Data driven regularization by projection

Andrea Aspri[1], Yury Korolev[2,*] and Otmar Scherzer[1,3]

[1] Johann Radon Institute for Computational and Applied Mathematics, Altenberger Straße 69, 4040 Linz, Austria
[2] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, United Kingdom
[3] Department of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria

E-mail: andrea.aspri@ricam.oeaw.ac.at, y.korolev@damtp.cam.ac.uk and otmar.scherzer@univie.ac.at

Abstract

We study linear inverse problems under the premise that the forward operator is not at hand but given indirectly through some input-output training pairs. We demonstrate that *regularization by projection* and *variational regularization* can be formulated by using the training data only and without making use of the forward operator. We study convergence and stability of the regularized solutions in view of Seidman (1980 *J. Optim. Theory Appl.* **30** 535), who showed that regularization by projection is *not convergent* in general, by giving some insight on the generality of Seidman's nonconvergence example. Moreover, we show, analytically and numerically, that regularization by projection is indeed capable of *learning* linear operators, such as the Radon transform.

Keywords: data driven regularization, variational regularization, regularization by projection, inverse problems, Gram–Schmidt orthogonalization

(Some figures may appear in colour only in the online journal)

1. Introduction

Linear inverse problems are concerned with the reconstruction of a quantity $u \in \mathcal{U}$ from indirect measurements $y \in \mathcal{Y}$ which are related by the linear *forward operator* $A : \mathcal{U} \to \mathcal{Y}$. A models the physics of data acquisition and may involve, for instance, integral transforms (such as

---

*Author to whom any correspondence should be addressed.

the Radon transform, see for instance [2, 3]) and partial differential equations (see for instance [4]).

Until recently the methods for solving inverse problems were *model driven*, meaning that the physics and chemistry of the measurement acquisition process was represented mathematically by the forward operator *A* (for some relevant applications see [2, 5, 6]). Nowadays, with the rise of the area of big data, methods that combine forward modeling with data driven techniques are being developed [7]. Some of these techniques build upon the similarity between deep neural networks and classical approaches to inverse problems such as iterative regularization [8, 9] and proximal methods [10]. Some are based on postprocessing of the reconstructions obtained by a simple inversion technique such as filtered backprojection [11]. Others use data driven regularizers in the context of variational regularization [12, 13] or use deep learning to learn a component of the solution in the null space of the forward operator [14, 15].

On the other hand, methods that discard forward modeling, that means knowledge of the operator *A*, have emerged. They are appealing because they do not require knowledge on the physics and chemistry of the data acquisition process, bypass the costly forward model evaluation and often yield results of superior visual quality. However, it has been demonstrated that for ill-posed inverse problems naïve applications of such methods can be unstable with respect to small perturbations in the measurement data [16, 17]. Moreover, there is currently no theory for purely data driven regularization in inverse problems, i.e. a theory in the setting when the forward operator is given only via training pairs

$$\{u^i, y^i\} \quad \text{such that} \quad Au^i = y^i. \tag{1.1}$$

In this paper we make a first step of an analysis for purely data driven regularization. Our goal is to develop and analyze stable algorithms for solving $Au = y$ *without* making use of the explicit knowledge of the forward operator *A* but having only

- approximate measurement y^δ such that $\|y - y^\delta\| \leqslant \delta$ and
- training pairs (1.1).

In particular, we demonstrate that regularization by projection [1, 18] and variational regularization [19] can be formulated in a data driven setting. We provide new analysis of regularization by projection and derive sufficient conditions for convergence that shed some new light on the classical nonconvergence example in [1]. Finally, in the numerical part of the paper we demonstrate that the methods we propose yield similar results to classical methods and require moderate amounts of training data. Note that we consider the idealized case when the training pairs (1.1) are noise-free. We leave out for future research the natural extension to the case of noisy training data.

There is a whole body of literature concerning regularization with various types of projections, typically onto subspaces spanned by some predefined bases such as polynomials or wavelets; a complete list of references would be too long to be presented here. Our approach differs from these works in the fact that our subspaces are defined by training data. This presents challenges since there is typically no relation between these subspaces and the properties of the forward operator (e.g., its eigenspaces) or the properties of the exact solution such as smoothness. It is hard, therefore, to make claims about optimality of the methods.

A connection between regularization by projection and machine learning has been observed by several authors, for instance, in the context of statistical learning [20, 21], where it was demonstrated that *subsampling* plays the role of regularization by (random) projections. From the statistical point of view, the idea to approximate an operator from input-output samples is closely related to statistical numerical approximation (see, e.g., [22] for a recent review) and in particular to optimal recovery [23].

Structure of the paper. In section 2 we summarize our assumptions and describe our setting and notation in more detail. Regularization by projection onto the space of input data is the topic of section 3. We derive sufficient conditions on the input data and the outputs data, under which projections result in a regularization. In section 4 we present another regularization method based on projections onto the space of output data, the *dual least squares* method [18], which is known to be convergent. We demonstrate, however, that it cannot be realized using the training pairs (1.1) and requires a different type of training data. In section 5 we study variational regularization and in section 6 we present numerical experiments with all three methods. We show that regularization by projection is indeed capable of *learning* linear operators, such as the Radon transform.

Main contributions. Our main goal in this article is to carry over some classical results in regularization theory to the model-free setting, with an overarching theme of using projections on subspaces defined by training data in the framework of regularization by projection. This perspective requires new, data driven regularity conditions, for which we show a relation to source conditions in special cases, whilst in general such relationship remains an open question. A classical nonconvergence result of Seidman [1] demonstrates the sharpness of this regularity condition.

   We also demonstrate that the amount of training data (which is also related the complexity of the learned model in regularization by projection) plays the role of a regularization parameter, hence for noisy training data the size of the training set should be chosen in agreement with the noise level. This is in accordance with the results in [24, theorem 4.2] on training neural networks from noisy data, where the number of neurons in the network plays a role similar to the number of training inputs in our setting.

2. Setting and main assumptions

We consider a linear inverse problem, consisting in solving

$$Au = y, \tag{2.1}$$

with a linear bounded forward operator $A : \mathcal{U} \to \mathcal{Y}$ acting between separable Hilbert spaces $\mathcal{U}$ and $\mathcal{Y}$. Different to the standard setting of linear inverse problems (see for instance [25]) we do not make use of the knowledge of the operator A but only assume knowledge of training pairs (1.1). This is beneficial if the modeling of the forward operator is rather incomplete, very uncertain, or the numerical evaluation is costly.

   We assume that A is injective and its inverse A^{-1} is unbounded, hence the problem of solving (2.1) is ill-posed. Instead of the exact measurement y we are given a noisy measurement y^δ, satisfying

$$\|y - y^\delta\| \leqslant \delta. \tag{2.2}$$

Throughout this paper, the solution of (2.1) with *exact data* $y \in \mathcal{R}(A)$ will be denoted by $u^\dagger$.

   We refer to elements of $\mathcal{U}$ as *inputs* and to elements of $\mathcal{Y}$ as *outputs*. Accordingly, speaking of the training pairs (1.1), we will refer to $\{u^i\}_{i=1,\dots,n}$ as *training inputs* and to $\{y^i\}_{i=1,\dots,n}$ as *training outputs*.

   We make the following assumptions on the training pairs throughout this paper.

Assumption 1 (Independence, uniform boundedness, sequentiality).

   Linear independence: For every $n \in \mathbb{N}$ the inputs $\{u^i\}_{i=1,\dots,n}$ are linearly independent.

   Uniform boundedness: There exist constants $c_u, C_u > 0$ such that $c_u \leqslant \|u^i\| \leqslant C_u$ for all $i \in \mathbb{N}$. Hence with no loss of generality we will assume that $\|u^i\| = 1$ for all $i \in \mathbb{N}$.

Sequentiality: The families of training pairs are nested, i.e. for every $n \in \mathbb{N}$

$$\left\{u^i, y^i\right\}_{i=1,\ldots,n+1} = \left\{u^i, y^i\right\}_{i=1,\ldots,n} \cup \left\{u^{n+1}, y^{n+1}\right\}. \tag{2.3}$$

If the training inputs $\{u^i\}_{i=1,\ldots,n}$ are linearly dependent, we will need to discard part them. Which ones to discard is an interesting question in itself, and a possible topic for future work.

Definition 1. We denote the spans of the inputs $\{u^i\}_{i=1,\ldots,n}$ and the outputs $\{y^i\}_{i=1,\ldots,n}$ by

$$\mathcal{U}_n := \operatorname{span}\{u^i\}_{i=1,\ldots,n} \quad \text{and} \quad \mathcal{Y}_n := \operatorname{span}\{y^i\}_{i=1,\ldots,n}. \tag{2.4}$$

Orthogonal projection operators onto $\mathcal{U}_n$ and $\mathcal{Y}_n$ are denoted by $P_{\mathcal{U}_n}$ and $P_{\mathcal{Y}_n}$, respectively.

Remark 2. From the assumption 1 and the hypotheses made on A we deduce that:

(a) Since the forward operator A is injective, the outputs $\{y^i\}_{i=1,\ldots,n}$ are also linearly independent.

(b) Since A is a bounded operator and the inputs $\{u^i\}_{i=1,\ldots,n}$ are uniformly bounded, the outputs $\{y^i\}_{i=1,\ldots,n}$ are also uniformly bounded and $\|y^i\| \leqslant \|A\|$, for all $i \in \mathbb{N}$.

From assumption 1 and (2.3), it follows that the subspaces $\mathcal{U}_n$, $\mathcal{Y}_n$ are nested, that is

$$\mathcal{U}_n \subset \mathcal{U}_{n+1}, \quad \mathcal{Y}_n \subset \mathcal{Y}_{n+1}, \quad \text{for all } n.$$

We also need to make an assumption that the training data are sufficiently rich in the sense that the collection of all inputs $\{u^i\}_{i=1,\ldots,n}$ is dense in $\mathcal{U}$.

Assumption 2 (Density). We assume that the subspaces spanned by the inputs $\{u^i\}_{i=1,\ldots,n}$ are dense in $\mathcal{U}$, that is

$$\overline{\bigcup_{n\in\mathbb{N}} \mathcal{U}_n} = \mathcal{U}.$$

As a consequence of the previous assumptions, we have

Proposition 3. *By assumption 2 the subspaces spanned by the training outputs $\{y^i\}_{i=1,\ldots,n}$ are dense in $\overline{\mathcal{R}(A)}$, i.e.,*

$$\overline{\bigcup_{n\in\mathbb{N}} \mathcal{Y}_n} = \overline{\mathcal{R}(A)}.$$

We carry out our analysis in the setting when the training pairs $\{u^i, y^i\}_{i=1,\ldots,n} \subset \mathcal{U} \times \mathcal{Y}$ are infinite-dimensional. The case when these pairs are discretized, i.e. $\{u^i, y^i\}_{i=1,\ldots,n} \subset \mathbb{R}^k \times \mathbb{R}^l$ and $k, l, n \to \infty$ at certain rates is beyond the scope of this paper.

3. Regularization by projection

Let $y \in \mathcal{R}(A)$ be the exact, noise-free right-hand side in (2.1) and consider the following projected problem

$$AP_{\mathcal{U}_n} u = y. \tag{3.1}$$

Our goal in this section is to provide an explicit representation formula for the minimum norm solution of (3.1) in terms of the training pairs (1.1). We first observe that the minimum norm solution of (3.1) is given by

$$u_n^{\mathcal{U}} = (AP_{\mathcal{U}_n})^\dagger y, \tag{3.2}$$

where $(AP_{\mathcal{U}_n})^\dagger$ denotes the Moore–Penrose inverse of $AP_{\mathcal{U}_n}$ (see [26]). The superscript $\mathcal{U}$ in $u_n^{\mathcal{U}}$ reflects the fact that the projection in (3.1) takes place in $\mathcal{U}$. The need for this notation will become clear in section 4, where will use a projection of the original equation (2.1) in the space $\mathcal{Y}$.

3.1. A reconstruction formula

The following result shows that in the injective case, a simple formula for $(AP_{\mathcal{U}_n})^\dagger$ exists.

Theorem 4. *Let $P_{\mathcal{U}_n}$ and $P_{\mathcal{Y}_n}$ as in definition 1. Then the Moore–Penrose inverse of $AP_{\mathcal{U}_n}$ is given by*

$$(AP_{\mathcal{U}_n})^\dagger = A^{-1}P_{\mathcal{Y}_n}.$$

Proof. First we observe that $\mathcal{R}(A^{-1}P_{\mathcal{Y}_n}) = \mathcal{U}_n = \mathcal{N}(AP_{\mathcal{U}_n})^\perp$. To see this, note that for any $z \in \mathcal{Y}$ we have

$$P_{\mathcal{Y}_n}z = \sum_{i=1}^n \lambda_i^n y^i$$

(where the expansion coefficients, in general, change with n) and hence for $u = A^{-1}P_{\mathcal{Y}_n}z$ we have by definition

$$u = A^{-1}\left(\sum_{i=1}^n \lambda_i^n y^i\right) = \sum_{i=1}^n \lambda_i^n u^i \in \mathcal{U}_n.$$

On the other hand, for any $u \in \mathcal{U}_n$ we can also find a representation in the form $u = A^{-1}P_{\mathcal{Y}_n}y$ for some $y \in \mathcal{Y}$. Thus the first identity is shown. For the second identity, let $v \in \mathcal{N}(AP_{\mathcal{U}_n})$. Since A is injective, this is equivalent to $v \in \mathcal{N}(P_{\mathcal{U}_n}) = \mathcal{U}_n^\perp$ and hence $\mathcal{N}(AP_{\mathcal{U}_n}) = \mathcal{U}_n^\perp$. Since $\mathcal{U}_n$ is finite-dimensional and hence closed, this implies the second identity.

Using the obvious identities

$$P_{\mathcal{Y}_n}AP_{\mathcal{U}_n} = AP_{\mathcal{U}_n} \quad \text{and} \quad P_{\mathcal{U}_n}A^{-1}P_{\mathcal{Y}_n} = A^{-1}P_{\mathcal{Y}_n},$$

we directly verify the Moore–Penrose equations (see for instance [26]):

- $AP_{\mathcal{U}_n}A^{-1}P_{\mathcal{Y}_n}AP_{\mathcal{U}_n} = AP_{\mathcal{U}_n}A^{-1}AP_{\mathcal{U}_n} = AP_{\mathcal{U}_n}$;
- $A^{-1}P_{\mathcal{Y}_n}AP_{\mathcal{U}_n}A^{-1}P_{\mathcal{Y}_n} = A^{-1}P_{\mathcal{Y}_n}AA^{-1}P_{\mathcal{Y}_n} = A^{-1}P_{\mathcal{Y}_n}$;
- $A^{-1}P_{\mathcal{Y}_n}AP_{\mathcal{U}_n} = A^{-1}AP_{\mathcal{U}_n} = P_{\mathcal{U}_n} = I - P_{\mathcal{U}_n^\perp} = I - P_{\mathcal{N}(AP_{\mathcal{U}_n})}$;
- $AP_{\mathcal{U}_n}A^{-1}P_{\mathcal{Y}_n} = AA^{-1}P_{\mathcal{Y}_n} = P_{\mathcal{Y}_n} = P_{\mathcal{R}(AP_{\mathcal{U}_n})}$.

Since the Moore–Penrose equations uniquely characterize the Moore–Penrose inverse, the assertion follows. $\qquad\square$

Combination of (3.2) and theorem 4 shows that the Moore–Penrose inverse of (3.1) is given by

$$u_n^{\mathcal{U}} = A^{-1}P_{\mathcal{Y}_n}y. \tag{3.3}$$

Gram–Schmidt orthogonalization in $\mathcal{Y}$. We start by applying the Gram–Schmidt process to the outputs $\{y^i\}_{i=1,\dots,n}$ to obtain an orthonormal basis of $\mathcal{Y}_n$ (if the training data were not linearly independent, this will be detected by the Gram–Schmidt algorithm and the redundant data will be dismissed, effectively reducing the size of the training set n). We denote this basis

by $\{\underline{y}^i\}_{i=1,\ldots,n}$. By solving $Au = \underline{y}^i$ for $i = 1, \ldots, n$, we obtain, in general, a non-orthogonal basis $\{\underline{u}^i\}_{i=1,\ldots,n}$ of $\mathcal{U}_n$. In a matrix form, we can write that

$$Y_n = \underline{Y}_n R_n \quad \text{and} \quad \underline{U}_n = U_n R_n^{-1}, \tag{3.4}$$

where Y_n, $\underline{Y}_n$ and $\underline{U}_n$ are composed of a finite number of infinite-dimensional functions $\{y^i\}_{i=1,\ldots,n}$, $\{\underline{y}^i\}_{i=1,\ldots,n}$ and $\{\underline{u}^i\}_{i=1,\ldots,n}$, respectively,

$$Y_n := (y^1, \ldots, y^n), \quad \underline{Y}_n = (\underline{y}^1, \ldots, \underline{y}^n), \quad \underline{U}_n = (\underline{u}^1, \ldots, \underline{u}^n) \tag{3.5}$$

and R_n is an upper triangular $n \times n$ transformation matrix

$$R_n := \begin{pmatrix} (y^1, \underline{y}^1) & (y^2, \underline{y}^1) & \cdots & (y^n, \underline{y}^1) \\ 0 & (y^2, \underline{y}^2) & \cdots & (y^n, \underline{y}^2) \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & (y^n, \underline{y}^n) \end{pmatrix}. \tag{3.6}$$

Taking the noise-free right-hand side $y \in \mathcal{R}(A)$ in (2.1) and expanding $P_{\mathcal{Y}_n} y \in \mathcal{Y}_n$ in the orthonormal basis $\{\underline{y}^i\}_{i=1,\ldots,n}$, we get

$$P_{\mathcal{Y}_n} y = \sum_{i=1}^{n} (y, \underline{y}^i) \underline{y}^i,$$

hence using theorem 4 we can write (3.2) as follows

$$u_n^{\mathcal{U}} = A^{-1} P_{\mathcal{Y}_n} y = \sum_{i=1}^{n} (y, \underline{y}^i) \underline{u}^i, \tag{3.7}$$

where $\{\underline{u}^i\}_{i=1,\ldots,n}$ are the transformed inputs satisfying $A\underline{u}^i = \underline{y}^i$, $i = 1, \ldots, n$.

The transformed inputs can be calculated as follows:

Remark 5. It can be easily verified that the transformed inputs $\{\underline{u}^i\}_{i=1,\ldots,n}$ satisfy

$$\underline{u}^n = \frac{u^n - \sum_{i=1}^{n-1} (y^n, \underline{y}^i) \underline{u}^i}{\|y^n - P_{\mathcal{Y}_{n-1}} y^n\|}.$$

Hence,

$$\|\underline{u}^n\| = \frac{\|u^n - \sum_{i=1}^{n-1} (y^n, \underline{y}^i) \underline{u}^i\|}{\|y^n - P_{\mathcal{Y}_{n-1}} y^n\|} \geqslant \frac{\|u^n - P_{\mathcal{U}_{n-1}} u^n\|}{\|y^n - P_{\mathcal{Y}_{n-1}} y^n\|}. \tag{3.8}$$

The next remark considers a very particular set of training pairs, namely the singular values of a compact operator. With such peculiar training data regularization by projection becomes a regularization method, as the following remark shows:

Remark 6. Let A be compact with singular value decomposition $\{\sigma^i, x^i, z^i\}_{i\in\mathbb{N}}$, and assume that the training pairs are given by $\{(u^i := x^i, y^i = \sigma^i z^i = Ax^i)\}_{i=1,\ldots,n}$ (see for instance [18]). Consequently, $\{\underline{y}^i\}_{i=1,\ldots,n} = \{z^i\}_{i\in\mathbb{N}}$ is an orthonormal system and accordingly we get $\{\underline{u}^i\}_{i=1,\ldots,n} = \left\{\frac{x^i}{\sigma^i}\right\}$. Then (3.7) becomes

$$u_n^{\mathcal{U}} = \sum_{i=1}^{n} \frac{1}{\sigma^i}(y, z^i) x^i = \sum_{i=1}^{n} \frac{1}{\sigma^i}(Au^\dagger, z^i) x^i = \sum_{i=1}^{n} (u^\dagger, x^i) x^i,$$

i.e. $u_n^{\mathcal{U}}$ is the projection of the exact solution $u^{\dagger}$ onto the span of the first n singular vectors $\{x^i\}_{i=1,\dots,n}$.

In fact this method is a regularization method (see [18]).

3.2. Behavior in the limit of infinite data $n \to \infty$

The nonconvergence example by Seidman [1] demonstrates that, **in general**, the minimum norm solution $u_n^{\mathcal{U}}$ of (3.1) **does not converge** to the exact solution $u^{\dagger}$ as $n \to \infty$, i.e. $A^{-1}P_{\mathcal{Y}_n}$ is not a regularization of A^{-1} (see (3.3)). 'In general' in the above sentence refers to particular cases such as when the training pairs are not singular functions, in which case the method becomes truncated SVD, which is indeed a regularization method as we have outlined in remark 6.

Below we analyze the convergence of the Moore–Penrose inverse $u_n^{\mathcal{U}}$ of (3.1) as represented in (3.7) in a general setting, in particular, when the training pairs are not spectral pairs as discussed in remark 6. Letting $n \to \infty$ and applying the Gram–Schmidt process to the sequence $\{y^i\}_{i\in\mathbb{N}}$, we obtain an orthonormal basis $\{\underline{y}^i\}_{i\in\mathbb{N}}$ of $\overline{\mathcal{R}(A)}$ and a corresponding sequence $\{\underline{u}^i\}_{i\in\mathbb{N}}$ such that $A\underline{u}^i = \underline{y}^i$.

An essential algorithmic step for implementing (3.7) is the orthonormalization of the set $\{y^i\}$, which we perform with the Gram–Schmidt algorithm. Therefore we analyze first the stability of this algorithm, which depends on the diagonal elements $(y^i, \underline{y}^i)$ of the matrices R_n, $i = 1, \dots, n$. It is clear that

$$|(y^i, \underline{y}^i)| = \|y^i - P_{\mathcal{Y}_{i-1}}y^i\| \quad \forall\, i \in \mathbb{N}. \tag{3.9}$$

If A is compact then the part of y^n that lies outside the span of the previous points $\{y^i\}_{i=1,\dots,n-1}$ will become arbitrary small as $n \to \infty$, which shows the instability of the Gram–Schmidt algorithm:

Proposition 7. *If A is compact then*

$$\liminf_{i\to\infty} \|y^i - P_{\mathcal{Y}_{i-1}}y^i\| = 0.$$

Proof. Since by assumption 1 all training images $\{u^i\}_{i\in\mathbb{N}}$ are uniformly bounded and A is compact, the sequence $\{y^i\}_{i\in\mathbb{N}}$ has a convergent subsequence (that we do not relabel), which, in particular, satisfies

$$\|y^i - y^{i-1}\| \to 0 \quad \text{as} \quad i \to \infty.$$

Then we obtain the following estimate

$$\|y^i - P_{\mathcal{Y}_{i-1}}y^i\| = \min_{y\in\mathcal{Y}_{i-1}} \|y^i - y\| \leqslant \|y^i - y^{i-1}\| \to 0, \tag{3.10}$$

which proves the assertion. $\qquad\qquad\square$

If the inputs $\{u^i\}_{i=1,\dots,n}$ are such that

$$\|u^n - P_{\mathcal{U}_{n-1}}u^n\| \geqslant C \quad \forall\, n \in \mathbb{N},$$

then clearly $\|\underline{u}^i\| \to \infty$ as $n \to \infty$ by proposition 7 and remark 5. Hence, the Gram–Schmidt process becomes unstable as $n \to \infty$ and the transformed inputs $\{\underline{u}^i\}_{i=1,\dots,n}$ may become

unbounded, as we see from (3.9) and (3.6). Since $\{\underline{u}^i\}_{i=1,\dots,n}$ are linearly independent, we can expand the projection $P_{\mathcal{U}_n} u^\dagger$ of the exact solution of (2.1) in the basis of $\mathcal{U}_n$ as follows

$$P_{\mathcal{U}_n} u^\dagger = \sum_{i=1}^n \alpha_i^n \underline{u}^i, \tag{3.11}$$

where the expansion coefficients might be varying with respect to n, i.e., $\alpha_i^n \neq \alpha_i^m$ for $n \neq m$, since $\{\underline{u}^i\}_{i=1,\dots,n}$ are not orthogonal.

The next result shows how far the coefficients of the expansion of the Moore–Penrose approximation $u_n^{\mathcal{U}}$ in (3.7), i.e., $(y, \underline{y}^i)$, deviate from the coefficients α_i^n in (3.11) of the best-approximating solution $P_{\mathcal{U}_n} u^\dagger$ of $u^\dagger$.

Proposition 8. *Let $P_{\mathcal{U}_n} u^\dagger$ be represented as in (3.11), then the following identity holds*

$$\sum_{i=1}^n ((y, \underline{y}^i) - \alpha_i^n)^2 = \|y - AP_{\mathcal{U}_n} u^\dagger\|^2 - \|y - P_{\mathcal{Y}_n} Au^\dagger\|^2.$$

Proof. Consider the residual

$$u^\dagger - P_{\mathcal{U}_n} u^\dagger = u^\dagger - \sum_{i=1}^n \alpha_i^n \underline{u}^i.$$

Applying A to both sides and expanding $y = Au^\dagger$ in the orthonormal basis $\{\underline{y}^i\}_{i \in \mathbb{N}}$, we get

$$A(u^\dagger - P_{\mathcal{U}_n} u^\dagger) = \sum_{i=1}^\infty (y, \underline{y}^i)\underline{y}^i - \sum_{i=1}^n \alpha_i^n \underline{y}^i$$

$$= \sum_{i=1}^n ((y, \underline{y}^i) - \alpha_i^n)\underline{y}^i + \sum_{i=n+1}^\infty (y, \underline{y}^i)\underline{y}^i$$

and hence

$$\|A(u^\dagger - P_{\mathcal{U}_n} u^\dagger)\|^2 = \sum_{i=1}^n ((y, \underline{y}^i) - \alpha_i^n)^2 + \|y - P_{\mathcal{Y}_n} y\|^2.$$

Rearranging terms, we get the assertion. $\square$

The first term on the right-hand side becomes 0 if $AP_{\mathcal{U}_n} = P_{\mathcal{Y}_n} A$, which is for instance the case if the training pairs consist of the singular value decomposition (see remark 6). Therefore, the better the subspaces $\mathcal{U}_n$ and $\mathcal{Y}_n$ agree with the spaces spanned by the spectral functions of A, the smaller the discrepancy between $(y, \underline{y}^i)$ and α_i^n for $i, n \in \mathbb{N}$ will be.

3.3. Convergence analysis

In the following we analyze weak and strong convergence of the Moore–Penrose approximation $u_n^{\mathcal{U}}$ of (3.1), also in the case of noisy data.

3.3.1. Weak convergence. The first result gives us weak convergence of $u_n^{\mathcal{U}}$ for $n \to \infty$ in the case of noise free data y (see (2.2)).

Theorem 9. *Let $y \in \mathcal{R}(A)$ be the exact right-hand side in (2.1) and $\{u^i, y^i\}_{i=1,\dots,n}$ the training pairs defined in (1.1). Let λ_i^n be the expansion coefficients of $P_{\mathcal{Y}_n} y \in \mathcal{Y}_n$ in the*

non-orthogonal basis of training outputs $\{y^i\}_{i=1,\ldots,n}$

$$P_{\mathcal{Y}_n} y = \sum_{i=1}^{n} \lambda_i^n y^i.$$

Then $u_n^{\mathcal{U}}$ *as defined in* (3.7) *is bounded uniformly in n and converges weakly to* $u^\dagger$

$$u_n^{\mathcal{U}} \rightharpoonup u^\dagger \quad as \quad n \to \infty$$

if and only if there exists a constant $C_\lambda < \infty$ *such that*

$$\left\| \sum_{i=1}^{n} \lambda_i^n u^i \right\| \leqslant C_\lambda \quad \forall\, n \in \mathbb{N}. \tag{3.12}$$

Proof. We rewrite (3.7) as follows

$$u_n^{\mathcal{U}} = A^{-1} P_{\mathcal{Y}_n} y = A^{-1} \sum_{i=1}^{n} \lambda_i^n y^i = \sum_{i=1}^{n} \lambda_i^n u^i,$$

hence $u_n^{\mathcal{U}}$ is uniformly bounded if and only if (3.12) holds. By [18, theorem 3.20], $u_n^{\mathcal{U}} \rightharpoonup u^\dagger$ if and only if it is uniformly bounded. $\qquad\square$

Below we provide *a priori* conditions on the training data (1.1) and the exact solution $u^\dagger$ that ensure boundedness of (3.7). Moreover, we discuss them in the context of Seidman's nonconvergence example [1] for convergence of regularization by projection.

3.3.2. Gram–Schmidt orthogonalization in $\mathcal{U}$. We will use Gram–Schmidt orthogonalization again, this time on the inputs $\{u^i\}_{i=1,\ldots,n}$. We denote the resulting orthonormal basis of $\mathcal{U}_n$ by $\{\overline{u}^i\}_{i=1,\ldots,n}$. Solving for $\{\overline{y}^i\}_{i=1,\ldots,n}$ such that $A\overline{u}^i = \overline{y}^i$ we obtain, in general, a non-orthogonal basis $\{\overline{y}^i\}_{i=1,\ldots,n}$ of $\mathcal{Y}_n$.

Letting $n \to \infty$ and applying the Gram–Schmidt process to the sequence $\{u^i\}_{i\in\mathbb{N}}$ we obtain an orthonormal basis $\{\overline{u}^i\}_{i\in\mathbb{N}}$ of $\mathcal{U}$ and a corresponding sequence $\{\overline{y}^i\}_{i\in\mathbb{N}}$ such that $A\overline{u}^i = \overline{y}^i$. Expanding the exact solution in the basis $\{\overline{u}^i\}_{i\in\mathbb{N}}$, we get

$$u^\dagger = \sum_{i=1}^{\infty} (u^\dagger, \overline{u}^i)\overline{u}^i \tag{3.13}$$

and hence for the exact data $y = Au^\dagger$ we get

$$y = \sum_{i=1}^{\infty} (u^\dagger, \overline{u}^i)\overline{y}^i. \tag{3.14}$$

To check that the latter series is convergent, we note that the partial sums

$$\sum_{i=1}^{n} (u^\dagger, \overline{u}^i)\overline{y}^i = A \sum_{i=1}^{n} (u^\dagger, \overline{u}^i)\overline{u}^i$$

are bounded since the operator A is bounded.

Let us recall that we consider two orthonormalization procedures in this paper, which can be easily confused:

(a) For $\{\underline{u}^i, \underline{y}^i\}_{i=1,\dots,n}$ from section 3.1 the outputs $\{\underline{y}^i\}_{i=1,\dots,n}$ are orthonormal and the inputs $\{\underline{u}^i\}_{i=1,\dots,n}$ are chosen to match these outputs, while

(b) for $\{\overline{u}^i, \overline{y}^i\}_{i=1,\dots,n}$ the inputs $\{\overline{u}^i\}_{i=1,\dots,n}$ are orthonormal and the outputs $\{\overline{y}^i\}_{i=1,\dots,n}$ are chosen to match these inputs.

Thus a bar at the bottom stands for orthonormalization in the range of A and the bar above stands for an orthonormalization in the domain of A.

Below we will study the effect of some regularity assumptions of the solution $u^\dagger$ on convergence of $u_n^{\mathcal{U}}$ for $n \to \infty$.

Assumption 3. Coefficients of the expansion (3.13) are in ℓ^1, i.e.

$$\sum_{i=1}^{\infty} |(u^\dagger, \overline{u}^i)| < \infty.$$

Assumption 3 is a data-driven regularity assumption on the exact solution $u^\dagger$. We show in section 3.3.3 that in Seidman's nonconvergence example assumption 3 is implied by a source condition. The question whether there is more general relationship between assumption 3 and source conditions remains open.

We emphasize that to prove even weak convergence of regularization by projection (3.7), additional regularity conditions are unavoidable, since, as demonstrated by Seidman's nonconvergence example, weak convergence fails in general.

Assumption 4. For every $n \in \mathbb{N}$ and any $i \geqslant n + 1$ consider the following expansion of $P_{\mathcal{Y}_n}\overline{y}^i \in \mathcal{Y}_n$

$$P_{\mathcal{Y}_n}\overline{y}^i = \sum_{j=1}^{n} \beta_j^{i,n} \overline{y}^j. \tag{3.15}$$

We assume that for every $n \in \mathbb{N}$

$$\sum_{j=1}^{n} (\beta_j^{i,n})^2 \leqslant C, \quad \text{for every} \quad i \geqslant n + 1, \tag{3.16}$$

where $C > 0$ is a constant independent of i and n.

We extend the definition of $\beta_j^{i,n}$ for $i \leqslant n$.

Definition 10. For every $i \leqslant n$, $\overline{y}^i \in \mathcal{Y}_n$ we define $\beta_j^{i,n} = \delta_{ij}$, where δ_{ij} is the Kronecker symbol.

From this definition it follows that

$$\sum_{j=1}^{n} (\beta_j^{i,n})^2 = 1, \quad \text{for every} \quad i \leqslant n. \tag{3.17}$$

Theorem 11. *Let assumptions 3 and 4 be satisfied. Then $u_n^{\mathcal{U}}$ as defined in (3.7) is uniformly bounded with respect to n.*

Proof. Applying $P_{\mathcal{Y}_n}$ to (3.14), we get

$$P_{\mathcal{Y}_n}y = \sum_{i=1}^{\infty} (u^\dagger, \overline{u}^i) P_{\mathcal{Y}_n}\overline{y}^i$$

and hence, by applying A^{-1} to both sides, we get that

$$u_n^{\mathcal{U}} = A^{-1}P_{\mathcal{Y}_n}y = \sum_{i=1}^{\infty} (u^{\dagger}, \overline{u}^i)A^{-1}P_{\mathcal{Y}_n}\overline{y}^i.$$

Using Hölder's inequality, we estimate

$$\|u_n^{\mathcal{U}}\| \leqslant \sum_{i=1}^{\infty} \left(|(u^{\dagger}, \overline{u}^i)|\right) \sup_{i=1,\dots,\infty} \|A^{-1}P_{\mathcal{Y}_n}\overline{y}^i\|.$$

The sum $\sum_{i=1}^{\infty} |(u^{\dagger}, \overline{u}^i)|$ is bounded by assumption 3. We further observe that

$$A^{-1}P_{\mathcal{Y}_n}\overline{y}^i = A^{-1}\sum_{j=1}^{n} \beta_j^{i,n}\overline{y}^j = \sum_{j=1}^{n} \beta_j^{i,n}\overline{u}^j$$

and, since $\{\overline{u}^i\}_{i=1,\dots,n}$ are orthonormal, we get that

$$\|A^{-1}P_{\mathcal{Y}_n}\overline{y}^i\|^2 = \sum_{j=1}^{n} (\beta_j^{i,n})^2,$$

which is bounded uniformly in n and i by assumption 4. Therefore, $\sup_{i=1,\dots,\infty}\|A^{-1}P_{\mathcal{Y}_n}\overline{y}^i\|$ is bounded uniformly in n and

$$\|u_n^{\mathcal{U}}\| \leqslant C < \infty \quad \text{uniformly in } n.$$

$\square$

3.3.3. Seidman's nonconvergence example. Next we discuss Seidman's example from [1] on nonconvergence of regularization by projection.

Example 12 (Seidman [1]). Let $\{e^i\}_{i\in\mathbb{N}}$ be any orthonormal basis in $\mathcal{U}$ and $A : \mathcal{U} \to \mathcal{U}$ be defined as follows

$$A : \sum_{i=1}^{\infty} \xi_i e^i \to \sum_{i=1}^{\infty} (a_i\xi_i + b_i\xi_1)e^i$$

with

$$b_i = \begin{cases} 0, & \text{if } i = 1, \\ i^{-1}, & \text{if } i \geqslant 2, \end{cases} \quad a_i = \begin{cases} i^{-1}, & \text{if } i \text{ is odd}, \\ i^{-\frac{5}{2}}, & \text{if } i \text{ is even}. \end{cases} \tag{3.18}$$

This operator is injective and compact (for details see also [18]). In our notation, $\{\overline{u}^i\}_{i\in\mathbb{N}} = \{e^i\}_{i\in\mathbb{N}}$ (since the inputs are already orthonormalized) and

$$\overline{y}^i = A\overline{u}^i = \begin{cases} \displaystyle\sum_{j=1}^{\infty} j^{-1}e^j, & i = 1, \\ a_i e^i, & i \geqslant 2. \end{cases}$$

Or in other words our training pairs are $(e^i = \overline{u}^i, \overline{y}^i)_{i\in\mathbb{N}}$.

For every $n \geqslant 1$ and every fixed $i \geqslant n+1$, we expand $P_{\mathcal{Y}_n}\bar{y}^i \in \mathcal{Y}_n$ as follows

$$P_{\mathcal{Y}_n}\bar{y}^i = \sum_{k=1}^{n} \gamma_k \bar{y}^k = \gamma_1 \sum_{j=1}^{\infty} j^{-1}e^j + \sum_{k=2}^{n} \gamma_k a_k e^k. \tag{3.19}$$

For all $k = 2, \ldots, n$ and $i \geqslant n+1$ we therefore have that

$$(P_{\mathcal{Y}_n}\bar{y}^i, e^k) = (\bar{y}^i, P_{\mathcal{Y}_n}e^k) = (\bar{y}^i, e^k) = a_i(e^i, e^k) = 0,$$

hence, taking a scalar product with e^k in (3.19), we conclude that $\gamma_1 k^{-1} + \gamma_k a_k = 0$ and

$$\gamma_k = -\gamma_1 \frac{k^{-1}}{a_k}. \tag{3.20}$$

Now, to compute γ_1, we note that it minimizes the following expression for $i \geqslant n+1$

$$\begin{aligned}
\|P_{\mathcal{Y}_n}\bar{y}^i - \bar{y}^i\|^2 &= \left\| \gamma_1 \sum_{j=1}^{\infty} j^{-1}e^j + \sum_{k=2}^{n} \gamma_k a_k e^k - a_i e^i \right\|^2 \\
&= \left\| \gamma_1 e^1 + \sum_{k=2}^{n} \left(\gamma_1 k^{-1} + \gamma_k a_k \right) e^k + \gamma_1 \sum_{j=n+1}^{\infty} j^{-1}e^j - a_i e^i \right\|^2 \\
&\underset{(3.20)}{=} \left\| \gamma_1 e^1 + \gamma_1 \sum_{\substack{j=n+1 \\ j \neq i}}^{\infty} j^{-1}e^j + \left(\gamma_1 i^{-1} - a_i \right) e^i \right\|^2 \\
&= \gamma_1^2 + \gamma_1^2 \sum_{\substack{j=n+1 \\ j \neq i}}^{\infty} j^{-2} + \left(\gamma_1 i^{-1} - a_i \right)^2 \\
&= \gamma_1^2 \left(1 + \sum_{j=n+1}^{\infty} j^{-2} \right) - 2\gamma_1 i^{-1} a_i + a_i^2.
\end{aligned}$$

The minimizer of this quadratic expression with respect to γ_1 is given by

$$\gamma_1 = \frac{i^{-1}a_i}{1 + \sum_{j=n+1}^{\infty} j^{-2}} := C_n i^{-1} a_i, \tag{3.21}$$

where $0 < \frac{1}{1 + \pi^2/6} \leqslant C_n := \frac{1}{1 + \sum_{j=n+1}^{\infty} j^{-2}} \leqslant 1$ is uniformly bounded from below and above with respect to n.

Hence we get from (3.20) that

$$\gamma_k = -\gamma_1 \frac{k^{-1}}{a_k} = -C_n i^{-1} a_i \frac{k^{-1}}{a_k}. \tag{3.22}$$

Since $0 < a_i \leqslant i^{-1}$ for all $i \in \mathbb{N}$ and $\frac{k^{-1}}{a_k} \leqslant k^{3/2}$ for all $k \in \mathbb{N}$ (see (3.18)), from (3.21) and (3.22) we find

$$|\gamma_1| \leqslant C_n i^{-2} \leqslant C_n i^{-\frac{1}{2}},$$

$$|\gamma_k| \leqslant C_n i^{-2} k^{\frac{3}{2}} \leqslant C_n i^{-\frac{1}{2}}, \quad \text{for} \quad k = 2, \ldots, n.$$

Therefore,

$$\sum_{k=1}^{n} \gamma_k^2 \leqslant n C_n^2 i^{-1} \leqslant C_n^2 \leqslant 1 \quad \text{for all } i \geqslant n+1$$

and assumption 4 is satisfied.

The nonconvergence example in [1] is obtained with the following exact solution

$$u^\dagger = \sum_{i=1}^{\infty} i^{-1} e^i.$$

Clearly, the expansion coefficients are in ℓ^2, hence $u^\dagger$ is well defined. However, assumption 3 fails since

$$\sum_{i=1}^{\infty} i^{-1} = \infty.$$

Therefore, the key to the nonconvergence example is the slow decay of the expansion coefficients in (3.13) and hence a violation of assumption 3. This assumption is indeed satisfied in our numerical experiments with natural images (photographs of people) in section 6.

Numerically validating assumption 4 can be complicated due to numerical errors. Our numerical experiments in section 6 do not give a definitive answer. However, as the above example demonstrates, this assumption can be satisfied for data generated by a compact operator.

Important for the analysis of regularization methods are source conditions see [25] and some references such as [27–29], indeed the l^1-condition from assumption 3 is related to the source condition as the following remark shows.

Remark 13. In Seidman's example, the l^1-condition from assumption 3 follows from a source condition

$$u^\dagger \in \mathcal{R}(A^*). \tag{3.23}$$

Indeed, suppose that $u = A^* v$ for some $v \in \mathcal{U}$. We get

$$\sum_{i=1}^{\infty} |(u, e^i)| = \sum_{i=1}^{\infty} |(A^* v, e^i)| = \sum_{i=1}^{\infty} |(v, A e^i)| = |(v, A e^1)| + \sum_{i=2}^{\infty} |(v, a_i e^i)|$$

$$= |(v, \bar{y}^1)| + \sum_{i=2}^{\infty} a_i |(v, e^i)| \leqslant \|v\| \|\bar{y}^1\| + \|v\| \sqrt{\sum_{i=2}^{\infty} a_i^2}.$$

Since $a_i \leqslant i^{-1}$ for all $i \in \mathbb{N}$, we get that

$$\sum_{i=1}^{\infty} |(u, e^i)| \leqslant \|v\| \|\bar{y}^1\| + \|v\| \sqrt{\sum_{i=2}^{\infty} i^{-2}} < \infty.$$

Therefore, for Seidman's nonconvergence example $u^\dagger$ violates this source condition, and in particular this means that for obtaining a nonconvergent sequence of regularized solutions $u_n^{\mathcal{U}}$, roughness of the solution $u^\dagger$ (meaning not satisfying a source condition) is required.

Concluding our discussion of Seidman's nonconvergence example, we recall that the nonconvergence of the regularized solution $u_n^{\mathcal{U}}$ would also not hold if we project onto eigenspaces of the operator A (cf. remark 6). This is independent of a smoothness assumption on $u^\dagger$.

3.3.4. Strong convergence. Thanks to the assumptions on the operator A and training pairs, expanding the exact data $y \in \mathcal{R}(A)$ of (2.1) in the basis of the orthonormalized training outputs $\{\underline{y}^i\}_{i=1,\ldots,n}$ and formally applying A^{-1}, we obtain the following formal expansion

$$u^\dagger = \sum_{i=1}^{\infty} (y, \underline{y}^i)\underline{u}^i, \tag{3.24}$$

where $\{\underline{u}^i\}_{i=1,\ldots,n}$ are the transformed inputs satisfying $A\underline{u}^i = \underline{y}^i$, $i = 1, \ldots, n$ (cf. section 3.1). The residual $u^\dagger - u_n^{\mathcal{U}}$ is given by

$$u^\dagger - u_n^{\mathcal{U}} = \sum_{i=n+1}^{\infty} (y, \underline{y}^i)\underline{u}^i.$$

A sufficient condition for convergence of this residual is absolute convergence of the series (3.24):

$$\sum_{i=1}^{\infty} |(y, \underline{y}^i)| \|\underline{u}^i\| < \infty. \tag{3.25}$$

By remark 5 we have a lower bound for $\|\underline{u}^i\|$

$$\|\underline{u}^i\| = \frac{\|A^{-1}(y^i - P_{\mathcal{Y}_{i-1}}y^i)\|}{\|y^i - P_{\mathcal{Y}_{i-1}}y^i\|} \geqslant \frac{\|u^i - P_{\mathcal{U}_{i-1}}u^i\|}{\|y^i - P_{\mathcal{Y}_{i-1}}y^i\|}, \quad \forall\, i \geqslant 2. \tag{3.26}$$

Hence it is clear that for this series to converge, the coefficients $(y, \underline{y}^i)$ would have to decay fast. To understand how fast they need to decay, we need an upper bound on the norm of $\underline{u}^i$. The following result shows that under assumption 4 the estimate (3.26) is actually sharp up to a constant factor.

Proposition 14. *Let assumption 4 be satisfied. Then the following estimate holds*

$$\|\underline{u}^i\| \leqslant \sqrt{C+1}\frac{\|u^i - P_{\mathcal{U}_{i-1}}u^i\|}{\|y^i - P_{\mathcal{Y}_{i-1}}y^i\|}, \quad \forall\; i \geqslant 2,$$

where C is the constant from assumption 4.

Proof. First we observe that for all $i \geqslant 2$ the following identity holds

$$y^i - P_{\mathcal{Y}_{i-1}}y^i = (\overline{y}^i - P_{\mathcal{Y}_{i-1}}\overline{y}^i)\|u^i - P_{\mathcal{U}_{i-1}}u^i\|, \tag{3.27}$$

where $\{u^i, y^i\}_{i=1,\ldots,n}$ are the training pairs from (1.1) and $\{\overline{u}^i, \overline{y}^i\}_{i=1,\ldots,n}$ are the transformed training pairs such that $\{\overline{u}^i\}_{i=1,\ldots,n}$ are orthonormal and $\{\overline{y}^i\}_{i=1,\ldots,n}$ satisfy $A\overline{u}^i = \overline{y}^i$ (cf. section 3.3.2). Indeed, the Gram–Schmidt process yields that for all $i \geqslant 2$

$$\overline{u}^i = \frac{u^i - P_{\mathcal{U}_{i-1}}u^i}{\|u^i - P_{\mathcal{U}_{i-1}}u^i\|} \quad \text{and} \quad \overline{y}^i = \frac{1}{\|u^i - P_{\mathcal{U}_{i-1}}u^i\|}A(u^i - P_{\mathcal{U}_{i-1}}u^i).$$

Hence, we get

$$(\bar{y}^i - P_{\mathcal{Y}_{i-1}}\bar{y}^i)\|u^i - P_{\mathcal{U}_{i-1}}u^i\| = A(u^i - P_{\mathcal{U}_{i-1}}u^i) - P_{\mathcal{Y}_{i-1}}A(u^i - P_{\mathcal{U}_{i-1}}u^i)$$
$$= y^i - P_{\mathcal{Y}_{i-1}}y^i - AP_{\mathcal{U}_{i-1}}u^i + P_{\mathcal{Y}_{i-1}}AP_{\mathcal{U}_{i-1}}u^i$$
$$= y^i - P_{\mathcal{Y}_{i-1}}y^i - AP_{\mathcal{U}_{i-1}}u^i + AP_{\mathcal{U}_{i-1}}u^i$$
$$= y^i - P_{\mathcal{Y}_{i-1}}y^i,$$

where we used the obvious identity $P_{\mathcal{Y}_{i-1}}AP_{\mathcal{U}_{i-1}} = AP_{\mathcal{U}_{i-1}}$. Now, using the expansion in assumption 4 with $n = i - 1$, we get

$$P_{\mathcal{Y}_{i-1}}\bar{y}^i = \sum_{j=1}^{i-1} \beta_j^{i,i-1}\bar{y}^j,$$

and hence

$$A^{-1}(\bar{y}^i - P_{\mathcal{Y}_{i-1}}\bar{y}^i) = \bar{u}^i - \sum_{j=1}^{i-1} \beta_j^{i,i-1}\bar{u}^j.$$

Since $\{\bar{u}^i\}_{i=1,\dots,n}$ are orthonormal, we get

$$\|A^{-1}(\bar{y}^i - P_{\mathcal{Y}_{i-1}}\bar{y}^i)\|^2 = 1 + \sum_{j=1}^{i-1} (\beta_j^{i,i-1})^2 \leqslant C + 1,$$

where the upper bound follows from assumption 4. Using (3.27), we get

$$\|A^{-1}(y^i - P_{\mathcal{Y}_{i-1}}y^i)\| = \|A^{-1}(\bar{y}^i - P_{\mathcal{Y}_{i-1}}\bar{y}^i)\|\|u^i - P_{\mathcal{U}_{i-1}}u^i\|$$
$$\leqslant \sqrt{C+1}\|u^i - P_{\mathcal{U}_{i-1}}u^i\|.$$

Finally, we obtain the desired estimate

$$\|\underline{u}^i\| = \frac{\|A^{-1}(y^i - P_{\mathcal{Y}_{i-1}}y^i)\|}{\|y^i - P_{\mathcal{Y}_{i-1}}y^i\|} \leqslant \sqrt{C+1}\frac{\|u^i - P_{\mathcal{U}_{i-1}}u^i\|}{\|y^i - P_{\mathcal{Y}_{i-1}}y^i\|}.$$

$\square$

Hence, the sufficient condition (3.25) together with (3.9) can be written as follows

$$\sum_{i=1}^{\infty} |(y,\underline{y}^i)|\frac{\|u^i - P_{\mathcal{U}_{i-1}}u^i\|}{\|y^i - P_{\mathcal{Y}_{i-1}}y^i\|} = \sum_{i=1}^{\infty} \frac{|(y,\underline{y}^i)|}{|(y^i,\underline{y}^i)|}|(u^i,\bar{u}^i)| < \infty, \tag{3.28}$$

where $\{\underline{y}^i\}_{i=1,\dots,n}$ are the orthonormalized outputs (cf. section 3.1) and $\{\bar{u}^i\}_{i=1,\dots,n}$ are the orthonormalized inputs (cf. section 3.3.2). Convergence of this series can be ensured either by sufficiently fast decay of the coefficients $(y,\underline{y}^i)$ of the expansion of the exact $y \in \mathcal{R}(A)$ in the orthonormal basis $\{\underline{y}^i\}_{i\in\mathbb{N}}$ or by sufficiently fast decay of the scalar products $(u^i,\bar{u}^i)$. Note that in $(u^i,\bar{u}^i)$, the index i is present at both positions in the scalar product and $(u^i,\bar{u}^i)$ are not expansion coefficients of any particular element of $\mathcal{U}$ in the basis $\{\bar{u}^i\}_{i\in\mathbb{N}}$. Rather, $|(u^i,\bar{u}^i)| = \|u^i - P_{\mathcal{U}_{i-1}}u^i\|$ is the part of the training input u^i that lies outside the span of the previous training vectors $\{u^i\}_{i=1,\dots,n-1}$.

For real data where the training inputs $\{u^i\}_{i=1,\dots,n}$ and the exact solution $u^\dagger$ are similar, it is likely that $|(y,\underline{y}^i)|$ and $|(y^i,\underline{y}^i)|$ will be of similar order, hence $\frac{|(y,\underline{y}^i)|}{|(y^i,\underline{y}^i)|}$ is unlikely to decay fast. Decay of the scalar products $(u^i,\overline{u}^i)$, however, can be controlled by choosing sufficiently similar training inputs $\{u^i\}_{i=1,\dots,n}$. Hence, we obtain the following sufficient condition.

Theorem 15. *Let $y \in \mathcal{R}(A)$ be the exact right-hand side in (2.1). Suppose that*

(a) the expansion coefficients $(y, \underline{y}^i)$ of $y \in \mathcal{R}(A)$ in the orthonormal basis $\{\underline{y}^i\}_{i\in\mathbb{N}}$ satisfy

$$\sup_{i=1,\dots,\infty} \frac{|(y,\underline{y}^i)|}{|(y^i,\underline{y}^i)|} \leqslant C < \infty, \tag{3.29}$$

where C is a constant independent of n and $\{\underline{y}^i\}_{i\in\mathbb{N}}$ are the orthonormalized outputs,
(b) the training inputs $\{u^i\}_{i\in\mathbb{N}}$ satisfy

$$\sum_{i=1}^{\infty} |(u^i,\overline{u}^i)| < \infty, \tag{3.30}$$

where $\{\overline{u}^i\}_{i\in\mathbb{N}}$ are the orthonormalized inputs,
(c) Assumption 4 is satisfied.

Then the solution obtained with regularization by projection (3.7) converges strongly to the exact solution of (2.1), i.e.

$$u_n^{\mathcal{U}} \to u^\dagger.$$

Proof. Using Hölder's inequity in (3.28), we get that

$$\sum_{i=1}^{\infty} \frac{|(y,\underline{y}^i)|}{|(y^i,\underline{y}^i)|}|(u^i,\overline{u}^i)| \leqslant \left(\sup_{i=1,\dots,\infty} \frac{|(y,\underline{y}^i)|}{|(y^i,\underline{y}^i)|}\right) \sum_{i=1}^{\infty} |(u^i,\overline{u}^i)| < \infty.$$

This implies (3.25), which is sufficient for strong convergence of the series (3.7). $\square$

Remark 16. We notice the similarity between the condition (3.30) and assumption 3. Both require ℓ^1 convergence of certain series; assumption 3 requires it for $(u^\dagger,\overline{u}^i)$, where $u^\dagger$ is the exact solution of (2.1), whilst (3.30) requires ℓ^1 convergence of $(u^i,\overline{u}^i)$. If the exact solution $u^\dagger$ and the training inputs $\{u^i\}_{i\in\mathbb{N}}$ are similar, both conditions are likely to be satisfied or violated simultaneously. Both conditions are satisfied in our numerical experiments with natural images (photographs) in section 6. Condition (3.30) seems also to be satisfied in these experiments.

3.4. Noisy data

Even if conditions of theorem 15 are satisfied and we have convergence for clean data $y \in \mathcal{R}(A)$, for noisy data $y^\delta \notin \mathcal{R}(A)$ convergence may fail. Noise in the data y^δ is understood as a deterministic perturbation $\Delta \notin \mathcal{R}(A)$ such that $\|\Delta\| \leqslant \delta$ and $y_\delta = y + \Delta$. Define

$$u_{n,\delta}^{\mathcal{U}} := A^{-1} P_{\mathcal{Y}_n} y^\delta. \tag{3.31}$$

Since $P_{\mathcal{Y}_n} y^\delta \in \mathcal{R}(A)$ for all n, this is well defined. Now we get that

$$u^\dagger - u_{n,\delta}^{\mathcal{U}} = A^{-1}y - A^{-1} P_{\mathcal{Y}_n}(y + \Delta) = A^{-1}(I - P_{\mathcal{Y}_n})Au^\dagger - A^{-1} P_{\mathcal{Y}_n}\Delta$$

and therefore

$$\|u^\dagger - u^{\mathcal{U}}_{n,\delta}\| \leqslant \|A^{-1}(I - P_{\mathcal{Y}_n})Au^\dagger\| + \|A^{-1}P_{\mathcal{Y}_n}\Delta\|. \tag{3.32}$$

We observe the typical *semi-convergence* behavior. While the first term converges to zero as $n \to \infty$ under assumptions of theorem 11, the second term clearly explodes. The amount of training pairs n therefore plays the role of a regularization parameter that balances the influence of the two error terms in (3.32).

Theorem 17.   *Let $n = n(\delta)$ be such that $n(\delta) \to \infty$ as $\delta \to 0$ and*

$$\delta\sqrt{n} \sup_{i=1,\dots,n} \frac{1}{|(y^i, \underline{y}^i)|} \to 0 \quad as \quad \delta \to 0.$$

Then (3.31) *together with the function $n(\delta)$ (which we refer to as an* a priori *parameter choice rule* [18]) *defines a convergent regularization, i.e.*

$$u^{\mathcal{U}}_{n(\delta),\delta} \to u^\dagger \quad as \quad \delta \to 0.$$

Proof.   All we need to do is estimate the growth of $\|A^{-1}P_{\mathcal{Y}_n}\|$ as $n \to \infty$. For an arbitrary $z \in \mathcal{Y}$ with $\|z\| \leqslant 1$ we expand $P_{\mathcal{Y}_n}z \in \mathcal{Y}_n$ in the orthonormal basis $\{y^i\}_{i=1,\dots,n}$ of $\mathcal{Y}_n$ (see section 3.1) and get

$$\|A^{-1}P_{\mathcal{Y}_n}z\| = \left\| A^{-1}\sum_{i=1}^{n}(z, \underline{y}^i)\underline{y}^i \right\| = \left\| \sum_{i=1}^{n}(z, \underline{y}^i)\underline{u}^i \right\|$$

$$\leqslant \sqrt{\sum_{i=1}^{n}(z, \underline{y}^i)^2}\sqrt{\sum_{i=1}^{n}\|\underline{u}^i\|^2} \leqslant \|\underline{U}_n\|_2,$$

where $\underline{U}_n$ is composed of the functions $\{\underline{u}^i\}_{i=1,\dots,n}$ as defined in (3.5). Using (3.4), we further estimate

$$\|\underline{U}_n\|_2 = \|U_nR_n^{-1}\|_2 \leqslant \|U_n\|_2\|R_n^{-1}\|_2,$$

where U_n is composed of the functions $\{u^i\}_{i=1,\dots,n}$ as defined in (3.5) and R_n is the upper triangular $n \times n$ transformation matrix defined in (3.6). Now, since R_n is upper triangular, its eigenvalues are given by its diagonal entries and $\|R_n^{-1}\|_2$ is the inverse of the smallest one, i.e.

$$\|R_n^{-1}\|_2 = \sup_{i=1,\dots,n} \frac{1}{|(y^i, \underline{y}^i)|}.$$

We further note that

$$\|U_n\|_2 = \sqrt{\sum_{i=1}^{n}\|u^i\|^2} = \sqrt{n}$$

by assumption 1 and finally we obtain

$$\|A^{-1}P_{\mathcal{Y}_n}z\| \leqslant \sqrt{n} \sup_{i=1,\dots,n} \frac{1}{|(y^i, \underline{y}^i)|}.$$

The rest follows from (3.32).   □

4. Dual least squares

Although projecting the equation (2.1) in $\mathcal{U}$ as in (3.1) does not yield a convergent solution in general, it is known that projecting (2.1) in $\mathcal{Y}$ yields convergent solutions. This method is also referred to as *dual least squares* [18].

The dual least squares method consists of finding the minimum norm solution (i.e. least squares solution with minimal norm) of the following problem

$$P_{\mathcal{Y}_n}Au = P_{\mathcal{Y}_n}y, \tag{4.1}$$

where $P_{\mathcal{Y}_n}$ is the orthogonal projector onto the span of the outputs $\{y^i\}_{i=1,\dots,n}$ and $y \in \mathcal{R}(A)$ is the exact right-hand side in (2.1). We denote the minimum norm solution of (4.1) by $u_n^{\mathcal{Y}}$, where the superscript $\mathcal{Y}$ emphasizes the fact that the projection in (4.1) takes place in $\mathcal{Y}$ (cf. section 3).

The following result shows that $u_n^{\mathcal{Y}}$ converges strongly to the exact solution $u^\dagger$ as $n \to \infty$.

Theorem 18 ([18, theorem 3.24]). *Let $y \in \mathcal{R}(A)$ be the exact data in* (2.1). *Then the minimum norm solution of* (4.1) *is given by*

$$u_n^{\mathcal{Y}} = P_{A^*\mathcal{Y}_n}u^\dagger, \tag{4.2}$$

where $P_{A^*\mathcal{Y}_n}$ is the orthogonal projector onto $A^*\mathcal{Y}_n$. Consequently,

$$u_n^{\mathcal{Y}} \to u^\dagger \quad as \quad n \to \infty.$$

Hence, taking projections in $\mathcal{Y}$ as in (4.1) seems a good alternative to projecting in the input space $\mathcal{U}$ as in (3.1). However, we will demonstrate that one cannot solve (4.1) using training pairs (1.1). Instead, one requires training data for the adjoint A^* of the form

$$\{v^i, y^i\}_{i=1,\dots,n} \quad \text{such that} \quad v^i = A^*y^i. \tag{4.3}$$

It is not clear whether this kind of training data can be obtained in practice, hence the relevance of the dual least squares method in the data driven setting is not clear.

A reconstruction formula. The following result gives a simple characterization of the Moore–Penrose inverse of $P_{\mathcal{Y}_n}A$, similarly to theorem 4, and highlights a connection between minimum norm solutions of projected problems (3.1) and (4.1).

Theorem 19. *Let A have a dense range, $P_{\mathcal{Y}_n}$ as in definition 1 and $P_{A^*\mathcal{Y}_n}$ as defined above. Then the Moore–Penrose inverse of $P_{\mathcal{Y}_n}A$ is given by*

$$(P_{\mathcal{Y}_n}A)^\dagger = P_{A^*\mathcal{Y}_n}A^{-1}.$$

Hence, the minimum norm solution $u_n^{\mathcal{Y}}$ of (4.1) *is given by*

$$u_n^{\mathcal{Y}} = P_{A^*\mathcal{Y}_n}A^{-1}P_{\mathcal{Y}_n}y = P_{A^*\mathcal{Y}_n}u_n^{\mathcal{U}}, \tag{4.4}$$

where $u_n^{\mathcal{U}}$ is the minimum norm solution of (3.1) *as defined in* (3.7).

Proof. First we observe that

$$(P_{\mathcal{Y}_n}A)^\dagger = ((A^*P_{\mathcal{Y}_n})^*)^\dagger = ((A^*P_{\mathcal{Y}_n})^\dagger)^*.$$

Since $\mathcal{R}(A)$ is dense in $\mathcal{Y}$, the adjoint A^* is injective, hence we can proceed similarly to theorem 4 and obtain

$$(A^*P_{\mathcal{Y}_n})^\dagger = (A^*)^\dagger P_{A^*\mathcal{Y}_n} = (A^\dagger)^*P_{A^*\mathcal{Y}_n} = (A^{-1})^*P_{A^*\mathcal{Y}_n},$$

since A is injective. Taking the adjoint, we get that

$$(P_{\mathcal{Y}_n}A)^\dagger = ((A^{-1})^*P_{A^*\mathcal{Y}_n})^* = P_{A^*\mathcal{Y}_n}A^{-1}.$$

The formula (4.4) follows from this expression. $\qquad\qquad\qquad\qquad\qquad\square$

Remark 20.   Comparing (4.2) and (4.4), we notice that

$$P_{A^*\mathcal{Y}_n}u_n^{\mathcal{U}} = P_{A^*\mathcal{Y}_n}u^\dagger,$$

i.e. $u_n^{\mathcal{U}}$ and $u^\dagger$ differ only on the orthogonal complement of $A^*\mathcal{Y}_n$.

Since the subspace $A^*\mathcal{Y}_n$ is given by

$$A^*\mathcal{Y}_n = \operatorname{span}\{A^*y^i\}_{i=1,\dots,n},$$

to compute the projection $P_{A^*\mathcal{Y}_n}$ one needs to know what A^*y^i are. Hence, training pairs of the form (4.3) are required, i.e. one needs training data for the *adjoint A^**.

Using the orthonormal basis $\{\underline{y}^i\}_{i=1,\dots,n}$ of $\mathcal{Y}_n$, we can write the projected equation (4.1) as follows

$$\sum_{i=1}^{n} (Au, \underline{y}^i)\underline{y}^i = \sum_{i=1}^{n} (y, \underline{y}^i)\underline{y}^i$$

and hence

$$(y, \underline{y}^i) = (Au, \underline{y}^i) = (u, A^*\underline{y}^i) = (u, \underline{v}^i), \quad i = 1, \dots, n,$$

with $\underline{v}^i := A^*\underline{y}^i$, $i = 1, \dots, n$. Hence, $u_n^{\mathcal{Y}}$ solves the following problem

$$\min_{u \in \mathcal{U}} \|u\|^2 \quad \text{s.t.} (u, \underline{v}^i) = (y, \underline{y}^i), \quad \text{for} \quad i = 1, \dots, n, \tag{4.5}$$

which does not require evaluating A and can be solved using only the training pairs for the adjoint A^* (4.3).

Noisy data. To complete the presentation of the dual least squares method, we consider reconstructions from noisy data $y^\delta \notin \mathcal{R}(A)$ such that $\|y - y^\delta\| \leqslant \delta$. In this case the size of the training set n also plays the role of the regularization parameter, as the following result demonstrates.

Theorem 21 ([18, theorem 3.26]).   *Let $y \in \mathcal{R}(A)$ and $y^\delta \in \mathcal{Y}$ s.t. $\|y - y^\delta\| \leqslant \delta$. Let μ_n be the smallest singular value of $P_{\mathcal{Y}_n}A$. If $n = n(\delta)$ is chosen such that $n(\delta) \to \infty$ and $\frac{\delta}{\mu_{n(\delta)}} \to 0$ as $n \to \infty$ then*

$$u_{n,\delta}^{\mathcal{Y}} := (Q_{n(\delta)}A)^\dagger Q_{n(\delta)}y^\delta \to u^\dagger \quad as \quad \delta \to 0. \tag{4.6}$$

Non-zero singular values of $P_{\mathcal{Y}_n}A$ coincide with those of $(P_{\mathcal{Y}_n}A)^* = A^*P_{\mathcal{Y}_n}$ and can be computed using the Gram matrix

$$(A^*\underline{y}^i, A^*\underline{y}^j)_{i,j=1,\dots,n} = (\underline{v}^i, \underline{v}^j)_{i,j=1,\dots,n},$$

which is available from (4.3), and a parameter choice rule similar to that in theorem 17 can be obtained.

5. Variational regularization

In this section we turn back to the projected problem (3.1) using projections in $\mathcal{U}$. If assumptions of theorem 9 (or theorem 11) are not satisfied and the minimum norm solution of (3.1) may become unbounded as $n \to \infty$, explicit regularization can be used to ensure boundedness and convergence. A straightforward way to approximate a minimum norm solution of (3.1) is Tikhonov regularization

$$\min_{u \in \mathcal{U}} \frac{1}{2} \|AP_{\mathcal{U}_n} u - y\|^2 + \alpha \|u\|^2, \tag{5.1}$$

where $y \in \mathcal{R}(A)$ is the exact right-hand side in (2.1), $P_{\mathcal{U}_n}$ is the projector onto the span of the inputs $\{u^i\}_{i=1,\dots,n}$ and $\alpha \in \mathbb{R}_+$ is a regularization parameter. Clearly, by setting $\alpha = 0$ we obtain a least-squares solution of (3.1).

More generally, the term $\|u\|^2$ in (5.1) can be replaced with an arbitrary regularizer $\mathcal{J} : \mathcal{U} \to \mathbb{R} \cup \{+\infty\}$ to obtain the following variational regularization problem

$$\min_{u \in \mathcal{U}} \frac{1}{2} \|AP_{\mathcal{U}_n} u - y\|^2 + \alpha \mathcal{J}(u). \tag{5.2}$$

Finally replacing the exact data y with a noisy measurement y^δ such that $\|y - y^\delta\| \leqslant \delta$, we obtain

$$\min_{u \in \mathcal{U}} \frac{1}{2} \|AP_{\mathcal{U}_n} u - y^\delta\|^2 + \alpha \mathcal{J}(u). \tag{5.3}$$

The restricted operator $AP_{\mathcal{U}_n}$ in (5.3) can be computed *without numerical access* to A. Indeed, for any $u \in \mathcal{U}$, we have

$$P_{\mathcal{U}_n} u = \sum_{i=1}^n (u, \overline{u}^i) \overline{u}^i \quad \text{and} \quad AP_{\mathcal{U}_n} u = \sum_{i=1}^n (u, \overline{u}^i) \overline{y}^i,$$

where $\{\overline{u}^i\}_{i=1,\dots,n}$ are the orthonormalized inputs and $\{\overline{y}^i\}_{i=1,\dots,n}$ the accordingly transformed outputs, cf. section 3.3.2. Therefore, solving (5.3) does not require direct evaluation of A and can be done just using the training pairs (1.1). Once the Gram–Schmidt process is complete, evaluating the operator $AP_{\mathcal{U}_n}$ becomes very fast; the main computational burden lies on the orthogonalization process, which, however, is done 'offline', i.e. before running any optimization algorithms on (5.3), and only once. Therefore, Gram–Schmidt orthogonalization can be regarded as 'training'. We note also that adding more training pairs to (1.1) does not require retraining, i.e. running the Gram–Schmidt process on the whole training set. Only the new portion of training data needs to be made orthogonal to the old data.

Clearly, the operator $AP_{\mathcal{U}_n}$ approximates A pointwise as $n \to \infty$; if A is compact then we also get approximation in the operator norm [30]. Hence, (5.3) can be regarded as an inverse problem with an inexact forward operator and known results on inverse problems with operator errors can be used (e.g. [31, 32]).

5.1. Convergence analysis

The goal of this section is to show existence of minimizers in (5.3) and obtain convergence rates under an appropriate parameter choice rule $\alpha = \alpha(n, \delta)$. There are no new results here as standard results are applicable, the main point being that by formulating (5.3) we are able to transfer these standard, model-based results into the purely data driven, model-free setting.

We emphasize that the size of the training set in this setting n controls the approximation quality of the forward operator and affects the choice of the regularization parameter α along with the noise level δ. We formulate the parameter choice rules in terms of the residual $\|(I - P_{\mathcal{U}_n})u^\dagger\|$ of the expansion of the exact solution in the basis $\{\overline{u}^i\}_{i\in\mathbb{N}}$. A more common way of deriving parameter choice rules (e.g., [31, 32]) is in terms of the approximation error in the operator norm h_n such that $\|A - AP_{\mathcal{U}_n}\| \leqslant h_n$, however, this is a global estimate that depends on how well the subspaces $\mathcal{U}_n$ agree with the operator A (the ideal choice would be, obviously, the eigenspaces of A corresponding to the n largest eigenvalues). The residual $\|(I - P_{\mathcal{U}_n})u^\dagger\|$ is a local quantity that shows how well the subspaces $\mathcal{U}_n$ approximate the particular solution $u^\dagger$ that we are looking for. Hence, even if the global approximation error $\|A - AP_{\mathcal{U}_n}\|$ is large, convergence can still be fast if the training data (1.1) are chosen well for a particular solution $u^\dagger$. In some sense, the choice of the training inputs $\{u^i\}_{i=1,\ldots,n}$ in (1.1) is a way of using *a priori* information about the solution $u^\dagger$ to solve the inverse problem (2.1).

Below we summarize the main assumptions and recall the required existence and convergence results.

Assumption 5. The regularization functional $\mathcal{J} : \mathcal{U} \to \mathbb{R}_+ \cup \{+\infty\}$ is proper, convex and lower-semicontinuous.

Assumption 6. For every $M, \alpha > 0$ and every $n \in \mathbb{N}$, the sets $\{u \in \mathcal{U} : \|AP_{\mathcal{U}_n}u - y^\delta\|^2 + \alpha J(u) \leqslant M\}$ are weakly sequentially compact.

Remark 22. If $\mathcal{J}$ is the total variation (TV), assumptions 5 and 6 are satisfied if $AP_{\mathcal{U}_n} : L^2 \to L^2$ does not annihilate constant functions, see [33].

We are now ready to state a-priori bounds for minimizers of (5.3).

Theorem 23. *Suppose that assumptions 5 and 6 are satisfied and the regularization parameter $\alpha = \alpha(\delta, n)$ is chosen such that*

$$\alpha \to 0 \quad and \quad \frac{(\delta + \|(I - P_{\mathcal{U}_n})u^\dagger\|)^2}{\alpha} \to 0 \quad as \quad \delta \to 0 \quad and \quad n \to \infty,$$

where $u^\dagger$ is the exact solution of (2.1). Then (5.3) admits a minimizer and for every minimizer $u_{\mathcal{J}}^{n,\delta}$ there exists a constant $C > 0$ independent of n and δ such that

$$\mathcal{J}(u_{\mathcal{J}}^{n,\delta}) \leqslant C \quad and \quad \|u_{\mathcal{J}}^{n,\delta}\| \leqslant C.$$

Proof. The proof is similar to [31, 32]. However, note that in [31, 32], the regularization solution of the functional $u \in U \to \|Au - y^\delta\|^2 + \alpha J(u)$ is approximated by the minimizer of the same functional over $\mathcal{U}_n$, while we solve the minimization on the infinite dimensional space. $\square$

In modern variational regularization, (generalized) Bregman distances are typically used to study convergence of approximate solutions [34]. We briefly recall the definition.

Definition 24. For a proper convex functional $\mathcal{J}$ the generalized Bregman distance between $u, v \in \mathcal{U}$ corresponding to the subgradient $p \in \partial\mathcal{J}(v)$ is defined as follows

$$D_{\mathcal{J}}^p(u, v) := J(u) - J(v) - (p, u - v).$$

Here $\partial\mathcal{J}(v)$ denotes the subdifferential of $\mathcal{J}$ at $v \in \mathcal{U}$.

To obtain convergence rates, an additional assumption on the regularity of the exact solution, called the *source condition*, needs to be made [18]. Several variants of the source condition exist (e.g., [18, 19, 34]); we use the following one [35].

Assumption 7 (Source condition). There exists an element $q \in \mathcal{Y}$ such that

$$A^*q \in \partial \mathcal{J}(u^\dagger).$$

Theorem 25. *Suppose that assumptions 5, 6 and 7 are satisfied. Then the following estimate for the Bregman distance between $u_{\mathcal{J}}^{n,\delta}$ and $u^\dagger$ corresponding to the subgradient A^*q from assumption 7 holds*

$$
\begin{aligned}
D_{\mathcal{J}}^{A^*q}(u_{\mathcal{J}}^{n,\delta}, u^\dagger) \leqslant & \frac{1}{2\alpha} \left(\delta + \|A\| \|(I - P_{\mathcal{U}_n})u^\dagger\| \right)^2 + \frac{\alpha}{2} \|q\|^2 \\
& + \left(\delta \|q\| + C \|(I - P_{\mathcal{U}_n})A^*q\| \right)
\end{aligned}
$$

for some constant $C > 0$.

If the regularization parameter $\alpha = \alpha(\delta, n)$ is chosen as in theorem 23 then

$$D_{\mathcal{J}}^{A^*q}(u_{\mathcal{J}}^{n,\delta}, u^\dagger) \to 0 \quad as \quad \delta \to 0 \quad and \quad n \to \infty.$$

For the particular choice

$$\alpha \sim (\delta + \max\{\|(I - P_{\mathcal{U}_n})u^\dagger\|, \|(I - P_{\mathcal{U}_n})A^*q\|\}) \tag{5.4}$$

we obtain the following estimate

$$D_{\mathcal{J}}^{A^*q}(u_{\mathcal{J}}^{n,\delta}, u^\dagger) \sim \alpha.$$

Proof. The proof is similar to [31, 32], with the same caveat as in theorem 23.　　□

The fact that the speed of convergence depends on how well the subspaces $\mathcal{U}_n$ spanned by the training inputs $\{u^i\}_{i=1,\dots,n}$ approximate the solution $u^\dagger$ is not surprising. However, from theorem 25 we conclude that it is equally important for the convergence rate that the subgradient A^*q from the source condition is well approximated by the subspaces $\mathcal{U}_n$.

Remark 26. The results of this section are applicable in a more general setting than the previous parts of the paper. In particular, the forward operator A does not have to be injective. In this case, with some minor (and standard) modifications of the assumptions, we would get convergence of minimizers of (5.3) to a solution of (2.1) with the minimal value of the regularizer $\mathcal{J}$ (a $\mathcal{J}$-minimizing solution).

6. Numerical experiments

In this section we present some numerical experiments with the data driven regularization by projection method based on formula (3.7), the dual least squares method (4.4) and data driven variational regularization (5.3). We explain the data preparation and describe the used data sets first—in particular, how we derive the training pairs for the Radon operator.
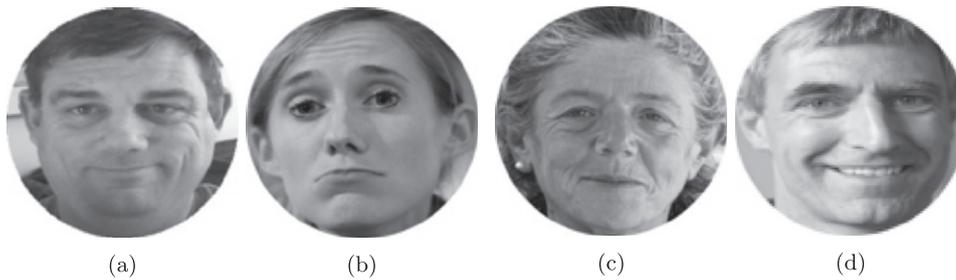
Figure 1. Samples from the 'Faces' dataset.

6.1. Setting

6.1.1. Dataset.   We use images from 'The 10k US Adult Faces Database' [36] (to which we refer as 'Faces'). This is a dataset of 10 168 natural face photographs of 256 pixels height and variable width (typically between 150 and 220 pixels). To see what happens as the size of the training set n gets close to the number of pixels in the image (the discrete analogue of $n \to \infty$), we resize all images to 100×100 pixels to match the size of our data base (10k images). The effective dimension of the discretized space $\mathcal{U}$, i.e. the number linearly independent images among the 10 000, is smaller than the total dimension, since the images have been cropped with an oval around the face. About 700 pixels have the value of exactly 1 (white) in all images, and further 1500 have values between 0.9 and 1 in all images. This makes the number of 'degrees of freedom' effectively around 7800. Sample images from the database are shown in figure 1.

As our training inputs $\{u^i\}_{i=1,\ldots,n}$ in (1.1) we choose n randomly selected photographs, with different $n \leqslant 10\,000$. As the ground truth $u^\dagger$ in (2.1) we also take photographs from the 'Faces' dataset (which are not contained in the training set $\{u^i\}_{i=1,\ldots,n}$).

Applying Gram–Schmidt orthogonalization to $\{u^i\}_{i=1,\ldots,n}$ (cf. section 3.3.2), we obtain an orthonormal system $\{\overline{u}^i\}_{i=1,\ldots,n}$. (We use the modified Gram-Schmidt algorithm [37].) Using this system, we can check assumption 3 from section 3.3.1 numerically by plotting the partial sums

$$\sum_{i=1}^{n} |(u^\dagger, \overline{u}^i)|$$

for different n. The result is shown in figure 2. The series seem to be bounded uniformly and hence the figure does not contradict assumption 3. We stress that assumption 3 is an assumption on the dataset and does not depend on the forward operator; it only measures how well a single element of the dataset (the exact solution $u^\dagger$) can be approximated with other elements of the dataset (the training inputs $\{u^i\}_{i=1,\ldots,n}$).

6.1.2. Gram–Schmidt orthogonalization vs Householder reflections.   It is well know that (even the modified) Gram–Schmidt algorithm is numerically unstable if the number of vectors is large and Householder reflections [37] provide a stable alternative. However, the decisive advantage of Gram–Schmidt is that it allows adding new training points without the need to re-orthogonalize the whole dataset, which is clearly advantageous in practice, since orthogonalization is the most time consuming part of the pipeline. A practical recommendation would be to use Householder reflections to orthogonalize the 'initial' dataset (say, the training pairs that are available as one starts using the method) and then use Gram–Schmidt to add new training points. Perhaps one could also use Householder reflections once in a while
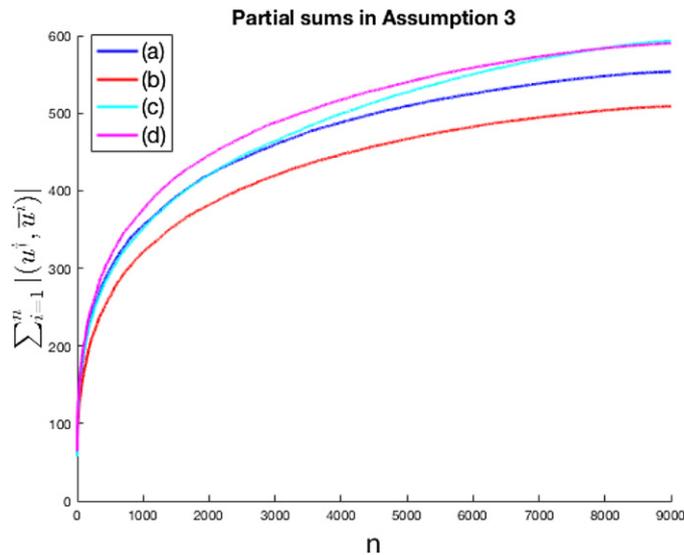
Figure 2. Partial sums $\sum_{i=1}^{n} |(u^{\dagger}, \overline{u}^i)|$ in assumption 3 for images $u^{\dagger}$ shown in figure 1 and $n = 1, \ldots, 9000$. The figure does not contradict assumption 3.

to re-orthogonalize the training set if the number of 'new' training points added after the last application of Householder reflections is sufficiently large.

However, we would like to emphasize that when regularization by projection is used with a noisy measurement y^{δ} in (2.1) (which is always the case in practice), the amount of training pairs should not be too large anyway as this would compromise stability (cf. theorems 17 and 21 as well as [24, theorem 4.2]). This is not the case, however, for variational regularization, where the noise is counteracted by increasing the regularization parameter α.

6.1.3. Forward operator.   We consider our images to be elements of $L^2(\Omega)$, where $\Omega \subset \mathbb{R}^2$ is the (bounded) image domain. As the forward operator, we take the Radon transform with a parallel beam geometry $\mathcal{R} : L^2(\Omega) \to L^2(\mathbb{R} \times [0, \pi])$. We use Matlab's implementation of the Radon transform `radon` with 70 projections uniformly distributed on the interval $[0, \pi]$. The size of the Radon data produced by `radon` is $145 \times 70 \sim 10k$ pixels. Applying the Radon transform to the training inputs $\{u^i\}_{i=1,\ldots,n}$, we obtain the outputs (sinograms) $\{y^i\}_{i=1,\ldots,n}$ and hence get the pairs (1.1). Sample training pairs are shown in figure 3 (left and central columns).

Assumption 4 depends both on the training inputs $\{u^i\}_{i=1,\ldots,n}$ and the forward operator A. Transforming the training outputs $\{y^i\}_{i=1,\ldots,n}$ to match the orthonormalized inputs $\{\overline{u}^i\}_{i=1,\ldots,n}$, we obtain the transformed sinograms $\{\overline{y}^i\}_{i=1,\ldots,n}$. Recall that assumption 4 was a condition on the expansion coefficients of $P_{\mathcal{Y}_n}\overline{y}^i$ for $i \geqslant n+1$ in the non-orthogonal basis $\{\overline{y}^i\}_{i=1,\ldots,n}$. To speed up computations, we check this condition on downsampled images of size 32×32. The result is shown in figure 4(a). The sums are not bounded, hence assumption 4 does not seem to be satisfied.

However, there is a significant numerical error accumulating in the Gram–Schmidt process as n increases and this error will have an effect on the computed expansion coefficients $\beta_j^{i,n}$. To find these coefficients, we need to solve a linear system whose matrix $\overline{Y}_n := (\overline{y}^1, \ldots, \overline{y}^n)$ is formed of the non-orthogonal vectors $\{\overline{y}^i\}_{i=1,\ldots,n}$. This matrix will become ill-conditioned for large n, since the vectors $\{\overline{y}^i\}_{i=1,\ldots,n}$ will become close to being linearly dependent (cf.
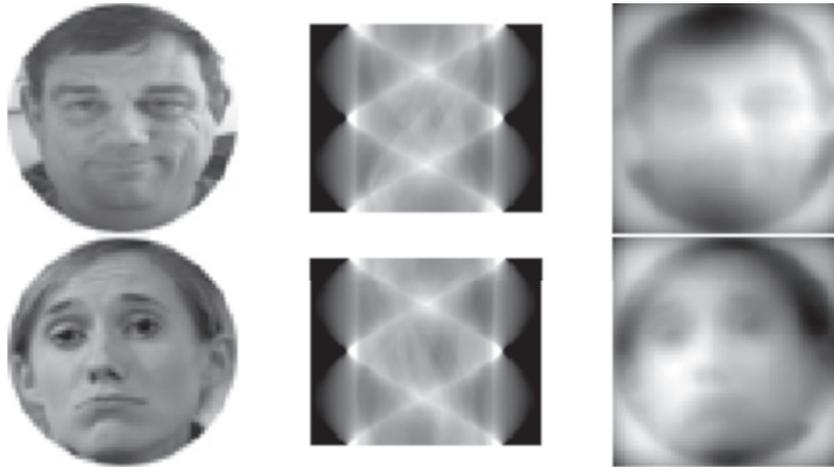
Figure 3. Samples u^i from the 'Faces' dataset (left column), their Radon transforms $y^i := Au^i$ (central column) and their adjoint transforms $v^i := A^*y^i = A^*Au^i$ (right column).
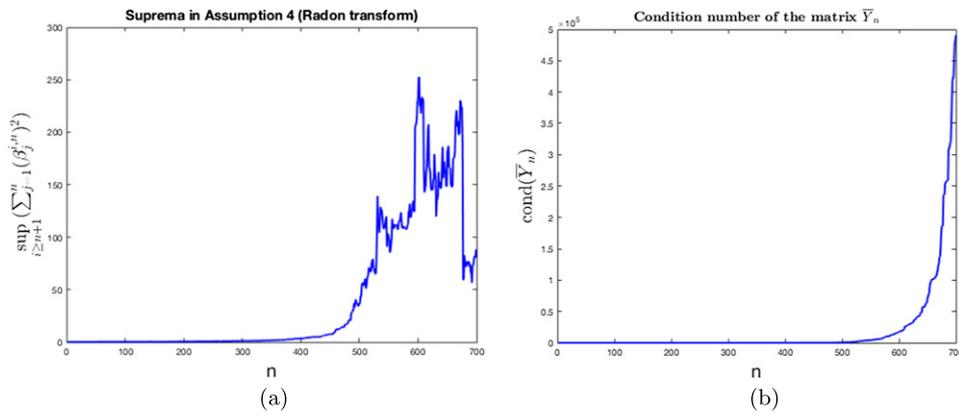

Figure 4. (a) The supremum over the squared two-norms of the expansion coefficients of $\overline{y}^i$, $i \geqslant n+1$, in the basis $\{\overline{y}^i\}_{i=1,\ldots,n}$ for $n = 1, \ldots, 700$ (computed on downsampled images of size 32×32). Since the basis is non-orthogonal, the expansion coefficients change with n. However, we note that the growth can be a numerical artifact: to compute the coefficients $\beta_j^{i,n}$, we need to invert an ill-conditioned matrix whose condition number as a function of n is shown in figure 4. (b) To compute the coefficients $\beta_j^{i,n}$ from assumption 4, we need to invert the matrix $\overline{Y}_n := (\overline{y}^1, \ldots, \overline{y}^n)$ composed of the transformed outputs $\{\overline{y}^i\}_{i=1,\ldots,n}$, cf. section 3.3.2. This matrix becomes ill-conditioned for large n, which may result in inaccuracies in the computation of the coefficients $\beta_j^{i,n}$. We emphasize, however, that inverting this matrix is only needed to numerically check assumption 4; the actual reconstruction algorithm (3.7) does not need it.

proposition 7, where this is proven for $\{y^i\}_{i\in\mathbb{N}}$; the proof for $\{\overline{y}^i\}_{i\in\mathbb{N}}$ is the same). The condition number of this matrix as a function of n is shown in figure 4(b) it starts exploding at $n \sim 500$. Therefore, in our opinion, the question whether assumption 4 can be satisfied for real data remains open.
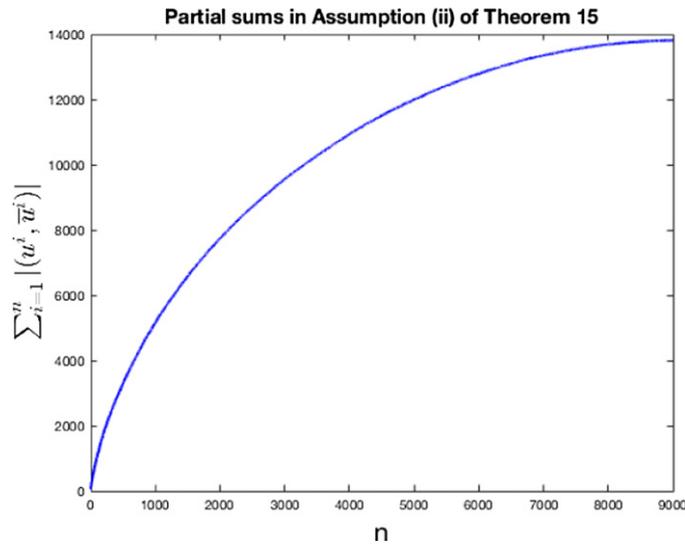
Figure 5. Partial sums in assumption (b) of theorem 15. The figure does not contradict assumption (b) of theorem 15. The behavior is similar to that in figure 2, but the scale is different ($\sim 14\,000$ here vs ~ 600 in figure 2).

We shall see in section 6.2, however, that the reconstructions (3.7) obtained from clean data $y \in \mathcal{R}(A)$ remain bounded as n increases (until numerical artifacts kick in).

Assumptions (a) and (b) of theorem 15 are satisfied. Since the training inputs $\{u^i\}_{i=1,\ldots,n}$ and the exact solution $u^\dagger$ are similar, we expect that $\frac{|(y,y^i)|}{|(y^i,y^i)|} \sim 1$, which is in fact what we observe numerically (we omit the plot). Hence, assumption (a) is satisfied. Assumption (b) is also satisfied as shown in figure 5.

6.1.4. Absence of inverse crime. The term 'inverse crime' was introduced in [38] to describe the situation when the same model is used to generate synthetic data and subsequently to solve an inverse problem involving this data. This procedure may artificially decrease the error introduced by a reconstruction algorithm and therefore should be avoided. In our experiments, we use the same model (Matlab's implementation of the Radon transform) to generate both the training data $\{y^i\}_{i=1,\ldots,n}$ and the data y that we invert. However, we do not use the same model to solve the inverse problem: we only use the training pairs $\{u^i, y^i\}_{i=1,\ldots,n}$ and the approximation of the forward operator or its inverse that we learn from these pairs. Hence, no inverse crime is committed here. Moreover, it is only natural to assume that in real applications the measurement y and the training data $\{y^i\}_{i=1,\ldots,n}$ will be generated under very similar conditions, perhaps even by the same measurement device.

6.2. Regularization by projection

We start by analyzing reconstructions from clean data y from (2.2). Figure 8a shows the norm of $u_n^{\mathcal{U}}$ defined in (3.7) as a function of n (blue solid line). After some oscillations at small n, the norm remains stable until $n \sim 7000$ and then it explodes. This may be due to the failure of assumption 4 and the instability of regularization by projection, however, this may be also caused by the numerical instability of Gram–Schmidt orthogonalization, or by both. We will
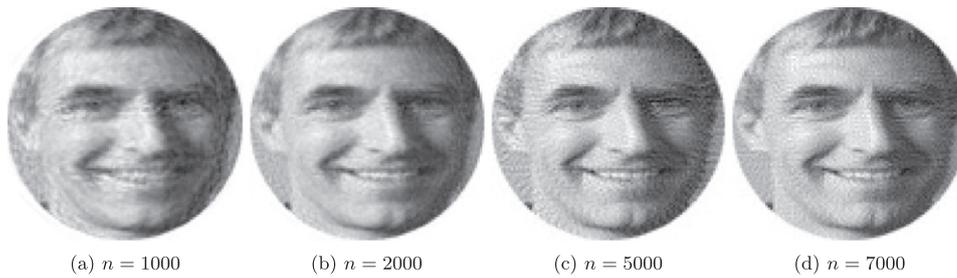
(a) $n = 1000$ (b) $n = 2000$ (c) $n = 5000$ (d) $n = 7000$

Figure 6. Reconstructions using regularization by projection (3.7) from clean data $y \in \mathcal{R}(A)$. Images develop oscillations as the size of the training set n increases, but generally the reconstructions remain stable for a large range of n, until $n \sim 7500$ when they explode.
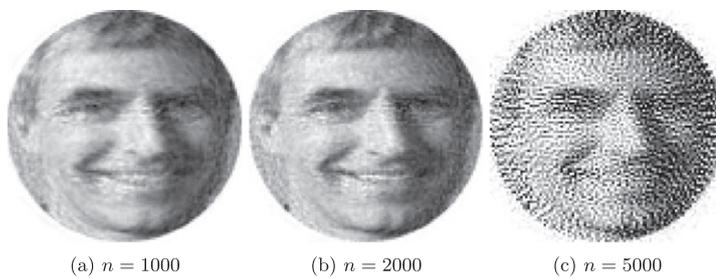

(a) $n = 1000$ (b) $n = 2000$ (c) $n = 5000$

Figure 7. Reconstructions using regularization by projection (3.31) from noisy data y^δ (1% noise). Reconstructions become unstable for relatively small n. The reconstruction for $n = 7000$ is completely unstable and not shown here.

discuss this in more detail in section 6.3. In any case, boundedness is satisfied for a wide range of n. Reconstructions for different values of n are shown in figure 6. As the size of the training set n increases, the reconstructions start developing oscillations, but generally remain stable for a wide range of n.

Adding noise to the data y changes the situation. Reconstructions from noisy data y^δ (cf. (3.31)) with 1% Gaussian noise are shown in figure 7. Reconstructions become unstable much earlier, reflecting the ill-posedness of the problem. Already for $n = 5000$ the reconstruction is highly oscillatory. Figure 8(a) (red dashed line) shows that the norm of the reconstruction $u_{n,\delta}^{\mathcal{U}}$ (3.31) grows with n (except for very small n where it oscillates) and around $n = 5000$ it explodes.

Figure 8(b) shows the relative reconstruction error as a function of the size of the training set n, averaged over a validation set of about 100 images. The reconstruction error from clean data ($\delta = 0$) first decreases monotonically as expected, but has an anomaly around $n = 5000$. This may again be due to numerical artifacts of the Gram–Schmidt algorithm. For all other noise levels $\delta = 10^{-4}; 10^{-3}; 10^{-2}; 10^{-1}$ we observe the *semiconvergence* behavior, i.e. the reconstruction error first decreases with n until a certain optimal value and then explodes. This behavior is expected from section 3.4.

These experiments demonstrate that for ill-posed problems increasing the size of the training set can have an adverse effect on the reconstructions.
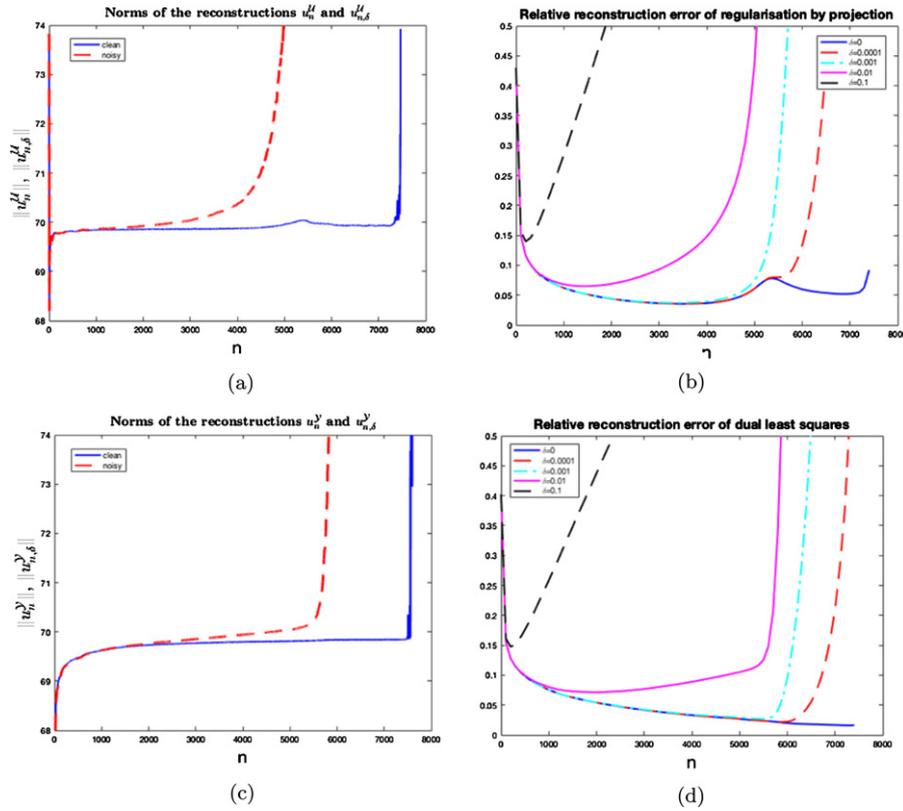
(a)

(b)

(c)

(d)

Figure 8. (a) Regularization by projection: the norm of reconstructions from clean data $y \in \mathcal{R}(A)$ and from noisy data y^δ, denoted by $u_n^{\mathcal{U}}$ (3.7) and $u_{n,\delta}^{\mathcal{U}}$ (3.31), respectively, as a function of n. After some oscillations for small n, the norm of $u_n^{\mathcal{U}}$ (clean data) stays almost constant until $n \sim 7500$. The blow-up observed later is caused by the common white oval in all training images. The number of white pixels (with value 1) in the oval is ca. 700, and ca. 1500 more have values between 0.9 and 1, which means that the number of linearly independent pixels in the training images is around 7800. (b) Relative error of regularization by projection as a function of the size of the training set n for different noise levels δ. The error has been averaged over a validation set of about 100 images. All curves demonstrate semi-convergence behavior; the larger the noise, the smaller the optimal n and the larger the error. (c) Dual least squares: the norm of reconstructions from clean data $y \in \mathcal{R}(A)$ and from noisy data y^δ, denoted by $u_n^{\mathcal{Y}}$ (4.4) and $u_{n,\delta}^{\mathcal{Y}}$ (4.6), respectively, as a function of n. The norm of $u_n^{\mathcal{Y}}$ (clean data) grows monotonically in the beginning and then stays almost constant. At $n \sim 7500$ it explodes, which is caused by the common white oval in all training images. The number of white pixels (with value 1) in the oval is ca. 700, and ca. 1500 more have values between 0.9 and 1, which means that the number of linearly independent pixels in the training images is around 7800. (d) Relative error of dual least squares as a function of the size of the training set n for different noise levels δ. The error has been averaged over a validation set of about 100 images. All curves except for $\delta = 0$ demonstrate semi-convergence behavior; the larger the noise, the smaller the optimal n and the larger the error. Reconstruction error from clean data ($\delta = 0$) decreases monotonically for all n.

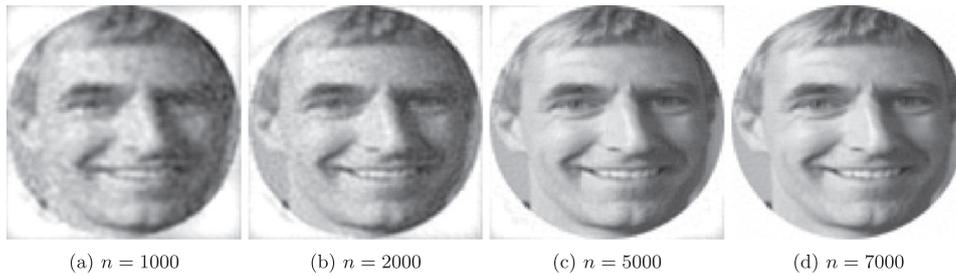(a) $n = 1000$ (b) $n = 2000$ (c) $n = 5000$ (d) $n = 7000$

Figure 9. Reconstructions using dual least squares (4.4) from clean data $y \in \mathcal{R}(A)$. Reconstructions remain stable and converge to the ground truth as $n \to \infty$.
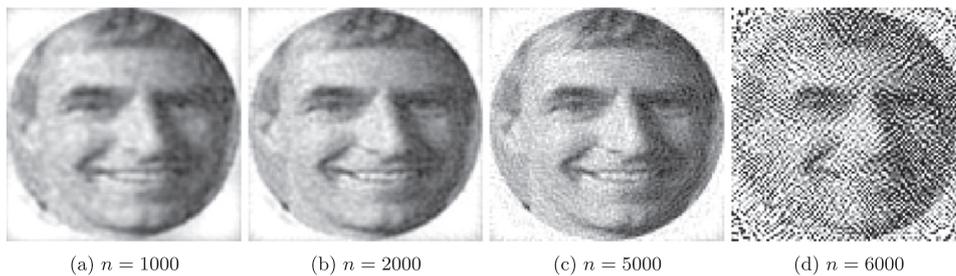

(a) $n = 1000$ (b) $n = 2000$ (c) $n = 5000$ (d) $n = 6000$

Figure 10. Reconstructions using dual least squares (4.6) from noisy data y^δ (1% noise). We observe a typical semi-convergence behavior: after improving initially, the reconstructions become unstable as $n \to \infty$ and eventually blow up around $n = 6000$. The reconstruction for $n = 7000$ is completely unstable and not shown here.

6.3. Dual least squares

The dual least squares approach requires a different type of training data—training data for the adjoint operator (4.3). It is not clear how this type of training data can be obtained in practice and hence the relevance of the dual least squares for learning is not obvious, but we still perform experiments with this approach for the sake of completeness. To generate $\{v^i\}_{i=1,\dots,n}$ such that $v^i = A^* y^i$ for $\{y^i\}_{i=1,\dots,n}$ as defined in (1.1), we apply the adjoint of the Radon transform to $\{y^i\}_{i=1,\dots,n}$. Sample training triplets (u^i, y^i, v^i) are shown in figure 3.

We start with reconstructions from clean data $y \in \mathcal{R}(A)$. The norm of the reconstruction (4.4) as a function of the size of training set n is shown in figure 8(c) (solid blue line). After a small initial period of monotonic growth, the norm stays constant until $n \sim 7500$ when it explodes. This is the same value as for the solution of regularization by projection $u_n^{\mathcal{U}}$ in figure 8(a).

To understand why this happens let us recall that we compute u_n^y by projecting the solution of regularization by projection $u_n^{\mathcal{U}}$ onto some subspace (cf. (4.4)). Since the dual least squares solution u_n^y must be stable as $n \to \infty$ (theorem 18), the instability must come from an error in computing $u_n^{\mathcal{U}}$ (cf. (3.7)), which is likely to be caused by the numerical instability of the Gram–Schmidt algorithm. This is also an indication that the instability of the solution of regularization by projection $u_n^{\mathcal{U}}$ that we see in figure 8(a) is likely to be due to numerical effects rather than the instability of regularization by projection.

Reconstructions from clean data $y \in \mathcal{R}(A)$ obtained using the dual least squares method (4.4) are shown in figure 9. They remain stable as n grows (until numerical instability of the

Gram–Schmidt algorithm kicks in at $n \sim 7500$) and converge to the ground truth. They do not develop oscillations that we have seen in figure 6 for regularization by projection.

For noisy data y^δ (figure 10) we observe the expected semi-convergence behavior: after improving initially, the reconstructions diverge as n increases and blow up somewhere between $n = 5000$ and $n = 6000$. Although both regularization by projection and dual least squares diverge eventually for noisy data y^δ, the dual least squares method remains stable for a larger n, cf. figures 8(a) vs 8(c) and figures 7 vs 10.

Figure 8(d) shows the relative reconstruction error as a function of the size of the training set n, averaged over a validation set of about 100 images. As expected, the reconstruction error from clean data ($\delta = 0$) decreases monotonically for all n, while the reconstruction error from noisy data with $\delta = 10^{-4}; 10^{-3}; 10^{-2}; 10^{-1}$ demonstrates the semi-convergence behavior. The reconstruction error first decreases with n until a certain optimal value and then explodes.

6.4. Variational regularization

In this section we assess the performance of projected variational regularization (5.3). We only show reconstructions from noisy data y^δ.

For convenience, we repeat the statement of projected variational regularization (5.3)

$$\min_{u \in \mathcal{U}} \frac{1}{2} \|AP_{\mathcal{U}_n} u - y^\delta\|^2 + \alpha \mathcal{J}(u). \tag{6.1}$$

Our goal is to compare reconstructions for different sizes of the training set n with a model-based reconstruction that has access to the forward operator A (the Radon transform)

$$\min_{u \in \mathcal{U}} \frac{1}{2} \|Au - y^\delta\|^2 + \alpha \mathcal{J}(u). \tag{6.2}$$

As in the previous sections, $\mathcal{U} = L^2(\Omega)$, where $\Omega \subset \mathbb{R}^2$ is the image domain. As a prototypical example of a regularization functional $\mathcal{J}$ we take TV [39], which we define as a functional on $L^2(\Omega)$ extending it with the value $+\infty$ on $L^2(\Omega)\backslash BV(\Omega)$. This is well defined, since $\Omega \subset \mathbb{R}^2$ and hence $BV(\Omega) \subset L^2(\Omega)$. This is a common setting in imaging [40].

TV is a proper, convex and lower semicontinuous functional on $L^2(\Omega)$ [41], hence assumption 5 is satisfied. Zeros of the TV functional consist of constant functions. Note that the Radon transform does not annihilate the constant functions, hence assumption 6 is also satisfied if $P_{\mathcal{U}_n}$ does not annihilate constant functions. For this, it is clearly sufficient that $P_{\mathcal{U}_1} \mathbb{1} \neq 0$, i.e. $(u^1, \mathbb{1}) \neq 0$ (in other words, u^1 does not have zero mean). This is clearly satisfied for the photographs in the 'Faces' dataset. Therefore theorems 23 and 25 hold.

To see how well the *learned* operator $AP_{\mathcal{U}_n}$ approximates the Radon transform A when evaluated at the ground truth image $u^\dagger$ (the same as in figures 9 and 10), we show the sinograms $AP_{\mathcal{U}_n} u^\dagger$ (learned) and $Au^\dagger$ (exact) in figure 11. Already for a moderate size of the training set $n = 1000$, the approximation is very good. The relative approximation error as a function of n is shown in figure 12 and seems to decrease exponentially with n.

To solve (6.1) and (6.2), we use the CVX package [42, 43]. To generate the differential operator needed to evaluate TV, we use the DIFFOP package [44].

Reconstructions from noisy data y^δ with 1% noise obtained using data driven variational regularization (6.1) are shown in figures 13(a)–(c) and the solution obtained with the standard model-based approach (6.2) is shown in figure 11. Already for a moderate size of the training set ($n = 1000$, which is 10% of the total number of pixels in each image), we obtain a reasonable reconstruction and as n increases, the reconstructions become closer and closer to
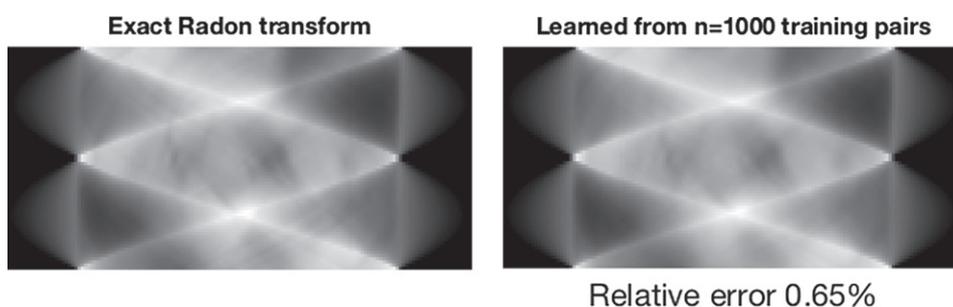
Figure 11. Left: The sinogram of the image in figure 3(d). Right: Learned approximation for $n = 1000$ training pairs. Already for a moderate amount of training pairs, the learned operator is able to approximate the Radon transform very well (on an input that is similar to the training inputs).
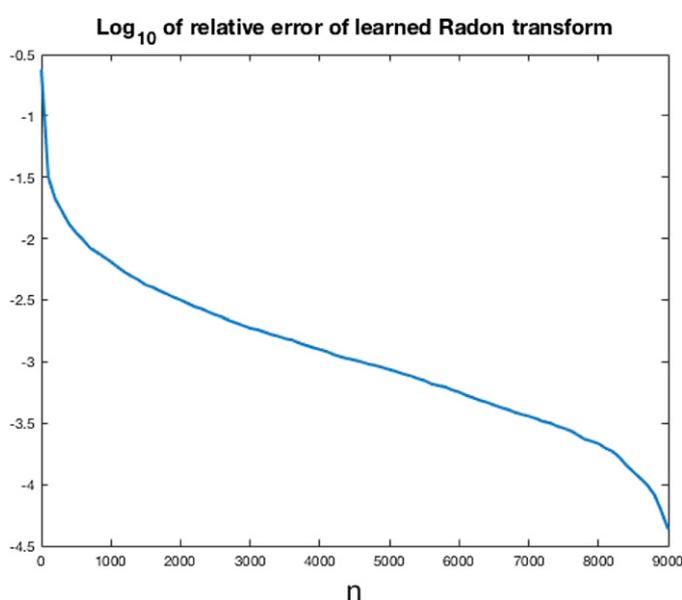

Figure 12. Reconstruction error of the learned Radon transform as a function of the size of training set n (on a log scale). The error seems to decrease exponentially with n.

the 'ideal' one obtained using explicit knowledge of the forward model (figure 13(d)). The regularization parameter α was the same in all these experiments.

Data driven reconstructions in figures 13(a)–(c) exhibit the same qualitative behavior as the model-based reconstruction in figure 13(d), e.g., we see characteristic for TV *staircasing* that becomes less apparent as n increases.

Relative reconstruction errors for the image shown in figure 13 are given in table 1. For computational reasons, we do not perform experiments on the whole validation set of ca. 100 images. Although this is not an entirely fair comparison, the reconstruction errors in table 1 are smaller than those in figures 8(b) (regularization by projection) and 8(d) (dual least squares). In the latter two methods the optimal error for $\delta = 0.01$ is around 7% (at approximately $n = 2000$), whereas for variational regularization it is smaller even for the same n and keeps
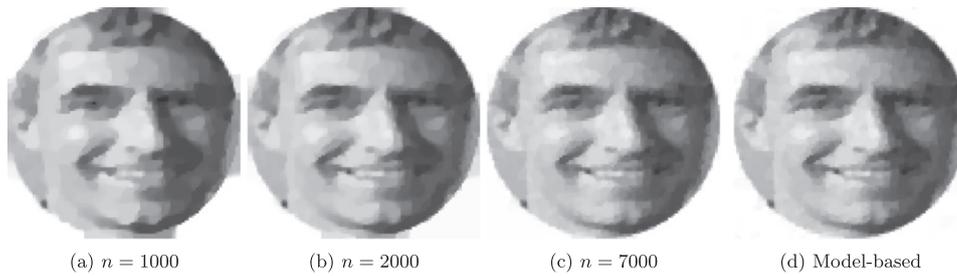
(a) $n = 1000$ (b) $n = 2000$ (c) $n = 7000$ (d) Model-based

Figure 13. Reconstructions using variational regularization from noisy data y^δ (1% noise): (a)–(c) data driven reconstructions (cf. (6.1)) for different sizes of the training set n and (d) a model based reconstruction that has access to the forward operator (cf. (6.2)). Even for a modest size of the training set $n = 1000$ (10% of the number of pixels in the image) the reconstruction is very reasonable; as n increases, the data driven reconstruction becomes almost indistinguishable from the model-based one.

Table 1. Relative reconstruction error of data driven variational regularization for different sizes of the training set n vs model based reconstruction. Numbers based on the image shown in figure 13. As n increases, the reconstruction quality approaches that of a model based method. Already for a modest size of the training set $n = 3000$ the reconstruction quality is comparable with the optimal one.

n	1000	2000	3000	4000	5000	6000	7000	8000	9000	Model
Rel. error	0.078	0.053	0.043	0.037	0.034	0.033	0.031	0.031	0.030	0.030

decreasing with growing n. The difference in the visual quality of the reconstructions is even more apparent, cf. figures 7, 10 and 13.

Comparing reconstructions from noisy data y^δ obtained with data driven variational regularization (figures 13(a)–(c)) with those obtained with regularization by projection (figure 7) and dual least squares (figure 10), we observe that variational regularization produces much better reconstructions. The typical relative error is also smaller, cf. table 1 and figure 8. The price to pay for is that variational regularization (6.1) requires solving a (potentially computationally costly) optimization problem, even if evaluating the projected forward operator $AP_{\mathcal{U}_n}$ becomes cheap once the Gram–Schmidt orthogonalization is complete, whilst regularization by projection (3.7) and dual least squares (4.4) only require taking a matrix-vector product (also once the Gram–Schmidt algorithm is complete).

6.5. Practical recommendations

Depending on the situation, different methods discussed in this paper will become preferable. If the amount of noise in the measurement y^δ in (2.1) is small then regularization by projection (section 3) is a good option since, once trained, it is very efficient computationally. Care should be taken, however, in checking the validity of our assumption, since this method is non-convergent in general.

The dual least squares method (section 4) is better since it does not require additional assumptions (except for an appropriate choice of the size of the training set, cf. theorem 21), however, collecting training data for the adjoint operator experimentally is not an obvious task. However, if the goal is to replace a computationally expensive model, then such training data can be collected, and this is the preferred option compared to regularization by projection.

Finally, if the amount of noise in the measurement y^δ is large, parameter choice rules in theorems 17 and 21 will require that the size of the training set is too small and, although stability will hold, the approximation quality will likely be not satisfactory. Variational regularization is in this case the method of choice, however, it has two drawbacks. Firstly, it is more expensive computationally since it requires solving an optimization problem, although even here projections might provide a speed-up compared to using an analytic model if this model is computationally expensive. Secondly, a large amount of noise will require a larger regularization parameter, which will make the effects of the regularizer mode apparent (such as staircasing with Total Variation).

7. Conclusions

We have seen that some results of model-based regularization theory can be extended to the purely data driven setting when the forward operator is given only through input-output training pairs. It has also been demonstrated that restrictions of the forward operator and (in the injective case) its inverse to the spans of the training data can be computed without numerical access to the forward operator. This was used to formulate data driven analogues of regularization by projection and variational regularization and carry over some classical results such as convergence rates of variational regularization. We have also seen that the role of the size of the training set is twofold: in variational regularization, it controls the approximation quality of the forward operator and hence having more training data is always better, while in regularization by projection, the size of the training set is a regularization parameter and hence using more training pairs than is allowed by the noise in the measurement will compromise stability. This is due to the ill-posed nature of the inverse problem and different from overfitting, where poor performance is typically a consequence of the training set being too small. The numerical studies should not be considered final. In particular the restrictiveness of assumption 4 in real world applications needs further studies.

Acknowledgments

ORCID iDs

Yury Korolev ⬥ https://orcid.org/0000-0002-6339-652X
Otmar Scherzer ⬥ https://orcid.org/0000-0001-9378-7452

References

[1] Seidman T I 1980 Nonconvergence results for the application of least-squares estimation to ill-posed problems *J. Optim. Theor. Appl.* **30** 535–47
[2] Natterer F and Wüubbeling F 2001 *Mathematical Methods in Image Reconstruction* (Philadelphia, PA: SIAM)

[3] Scherzer O (ed) 2015 Tomography *Handbook of Mathematical Methods in Imaging* 2nd edn (BerlinSpringer) pp 691–733

[4] Isakov V 2017 *Inverse Problems for Partial Differential Equations* (Berlin: Springer)

[5] Ellerbroek B L and Vogel C R 2009 Inverse problems in astronomical adaptive optics *Inverse Problems* **25** 063001

[6] Symes W W 2009 The seismic reflection inverse problem *Inverse Problems* **25** 123008

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

[8] Adler J and Öktem O 2017 Solving ill-posed inverse problems using iterative deep neural networks *Inverse Problems* **33** 124007

[9] Aspri A, Banert S, Öktem O and Scherzer O 2020 A data-driven iteratively regularized Landweber iteration *Numer. Funct. Anal. Optim.* **41** 1190–227

[10] Kobler E, Klatzer T, Hammernik K and Pock T 2017 Variational networks: connecting variational methods and deep learning *German Conf. on Pattern Recognition*

[11] Jin K H, McCann M T, Froustey E and Unser M 2017 Deep convolutional neural network for inverse problems in imaging *IEEE Trans. Image Process.* **26** 4509–22

[12] Li H, Schwab J, Antholzer S and Haltmeier M 2020 NETT: solving inverse problems with deep neural networks *Inverse Problems* **36** 065005

[13] Lunz S, Öktem O and Schönlieb C-B 2018 Adversarial regularizers in inverse problems *Proc. 32nd Int. Conf. on Neural Information Processing Systems* pp 8516–25

[14] Schwab J, Antholzer S and Haltmeier M 2019 Deep null space learning for inverse problems: convergence analysis and rates *Inverse Problems* **35** 025008

[15] Bubba T A *et al* 2019 Learning the invisible: a hybrid deep learning-shearlet framework for limited angle computed tomography *Inverse Problems* **35** 064002

[16] Antun V, Renna F, Poon C, Adcock B and Hansen A C 2019 On instabilities of deep learning in image reconstruction-does AI come at a cost? (arXiv:1902.05300)

[17] Maass P 2019 Deep learning for trivial inverse problems *Compressed Sensing and its Applications* (*Applied and Numerical Harmonic Analysis*) eds H Boche *et al* (Basel: Birkhäuser)

[18] Engl H, Hanke M and Neubauer A 1996 *Regularization of Inverse Problems* (Berlin: Springer)

[19] Scherzer O, Grasmair M, Grossauer H, Haltmeier M and Lenzen F 2009 *Variational Methods in Imaging* (Berlin: Springer)

[20] Rudi A, Camoriano R and Rosasco L 2015 Less is more: Nyström computational regularization *Proc. of the 29th Int. Conf. on Neural Information Processing Systems*

[21] Kriukova G, Pereverzyev S and Tkachenko P 2017 Nyström type subsampling analyzed as a regularized projection *Inverse Problems* **33** 074001

[22] Owhadi H, Scovel C and Schäfer F 2019 Statistical numerical approximation *Not. AMS* **66** 1608–17

[23] Micchelli C and Rivlin T 1977 A survey of optimal recovery *Optimal Estimation in Approximation Theory* (*The IBM Research Symposia Series*) (Berlin: Springer)

[24] Burger M and Engl H W 2000 Training neural networks with noisy data as an ill-posed problem *Adv. Comput. Math.* **13** 335–54

[25] Groetsch C 1984 *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind* (Boston, MA: Pitman)

[26] Nashed M (ed) 1976 A unified operator theory of generalized inverses *Generalized Inverses and Applications* (New YorkAcademic) pp 1–109

[27] Böttcher A, Hofmann B, Tautenhahn U and Yamamoto M 2006 Convergence rates for Tikhonov regularization from different kinds of smoothness conditions *Appl. Anal.* **85** 555–78

[28] Flemming J, Hofmann B and Mathé P 2011 Sharp converse results for the regularization error using distance functions *Inverse Problems* **27** 025006

[29] Andreev R, Elbau P, de Hoop M V, Qiu L and Scherzer O 2015 Generalized convergence rates results for linear inverse problems in Hilbert spaces *Numer. Funct. Anal. Optim.* **36** 549–66

[30] Conway J B 1985 *A Course in Functional Analysis* (Berlin: Springer)

[31] Neubauer A and Scherzer O 1990 Finite-dimensional approximation of Tikhonov regularized solutions of nonlinear ill-posed problems *Numer. Funct. Anal. Optim.* **11** 85–99

[32] Pöschl C, Resmerita E and Scherzer O 2010 Discretization of variational regularization in Banach spaces *Inverse Problems* **26** 105017

[33] Aubert G and Vese L 1997 A variational method in image recovery *SIAM J. Numer. Anal.* **34** 1948–79

[34] Benning M and Burger M 2018 Modern regularization methods for inverse problems *Acta Numer.* **27** 1–111

[35] Burger M and Osher S 2004 Convergence rates of convex variational regularization *Inverse Problems* **20** 1411

[36] Bainbridge W A, Isola P and Oliva A 2013 The intrinsic memorability of face photographs *J. Exp. Psychol. Gen.* **142** 1323–34

[37] Trefethen L N and Bau D III 1997 *Numerical Linear Algebra* (Philadelphia, PA: SIAM)

[38] Colton D and Kress R 1992 *Inverse Acoustic and Electromagnetic Scattering Theory* (Berlin: Springer)

[39] Rudin L I, Osher S and Fatemi E 1992 Nonlinear total variation based noise removal algorithms *Physica* D **60** 259–68

[40] Chambolle A and Lions P-L 1997 Image recovery via total variation minimization and related problems *Numer. Math.* **76** 167–88

[41] Acar R and Vogel C R 1994 Analysis of bounded variation penalty methods for ill-posed problems *Inverse Problems* **10** 1217–29

[42] Grant M and Boyd S 2014 CVX: Matlab software for disciplined convex programming, version 2.1 http://cvxr.com/cvx

[43] Grant M and Boyd S 2008 Graph implementations for nonsmooth convex programs *Recent Advances in Learning and Control* (*Lecture Notes in Control and Information Sciences*) eds V Blondel, S Boyd and H Kimura (Berlin: Springer) pp 95–110

[44] Lellmann J 2014 Diffop-differential operators in MATLAB without the pain https://lellmann.net/work/software/start