Nr. 77

universität innsbruck
Institut für Mathematik

Preprint-Series: Department of Mathematics - Applied Mathematics

Discretization of learned NETT regularization for solving inverse problems

S. Antholzer, M. Haltmeier

AppliedMathematics

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

Discretization of learned NETT regularization for solving inverse problems

Stephan Antholzer

Department of Mathematics, University of Innsbruck
Technikerstrasse 13, 6020 Innsbruck, Austria
stephan.antholzer@uibk.ac.at

Markus Haltmeier

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

Abstract

Deep learning based reconstruction methods deliver outstanding results for solving inverse problems and are therefore becoming increasingly important. A recently invented class of learning-based reconstruction methods is the so-called NETT (for Network Tikhonov Regularization), which contains a trained neural network as regularizer in generalized Tikhonov regularization. The existing analysis of NETT considers fixed operator and fixed regularizer and analyzes the convergence as the noise level in the data approaches zero. In this paper, we extend the frameworks and analysis considerably to reflect various practical aspects and take into account discretization of the data space, the solution space, the forward operator and the neural network defining the regularizer. We show the asymptotic convergence of the discretized NETT approach for decreasing noise levels and discretization errors. Additionally, we derive convergence rates and present numerical results for a limited data problem in photoacoustic tomography.

1 Introduction

In this paper, we are interested in neural network based solution of inverse problems of the form

$$\text{Find } x \text{ from data} \quad y^\delta = \mathbf{A}x + \eta. \tag{1}$$

Here $\mathbf{A}$ is a potentially non-linear operator between Banach spaces $\mathbb{X}$ and $\mathbb{Y}$, y^δ are the given noisy data, x is the unknown to be recovered, η is the unknown noise perturbation and $\delta \geq 0$ indicates the noise level. Numerous image reconstruction problems, parameter identification tasks or geophysical applications can be stated as such inverse problems [8, 28, 19, 35]. Special challenges in solving inverse problems are the non-uniqueness of the solutions and the instability of the solutions with respect to the given

data. To overcome these issues, regularization methods are needed, which are used as criteria for selecting specific solutions and at the same time stabilize the inversion process.

Reconstruction with learned regularizers

One of the most established class of methods for solving inverse problems is variational regularization where regularized solutions are defined as minimizers of the Tikhonov functional [28, 17, 30]

$$\mathcal{T}_{y^\delta,\alpha}\colon \mathbb{X} \to [0,\infty]\colon x \mapsto \mathcal{D}(\mathbf{A}x, y^\delta) + \alpha\mathcal{R}(x)\,. \tag{2}$$

Here $\mathcal{D}$ is a distance like function measuring closeness of the data, $\mathcal{R}$ a regularization term enforcing regularity of the minimizer and α is the regularization parameter. In the case that $\mathcal{D}$ and the regularizer are defined by the Hilbert space norms, (2) is classical Tikhonov regularization for which the theory is quite complete [8, 11]. In particular, in this case, convergence rates, which name quantitative estimates for the distance between the true and regularized solutions are well known. Convergence rates for non-convex regularizers are derived in [9].

Typical regularization techniques are based on simple hand crafted regularization terms such as the total variation $\|f\|_{\mathrm{TV}} = \int|\nabla f|$ or quadratic Sobolev norms $\|\nabla f\|_2^2 = \int|\nabla f|^2$ on some function space. However, these regularizers are quite simplistic and might not well reflect the actual complexity of the underlying class of functions. Therefore, recently, it has been proposed and analyzed in [14] to use machine learning to construct regularizers in a data driven manner. The strategy in [14] is to construct a data-driven regularizer via the following consecutive steps:

(T1) Choose a family of desired reconstructions $(x_i)_{i=1}^n$.

(T2) For some $\mathbf{B}\colon \mathbb{Y} \to \mathbb{X}$, construct undesired reconstructions $(\mathbf{B}\mathbf{A}x_i)_{i=1}^n$.

(T3) Choose a class $(\mathbf{\Phi}_\theta)_{\theta\in\Theta}$ of functions (networks) $\mathbf{\Phi}_\theta\colon \mathbb{X} \to \mathbb{X}$.

(T4) Determine $\theta^\star \in \Theta$ with $\mathbf{\Phi}_{\theta^\star}(x_i) \simeq x_i \wedge \mathbf{\Phi}_{\theta^\star}(\mathbf{B}\mathbf{A}x_i) \simeq x_i$.

(T5) Define $\mathcal{R}(x) = r(x, \mathbf{\Phi}(x))$ with $\mathbf{\Phi} = \mathbf{\Phi}_{\theta^\star}$ for some $r\colon \mathbb{Y}\times\mathbb{Y} \to [0,\infty]$.

For imaging applications, the function class $(\mathbf{\Phi}_\theta)_{\theta\in\Theta}$ can be chosen as convolutional neural networks which have demonstrated to give powerful classes of mappings between image spaces. The function r measures distance between a potential reconstruction x and the output of the network $\mathbf{\Phi}(x)$, and possibly adds additional regularization [20, 21]. According to the training strategy in item (T4) the value of the regularizer will be small if the reconstruction is similar to elements in $(x_i)_{i=1}^n$ and large for elements in $(\mathbf{B}\mathbf{A}x_i)_{i=1}^n$. A simple example that we will use for our numerical results is the learned regularizer $\mathcal{R}(x) = \|x - \mathbf{\Phi}(x)\|^2 + \|x\|_{\mathrm{TV}}$.

Convergence analysis and convergence rates for NETT as well as training strategies have been established in [10, 14, 21]. A different training strategy for learning a regularizer has been proposed in [15, 18]. Note that learning the regularizer first and

then minimizing the Tikhonov functional is different from variational and iterative networks [2, 3, 7, 12, 34] where an iterative scheme is applied to enroll the functional $\mathcal{D}_\theta(\mathbf{A}x, y^\delta) + \alpha\mathcal{R}_\theta(x)$ which is then trained in an end to end fashion. Training the regularizer first has the advantage of being more modular, sharing some similarity with plug and play techniques [26], and the network training is independent of the forward operator $\mathbf{A}$. Moreover, it enables to derive a convergence analysis as the noise level tends to zero and therefore comes with theoretical recovery guarantees.

Discrete NETT

The existing analysis of NETT considers minimizers of the Tikhonov functional (2) with regularizer of the form $\mathcal{R}(x) = r(x, \mathbf{\Phi}(x))$ before discretization, typically in an infinite dimensional setting. However, in practice, only finite dimensionale approximations of the unknown, the operator and the neural network are given. To address these issues, in this paper, we study discrete NETT regularization which considers minimizers of

$$\mathcal{T}_{y^\delta,\alpha,n} \colon \mathbb{X}_n \to \mathbb{Y} \colon x \mapsto \mathcal{D}(\mathbf{A}_n z, y^\delta) + \alpha\mathcal{R}_n(z)\,. \tag{3}$$

Here $(\mathbb{X}_n)_{n\in\mathbb{N}}$, $(\mathbf{A}_n)_{n\in\mathbb{N}}$ and $(\mathcal{R}_n)_{n\in\mathbb{N}}$ are families of subspaces of $\mathbb{X}_n \subseteq \mathbb{X}$, mappings $\mathbf{A}_n \colon \mathbb{X} \to \mathbb{Y}$ and regularizers $\mathcal{R}_n \colon \mathbb{X} \to [0, \infty]$, respectively, which reflect discretization of all involved operations. We present a full convergence analysis as the noise level δ converges to zero and n, α are chosen accordingly. Discretization of variational regularization has studied in [25] for the case that $\mathcal{D}$ is given by the norm distance and the regularizer $\mathcal{R}$ is taken convex and fixed. However, in the case of discrete NETT regularization it is natural to consider the case where the regularization depends on the discretization as regularization is learned in a discretized setting based on actual data. For that purpose our analysis includes non-convex regularizers that are allowed to depend on the discretization and the noise level.

Outline

The convergence analysis including convergence rates is presented in Section 2. In Section 3 we will present numerical results for a non-standard limited data problem in photoacoustic tomography that can be considered as simultaneous inpainting and artifact removal problem. We conclude the paper with a short summary and conclusion presented in Section 4.

2 Convergence analysis

In this section we study the convergence of (3) and derive convergence rates.

2.1 Well-posedness

First we state the assumptions that we will use for well-posedness (existence and stability of minimizing NETT).

Assumptions 2.1 (Conditions for well-posedness).

(W1) $\mathbb{X}$, $\mathbb{Y}$ are Banach spaces, $\mathbb{X}$ reflexive, $\mathbb{D} \subseteq \mathbb{X}$ weakly sequentially closed.

(W2) The distance measure $\mathcal{D}\colon \mathbb{Y} \times \mathbb{Y} \to [0, \infty]$ satisfies

- (a) $\exists \tau \geq 1\colon \forall y_1, y_2, y_3 \in \mathbb{Y}\colon \mathcal{D}(y_1, y_2) \leq \tau \mathcal{D}(y_1, y_3) + \tau \mathcal{D}(y_3, y_2)$.
- (b) $\forall y_1, y_2 \in \mathbb{Y}\colon \mathcal{D}(y_1, y_2) = 0 \Leftrightarrow y_1 = y_2$.
- (c) $\forall y, \tilde{y} \in \mathbb{Y}\colon \mathcal{D}(y, \tilde{y}) < \infty \wedge \|\tilde{y} - y_k\| \to 0 \Rightarrow \mathcal{D}(y, y_k) \to \mathcal{D}(y, \tilde{y})$.
- (d) $\forall y \in \mathbb{Y}\colon \|y_k - y\| \to 0 \Rightarrow \mathcal{D}(y_k, y) \to 0$.
- (e) $\mathcal{D}$ is weakly sequentially lower semi-continuous (wslsc).

(W3) $\mathcal{R}\colon \mathbb{X} \to [0, \infty]$ is proper and wslsc.

(W4) $\mathbf{A}\colon \mathbb{D} \subseteq \mathbb{X} \to \mathbb{Y}$ is weakly sequentially continuous.

(W5) $\forall y, \alpha, C\colon \{x \in \mathbb{X} \mid \mathcal{T}_{y,\alpha} \leq C\}$ is nonempty and bounded.

(W6) $(\mathbb{X}_n)_{n\in\mathbb{N}}$ is a sequence of subspaces of $\mathbb{X}$.

(W7) $(\mathbf{A}_n)_{n\in\mathbb{N}}$ is a family of weakly sequentially continuous $\mathbf{A}_n\colon \mathbb{D} \to \mathbb{Y}$.

(W8) $(\mathcal{R}_n)_{n\in\mathbb{N}}$ is a family of proper wslsc regularizers $\mathcal{R}_n\colon \mathbb{X} \to [0, \infty]$.

(W9) $\forall y, \alpha, C, n\colon \{x \in \mathbb{X}_n \mid \mathcal{T}_{y,\alpha,n} \leq C\}$ is nonempty and bounded.

Conditions (W2)-(W5) are quite standard for Tikhonov regularization in Banach spaces to guarantee the existence and stability of minimizers of the Tikhonov functional and the given conditions are similar to [9, 10, 14, 20, 24, 28, 31]. In particular, (W2) describes the properties that the distance measure $\mathcal{D}$ should have. Clearly, the norm distance on $\mathbb{Y}$ fulfills these properties. Item (W2c) is the continuity of $\mathcal{D}(y, \cdot)$ while (W2d) considers the continuity of $\mathcal{D}(\cdot, y)$ at y. While (W2c) is not needed for existence and convergence of NETT it is required for the stability result as shown in [20, Example 2.7]. Assumption (W5) is a coercivity condition; see [14, Remark 2.4f.] on how to achieve this for a regularizer defined by neural networks. Note that for convergence and convergence rates we will require additional conditions that concern the discretization of the reconstruction space, the forward operator and regularizer.

The references [9, 14, 20, 24] all consider general distance measures and allow non-convex regularizers. However, existence and stability of minimizing (2) are shown under assumptions slightly different from (W1)-(W5). Below we therefore give a short proof of the existence and stability results.

Theorem 2.2 (Existence and Stability). *Let Assumption 2.1 hold. Then for all $y^\delta \in \mathbb{Y}$, $\alpha > 0$, $n \in \mathbb{N}$ the following assertions hold true:*

(a) $\operatorname{argmin} \mathcal{T}_{y,\alpha,n} \neq \emptyset$.

(b) Let $(y_k)_{k\in\mathbb{N}} \in \mathbb{Y}^{\mathbb{N}}$ with $y_k \to y$ and consider $x_k \in \operatorname{argmin} \mathcal{T}_{y_k,\alpha,n}$.

- *$(x_k)_{k\in\mathbb{N}}$ has at least one weak accumulation point.*

- *Every weak accumulation point $(x_k)_{k\in\mathbb{N}}$ is a minimizer of $\mathcal{T}_{y,\alpha,n}$.*

(c) The statements in (a),(b) also hold for $\mathcal{T}_{y,\alpha}$ in place of $\mathcal{T}_{y,\alpha,n}$,

Proof. Since (W1), (W6)-(W9) for $\mathcal{T}_{y,\alpha,n}$ when $n \in \mathbb{N}$ are fixed give the same assumption as (W1), (W3)-(W5) for the non-discrete counterpart $\mathcal{T}_{y,\alpha}$, it is sufficient to verify (a), (b) for the latter. Existence of minimizers follows from (W1), (W2e), (W3)-(W5), because these items imply that the $\mathcal{T}_{y,\alpha}$ is a wslsc coercive functional defined on a nonempty weakly sequentially closed subset of a reflexive Banach space. To show stability one notes that according to (W2a) for all $x \in \mathbb{X}$ we have

$$\begin{aligned}\mathcal{D}(\mathbf{A}x_k, y) + \alpha\mathcal{R}(x_k) &\leq \tau(\mathcal{D}(\mathbf{A}x_k, y_k) + \alpha\mathcal{R}(x_k)) + \tau\mathcal{D}(y, y_k)\\ &\leq \tau(\mathcal{D}(\mathbf{A}x, y_k) + \alpha\mathcal{R}(x)) + \tau\mathcal{D}(y, y_k)\,.\end{aligned}$$

According to (W2c), (W2d), (W5) there exists $x \in \mathbb{X}$ such that the right hand side is bounded, which by (W5) shows that $(x_k)_k$ has a weak accumulation point. Following the standard proof [28, Theorem 3.23] shows that weak accumulation points satisfy the claimed properties. □

In the following we write $x^\delta_{\alpha,n}$ for minimizers of $\mathcal{T}_{y^\delta,\alpha,n}$. For $y \in \mathbb{Y}$ we call $x^+ \in \operatorname{argmin}\{\mathcal{R}(x) \mid x \in \mathbb{X} \wedge \mathbf{A}x = y\}$ an $\mathcal{R}$-minimizing solution of $\mathbf{A}x = y$.

Lemma 2.3 (Existence of $\mathcal{R}$-minimizing solutions)**.** *Let Assumption 2.1 hold. For any $y \in \mathbf{A}(\mathbb{D})$ an $\mathcal{R}$-minimizing solution of $\mathbf{A}x = y$ exists. Likewise, if $n \in \mathbb{N}$ and $y \in \mathbf{A}_n(\mathbb{D})$ an $\mathcal{R}_n$-minimizing solution of $\mathbf{A}_n x = y$ exists.*

Proof. Again is is sufficient the verify the claim for $\mathcal{R}$-minimizing solution. Because $y \in \mathbf{A}(\mathbb{D})$, the set $\mathbf{A}^{-1}(\{y\}) = \{x \in \mathbb{X} \mid \mathbf{A}x = y\}$ is non-empty. Hence we can choose a sequence $(x_k)_{k\in\mathbb{N}}$ in $\mathbf{A}^{-1}(\{y\})$ with $\mathcal{R}(x_k) \to \inf\{\mathcal{R}(x) \mid x \in \mathbb{X} \wedge \mathbf{A}x = y\}$. Due to (W2b), $(x_k)_{k\in\mathbb{N}}$ is contained in $\{x \in \mathbb{X} \mid \mathcal{D}(\mathbf{A}(x), y) + \alpha\mathcal{R}(x) \leq C\}$ for some $C > 0$ which is bounded according to (W5). By (W1) $\mathbb{X}$ is reflexive and therefore $(x_k)_{k\in\mathbb{N}}$ has a weak accumulation point x^+. From (W1), (W4), (W3) we conclude that x^+ is an $\mathcal{R}$-minimizing solution of $\mathbf{A}x = y$. The case of $\mathcal{R}_n$-minimizing solutions follows analogous. □

2.2 Convergence

Next we proof that discrete NETT converges as the noise level goes to zero and the discretization as well as the regularization parameter are chosen properly. We write $\mathbb{D}_{n,M} := \{x \in \mathbb{D} \cap \mathbb{X}_n \mid \mathcal{R}_n(x) \leq M\}$ and formulate the following approximation conditions for obtaining convergence.

Assumptions 2.4 (Conditions for convergence)**.**
Element $x^+ \in \mathbb{D}$ satisfies the following for all $M > 0$:

(C1) $\exists(z_n) \in \prod_{n\in\mathbb{N}}(\mathbb{D} \cap \mathbb{X}_n)$ with $\lambda_n := |\mathcal{R}_n(z_n) - \mathcal{R}(x^+)| \to 0$.

(C2) $\rho_n := \sup_{x\in\mathbb{D}_{n,M}} |\mathcal{R}_n(x) - \mathcal{R}(x)| \to 0$.

(C3) $\gamma_n := \mathcal{D}(\mathbf{A}_n z_n, \mathbf{A}x^+) \to 0$.

(C4) $a_n := \sup_{x \in \mathbb{D}_{n,M}} |\mathcal{D}(\mathbf{A}_n x, \mathbf{A}x^+) - \mathcal{D}(\mathbf{A}x, \mathbf{A}x^+)| \to 0$.

Conditions (C1) and (C3) concerns the approximation of the true unknown x with elements in the discretization space, that is compatible with the discretization of the forward operator and regularizer. Conditions (C2) and (C4) are uniform approximation properties of the operator and the regularizer on $\mathcal{R}_n$-bounded sets.

Theorem 2.5 (Convergence). *Let (W1)-(W9) hold, $y \in \mathbf{A}(\mathbb{D})$ and let x^+ be an $\mathcal{R}$-minimizing solution of $\mathbf{A}x = y$ that satisfies (C1)-(C4). Moreover, suppose $(\delta_k)_{k\in\mathbb{N}} \in (0,\infty)^{\mathbb{N}}$ converges to zero and $(y_k)_{k\in\mathbb{N}} \in \mathbb{Y}^{\mathbb{N}}$ satisfies $\mathcal{D}(y, y_k) \le \delta_k$. Choose $(\alpha_k)_{k\in\mathbb{N}}$ and $(n_k)_{k\in\mathbb{N}}$ such that as $k \to \infty$ we have*

$$\alpha_k \to 0 \tag{4}$$

$$n_k \to \infty \tag{5}$$

$$(\delta_k + \mathcal{D}(\mathbf{A}_{n_k} z_{n_k}, y))/\alpha_k \to 0\,. \tag{6}$$

Then for $x_k \in \operatorname{argmin} \mathcal{T}_{y_k,\delta_k,n_k}$ the following hold:

(a) $(x_k)_{k\in\mathbb{N}}$ has a weakly convergent subsequence $(x_{\sigma(k)})_{k\in\mathbb{N}}$

(b) The weak limit of $(x_{\sigma(k)})_{k\in\mathbb{N}}$ is an $\mathcal{R}$-minimizing solution of $\mathbf{A}x = y$.

(c) $\mathcal{R}_{\sigma(k)}(x_{\sigma(k)}) \to \mathcal{R}(x^\star)$, where $x^\star$ is the weak limit of $(x_{\sigma(k)})_{k\in\mathbb{N}}$.

(d) If the $\mathcal{R}$-minimizing solution of $\mathbf{A}x = y$ is unique, then $(x_k)_{k\in\mathbb{N}} \rightharpoonup x^+$.

Proof. For convenience and some abuse of notation we use the abbreviations $\mathcal{R}_k := \mathcal{R}_{n_k}$, $\mathbf{A}_k := \mathbf{A}_{n_k}$, $a_k := a_{n_k}$, $z_k := z_{n_k}$ and $\rho_k := \rho_{n_k}$. Because x_k is a minimizer of the discrete NETT functional $\mathcal{T}_{y_k,\delta_k,n_k}$ by (W2) we have

$$\begin{aligned}\mathcal{D}(\mathbf{A}_k x_k, y_k) + \alpha_k \mathcal{R}_k(x_k) &\le \mathcal{D}(\mathbf{A}_k z_k, y_k) + \alpha_k \mathcal{R}_k(z_k) \\ &\le \tau \mathcal{D}(\mathbf{A}_k z_k, y) + \tau \mathcal{D}(y, y_k) + \alpha_k \mathcal{R}_k(z_k) = \tau \mathcal{D}(\mathbf{A}_k z_k, y) + \tau\delta_k + \alpha_k \mathcal{R}_k(z_k)\end{aligned}$$

According to (C1), (4), we get

$$\mathcal{D}(\mathbf{A}_k x_k, y_k) \le \tau(\mathcal{D}(\mathbf{A}_k z_k, y) + \delta_k)\,, \tag{7}$$

$$\mathcal{R}_k(x_k) \le \tau \cdot \frac{\mathcal{D}(\mathbf{A}_k z_k, y_k) + \delta_k}{\alpha_k} + \mathcal{R}_k(z_k)\,. \tag{8}$$

According to (C1), (C3), (5), (6) the right hand side in (7) converges to zero and the right hand side in (8) to $\mathcal{R}(x_+)$. Together with (C2) we obtain $\mathcal{R}_k(x) \le \mathcal{R}_k(x_k) + \rho_k \to \mathcal{R}(x_+)$ and $\mathcal{D}(\mathbf{A}x_k, y) \le \tau \mathcal{D}(\mathbf{A}_k x_k, y_k) + \tau\delta_k \le \tau \mathcal{D}(\mathbf{A}x_k, y) + \tau a_k + \tau\delta_k \to 0$. This shows that $(\mathcal{D}(\mathbf{A}x_k, y) + \mathcal{R}(x_k))_{k\in\mathbb{N}}$ is bounded and by (W1), (W9) there exists a weakly convergent subsequence $(x_{\sigma(k)})_{k\in\mathbb{N}}$. We denote the weak limit by $x^\star \in \mathbb{X}$. From (W2), (W4) we obtain $\mathbf{A}x = y$. The weak lower semi-continuity of $\mathcal{R}$ assumed in (W3) shows

$$\begin{aligned}\mathcal{R}(x^\star) \le \liminf_k \mathcal{R}(x_{\sigma(k)}) \le \limsup_k \mathcal{R}(x_{\sigma(k)}) \\ \le \limsup_k (\mathcal{R}_{\sigma(k)}(x_{\sigma(k)}) + \rho_k) \le \mathcal{R}(x^+)\,.\end{aligned}$$

Consequently, $x^\star$ is an $\mathcal{R}$-minimizing solution of $\mathbf{A}x = y$ and $\mathcal{R}(x_{\sigma(k)}) \to \mathcal{R}(x^\star)$. If the $\mathcal{R}$-minimizing solution is unique then x^+ is the only weak accumulation point of $(x_k)_{k\in\mathbb{N}}$ which concludes the proof. □

2.3 Convergence rates

Next we derive quantitative error estimates (convergence rates) in terms of the absolute Bregman distance. Recall that a function $\mathcal{R}\colon \mathbb{X} \to [0,\infty]$ is Gâteaux differentiable at some $x^\star \in \mathbb{X}$ if the directional derivative $\mathcal{R}'(x^\star)(h) := (\mathcal{R}(x^\star + th) - \mathcal{R}(x^\star))/t$ exist for every $h \in \mathbb{X}$. We denote by $\mathcal{R}'(x^\star)$ the Gâteaux derivative of $\mathcal{R}$ at x. In [14] we introduced the absolute Bregman distance $\mathcal{B}_\mathcal{R}(\cdot, x^\star)\colon \mathbb{X} \to [0,\infty]$ of a Gâteaux differentiable functional $\mathcal{R}\colon \mathbb{X} \to [0,\infty]$ at $x^\star \in \mathbb{X}$ with respect to $\mathcal{R}$ defined by

$$\forall x \in \mathbb{X}\colon \mathcal{B}_\mathcal{R}(x, x^\star) := \left|\mathcal{R}(x) - \mathcal{R}(x^\star) - \mathcal{R}'(x^\star)(x - x^\star)\right|. \tag{9}$$

We write $\sup_{y^\delta} H(y^\delta) := \sup\{H(y^\delta) \mid y^\delta \in \mathbb{X} \wedge \mathcal{D}(\mathbf{A}x^+, y^\delta) \le \delta\}$. Convergence rates in terms of the Bregman distance are derived under a smoothness assumption on the true solution in the form of a certina variational inequality. More precisely we assume the following:

Assumptions 2.6 (Conditions for convergence rates).
Element $x^+ \in \mathbb{D}$ satisfies the following for all $M, \delta > 0$:

(R1) Items (C1), (C2) hold.

(R2) $\gamma_{n,\delta} := \sup_{y^\delta} |\mathcal{D}(\mathbf{A}_n z_n, y^\delta) - \mathcal{D}(\mathbf{A}x^+, y^\delta)| \to 0$.

(R3) $a_{n,\delta} := \sup_{y^\delta} \sup_{x \in \mathbb{D}_{n,M}} |\mathcal{D}(\mathbf{A}_n x, y^\delta) - \mathcal{D}(\mathbf{A}x, y^\delta)| \to 0$.

(R4) $\mathcal{R}$ is Gâteaux differentiable at x^+

(R5) There exist a concave, continuous, strictly increasing $\varphi\colon [0,\infty) \to [0,\infty)$ with $\varphi(0) = 0$ and $\epsilon, \beta > 0$ such that for all $x \in \mathbb{X}$

$$|\mathcal{R}(x) - \mathcal{R}(x^+)| \le \epsilon \Rightarrow \beta \mathcal{B}_\mathcal{R}(x, x^+) \le \mathcal{R}(x) - \mathcal{R}(x^+) + \varphi(\mathcal{D}(\mathbf{A}x, \mathbf{A}x^+)).$$

According to (R5) the inverse function $\varphi^{-1}\colon [0,\infty) \to [0,\infty)$ exists and is convex. We denote by $\varphi^{-*}(s) := \sup\{s\,t - \varphi^{-1}(t) \mid t \ge 0\}$ its Fenchel conjugate.

Proposition 2.7 (Error estimates). *Let $y \in \mathbf{A}(\mathbb{D})$ and x^+ be an $\mathcal{R}$-minimizing solution of $\mathbf{A}x = y$ such that (W1)-(W9) and (R1)-(R5) are satisfied. For $y^\delta \in \mathbb{Y}$ with $\mathcal{D}(y, y^\delta) \le \delta$ let $x^\delta_{\alpha,n} \in \operatorname{argmin} \mathcal{T}_{y^\delta,\alpha,n}$. Then for sufficient small $\delta, \alpha > 0$ and sufficiently large $n \in \mathbb{N}$, we have the error estimate*

$$\mathcal{B}_\mathcal{R}(x^\delta_{\alpha,n}, x^+) \le \frac{a_{n,\delta} + \gamma_{n,\delta} + \delta}{\alpha} + \rho_n + \lambda_n + \varphi(\tau\delta) + \frac{\varphi^{-*}(\tau\alpha)}{\tau\alpha}. \tag{10}$$

Proof. We have $\mathcal{D}(\mathbf{A}_n x^\delta_{\alpha,n}, y^\delta) + \alpha\mathcal{R}_n(x^\delta_{\alpha,n}) \le \mathcal{D}(\mathbf{A}_n z_n, y^\delta) + \alpha\mathcal{R}_n(z_n)$. According to Theorem 2.5 we can assume $|\mathcal{R}(x^\delta_{\alpha,n}) - \mathcal{R}(x^+)| < \epsilon$ and with (R5) we obtain

$$\begin{aligned}
&\alpha\beta\mathcal{B}_\mathcal{R}(x^\delta_{\alpha,n}, x^+) \\
&\le \alpha\mathcal{R}(x^\delta_{\alpha,n}) - \alpha\mathcal{R}(x^+) + \alpha\varphi(\mathcal{D}(\mathbf{A}x^\delta_{\alpha,n}, y)) \\
&\le \alpha\mathcal{R}_n(x^\delta_{\alpha,n}) - \alpha\mathcal{R}(z_n) + \alpha\rho_n + \alpha\lambda_n + \alpha\varphi(\mathcal{D}(\mathbf{A}x^\delta_{\alpha,n}, y)) \\
&\le \mathcal{D}(\mathbf{A}_n z_n, y^\delta) - \mathcal{D}(\mathbf{A}_n x^\delta_{\alpha,n}, y^\delta) + \alpha\rho_n + \alpha\lambda_n + \alpha\varphi(\mathcal{D}(\mathbf{A}x^\delta_{\alpha,n}, y))
\end{aligned}$$

$$\leq \delta - \mathcal{D}(\mathbf{A}x^\delta_{\alpha,n}, y^\delta) + \gamma_{n,\delta} + a_{n,\delta} + \alpha\rho_n + \alpha\lambda_n + \alpha\varphi(\tau\delta) + \alpha\varphi(\tau\mathcal{D}(\mathbf{A}x^\delta_{\alpha,n}, y^\delta))$$
$$\leq \delta + \gamma_{n,\delta} + a_{n,\delta} + \alpha\rho_n + \alpha\lambda_n + \alpha\varphi(\tau\delta) + \tau^{-1}\varphi^{-*}(\tau\delta)\,.$$

where we used Young's inequality $\alpha\varphi(\tau t) \leq t + \tau^{-1}\varphi^{-*}(\tau\alpha)$ for the last step. □

Remark 2.8. The error estimate (10) includes the approximation quality of the discrete or inexact forward operator $\mathbf{A}_n$ and the discrete or inexact regularizer $\mathcal{R}_n$ described by $a_{n,\delta}$ and ρ_n, respectively. What might be unexpected at first is the inclusion of two new parameters λ_n and $\gamma_{n,\delta}$. These factors both arise from the approximation of $\mathbb{X}$ by the finite dimensional spaces $\mathbb{X}_n$, where $\gamma_{n,\delta}$ reflects approximation accuracy in the image of the operator $\mathbf{A}$ and λ_n approximation accuracy with respect to the true regularization functional $\mathcal{R}$. Note that in the case where the forward operator, the regularizer and the solution space $\mathbb{X}$ are given precisely, we have $a_{n,\delta} = \gamma_{n,\delta} = \lambda_n = \rho_n = 0$. In this particular case we recover the estimate derived for the NETT in [14].

Theorem 2.9 (Convergence rates)**.** *Let the assumptions of Proposition 2.7 hold and consider the parameter choice rule $\alpha(\delta) \asymp \delta/\varphi(\delta)$ and let the approximation errors satisfy $a_{n,\delta} + \gamma_{n,\delta} = \mathcal{O}(\delta)$, $\rho_n + \lambda_n = \mathcal{O}(\varphi(\tau\delta))$. Then we have the convergence rate*

$$\mathcal{B}_\mathcal{R}(x^\delta_{\alpha(\delta),n(\delta)}, x^\dagger) = \mathcal{O}(\varphi(\tau\delta))\,. \tag{11}$$

Proof. Noting that $\varphi^{-*}(\tau\delta/\varphi(\tau\delta))/\delta$ remains bounded as $\delta \to 0$, this directly follows from Proposition 2.7 □

Next we verify that a variational inequality of the form (R5) is satisfied with $\varphi(t) = c\sqrt{t}$ under a typical source like condition.

Lemma 2.10 (Variational inequality under source condition)**.** *Let $\mathcal{R}$, $\mathbf{A}$ be Gâteaux differentiable at $x^\dagger \in \mathbb{X}$, consider the distance measure $\mathcal{D}(y_1, y_2) = \|y_1 - y_2\|^2$ and assume there exist $\eta \in \mathbb{X}^\star$ and $c_1, c_2, \epsilon > 0$ with $c_1\|\eta\| < 1$ such that for all $x \in \mathbb{X}$ with $|\mathcal{R}(x) - \mathcal{R}(x^\dagger)| \leq \epsilon$ we have*

$$\begin{aligned} &\mathcal{R}'(x^\dagger) = \mathbf{A}'(x^\dagger)^*\eta \\ &\|\mathbf{A}x - \mathbf{A}x^\dagger - \mathbf{A}'(x^\dagger)(x - x^\dagger)\| \leq c_1\mathcal{B}_\mathcal{R}(x, x^\dagger) \\ &\mathcal{R}(x^\dagger) - \mathcal{R}(x) \leq c_2\|\mathbf{A}x - \mathbf{A}x^\dagger\|\,. \end{aligned} \tag{12}$$

Then (R5) holds with $\varphi(t) = (\|\eta\| + 2c_2)\sqrt{t}$ and $\beta = 1 - c_1\|\eta\|$.

Proof. Let $x \in \mathbb{X}$ with $|\mathcal{R}(x) - \mathcal{R}(x^\dagger)| \leq \epsilon$. Using the Cauchy-Schwarz inequality and equation (12), we can estimate

$$\begin{aligned} |\langle \mathcal{R}'(x^\dagger), x - x^\dagger\rangle| \leq& \|\mathbf{A}'(x^\dagger)(x - x^\dagger)\|\|\eta\| \\ \leq& \|\mathbf{A}x - \mathbf{A}x^\dagger\|\|\eta\| + \|\mathbf{A}x - \mathbf{A}x^\dagger - \mathbf{A}'(x^\dagger)(x - x^\dagger)\|\|\eta\| \\ \leq& \|\mathbf{A}x - \mathbf{A}x^\dagger\|\|\eta\| + c_1\|\eta\|\mathcal{B}_\mathcal{R}(x, x^\dagger)\,. \end{aligned}$$

Additionally, if $\mathcal{R}(x) \geq \mathcal{R}(x^\dagger)$, we have $|\mathcal{R}(x) - \mathcal{R}(x^\dagger)| = \mathcal{R}(x) - \mathcal{R}(x^\dagger)$, and on the other hand if $\mathcal{R}(x) < \mathcal{R}(x^\dagger)$, we have $|\mathcal{R}(x) - \mathcal{R}(x^\dagger)| \leq \mathcal{R}(x) - \mathcal{R}(x^\dagger) + 2(\mathcal{R}(x^\dagger) - \mathcal{R}(x)) \leq \mathcal{R}(x) - \mathcal{R}(x^\dagger) + 2c_2\|\mathbf{A}x - \mathbf{A}x^\dagger\|$. Putting this together we get

$$\begin{aligned}\mathcal{B}_{\mathcal{R}}(x, x^+) &\leq |\mathcal{R}(x) - \mathcal{R}(x^+)| + |\langle \mathcal{R}'(x^+), x - x^+ \rangle| \\ &\leq \mathcal{R}(x) - \mathcal{R}(x^+) + (\|\eta\| + 2c_2)\|\mathbf{A}x - \mathbf{A}x^+\| + c_1\|\eta\|\mathcal{B}_{\mathcal{R}}(x, x^+)\,,\end{aligned}$$

and thus $(1 - c_1\|\eta\|)\mathcal{B}_{\mathcal{R}}(x, x^+) \leq \mathcal{R}(x) - \mathcal{R}(x^+) + (\|\eta\| + 2c_2)\|\mathbf{A}x - \mathbf{A}x^+\|$. □

Corollary 2.11 (Convergence rates under source condition). *Let the conditions of Lemma 2.10 hold and suppose*

$$\begin{aligned}&\alpha(\delta) \asymp \sqrt{\delta} \\ &|\mathcal{R}_{n(\delta)}(z_{n(\delta)}) - \mathcal{R}(x^+)| = \mathcal{O}(\sqrt{\delta}) \\ &\sup\{|\mathcal{R}_{n(\delta)}(x) - \mathcal{R}(x)| \mid x \in \mathbb{D}_{n(\delta),M}\} = \mathcal{O}(\sqrt{\delta}) \\ &\|\mathbf{A}_{n(\delta)}z_{n(\delta)} - \mathbf{A}x^+\| = \mathcal{O}(\sqrt{\delta}) \\ &\sup\{\|\mathbf{A}_{n(\delta)}x - \mathbf{A}x\| \mid x \in \mathbb{D}_{n(\delta),M}\} = \mathcal{O}(\sqrt{\delta}) \\ &\sup\{\|\mathbf{A}_{n(\delta)}x\| \mid x \in \mathbb{D}_{n(\delta),M}\} < \infty\,.\end{aligned}$$

Then we have the convergence rates result

$$\mathcal{B}_{\mathcal{R}}(x^{\delta}_{\alpha(\delta),n(\delta)}, x^+) = \mathcal{O}(\sqrt{\delta})\,. \tag{13}$$

Proof. Follows from Theorem 2.9 and Lemma 2.10. Note that we use $\|\cdot\|$ in the theorem, while $\mathcal{D}(y_1, y_1) = \|y_1 - y_2\|^2$ uses the squared norm $\|\cdot\|^2$ and thus the approximation rates for the terms concerning $\mathbf{A}_{n(\delta)}$ are order $\sqrt{\delta}$ instead of δ as in Theorem 2.9. □

In Corollary 2.11, the approximation quality of the discrete operator $\mathbf{A}_n$ and the discrete and inexact regularization functional $\mathcal{R}_n$ need to be of the same order.

3 Application to a limited data problem in PAT

Photoacoustic Tomography (PAT) is an emerging non-invasive coupled-physics biomedical imaging technique with high contrast and high spatial resolution [13, 22]. It works by illuminating a semi-transparent sample with short optical pulses which causes heating of the sample followed by expansion and the subsequent emission of an acoustic wave. Sensors on the outside of the sample measure the acoustic wave and these measurements are then used to reconstruct the initial pressure $f\colon \mathbb{R}^d \to \mathbb{R}$, which provides information about the interior of the object. The cases $d = 2$ and $d = 3$ are relevant for applications in PAT. Here we only consider the case $d = 2$ and assume a circular measurement geometry. The 2D case arises for example when using integrating line detectors in PAT [22].

3.1 Discrete forward operator

The pressure data $p\colon \mathbb{R}^2 \times [0, \infty) \to \mathbb{R}$ satisfies the wave equation $(\partial_t^2 - \Delta)p(\mathbf{r}, t) = 0$ for $(\mathbf{r}, t) \in \mathbb{R}^2 \times (0, \infty)$ with initial data $p(\cdot\,, 0) =$ and $\partial_t p(\cdot\,, 0) = 0$. In the case of circular measurement geometry one assumes that f vanishes outside the unit disc

$D_1 := \{\mathbf{r} \in \mathbb{R}^2 \mid \|\mathbf{r}\| < 1\}$ and the measurement sensors are located on the boundary $\partial D_1 = \mathbb{S}^1$. We assume that the phantom will not generate any data for some region $I \subseteq D_1$, for example when the acoustic pressure generated inside I is too small to be recorded. This masked PAT problem consists in the recovery of the function f from sampled noisy measurements of $g = \mathcal{W}(\mathbb{1}_{I^c} f)$ where $\mathcal{W}$ denotes the solution operator of the wave equation and $\mathbb{1}_{I^c}$ the indicator function on $I^c := \mathbb{R}^2 \setminus I$. Note that the resulting inverse problem can be seen of the combination of an inpainting problem and in inverse problems for the wave equation.

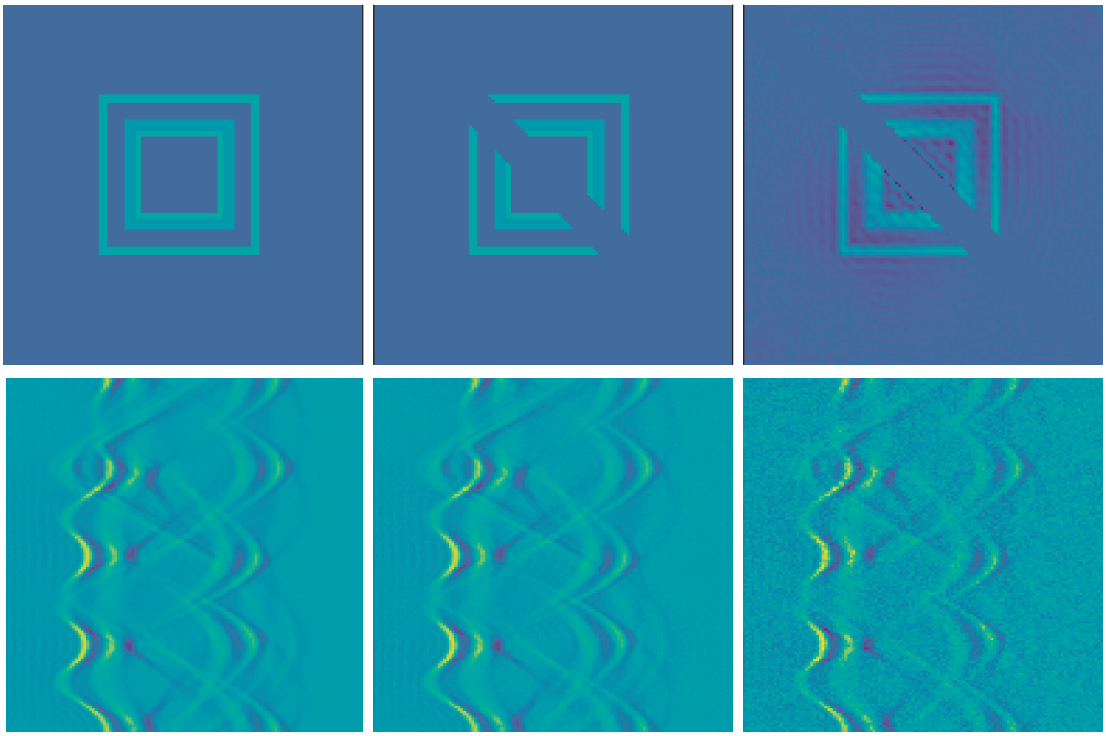

Figure 1: Top from left to right: phantom, masked phantom and initial reconstruction $\mathbf{A}^+\mathbf{A}x$. Bottom from left to right: data without noise, low noise $\sigma = 0.01$ and high noise $\sigma = 0.1$.

In order to implement the PAT forward operator we use a basis ansatz $f(\mathbf{r}) = \sum_{i=1}^{N\times N} x_i \psi(\mathbf{r} - \mathbf{r}_i)$ where $x_i \in \mathbb{R}$ are basis coefficients and $\psi \colon \mathbb{R}^2 \to \mathbb{R}$ a generalized Kaiser-Bessel (KB) and $\mathbf{r}_i = (i-1)/N$ with $i = (i_1, i_2) \in \{1, \dots, N\}^2$. The generalized KB functions are popular in tomographic inverse problems [16, 29, 32, 33] and denote radially symmetric functions with support in D_R defined by

$$\psi(\mathbf{r}) := \left(1 - \|\mathbf{r}\|^2/R^2\right)^{m/2} \frac{I_m\left(\gamma\sqrt{1 - \|\mathbf{r}\|^2/R^2}\right)}{I_m(\gamma)} \quad \text{for} \|\mathbf{r}\| \leq R\,. \tag{14}$$

Here I_m is the modified Bessel function of the first kind of order $n \in \mathbb{N}$ and the parameters $\gamma > 0$ and R denote the window taper and support radius, respectively. Since $\mathcal{W}$ is linear we have $\mathcal{W}f = \sum_{i=1}^{N\times N} x_i \mathcal{W}(\psi(\cdot - \mathbf{r}_i))$. For convenience we will use a pseudo-3D approach where use the 3D solution of $\mathcal{W}\psi$ for which there exists an analytical representation [32]. Denote by $\mathbf{s}_k$ uniformly spaced sensor locations on $\mathbb{S}^1$ and by $t_j > 0$ uniformly sampled measurement times in $[0, 2]$. Define the $N_t N_s \times N^2$ model

matrix by $\mathbf{W}_{N_t(k-1)+j,N(i_1-1)+i_2} = \mathcal{W}(\psi(\,\cdot\,-\mathbf{r}_i))(\mathbf{s}_k, t_j)$ and an $N^2 \times N^2$ diagonal matrix by $(\mathbf{M}_I)_{N(i_1-1)+i_2,N(i_1-1)+i_2} = 1$ if $\mathbf{r}_i \in I^c$ and zero otherwise. Let $\mathbf{W}\mathbf{M}_I = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\intercal$ be the singular valued decomposition. We then consider the discrete forward matrix $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}_\star\mathbf{V}^\intercal$ where $\mathbf{\Sigma}_\star$ is the diagonal matrix derived from $\mathbf{\Sigma}$ by setting singular values smaller than some $\sigma_\star$ to zero. In our experiments we use $N = N_t = 128$, $N_s = 150$ and take I fixed as a diagonal stripe of width 0.34.

Algorithm 1: NETT optimization.

Input: $y \in \mathbb{Y}$, $x_0 \in \mathbb{X}$, $\alpha, s > 0$.
Output: reconstruction x
for $\ell = 1, \dots, N_{\text{iter}}$ **do**
$x_{\ell+1/2} = x_{\ell-1} - s\alpha\nabla\mathcal{R}_n(x_{\ell-1})$
$x_\ell = (\mathbf{A}^\intercal\mathbf{A} - s\,\mathrm{Id})^{-1}\left(\mathbf{A}^\intercal y + s x_{\ell+1/2}\right)$
end

3.2 Discrete NETT

We consider the discrete NETT with discrepancy term $\mathcal{D}(\mathbf{A}x, y^\delta) = \|\mathbf{A}x - y^\delta\|_2^2/2$ and regularizer given by

$$\mathcal{R}^{(m)}(x) = \|x - \mathbf{\Phi}^{(m)}(x)\|_2^2 + \beta\|\nabla x\|_{1,\epsilon}\,, \tag{15}$$

where $\|\nabla x\|_{1,\epsilon} := \sum_{i_1,i_2=1}^{128}\sqrt{|\mathbf{r}_{i_1+1,i_2} - \mathbf{r}_i|^2 + |\mathbf{r}_{i_1,i_2+1} - \mathbf{r}_{i_1,i_2}|^2 + \epsilon^2}$ with $\epsilon > 0$ is a smooth version of the total variation [1]and $\mathbf{\Phi}^{(m)}$ is a learnable network. We take $\mathbf{\Phi}^{(m)}$ as the U-Net [27] with residual connection, which has first been applied to PAT image reconstruction in [4]. We generate training data that consist of square shaped rings with random profile and random location. See Figure 1 for an example of one such phantom (note that all plots in signal space use the same colorbar) and the corresponding data. We get a set of phantoms $x_1, \dots, x_{1000}$ and corresponding basic reconstructions $h_a := \mathbf{A}^+(\mathbf{A}x_a + \eta_a)$, where $\mathbf{A}^+$ is the pseudo-inverse and η_a is Gaussian noise with standard deviation of $\sigma\|\mathbf{A}x_a\|_\infty$ with $\sigma = 0.01$. The networks are trained by minimizing $\sum_{a=1}^{1000}\|\mathbf{\Phi}^{(m)}(h_a) - x_a\|_1 + \gamma\|\mathbf{\Phi}^{(m)}(x_a) - x_a\|_1$ where we used the Adam optimizer with learning rate 0.01 and $\gamma = 0.1$. The considered loss is that we want the trained regularizer to give small values for x_a and large values for h_a. The strategy is similar to [14] but we use the final output of the network for the regularizer as proposed in [5]. To minimize (15) we use Algorithm 1 which implements a forward-backward scheme [6].

3.3 Numerical results

For the numerical results we train two regularizers $\mathcal{R}^{(1)}$ and $\mathcal{R}^{(3)}$ as described in Section 3.2. The networks are implemented using PyTorch [23]. We also use PyTorch in order to calculate the gradient $\nabla_x\mathcal{R}^{(m)}$. We take $N_{\text{iter}} = 15$, $s = 0.25$ and $x_0 = \mathbf{\Phi}^{(m)}\mathbf{A}^\intercal y$ in Algorithm 1 and compute the inverse $(\mathbf{A}^\intercal\mathbf{A} - s\,\mathrm{Id})^{-1}$ only once and then use it for all examples. We set $\alpha = 0.015$ for the noise-free case, $\alpha = 0.016$ for the low noise

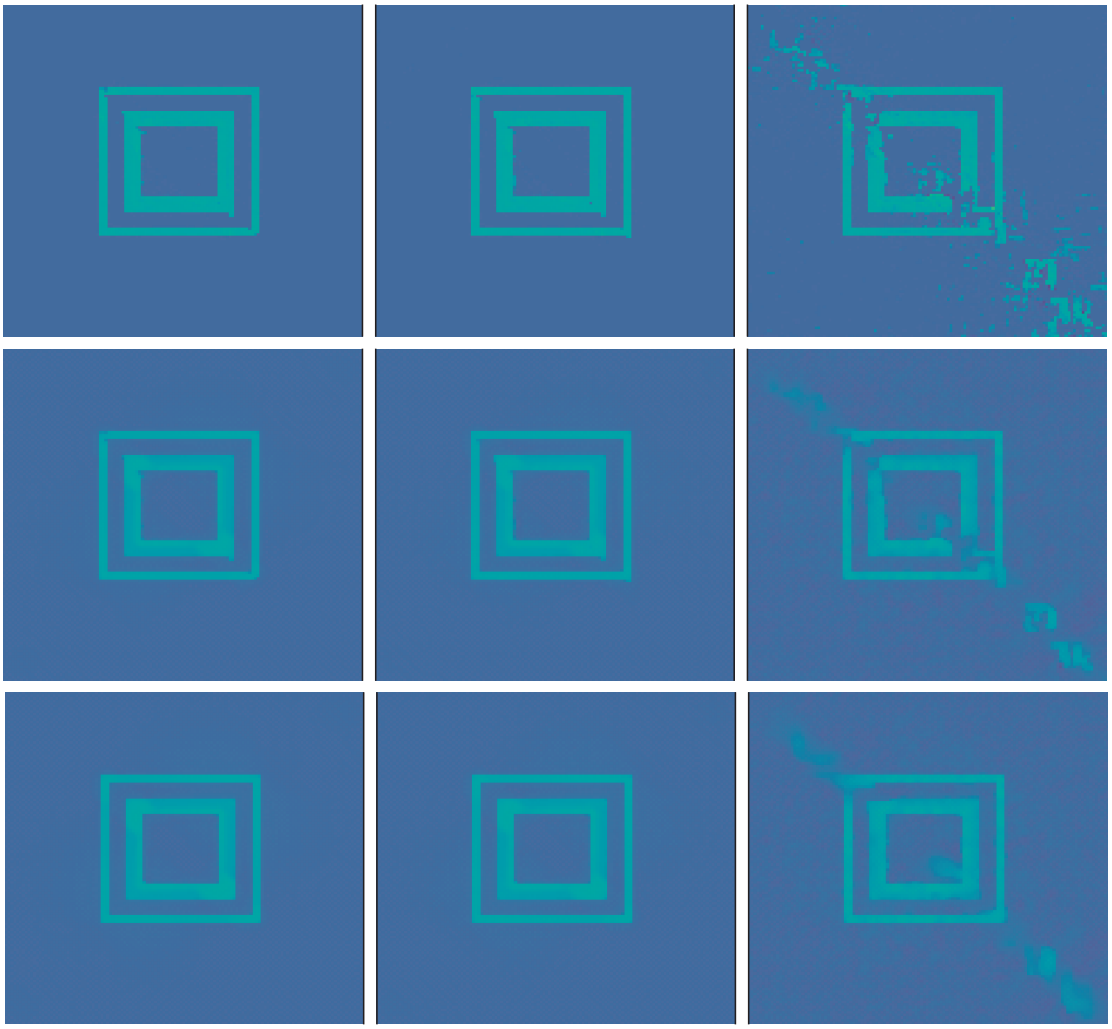

Figure 2: Top row: reconstructions using post-processing network $\Phi^{(1)}$. Middle row: NETT reconstructions using $\mathcal{R}^{(1)}$. Bottom row: NETT reconstructions using $\mathcal{R}^{(3)}$. From Left to Right: Reconstructions from data without noise, low noise ($\sigma = 0.01$) and high noise ($\sigma = 0.1$).

case and $\alpha = 0.02$ for the high noise cases, respectively, and selected a fixed $\beta = 15$. We expect that the NETT functional will yield better results due to data consistency, which is mainly helpful outside the masked center diagonal.

First we use the phantom from the testdata shown in Figure 1. The results using post processing and NETT are shown in Figure 2. One sees that all results with higher noise than used during training are not very good. This indicates that one should use similar noise as in the later applications even for the NETT. Figure 3 shows the average error using 10 test phantoms similar to the on in Figure 1. Careful numerical comparison of the numerical convergence rates and the theoretical results of Theorem 2.11 is an interesting aspect of further research. To investigate the stability of our method with respect to phantoms that are different from the training data we create a phantom with different structures as seen in Figure 4. As expected, the post processing network $\Phi^{(3)}$ is not really able to reconstruct the circles object, since it is quite different from the

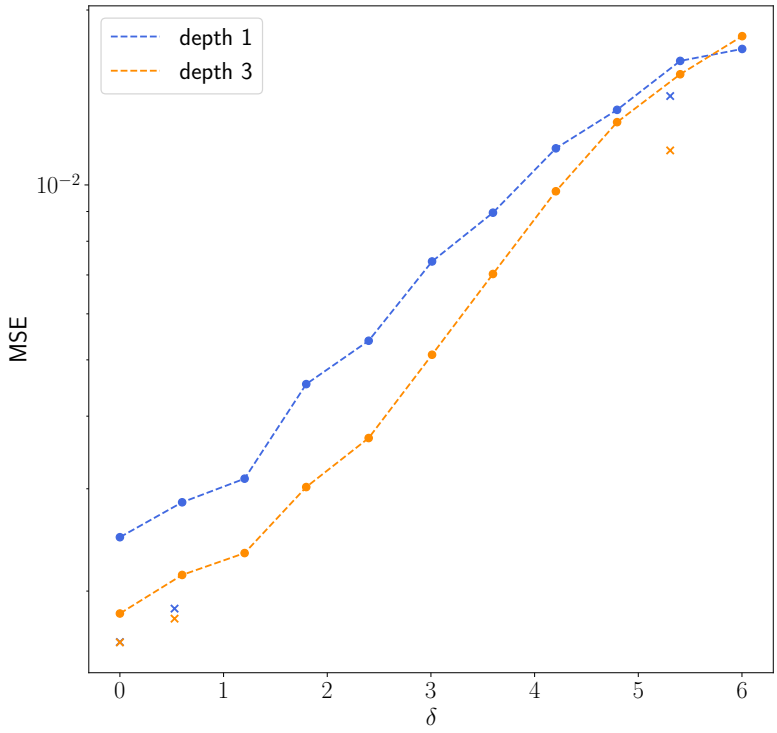

Figure 3: Semilogarithmic plot of the mean squared errors of the NETT using $\mathcal{R}^{(1)}$ and $\mathcal{R}^{(3)}$ depending on the noise level. The crosses are the values for the phantoms in Figure 2.

training data, but it also does not break down completely. On the other hand, the NETT approach yields good results due to data consistency.

4 Conclusion

We have analyzed the convergence a discretized NETT approach and derived the convergence rates under certain assumptions on the approximation quality of the involved operators. We performed numerical experiments using a limited data problem for PAT that is the combination of an inverse problem for the wave equation and an inpainting problem. To the best of our knowledge this is the first such problem studied with deep learning. The NETT approach yields better results that post processing for phantoms different from the training data. NETT still fails to recover some missing parts of the phantom in cases the data contains more noise than the training data. This highlights the relevance of using different regularizers for different noise levels.

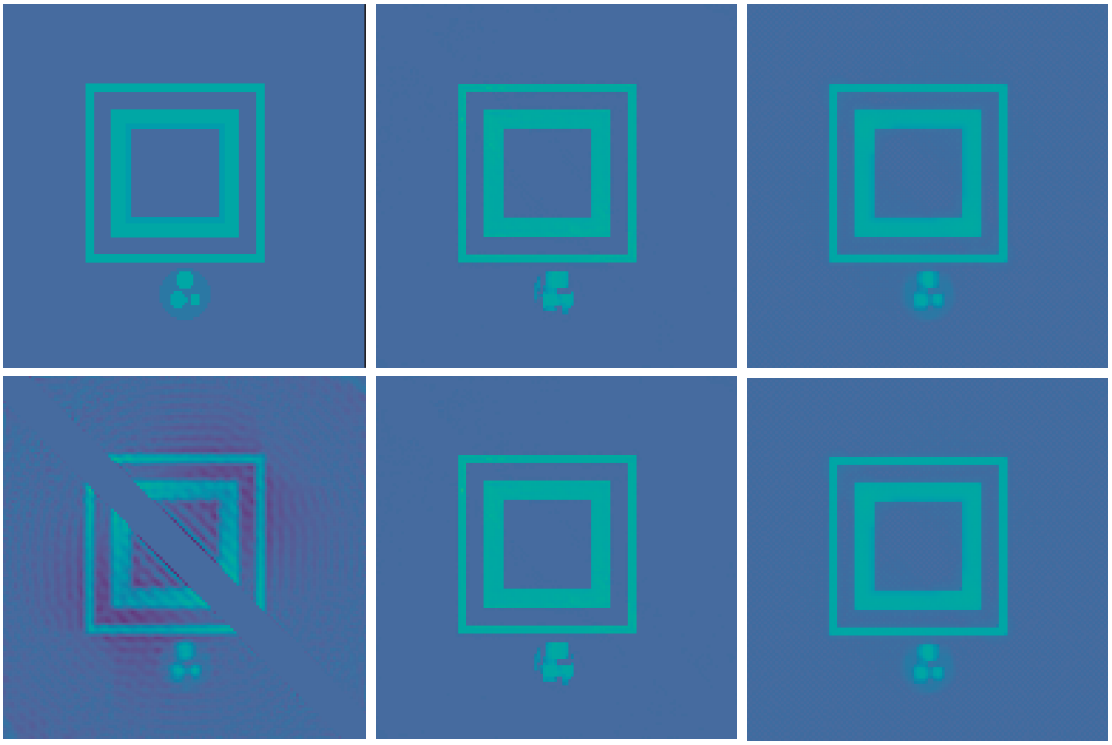

Figure 4: Left column: phantom with a structure not contained in the training data (top) and pseudo inverse reconstruction (bottom). Middle column: Post-processing reconstructions using exact (top) and noisy data (bottom). Right column: NETT reconstructions using exact (top) and noisy data (bottom).

Acknowledgments

S.A. and M.H. acknowledge support of the Austrian Science Fund (FWF), project P 30747-N32.

References

[1] R. Acar and C. R. Vogel. Analysis of bounded variation penalty methods for ill-posed problems. *Inverse Probl.*, 10(6):1217, 1994.

[2] J. Adler and O. Öktem. Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Probl.*, 33(12):124007, 2017.

[3] H. K. Aggarwal, M. P. Mani, and M. Jacob. MoDL: model-based deep learning architecture for inverse problems. *IEEE Trans. Med. Imaging*, 38(2):394–405, 2018.

[4] S. Antholzer, M. Haltmeier, and J. Schwab. Deep learning for photoacoustic tomography from sparse data. *Inverse Probl. Sci. Eng.*, 27(7):987–1005, 2019.

[5] S. Antholzer, J. Schwab, J. Bauer-Marschallinger, P. Burgholzer, and M. Haltmeier. Nett regularization for compressed sensing photoacoustic tomography. In *Photons Plus Ultrasound: Imaging and Sensing 2019*, volume 10878, page 108783B, 2019.

[6] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

[7] M. V. de Hoop, M. Lassas, and C. A. Wong. Deep learning architectures for nonlinear operator functions and nonlinear inverse problems. *arXiv:1912.11090*, 2019.

[8] 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.

[9] M. Grasmair. Generalized Bregman distances and convergence rates for non-convex regularization methods. *Inverse Probl.*, 26(11):115014, 2010.

[10] M. Haltmeier and L. V. Nguyen. Regularization of inverse problems by neural networks. *arXiv:2006.03972*, 2020.

[11] V. K. Ivanov, V. V. Vasin, and V. P. Tanana. *Theory of linear ill-posed problems and its applications*. Inverse and Ill-posed Problems Series. VSP, Utrecht, second edition, 2002.

[12] E. Kobler, T. Klatzer, K. Hammernik, and T. Pock. Variational networks: connecting variational methods and deep learning. In *German conference on pattern recognition*, pages 281–293. Springer, 2017.

[13] R. Kruger, P. Lui, Y. Fang, and R. Appledorn. Photoacoustic ultrasound (paus) – reconstruction tomography. *Med. Phys.*, 22(10):1605–1609, 1995.

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

[15] S. Lunz, O. Öktem, and C.-B. Schönlieb. Adversarial regularizers in inverse problems. In *NIPS*, pages 8507–8516, 2018.

[16] S. Matej and R. M. Lewitt. Practical considerations for 3-d image reconstruction using spherically symmetric volume elements. *IEEE Trans. Med. Imaging*, 15(1):68–78, 1996.

[17] V. A. Morozov. *Methods for Solving Incorrectly Posed Problems*. Springer Verlag, 1984.

[18] S. Mukherjee, S. Dittmer, Z. Shumaylov, S. Lunz, O. Öktem, and C.-B. Schönlieb. Learned convex regularizers for inverse problems. *arXiv:2008.02839*, 2020.

[19] F. Natterer and F. Wübbeling. *Mathematical Methods in Image Reconstruction*, volume 5 of *Monographs on Mathematical Modeling and Computation*. SIAM, Philadelphia, PA, 2001.

[20] D. Obmann, L. Nguyen, J. Schwab, and M. Haltmeier. Sparse ℓ^q-regularization of inverse problems using deep learning. *arXiv:1908.03006*, 2019.

[21] Obmann, D.; Nguyen, L.; Schwab, J.; Haltmeier, M. Augmented NETT regularization of inverse problems. *J. Phys. Commun.*, 5:105002, 2021.

[22] G. Paltauf, R. Nuster, M. Haltmeier, and P. Burgholzer. Photoacoustic tomography using a Mach-Zehnder interferometer as an acoustic line detector. *Appl. Opt.*, 46(16):3352–3358, 2007.

[23] A. Paszke, S. Gross, et al. Pytorch: An imperative style, high-performance deep learning library. In *NIPS*, pages 8024–8035. 2019.

[24] C. Pöschl. *Tikhonov regularization with general residual term*. Thesis, 2008.

[25] C. Pöschl, E. Resmerita, and O. Scherzer. Discretization of variational regularization in banach spaces. *Inverse Probl.*, 26(10):105017, 2010.

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

[27] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *MICCAI 2015*, pages 234–241, Cham, 2015. Springer.

[28] 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.

[29] J. Schwab, S. Pereverzyev Jr, and M. Haltmeier. A galerkin least squares approach for photoacoustic tomography. *SIAM J. Numer. Anal.*, 56(1):160–184, 2018.

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

[31] A. N. Tikhonov, A. S. Leonov, and A. G. Yagola. *Nonlinear ill-posed problems. Vol. 1, 2*, volume 14 of *Applied Mathematics and Mathematical Computation*. Chapman & Hall, London, 1998. Translated from the Russian.

[32] K. Wang, R. W. Schoonover, R. Su, A. Oraevsky, and M. A. Anastasio. Discrete imaging models for three-dimensional optoacoustic tomography using radially symmetric expansion functions. *IEEE Trans. Med. Imaging*, 33(5):1180–1193, 2014.

[33] K. Wang, R. Su, A. A. Oraevsky, and M. A. Anastasio. Investigation of iterative image reconstruction in three-dimensional optoacoustic tomography. *Phys. Med. Biol.*, 57(17):5399, 2012.

[34] Y. Yang, J. Sun, H. Li, and Z. Xu. Deep ADMM-net for compressive sensing MRI. In *Proc. 30th International Conference on Neural Information Processing Systems*, pages 10–18, 2016.

[35] M. S. Zhdanov. *Geophysical inverse theory and regularization problems*, volume 36. Elsevier, 2002.