)\|/(1+L)$.

Proof. Item (1) is an immediate consequence of Theorem 3.2. If $\mathbf{G}(\mathbb{S}) \subseteq \ker(\mathbf{A})^\perp$ and $x_1, x_2 \in \mathbb{S}$, then $L\,\|x_1 - x_2\| \geq \|(\mathbf{G} - \mathrm{Id})(x_1) - (\mathbf{G} - \mathrm{Id})(x_2)\| \geq \left\|\mathbf{P}_{\ker(\mathbf{A})}(x_1 - x_2)\right\|$ which gives (2). If $\mathbb{S} \subseteq \mathbb{V}$ where $\mathbb{V}$ is a closed subspace of $\mathbb{X}$ with $\left\|\mathbf{P}_{\ker(\mathbf{A})}P_{\mathbb{V}}\right\| < 1$, then $\mathbf{G} = \mathrm{Id} - \mathbf{P}_{\ker(\mathbf{A})}\mathbf{P}_{\mathbb{V}}$ satisfies (3.1) and $\mathbf{G}(\mathbb{S}) \subseteq \ker(\mathbf{A})^\perp$, which gives (3). Finally (4) follows from the Lipschitz-continuity of $\mathbf{G}$. It holds $(L+1)\,\|\mathbf{R}_0(\mathbf{A}x) - x\| \geq \|\mathbf{G}(\mathbf{R}_0(\mathbf{A}x)) - \mathbf{G}(x)\| \geq \|\mathbf{P}_{\ker(\mathbf{A})}\mathbf{G}(x)\|$ because $\mathbf{G}(\mathbf{R}_0(\mathbf{A}x)) \in \ker(\mathbf{A})^\perp$. □

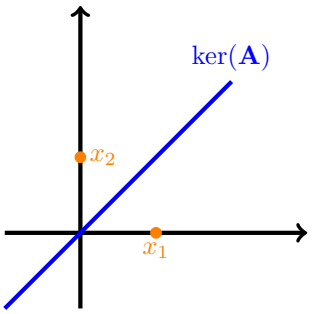

Figure 3.2: Subspace condition may not be satisfied even in simple cases, if the underlying signal-class $\mathbb{S}$ is close to being full-dimensional.

Remark 3.10. Condition (2) gives a lower bound on the contraction constant $L \geq L_*$. For any $\mathbf{G}$ satisfying (3.1) with $L < L_*$ there is at least one element $x \in \mathbb{S}$ which is not recovered by the limiting problem $\mathbf{R}_0$. Notably, this constant solely depends on $\mathbb{S}$ and $\ker(\mathbf{A})$ and thus can be estimated before choosing $\mathbf{G}$. While the subspace condition in Item (3) is sufficient for the existence of a regularization operator for $\mathbb{S}$ it may fail even in simple examples. Consider for example Figure 3.2, where the set $\mathbb{S} \subseteq \mathbb{R}^2$ consists of only two points, but the smallest subspace $\mathbb{V}$ containing $\mathbb{S}$ is the whole space. It is, however, clear that the condition $\|\mathbf{P}_{\ker(\mathbf{A})}(x_1 - x_2)\| \leq L\|x_1 - x_2\|$ can be satisfied, with reasonably small L. This example shows, that the assumption $\mathbb{S} \subseteq \mathbb{V}$ for some closed subspace $\mathbb{V}$ satisfying $\|\mathbf{P}_{\ker(\mathbf{A})}\mathbf{P}_{\mathbb{V}}\| < 1$ is in general too strict and suggests that non-linear operators $\mathbf{G}$ should be chosen.

Item (4) in Theorem 3.9 gives a bound on how well a non-recoverable signal x can be approximated by solutions of (1.2) for asymptotic case as $\delta \to 0$. We next consider the non-asymptotic case.

Corollary 3.11 (Exact recovery is not possible)**.** *Let $x \in \mathbb{X}$ and $\alpha > 0$. Then, x is a solution of* (1.2) *with data $y = \mathbf{A}x$ if and only if $\mathbf{G}(x) = 0$. In particular, at most one element $x \in \mathbb{S}$ can be a solution of equation* (1.2) *with exact data.*

Proof. The first part is clear. The second part follows, because, according to Lemma 3.1, $\mathbf{G}$ is one-to-one. □

Corollary 3.11 shows that exact recovery using (1.2) and noise-free data is possible for at most one element. If instead we consider the noisy case where $y^\delta = \mathbf{A}x + z_\delta$, then the condition is given by $\mathbf{A}^* z_\delta = -\alpha \mathbf{G}(x)$. Clearly, this is only satisfied for specific values of z_δ and hence exact recovery is in general not possible for arbitrary noise. Instead, we will always incur some error depending on x and α.

Proposition 3.12 (Lower bound)**.** *For any $x \in \mathbb{X}$ and $\alpha > 0$ we have*

$$\|x - \mathbf{R}_\alpha \mathbf{A}x\| \geq \frac{\alpha}{\|\mathbf{A}\|^2 + \alpha(1+L)} \|\mathbf{G}(x)\| \,.$$

Proof. The mapping $x \mapsto \mathbf{T}_\alpha(x, y)$ is Lipschitz with $\mathrm{Lip}(\mathbf{T}_\alpha) \leq \|\mathbf{A}\|^2 + \alpha(1+L)$. Hence, $\big(\|\mathbf{A}\|^2 + \alpha(1+L)\big) \|x - \mathbf{R}_\alpha \mathbf{A}x\| \geq \|\mathbf{T}_\alpha(\mathbf{R}_\alpha \mathbf{A}x, \mathbf{A}x) - \mathbf{T}_\alpha(x, \mathbf{A}x)\| = \alpha \|\mathbf{G}(x)\|$. □

Proposition 3.12 uses the assumption that regularized solutions solve (1.2) exactly. In practical applications one typically only has access to near-equilibrium points. Consequently, these lower bounds may be improved. A reasonable question is then how to choose the accuracy of nearly solving the equilibrium equation (1.2) in order to get the best possible results. While we cannot answer this question at this point in time, we briefly discuss a simple heuristic for the choosing accuracy. If $\|y^\delta - \mathbf{A}x\| \leq \delta$, then $\|\mathbf{T}_\alpha(x, y^\delta)\| \leq C(\delta + \alpha)$ which suggests that one can stop at an accuracy $\mathcal{O}(\delta + \alpha)$. Hence, with the choice $\alpha \asymp \delta$ (motivated by Theorem 3.2) we find, that an accuracy on the order of the noise-level δ should be chosen. Interestingly, this differs from the choice according to Theorem 3.7.

3.4 A novel loss-function

Next we address the issue of constructing a regularization operator $\mathbf{G}$. For that purpose we assume that there is some underlying distribution $x \sim \pi_{\mathbb{S}}$ supported some set $\mathbb{S} \subseteq \mathbb{X}$ according to which the signal $x \in \mathbb{S}$ of interest are drawn. Existing equilibrium methods take $\mathbf{G}$ as (near) minimizers of

$$\mathrm{MSE}_\alpha(\mathbf{G}) \coloneqq \mathbb{E}_{x, z_\delta} \left\| x - \mathbf{R}_\alpha(\mathbf{A}x + z_\delta; \mathbf{G}) \right\|^2 , \tag{3.6}$$

where the dependence of $\mathbf{R}_\alpha$ on the regularization operator $\mathbf{G}$ is now made explicit. Below we discuss that the analysis in Section 3 guides the design of alternative loss-functions.

We have seen above that the limiting solutions $\mathbf{R}_0(\mathbf{A}x)$ of the equilibrium points $\mathbf{R}_\alpha(y^\delta)$ are uniquely determined. This means that in order to get regularized solutions closed to a desired signal $x \in \mathbb{S}$ it is sufficient to enforce $\mathbf{R}_0(\mathbf{A}x) = x$. According to Theorem 3.2, we have to guarantee that $\mathbf{G}(x) \in \ker(\mathbf{A})^\perp$. This in turn is equivalent to $\|\mathbf{P}_{\ker(\mathbf{A})} \mathbf{G}(x)\|^2 = 0$. These considerations can be translated to a loss function

$$\mathcal{L}_0(\mathbf{G}) \coloneqq \mathbb{E}_x \|\mathbf{P}_{\ker(\mathbf{A})} \mathbf{G}(x)\|^2 \,, \tag{3.7}$$

that is minimized over a class of of weakly continuous mappings $\mathbf{G}\colon \mathbb{X} \to \mathbb{X}$ that satisfy (3.1).

Remark 3.13. Opposed to (3.6), the proposed loss function (3.7) does not depend on the regularized reconstruction $\mathbf{R}_\alpha$ and as a consequence can avoid costly iterative algorithms for the fixed point computations. Additionally, it enforces the equilibrium point method to actually approximate desired solutions in $\mathbb{S}$; compare Theorem 3.9. Finally, the loss-function is independent of the noise-model. This means, that we can avoid choosing a specific noise-model during training which for practical applications is typically not known anyway.

The loss function (3.7) only restricts the behavior of $\mathbf{P}_{\ker(\mathbf{A})}\mathbf{G}$ but not $\mathbf{P}_{\ker(\mathbf{A})^\perp}\mathbf{G}$. In order to get appropriate behavior on the latter one can simply add a regularization term. In view of Lemma 3.12 this could for example be done by restricting the norm of $\mathbf{G}(x)$ thus arriving at

$$\mathcal{L}_\lambda(\mathbf{G}) \coloneqq \mathbb{E}_x \left[\|\mathbf{P}_{\ker(\mathbf{A})}\mathbf{G}(x)\|^2 + \lambda \|\mathbf{G}(x)\|^2 \right] , \tag{3.8}$$

where $\lambda \geq 0$ is an appropriately chosen trade-off parameter.

We now want to briefly discuss the connection of (3.8) and (3.6) for $x \in \mathbb{S}$. With the data $y = \mathbf{A}x$ and $\|y^\delta - y\| \leq \delta$ we have

$$\begin{aligned} &\langle \mathbf{T}_\alpha(\mathbf{R}_\alpha(y), y) - \mathbf{T}_\alpha(x, y), \mathbf{R}_\alpha y) - x \rangle \\ &\qquad \geq \alpha \langle \mathbf{G}(\mathbf{R}_\alpha(y)) - \mathbf{G}(x), \mathbf{R}_\alpha(y) - x \rangle \geq \alpha(1 - L)\|x - \mathbf{R}_\alpha(y)\|^2 . \end{aligned}$$

Using the Cauchy-Schwartz inequality and the Lipschitz-continuity of $\mathbf{T}_\alpha(\cdot, y)$ and the identity $\mathbf{T}_\alpha(\mathbf{R}_\alpha(y), y) = 0$ we find $\|\mathbf{T}_\alpha(\mathbf{R}_\alpha(y), y) - \mathbf{T}_\alpha(x, y)\| \geq \alpha(1 - L)\|x - \mathbf{R}_\alpha(y)\|$ and $\|\mathbf{T}_\alpha(\mathbf{R}_\alpha(y), y) - \mathbf{T}_\alpha(x, y)\| = \|\mathbf{T}_\alpha(x, y)\|$. Moreover, since $y = \mathbf{A}x$ we get $\mathbf{T}_\alpha(x, y) = \alpha\mathbf{G}(x)$ and thus $\|\mathbf{G}(x)\| \geq (1 - L)\|x - \mathbf{R}_\alpha(y)\|$. Thus, we find that the loss in (3.8) is a combination of a term enforcing recoverability conditions and a term which acts as a surrogate for (3.6) with exact data.

Example 3.14 (Linear case)**.** Consider the form $\mathbf{G} = \mathrm{Id} - \mathbf{W}$ where $\mathbf{W}$ is linear and bounded and that $\mathbb{S}$ satisfies $\mathbb{S} \subseteq \mathbb{V}$ for a close subspace $\mathbb{V}$ with $\|\mathbf{P}_{\ker(\mathbf{A})}\mathbf{P}_\mathbb{V}\| < 1$. This guarantees $\mathbf{G}$-recoverability of $\mathbb{S}$ by the explicit choice in Theorem 3.9. Assume for the time being, that we do not know about this explicit construction and consider minimizing (3.7). Then

$$\mathbb{E}_x \left\| \mathbf{P}_{\ker(\mathbf{A})}(\mathbf{G}x) \right\|^2 = \mathbb{E}_x \left\| \mathbf{P}_{\ker(\mathbf{A})}\mathbf{P}_\mathbb{V} x - \mathbf{P}_{\ker(\mathbf{A})}\mathbf{W}x \right\|^2 .$$

For this to be minimized we have $\mathbf{P}_{\ker(\mathbf{A})}\mathbf{W}x = \mathbf{P}_{\ker(\mathbf{A})}\mathbf{P}_\mathbb{V}x$ for all $x \in \mathbb{S}$. Thus, the loss-function (3.7) yields a construction similar to the one in Theorem 3.9. It might differ in the chosen subspace in the sense that a projection onto $\mathbb{V}_0 \subseteq \mathbb{V}$ might be selected, but its restriction on $\mathbb{S}$ has to be given by $\mathbf{P}_{\ker(\mathbf{A})}\mathbf{P}_\mathbb{V}$. This shows, that the proposed loss-function indeed enforces recoverability of the set $\mathbb{S}$, compare Theorem 3.9.

In the context of Example 3.14 it is also interesting to note, that $\mathbf{G} = \mathrm{Id} - \mathbf{P}_{\ker(\mathbf{A})}\mathbf{P}_{\mathbb{V}}$ is in general not self-adjoint, since the projections $\mathbf{P}_{\ker(\mathbf{A})}$ and $\mathbf{P}_{\mathbb{V}}$ in general do not commute. In particular, $\mathbf{G}$ is not of gradient form and thus the methods presented in this paper are different from variational regularization even in the linear case using (3.7).

4 Numerical simulations

In this section we present numerical results supporting the theory derived in this paper using simple practical examples. While extensive numerical simulations, comparison and further investigations of the proposed loss (3.7) are clearly of interest these are out of the scope of this paper.

4.1 General setup

We derive $\mathbb{S}$ from the the MNIST dataset [8] which consists of 70 000 gray-scale images of digits of size 28×28. We choose 10 000 of these digits for training and another 10 000 for testing. The training-set is used exclusively for constructing the operator $\mathbf{G}$ whereas the test-set is used to perform the numerical simulations. Throughout we restrict to linear $\mathbf{G}$ chosen to minimize the proposed loss (3.7) which is done by the explicit construction discussed in Example 3.14. To enforce $\mathbb{S} \subseteq \mathbb{V}$ we project the MNIST dataset $\mathbb{S}_0$ onto an appropriate subspace $\mathbb{V}$ and set $\mathbb{S} = \mathbf{P}_{\mathbb{V}}(\mathbb{S}_0)$. To this end, we perform a principal component analysis of the training-set and choose $q \in \mathbb{N}$ such that the first q principal components span a subspace $\mathbb{V}$ with $\|\mathbf{P}_{\ker(\mathbf{A})}\mathbf{P}_{\mathbb{V}}\| < 1$. We set $\mathbf{G} = \mathrm{Id} - \mathbf{P}_{\ker(\mathbf{A})}\mathbf{P}_{\mathbb{V}}$ where the evaluation of $\mathbf{P}_{\ker(\mathbf{A})}$ is discussed below for different choices of $\mathbf{A}$.

Remark 4.1. Choosing $\mathbf{G}$ linear allows for easier testing of the theory. However, this restriction has a notable practical caveat. Following Remark 3.3 and the discussion in Section 3.3, an ideal choice of $\mathbf{G}$ would essentially correspond to a parametrization of the given signal-class using $\mathbf{G}^{-1}$. With linear $\mathbf{G}$ we parametrize linear manifolds. It is, however, also clear that even simple signal-classes such as MNIST are highly non-linear and as a consequence any linear $\mathbf{G}$ is in a sense sub-optimal. Nevertheless as the examples below demonstrate that even linear operators perform well when properly trained.

Noise-free data $\mathbf{A}x$ are computed for $x \in \mathbb{S}$ and noisy data y^δ are constructed by adding Gaussian noise with varying standard-deviations to get a desired noise-level $\delta = \|y^\delta - \mathbf{A}x\|$. We choose $\alpha(\delta) = \delta$ and solve (1.2) with the built-in PyTorch [14] function `torch.linalg.solve` to obtain regularized solutions $x_\alpha^\delta = \mathbf{R}_{\alpha(\delta)}(y^\delta)$. Due to practical limitations we solve equation (1.2) only up to a certain accuracy; compare Lemma 3.5 and Remark 3.6. According to Theorem 3.2, the limiting solution $x_+ = \mathbf{R}_0(\mathbf{A}x)$ is uniquely determined by $x_+ \in \mathbf{G}^{-1}(\ker(\mathbf{A})^\perp) \cap \mathbf{A}^{-1}(\mathbf{A}x)$. Since any solution $z \in \mathbf{A}^{-1}(xx)$ is given by $z = x + z_0$ where $z_0 \in \ker(\mathbf{A})$, we have that x_+ is the unique $z_0 \in \ker(\mathbf{A})$ with $\mathbf{G}(x + z_0) \in \ker(\mathbf{A})^\perp$.

(a) Inpainting. (b) Deblurring.

Figure 4.1: Example reconstructions for inpainting and deblurring. Each subfigure shows the original signal x (top left), exact data $\mathbf{A}x$ (top right), the limiting solution x_+ (bottom left) and an example regularized solution (bottom right).

Clearly, this is equivalent to $\mathbf{P}_{\ker(\mathbf{A})}\mathbf{G}(x + z_0) = 0$ which is another linear equation that can be solved for z_0. In our numerical simulations x_+ is constructed in this way and convergence of $x_\alpha^\delta = \mathbf{R}_\alpha(y^\delta)$ to this solution is tested. The forward operators are described below and closely follow [16].

4.2 Inpainting

The first problem we consider is inpainting where the forward $\mathbf{A}$ simply sets the first 10 rows of any image to 0. In this case, $\mathbf{P}_{\ker(\mathbf{A})} = \mathrm{Id} - \mathbf{A}$. To define $\mathbb{V}$ we use the first $q = 256$ principal components arriving at a relatively large value $\|\mathbf{P}_{\ker(\mathbf{A})}\mathbf{P}_{\mathbb{V}}\| = 0.99997$. Simulations are performed with this choice and the results are shown in Figure 4.1a and Figure 4.2. Figure 4.1a shows the original signal $x \in \mathbb{S}$ (top left), the simulated noise-free data (top right), the limiting solution x_+ (bottom left) and an example of a regularized solution (bottom right). One can see, that the original signal and the limiting solution are indeed the same; see the discussions in Section 3.3.

Figure 4.2 shows numerical convergence rates for the symmetric Bregman-distance $\langle \mathbf{G}(x_\alpha^\delta) - \mathbf{G}(x_+), x_\alpha^\delta - x_+ \rangle$ (left) and the residual $\|\mathbf{A}x_\alpha^\delta - y^\delta\|$ (right). In both images we display the reference line $\delta \mapsto \delta$ in black and $\mu(\delta) \pm \sigma(\delta)$ where $\mu(\delta)$ is the average and $\sigma(\delta)$ the standard-deviation of these quantities calculated on the test-set (note that due to the small standard-deviation only the mean value can be seen). One can see that the residual closely follows the reference line as expected. The convergence rate for the symmetric Bregman-distance, however, shows an even better convergence rate than the one given by Theorem 3.2. Note that the symmetric Bregman-distance is essentially the square of a

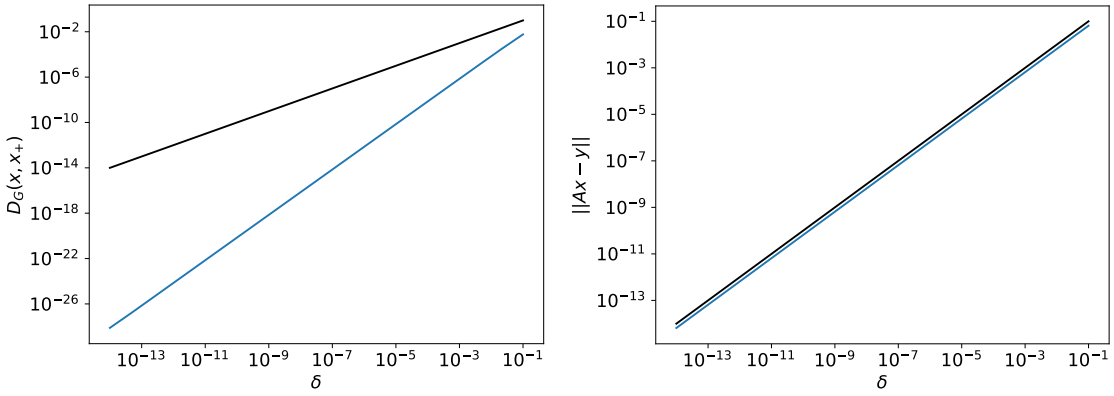

Figure 4.2: Average convergence rates plus/minus standard-deviation of the rates computed over the test-set (blue line; note that due to the small standard-deviation only the mean value can be seen) and a reference line (black) for the inpainting task. Left: Symmetric Bregman-distance. Right: Residual.

norm (see Example 2.4) and hence we numerically find, that convergence in this norm is of order $\mathcal{O}(\delta)$ improving upon the theoretically proven rate $\mathcal{O}(\delta^{1/2})$.

4.3 Deblurring

The second problem we consider is deblurring where similar to [16] we take the convolution operator to mimic a diagonal motion blur, for which we choose the 2D convolution kernel as 5×5 identity matrix. To get an approximation of $\mathbf{P}_{\ker(\mathbf{A})}$ we consider the singular-value decomposition of $\mathbf{A}$ and define $\mathbf{P}_{\ker(\mathbf{A})}$ to be the projection onto the subspace spanned by the singular vectors corresponding to singular values below 10^{-12}. We define the subspace $\mathbb{V}$ by the first $q = 512$ principal components which results in a value of $\|\mathbf{P}_{\ker(\mathbf{A})}\mathbf{P}_{\mathbb{V}}\| = 0.8958$.

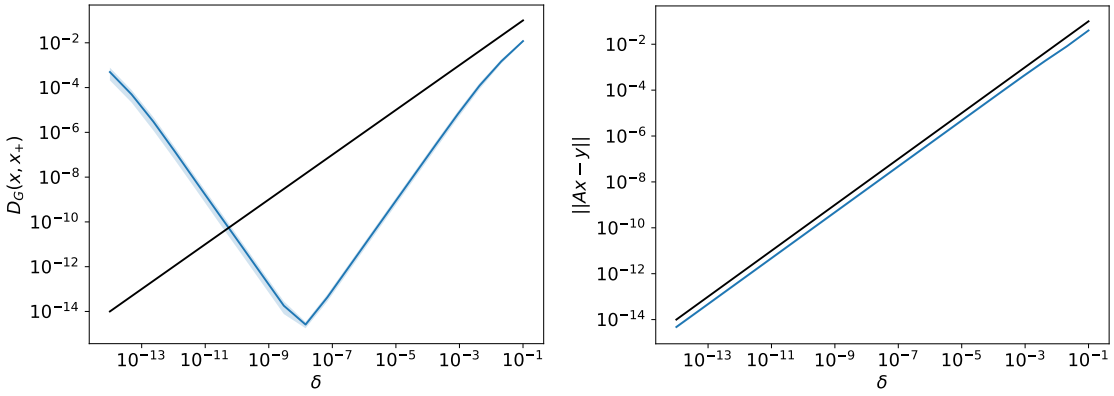

Figure 4.3: Average convergence rates plus/minus standard-deviation of the rates calculated over the test-set (blue line) and a reference line (black) for the deblurring task. Left: Symmetric Bregman-distance. Right: Residual.

Simulation results are shown in Figure 4.1b and Figure 4.3. Figure 4.1b shows the original signal $x \in \mathbb{S}$ (top left), the simulated noise-free data (top right), the limiting solution x_+

(bottom left) and an example of a regularized solution (bottom right). Similar to the inpainting example, the original signal x and the limiting solution x_+ coincide. Figure 4.3 visualizes convergence rates in the symmetric Bregman-distance $\langle \mathbf{G}(x_\alpha) - \mathbf{G}(x_+), x_\alpha - x_+ \rangle$ (left) and the residual $\left\| \mathbf{A}x_\alpha - y^\delta \right\|$ (right). As in the inpainting example we show a reference line $\delta \mapsto \delta$ in black and $\delta \mapsto \mu(\delta) \pm \sigma(\delta)$ where $\mu(\delta)$ is the average distance measure and $\sigma(\delta)$ is the standard-deviation of these distance for the test-set. One can see that the residual closely follows the reference line as expected. However, in the symmetric Bregman-distance we obtain a semi-convergent behavior . This might be due to the fact that we solve equation (1.2) with fixed finite accuracy only and hence convergence cannot be expected. It is again interesting to note that numerically, we find in the convergent regime ($\delta \geq 10^{-8}$) a faster numerical convergence rate than the one proven in Theorem 3.2 can be observed.

5 Conclusion and outlook

We presented a convergence analysis for solving inverse problems with the equilibrium equation (1.2) including the derivation of stability estimates and convergence rates. This analysis was strengthened in the case where the regularization operator $\mathbf{G}$ satisfies (3.1). In particular we derived the limiting problem (1.5) for $\alpha \to 0$ which particular lead to a new loss function for training the regularization operator $\mathbf{G}$. We have further shown for finite α, the equilibrium equation has limited performance on any given signal-class $\mathbb{S}$. The results in our paper raise many new questions for future research. One such potential direction of research is to weaken (3.1) that does not suffer from the lower-bound. Other directions could include the analysis of the novel loss-function provided in Section 3.4, its extension and the comparison with existing ones. Deriving regularization methods different from (3.1) based on (1.5) is another promising research direction.

References

[1] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019.

[2] D. Chen, M. Davies, M. J. Ehrhardt, C.-B. Schönlieb, F. Sherry, and J. Tachella. Equivariant deep learning: From unrolled network design to fully unsupervised learning. *IEEE Signal Processing Magazine*, 40(1):134–147, 2023.

[3] A. Ebner and M. Haltmeier. Plug-and-play image reconstruction is a convergent regularization method. *arXiv:2212.06881*, 2022.

[4] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

[5] D. Gilton, G. Ongie, and R. Willett. Deep equilibrium architectures for inverse problems in imaging. *IEEE Transactions on Computational Imaging*, 7:1123–1133, 2021.

[6] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.

[7] U. S. Kamilov, C. A. Bouman, G. T. Buzzard, and B. Wohlberg. Plug-and-play methods for integrating physical and learned models in computational imaging: Theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 40(1):85–97, 2023.

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[9] H. Li, J. Schwab, S. Antholzer, and M. Haltmeier. NETT: Solving inverse problems with deep neural networks. *Inverse Probl.*, 36(6):065005, 2020.

[10] S. Lunz, O. Öktem, and C.-B. Schönlieb. Adversarial regularizers in inverse problems. In *Advances in Neural Information Processing Systems*, pages 8507–8516, 2018.

[11] M. T. McCann, K. H. Jin, and M. Unser. Convolutional neural networks for inverse problems in imaging: A review. *IEEE Signal Processing Magazine*, 34(6):85–95, 2017.

[12] T. Meinhardt, M. Moller, C. Hazirbas, and D. Cremers. Learning proximal operators: Using denoising networks for regularizing inverse imaging problems. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1781–1790, 2017.

[13] D. Obmann and M. Haltmeier. Convergence rates for critical point regularization. *arXiv:2302.08830*, 2023.

[14] A. Paszke, S. Gross, and F. e. a. Massa. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[15] A. Pramanik and M. Jacob. Stable and memory-efficient image recovery using monotone operator learning (MOL). *arXiv:2206.04797*, 2022.

[16] D. Riccio, M. J. Ehrhardt, and M. Benning. Regularization of inverse problems: Deep equilibrium models versus bilevel learning. *arXiv:2206.13193*, 2022.

[17] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

[18] Y. Romano, M. Elad, and P. Milanfar. The little engine that could: Regularization by denoising (red). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017.

[19] E. Ryu, J. Liu, S. Wang, X. Chen, Z. Wang, and W. Yin. Plug-and-play methods provably converge with properly trained denoisers. In *International Conference on Machine Learning*, pages 5546–5557. PMLR, 2019.

[20] E. K. Ryu and S. P. Boyd. A primer on monotone operator methods. 2015.

[21] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen. *Variational methods in imaging*, volume 167 of *Applied Mathematical Sciences*. Springer, New York, 2009.

[22] J. Schwab, S. Antholzer, and M. Haltmeier. Deep null space learning for inverse problems: convergence analysis and rates. *Inverse problems*, 35(2):025008, 2019.

[23] Y. Sun, B. Wohlberg, and U. S. Kamilov. An online plug-and-play algorithm for regularized image reconstruction. *IEEE Transactions on Computational Imaging*, 5(3):395–408, 2019.

[24] H. Y. Tan, S. Mukherjee, J. Tang, and C.-B. Schönlieb. Provably convergent plug-and-play quasi-Newton methods. *arXiv:2303.07271*, 2023.