Foundations of Computational Mathematics
https://doi.org/10.1007/s10208-024-09654-x

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Convergent Regularization in Inverse Problems and Linear Plug-and-Play Denoisers

Andreas Hauptmann[1,2] · Subhadip Mukherjee[3] · Carola-Bibiane Schönlieb[4] · Ferdia Sherry[4]

Received: 14 December 2023 / Revised: 25 March 2024 / Accepted: 25 March 2024

Abstract

Regularization is necessary when solving inverse problems to ensure the well-posedness of the solution map. Additionally, it is desired that the chosen regularization strategy is convergent in the sense that the solution map converges to a solution of the noise-free operator equation. This provides an important guarantee that stable solutions can be computed for all noise levels and that solutions satisfy the operator equation in the limit of vanishing noise. In recent years, reconstructions in inverse problems are increasingly approached from a data-driven perspective. Despite empirical success, the majority of data-driven approaches do not provide a convergent regularization strategy. One such popular example is given by iterative plug-and-play (PnP) denoising using off-the-shelf image denoisers. These usually provide only convergence of the PnP iterates to a fixed point, under suitable regularity assumptions on the denoiser, rather than convergence of the method as a regularization technique, thatis under van-

Communicated by Teresa Krick and Hans Munthe-Kaas.

Invited paper associated to the FoCM 2021 Online Seminar lecture *Machine Learned Regularization for Solving Inverse Problems* presented by Carola-Bibiane Schönlieb in April 2021.

✉ Carola-Bibiane Schönlieb
cbs31@cam.ac.uk

Andreas Hauptmann
Andreas.Hauptmann@oulu.fi

Subhadip Mukherjee
smukherjee@ece.iitkgp.ac.in

Ferdia Sherry
fs436@cam.ac.uk

1 Research Unit of Mathematical Sciences, University of Oulu, Oulu, Finland

2 Department of Computer Science, University College London, London, UK

3 Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology Kharagpur, Kharagpur, India

4 Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK

Published online: 03 June 2024

ishing noise and regularization strength. This paper serves two purposes: first, we provide an overview of the classical regularization theory in inverse problems and survey a few notable recent data-driven methods that are provably convergent regularization schemes. We then continue to discuss PnP algorithms and their established convergence guarantees. Subsequently, we consider PnP algorithms with learned linear denoisers and propose a novel spectral filtering technique of the denoiser to control the strength of regularization. Further, by relating the implicit regularization of the denoiser to an explicit regularization functional, we are the first to rigorously show that PnP with a learned linear denoiser leads to a convergent regularization scheme. The theoretical analysis is corroborated by numerical experiments for the classical inverse problem of tomographic image reconstruction.

Keywords Inverse problems · Variational regularization · Data-driven learning · Plug-and-play denoising

Mathematics Subject Classification 47A52 · 46N10 · 65F22

1 Introduction

Inverse problems deal with the estimation of an unknown model parameter $x^* \in X$ from its noisy and indirect measurement $y^\delta \in Y$ given by

$$y^\delta = Ax^* + e. \tag{1}$$

We consider the case where X and Y are (potentially infinite dimensional) separable Hilbert spaces and $A : X \to Y$ is a bounded linear operator. X and Y are endowed with inner products $\langle \cdot, \cdot \rangle_X$ and $\langle \cdot, \cdot \rangle_Y$, inducing the norms $\| \cdot \|_X$ and $\| \cdot \|_Y$, respectively. The measurement noise level is bounded by δ, i.e., $\|e\|_Y \leq \delta$. The clean measurement is denoted by y^0.

The inverse problem in (1) is considered ill-posed in the sense of Hadamard, if either *injectivity* or *surjectivity* of the forward operator, or *stability* of the solution map is violated. For instance, if A is a compact operator with an infinite-dimensional range, then surjectivity and stability are not satisfied. This is, for example, the case for the ray transform operator that underlies many applications in medical imaging, such as computed tomography (CT) and positron emission tomography (PET) [35, 36]. The study of inverse problems usually assumes ill-posedness, as we will also do in the following.

To address ill-posedness, one needs to introduce a general concept for stable and unique solvability for an inverse problem of the form (1). Due to the aforementioned ill-posedness, we can not guarantee the recovery of the true solution x^* for all measurements and hence we first need the concept of a generalized solution. A common approach is to search for solutions that are closest to the measured data with respect to a suitable data discrepancy term $f : Y \times Y \to \mathbb{R}_+$, such as the (squared) distance

in the norm, i.e., $f(Ax, y^\delta) = \|Ax - y^\delta\|_Y^2$. Then we search for $\widetilde{x} \in X$ such that

$$f(A\widetilde{x}, y^\delta) \leq f(Ax, y^\delta) \qquad \text{for all } x \in X. \tag{2}$$

(2) implies that $\widetilde{x}$ is closest to the measured data with respect to f, which deals with the violation of surjectivity by disregarding components of y^δ in the co-kernel of A. Furthermore, if A has a non-trivial null space, then $\widetilde{x}$ is not unique. To obtain a unique solution, one can define the minimum norm solution as

$$x^\dagger = \arg\min_{x \in X} \{\|x\|_X : x \text{ minimizes } f(Ax, y^\delta)\}. \tag{3}$$

The element $x^\dagger$ can now be considered a desirable generalized solution to (1). When f and $\|\cdot\|_X$ are given by the squared L^2-norm, we call $x^\dagger$ the least-squares minimum-norm solution and can define a mapping $A^\dagger : Y \to X$, such that $x^\dagger = A^\dagger y^\delta$. In fact, the mapping $A^\dagger$ defines what is referred to as the Moore-Penrose pseudo-inverse. Unfortunately, if the operator A is compact, then $A^\dagger$ will be unbounded and as such does not take care of the stability problem in the presence of noise in the data. This is where the concept of *regularization* becomes important, as we will discuss next.

Regularization theory considers specifically designed solution maps to deal with the stability issue. Such a solution map $\mathcal{R}(\cdot; \lambda) : Y \to X$, also called a *reconstruction operator*, is expressed as a parametric map that produces a solution estimate of x^* given y^δ. Here, the parameter λ depends on the noise level δ and the measured data y^δ, which we denote explicitly by the mapping $\lambda = \lambda(\delta, y^\delta)$. In this paper, we are specifically interested in the notion of *convergent regularization* which can be understood as *convergence* of the reconstruction operator when the noise level δ tends to zero. More specifically, we want that when the noise level $\delta \to 0$, then $\lambda(\delta, y^\delta) \to \lambda_0 \geq 0$, and the reconstruction operator $\mathcal{R}(y^\delta; \lambda)$ converges to a generalized solution of the noiseless operator equation

$$Ax = y^0. \tag{4}$$

A family of such reconstruction operators $\mathcal{R}(y^\delta; \lambda)$ can be formulated in the framework of variational regularization (see Sect. 2.1.3 for more details) by defining them as the mapping to the minimizer of a variational energy function

$$f(Ax, y^\delta) + \lambda g(x). \tag{5}$$

Here, the first term ensures data consistency as in (2) and the second term acts as the regularizer to make the problem well-posed. It is often the case (although not always) that f is convex and smooth in x, while the regularizer g is convex but potentially non-smooth, e.g., sparsity-promoting regularizers involving the L^1-norm [6]. Therefore, to compute a minimizer of the variational problem, one utilizes non-smooth convex optimization algorithms to iteratively estimate the solution map $\mathcal{R}(y^\delta; \lambda)$. In particular, proximal splitting schemes are often used, such as forward-backward splitting (FBS), also referred to as proximal gradient descent, and the alternating directions method

of multipliers (ADMM), which involve applying the proximal operator of g (see (21) for definition) in each iteration to refine the solution. For instance, the FBS scheme iteratively updates the solution as

$$x_{k+1} = \mathrm{prox}_{\lambda g}\left(x_k - \eta \nabla f(Ax_k, y^\delta)\right), \tag{6}$$

starting from an initialization x_0, where η is the step-size and $\mathrm{prox}_{\lambda g}$ is the proximal operator of $\lambda g(x)$ (see (21) for definition). Note that (6) retains the modularity of the variational framework, in the sense that the proximal operator enforces prior knowledge about the image and its argument depends entirely on the forward operator and the fidelity loss. This decoupling of the fidelity and the prior in proximal splitting algorithms forms the basis of the so-called *plug-and-play* (PnP) denoising algorithms.

The PnP approach, pioneered by Venkatakrishnan et al. [49], noted that the proximal step can be interpreted as a denoising step and suggests replacing the proximal operator with an off-the-shelf (Gaussian) image denoiser. While the initial PnP methods utilized classical model-based denoisers (such as BM3D, non-local means, KSVD with a learned basis, etc.), more recent PnP schemes have leveraged deep denoisers that outperform their classical counterparts in terms of denoising quality.

Figure 1 shows a comparison of the performance of variational regularization methods and PnP methods on an image deblurring task. We compare linear reconstructions (corresponding to a quadratic regularization functional and linear denoiser) in (b) and (e) and non-linear reconstructions (with a non-quadratic regularization functional and non-linear denoiser) in (c) and (f). Notwithstanding the impressive empirical performance of data-driven PnP algorithms for imaging problems, their analysis as a convergent regularization scheme has largely remained unaddressed [15, 33]. In this paper, we address the question of convergent data-driven regularization, provide a survey of existing approaches and establish a convergence result for learned PnP denoisers, based on a linearity assumption of the denoiser (corresponding to the setting of (e) in Fig. 1) and a novel spectral filtering to control regularization strength.

Organization and contributions: We will first review classical approaches to regularization in Sect. 2 and why inverse problems necessitate a generalized notion of *solvability*. We will then continue to discuss how this classical approach can be combined with modern data-driven methods. In particular, we will devote special attention to plug-and-play (PnP) approaches in Sect. 3. We discuss, that depending on the regularity of the denoiser, different convergence guarantees of the PnP iterations can be established, such as fixed point [41] and objective convergence [22, 34]. Nevertheless, there is still a gap in the literature on whether PnP can provide a convergent regularization for general denoisers.

In Sect. 4 we will provide an important step forward to fill this gap and consider PnP algorithms with data-driven linear denoisers. We propose a novel spectral filtering technique to control the regularization strength of the denoiser, which allows us to establish that PnP with a learned linear denoiser leads to a convergent regularization scheme. More specifically, we prove that in the limit as the noise vanishes, the PnP reconstruction converges to the minimizer of a regularizing potential subject to the solution satisfying the noiseless operator equation (4). In the numerical experiments we examine the spectral filtering approach for the classical inverse problem of X-ray

(a) ground-truth (b) H^1, PSNR 25.1 dB (c) TV, PSNR 25.3 dB

(d) data, PSNR 22.1 dB (e) $\mathrm{PnP_l}$, PSNR 25.7 dB (f) $\mathrm{PnP_{nl}}$, PSNR 26.3 dB

Fig. 1 Image reconstruction with different regularizers from images corrupted by Gaussian blurring. The data in (**d**) is generated by convolving the ground-truth image (**a**) with a Gaussian kernel and adding white Gaussian noise. **b** and **c** correspond to variational reconstructions with the H^1 seminorm, $g(x) = \|\nabla x\|_2^2$, and total variation (TV) seminorm, $g(x) = \|\nabla x\|_1$, respectively. On the other hand, **e** and **f** represent plug-and-play (PnP) reconstructions (see Sect. 3) with a linear denoiser and non-linear gradient-step denoiser respectively. For **b** and **c** the regularization parameters were chosen to optimize the peak signal-to-noise ratio (PSNR). Both **b** and **e** correspond to linear reconstructions, whereas **c** and **f** are non-linear reconstructions

tomography and show that the proposed method does indeed provide a numerically verifiable convergent regularization. Finally, in Sect. 5 we provide concluding remarks and discuss directions forward.

2 Regularization for Inverse Problems and Data-Driven Methods

Regularization theory has been a rich and successful field in inverse problems for several decades. The primary motivation is to formulate a well-posed and stable inversion procedure that converges provably to a solution of the noiseless operator equation (4). The emergence of data-driven methods has given the field of inverse problems a new direction: by using large quantities of data we can significantly improve reconstruction results. However, the underlying question of a convergent regularization remains: does the obtained reconstruction solve the underlying operator equation?

Indeed, there exist a few methods that are provably convergent regularization methods, we refer to [33] for a survey. In the following, we will give a short overview of the

regularization theory and existing data-driven approaches that are provably convergent regularization methods in this context.

2.1 Classical Regularization Theory

Stable solutions to inverse problems need a way to handle varying noise levels. For this purpose, the concept of regularization has proven highly useful. Roughly, regularization can be understood as a convergence requirement to a unique solution, e.g., the minimum norm solution $x^\dagger$, where convergence depends on the noise level δ. That is, formally we consider the previously discussed reconstruction operator $\mathcal{R}_\lambda := \mathcal{R}(\cdot, \lambda)$, which provides a parameterized family of continuous operators $\mathcal{R}_\lambda : Y \to X$. The parameter λ depends on the noise level $\delta > 0$, where $\|y^\delta - y^0\| \leq \delta$ and $y^0 := Ax^*$ denotes noise-free data. We say that the family of reconstruction operators is a convergent regularization method if there exists a parameter choice rule $\lambda = \lambda(\delta, y^\delta)$ such that reconstructions $x^\delta := \mathcal{R}_{\lambda(\delta, y^\delta)}(y^\delta)$ converge to the solution $x^\dagger := A^\dagger y^0$ given by the pseudo-inverse as noise vanishes, in the sense that

$$\limsup_{\delta \to 0} \left\| x^\delta - x^\dagger \right\|_X = 0 \quad \text{as} \quad \limsup_{\delta \to 0} \{\lambda(\delta, y^\delta)\} = 0. \tag{7}$$

In other words, we have point-wise convergence of the reconstruction operators to the pseudo-inverse, i.e., $\mathcal{R}_{\lambda(\delta, y^\delta)}(y^\delta) \to A^\dagger y^0$ as $\delta \to 0$. We refer interested readers to [16] for a detailed discussion. This is, of course, quite restrictive and only considers convergence to the least-squares minimum-norm solution. Nevertheless, this can already be used as an important tool to design learned regularization methods, i.e., learned reconstruction approaches that formally satisfy the above convergence criteria, as we will discuss in the following.

2.1.1 Direct Regularization

Motivated by the convergence to the pseudo-inverse solution, one can obtain a regularization method by mimicking the construction of the pseudo-inverse. In finite dimensions, this can be achieved by the singular value decomposition (SVD) $A = USV^\top$ of the forward operator. The pseudo-inverse can then be simply obtained by $A^\dagger = VS^\dagger U^\top$, where $S^\dagger$ is the transposed singular value matrix with inverted singular values. A regularization method is now obtained by filtering the singular values with a noise-dependent filter function, or a noise level-dependent truncation.

Similarly, direct reconstruction methods that apply a regularized inverse of the forward operator can be shown to be convergent regularization methods. The most prominent example of such methods is the filtered back-projection (FBP) for X-ray CT, which is, in fact, still relevant in clinical practice. Here, the filtering operation removes high-frequency components in Fourier space to regularize the reconstructions. If the filtering is interpreted as a noise-dependent mollifier, one obtains the general class of approximate inverse [44] with convergence as noise vanishes.

A popular approach in data-driven methods is to formulate a learned reconstruction operator as the composition of a regularized reconstruction operator $\mathcal{R}_\lambda : Y \to X$

with a data-driven component $C_\theta : X \to X$. That is, the reconstruction operator is parameterized as $\mathcal{R}_{(\theta,\lambda)} := C_\theta \circ \mathcal{R}_\lambda$, where the data-driven component C_θ, usually parameterized using a deep convolutional neural network (CNN), is designed to improve the reconstruction by removing noise or undersampling artifacts [23, 26]. These approaches are also popularly referred to as *post-processing methods*.

Such one-step post-processing approaches are especially popular due to their simplicity, as C_θ can be efficiently trained when supervised pairs of high and low-quality reconstructions are available. Unfortunately, there are very few results on reconstruction guarantees for such methods. Specifically, the problem formulation as a composition of a regularized reconstruction followed by the data-driven component causes the reconstruction to often violate the so-called *data-consistency criterion*. That is, even if the data-fidelity $f\big(\big(A \circ \mathcal{R}_\lambda\big)(y^\delta), y^\delta\big)$ is small, it does not necessarily imply a small value of $f\Big(\big(A \circ C_\theta \circ \mathcal{R}_\lambda\big)(y^\delta), y^\delta\Big)$ corresponding to the output of the post-processing network C_θ. Thus, such schemes do not satisfy the convergence of the data fidelity and hence fail to be a convergent regularization strategy.

Nevertheless, as proposed in [45], this approach can be reformulated by constructing the post-processing network as $C_\theta = \mathrm{id} + \big(\mathrm{id} - A^\dagger A\big)\, Q_\theta$, where Q_θ is a Lipschitz-continuous deep neural network (DNN) and id denotes the identity operator on X. Here, $\big(\mathrm{id} - A^\dagger A\big)$ is the projection operator onto the null-space of A and hence the operator C_θ (referred to as null-space network) always satisfies $\big(A \circ C_\theta \circ \mathcal{R}_\lambda\big)(y^\delta) = \big(A \circ \mathcal{R}_\lambda\big)(y^\delta)$, ensuring that the output of C_θ explains the observed data. More importantly, the null-space network maintains the regularizing properties of the reconstruction method $\mathcal{R}_\lambda$ and hence provides a convergent regularization scheme [45] in the sense of direct regularization. See [7] for a recent extension of null-space networks to non-linear inverse problems.

2.1.2 Iterative Regularization

Iterative techniques constitute another important class of regularization approach in the classical literature [24]. Here, a regularized solution is obtained by applying early stopping on an iterative algorithm based on a discrepancy principle. Landweber iteration is a such a classical iterative regularization approach, in which iterations of the form $x_{k+1} = x_k - A^\top(Ax_k - y^\delta)$ are terminated after K steps, where K is the smallest integer such that $\|Ax_K - y^\delta\| \leq \delta$ is satisfied. Landweber iteration can be modified to include a damping term (see [43]), resulting in iterations of the form

$$x_{k+1} = x_k - A^\top(Ax_k - y^\delta) - \lambda_k(x_k - x^{(0)}),$$

where $x^{(0)}$ is an initial guess that encodes prior knowledge about the solution. The modified iteration converges to a solution that is closest to $x^{(0)}$, thereby introducing further stability. Aspri et al. [5] considered the modified Landweber scheme for non-linear inverse problems (with a forward operator F) and proposed a data-driven variant of it by using a learned damping term, leading to an iterative regularization scheme of the form

$$x_{k+1} = x_k - \nabla F(x_k)^\top(F(x_k) - y^\delta) - \lambda_k^\delta \tilde{A}^\top(\tilde{A}x_k - x^{(0)}),$$

where $\tilde{A}$ is a learned linear operator introduced in the damping term. The authors are able to prove strong convergence and stability in infinite dimensional Hilbert spaces. We refer interested readers to [5] for a detailed exposition on the learning strategy for $\tilde{A}$ and the analysis of the resulting data-driven iterative regularization approach.

2.1.3 Variational Regularization

The classical regularization theory, which defines convergent regularization by convergence to the pseudo-inverse solution as defined in (7) limits possible solutions. Therefore, one can consider more general variational approaches to inverse problems, which have been particularly popular due to their flexibility in incorporating prior knowledge and dealing with varying noise distributions. In the variational regularization framework, solutions are computed by minimizing a composite objective consisting of the data-consistency term and a regularization term. In particular, the solutions are given by

$$\mathcal{R}(y^\delta; \lambda) \in \arg\min_{x \in X} f(Ax, y^\delta) + \lambda g(x). \tag{8}$$

The loss functional $f : Y \times Y \to \mathbb{R}^+$ measures data fidelity and is not restricted anymore to be the squared L^2-norm. The regularization functional $g : X \to \mathbb{R}$ encodes prior belief about the ground-truth x^* and effectively restricts the null space of A. Here, $\lambda > 0$ is a simple weighting parameter to balance between the two terms of the composite objective in (8), but more generally could be a parameter of the functional itself, in which case we will write g_λ instead. The choice of a suitable regularizer g is governed by the need to balance two important factors: desirable analytical features and the encoded prior belief. For instance, an analytically favorable choice is given by the squared L^2-norm, which, in combination with a squared L^2-norm for the data fidelity, provides a closed-form solution. Unfortunately, the obtained solutions corresponding to this choice of the regularizer will be smooth, which may not be suitable for many imaging applications. Consequently, more advanced sparsity-promoting priors have been favored, most commonly the L^1-norm for sparse signals and total variation (TV) for sparse gradients, i.e., piece-wise constant functions. These regularizers are non-differentiable and hence need more advanced non-smooth optimization techniques to compute a minimizer [6], but they typically lead to a better reconstruction than the simple squared L^2-norm-based regularization. See the top row of Fig. 1 for a comparison of some handcrafted regularizers in the context of the inverse problem of image deblurring.

Notably, the role of the two terms in (8) is conceptually similar to the general formulation in (3) of a minimum-norm solution. Nevertheless, the variational formulation provides more flexibility and also necessitates a broader concept of regularization. This is because we can not always guarantee convergence to the minimum-norm solution, but we have to consider convergence with respect to the chosen regularization functional g [42]. The formal definition of a convergent regularization scheme is given in Definition 1. The primary differences to the classical formulation here are, that the

minimizer of the regularizing functional g is not necessarily unique and the regularization parameter is not required to converge to 0.

Definition 1 (Convergent regularization scheme) Let $x_\lambda \in X$ be a minimizer to the objective in (8) for a given λ with data $y^\delta \in Y$ and noise level $\|y^\delta - y^0\|_Y < \delta$. Assume that there is a corresponding parameter choice rule $\lambda = \lambda(\delta, y^\delta)$ such that $\lambda \to \lambda_0$ as $\delta \to 0$. The variational model defined by (8) is then said to *converge to a g-minimizing solution* if $x_{\lambda(\delta, y^\delta)} \to \hat{x}$ as $\delta \to 0$. Here, $\hat{x} \in X$ solves the variational model that corresponds to (8) with noise-free data $y^0 \in Y$, i.e.,

$$\hat{x} \in \operatorname*{arg\,min}_{x \in X} \; g(x) \text{ subject to } y^0 = A\hat{x} \text{ and where } \lambda_0 := \lim_{\delta \to 0} \lambda(\delta, y^\delta). \quad (9)$$

Let us remark to this end, that it is desirable to formulate a regularizer that has small values for the desired images, i.e., it penalizes undesired solutions but is also analytically or computationally tractable. It is important to note at this point that different regularizers g which provide a convergent regularization, will still produce different reconstruction results as illustrated in Fig. 1, as not all choices of g are a good representation of the desired ground-truth image. Here, learned regularizers have proven very successful, as the data itself can now be used to represent the regularizer and hence naturally offer a good representation of the desired features. Depending on the choice of representation, analysis of the learned regularizer may become more involved. In the following, we will discuss several choices for learned data-driven regularizers and how these can be used within the realm of variational regularization.

2.2 Learning a Regularizer

The idea of learning a regularizer from data, rather than the classical approach of modeling it from first principles as outlined above, has appeared in the literature in various forms. We outline here a few such approaches, ranging from relatively older yet widely popular ideas like dictionary learning to the more recent approaches of learning regularizers using deep neural networks.

2.2.1 Learning Sparsity-Promoting Dictionaries

We start with the concept of dictionary learning, which nicely illustrates how data can be used to learn a representation of the desired images. Here, we will use the concept of sparsity, which has long been important for modeling prior knowledge of solutions, to regularize inverse problems. Assuming that the reconstruction possesses a sparse representation in a given dictionary $\mathbb{D}$, one can develop sparse recovery strategies, associated computational approaches, and error estimates for the reconstruction. Instead of working with a given dictionary, the key idea is to *learn* a dictionary either *a-priori* or *jointly* with the reconstruction. Notably, almost all work on dictionary learning in sparse models has been carried out in the context of denoising, i.e., with $A = \mathrm{id}$.

Learning the dictionary separately to solve the reconstruction problem is usually done using a sparsity assumption on the representation given by the dictionary. Let $L_X : X \times X \to \mathbb{R}$ be a given loss function (e.g. the L^2- or L^1-norm). Further, let $x_1, \ldots, x_N \in X$ be the given unsupervised training data, $\mathbb{D} = \{\phi_i\} \subset X$ a dictionary, and the synthesis operator $\mathcal{E}^*_{\mathbb{D}} : \Xi \to X$ acting on the encoder space Ξ given as $\mathcal{E}^*_{\mathbb{D}}(\xi) = \sum_i \xi_i \phi_i$ for $\xi \in \Xi$. One approach in dictionary learning is given by

$$(\widehat{\mathbb{D}}, \hat{\xi}_i) = \underset{\xi_i \in \Xi, \mathbb{D} \subset X}{\arg\min} \sum_{i=1}^{N} [L_X(x_i, \mathcal{E}^*_{\mathbb{D}}(\xi_i)) + \theta \|\xi_i\|_0]. \tag{10}$$

Here, (10) is posed in terms of the L^0-norm and is an NP-hard problem. This suggests the use of *convex relaxation*, by replacing $\|\xi_i\|_0$ with $\|\xi_i\|_1$ in (10). This relaxation turns (10) into a *bi-convex problem* (convex in each variable when the others are kept fixed) subject to usual choices for L_X, and one can apply alternating minimization approaches for obtaining an approximate solution. Seminal work on sparse dictionary learning includes the K-SVD approach [2], geometric multi-resolution analysis (GMRA) [3], and online dictionary learning [30]. See also [40] and references therein for a thorough discussion on sparse dictionary learning approaches.

While dictionary learning in the context of sparse coding has been very popular and successful, there are still several issues with it related to the locality of learned structures and the computational effort needed, for instance when sparse coding is performed over a large number of images or image patches. Aiming for a computationally more feasible approach, convolutional dictionaries have been introduced. Here, the dictionary atoms are given by convolution kernels that act on signal features via convolution and hence provide computationally feasible shift-invariant dictionaries, where the atoms depend on the entire signal/image, see for instance [17].

The dictionary can also be learned jointly with the reconstruction, by formulating a joint optimization problem. An example of such an approach is the adaptive dictionary-based statistical iterative reconstruction (ADSIR) [52], and its variants [11, 51]. A joint problem could be formulated as:

$$\min_{x \in X, \xi_i \in \Xi, \mathbb{D}} \{f(Ax, y) + g_\lambda(x, \xi_1, \ldots, \xi_N, \mathbb{D})\}, \tag{11}$$

where

$$g_\lambda(x, \xi_1, \ldots, \xi_N, \mathbb{D}) := \sum_{j=1}^{N} \left[L_X(x_i, \mathcal{E}^*_{\mathbb{D}}(\xi_i)) + \lambda \|\xi_j\|_p^p \right], \tag{12}$$

while $\mathcal{E}^*_{\mathbb{D}} : \Xi \to X$ being the synthesis operator associated with $\mathbb{D}$.

A convergent regularization could now be obtained under suitable conditions on g_λ following the variational regularization framework in Sect. 2.1.3. Finally, a formulation in infinite dimensional spaces is studied in [9], proposing a convex variational model for joint reconstruction and dictionary learning, that applies to inverse problems and allows to establish existence and stability guarantees for the reconstruction.

2.2.2 Bilevel Learning

Starting from variational regularization methods where the reconstruction operator $\mathcal{R}_\lambda : Y \to X$ is defined as the solution map for (8), one can formulate a generic setup for learning selected components of (8) utilizing supervised training data and a suitable loss function $L_X : X \times X \to \mathbb{R}$. This setup can be tailored towards learning the regularization functional g_λ [13, 14], the data fidelity term f, or even an appropriate component in the forward operator A, e.g., in blind image deconvolution [21]. Notably, the joint dictionary learning problem (11) can also be formulated as a *bilevel learning problem*.

First, we generalize the regularizer g_λ consisting of a single regularization parameter λ to a set of parameters θ (vector-valued). Subsequently, we define the reconstruction operator as

$$\mathcal{R}_\theta(y) := \operatorname*{arg\,min}_{x \in X} \{ f(Ax, y) + g_\theta(x) \} \quad \text{for} y \in Y. \tag{13}$$

Given paired training data $(x_i, y_i) \in X \times Y$ that are i.i.d. samples of the $(X \times Y)$-valued random variable $(\mathbf{x}, \mathbf{y}) \sim \pi_{\text{joint}}$, we can formulate the following *bilevel learning* problem:

$$\begin{cases} \widehat{\theta} \in \operatorname*{arg\,min}_{\theta} \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim \pi_{\text{joint}}} [L_X(\mathcal{R}_\theta(\mathbf{y}), \mathbf{x})], \text{ where} \\ \mathcal{R}_\theta(y) := \operatorname*{arg\,min}_{x \in X} \{ f(A(x), y) + g_\theta(x) \}. \end{cases} \tag{14}$$

Note that $\widehat{\theta}$ is, by definition, a Bayes estimator [27, Chapter 4]: a set of parameters that minimizes the risk over the distribution π_{joint}. However, the true joint distribution π_{joint} is typically unknown and is replaced by its empirical counterpart given by the training data, in which case $\widehat{\theta}$ corresponds to *empirical risk minimization*.

In the bilevel optimization literature, as in the optimization literature as a whole, there are two main and mostly distinct approaches. In the discrete approach, one first discretizes the problem (13) and subsequently optimizes its parameters. In this way, optimality conditions and their well-posedness are derived in finite dimensions. Alternatively, $\mathcal{R}$ and its parameter θ in (14) are optimized in the continuum (i.e., appropriate infinite-dimensional function spaces) and then discretized. It should be noted that the resulting problems present several difficulties due to the frequent non-smoothness of the lower-level problem (think of TV regularization), which, in general, makes it impossible to verify Karush–Kuhn–Tucker constraint qualification conditions. This issue has led to the development of alternative analytical approaches to obtain first-order necessary optimality conditions [12, 20].

2.2.3 Adversarial Regularization

Another notable alternative approach to include a data-driven regularization in the reconstruction process is to learn an explicit regularization term in (8) and solve the

variational problem subsequently. One such option is to learn *adversarial regularizers* as first proposed in [29] and further developed in [32]. Here, the construction of data-driven regularization is inspired by how discriminative networks (also referred to as *critics*) are trained using modern Generative Adversarial Network (GAN) architectures.

To train such an adversarial regularizer, we assume to have $\{x_i\}_{i=1}^{n_1} \in X$ and $\{y_i\}_{i=1}^{n_2} \in Y$, which are i.i.d. samples from the marginal distributions π_x and π_y of ground-truth images and measurement data, respectively. It is important to note here that the training samples are unpaired, i.e., y_i does not necessarily correspond to the noisy measurement of x_i, unlike, for instance, a supervised approach such as the learned primal-dual (LPD) method [1]. Additionally, we assume that there exists a (potentially regularizing) pseudo-inverse $A^\dagger\colon Y \to X$ to the forward operator A and define the measure $\pi_\dagger \in \mathbb{P}_X$ as $\pi_\dagger := A^\dagger_\#(\pi_{\text{data}})$ for $\pi_{\text{data}} \in \mathbb{P}_Y$.

Then, the idea is to train a regularizer g_θ, parameterized by a neural network, to discriminate between the distributions π_x and $\pi_\dagger$, the latter representing the distribution of imperfect solutions $A^\dagger y_i$. More concretely, we compute

$$g_{\widehat{\theta}}\colon X \to \mathbb{R} \quad \text{where} \quad \widehat{\theta} \in \arg\min_\theta L(\theta), \tag{15}$$

where $L(\theta)$ is chosen to be a Wasserstein-flavored loss functional [29]. In particular, one minimizes

$$L(\theta) := \mathbb{E}_{\mathbf{x}\sim\pi_x}\left[g_\theta(\mathbf{x})\right] - \mathbb{E}_{\mathbf{x}\sim\pi_\dagger}\left[g_\theta(\mathbf{x})\right] + \lambda\,\mathbb{E}_{\mathbf{x}\sim\widetilde{\pi}}\left[\left(\|\nabla g_\theta(\mathbf{x})\| - 1\right)_+^2\right]. \tag{16}$$

Here, $\widetilde{\pi}$ denotes the distribution of the random variable $\mathbf{u} = \epsilon\,\mathbf{x} + (1-\epsilon)\mathbf{z}$, where $\mathbf{x} \sim \pi_x$, $\mathbf{z} \sim \pi_\dagger$, and ϵ is drawn uniformly at random from $[0, 1]$. The heuristic behind this choice is that a regularizer trained this way will penalize noise and artifacts generated by the pseudo-inverse (and contained in $\pi_\dagger$). The term penalizing the gradient norm of g_θ in (16) encourages g_θ to be approximately 1-Lipschitz, which is required for the well-posedness of (16) and the stability of the variational solution obtained using the regularizer resulting from (15). When used as a regularizer, it will, therefore, prevent these undesirable features from occurring as a result of adversarial training. The resulting regularizer $g_{\widehat{\theta}}$ is called an adversarial regularizer (AR). Note that in practical applications, the measures $\pi_x, \pi_\dagger \in \mathbb{P}_X$ are replaced with their empirical counterparts given by training data x_i and $A^\dagger y_i$, respectively.

Suppose, one computes a gradient step on the learned regularizer, given by $x_\eta = x - \eta\,\nabla_{\mathbf{x}} g_{\widehat{\theta}}(\mathbf{x})$, starting from $x \sim \pi_\dagger$. Let $\pi_\dagger^\eta$ be the distribution of x_η. Under appropriate regularity assumptions on the Wasserstein distance $\mathcal{W}(\pi_\dagger^\eta, \pi_x)$ (see [29, Theorem 1]), one can show that

$$\frac{d}{d\eta}\mathcal{W}(\pi_\dagger^\eta, \pi_x)|_{\eta=0} = -\mathbb{E}_{\mathbf{x}\sim\pi_\dagger}\|\nabla_{\mathbf{x}} g_{\widehat{\theta}}(\mathbf{x})\|^2.$$

This ensures that by taking a small enough gradient step, one can reduce the Wasserstein distance from the ground truth π_x. This is a good indicator that using $g_{\widehat{\theta}}$ as

a variational regularization term and consequently penalizing it indeed introduces the highly desirable incentive to align the distribution of regularized solutions with the distribution π_x of ground truth samples. Further, one can show that if the AR is Lipschitz-continuous,[1] then a minimizer of the following variational problem exists

$$f(y^\delta, Ax) + \lambda \left(g_{\widehat{\theta}}(x) + \epsilon \|x\|_X^2 \right), \tag{17}$$

where the squared norm on x is needed to enforce coercivity. In this setting, convergence of the regularization procedure in the weak topology of X can be ensured.

Additionally, we can enforce (strong) convexity on g_θ, leading to the *adversarial convex regularizer* (ACR), to achieve stronger forms of convergence while precluding discontinuities in the reconstruction operator. This necessitates a suitable parameterization of the learned regularizer. One such option is given by input convex neural networks for imposing convexity [4] on $g_{\widehat{\theta}}$. Given a so-constructed (ACR) $g_{\widehat{\theta}}$ that is convex in x, we then consider a similar regularization functional of the form

$$g(x) = g_{\widehat{\theta}}(x) + \epsilon \, \|x\|_X^2 \,, \tag{18}$$

where $g_{\widehat{\theta}} : X \to \mathbb{R}$ is the trained (ACR) which we assume to be 1-Lipschitz and convex in x. The corresponding variational regularization problem then consists in minimizing

$$f(y^\delta, Ax) + \lambda g(x), \tag{19}$$

with respect to $x \in X$. In this setting, we get the following set of improved theoretical guarantees for the ACR, by following standard arguments in variational calculus for the proofs.

Theorem 1 (Properties of Adversarial Convex Regularizer [32])

i. Existence and uniqueness: The functional in (19) is strongly convex in x and has a unique minimizer $\widehat{x}_\lambda(y)$ for every $y \in Y$ and $\lambda > 0$.

ii. Stability: The optimal solution $\widehat{x}_\lambda(y)$ is continuous in y.

iii. Convergence: For $\delta \to 0$ and $\lambda(\delta) \to 0$ such that $\dfrac{\delta}{\lambda(\delta)} \to 0$, we have that $\widehat{x}_\lambda\left(y^\delta\right)$ converges to the g-minimizing solution $x^\dagger$ given in (9).

Theoretical guarantees notwithstanding, the numerical experiments in [32] (especially, for sparse-view CT reconstruction) indicate a lack of expressive power of ACRs as compared to their non-convex counterpart AR. This underscores the need to develop techniques that achieve a better compromise between empirical performance and theoretical certificates.

[1] 1-Lipschitz continuity is approximately enforced by the gradient penalty term in (16), which does not guarantee, however, that the (AR) is Lipschitz continuous. This property can be enforced by choosing the right network architecture. Indeed, all convolutional neural networks with ReLU activations are Lipschitz continuous for some Lipschitz constant L, which might be arbitrarily large.

2.2.4 The Network Tikhonov (NETT) Approach

Traditionally, regularizers are often chosen as sparsifying transforms with respect to certain features. For instance, total variation (TV) is sparsifying for piecewise constant functions. Similarly, neural networks are often trained in an encoder-decoder (autoencoder) structure, where the encoder is trained to represent the input signal in a low-dimensional space or to find a more efficient, i.e., a sparse structure. The approach proposed as Network Tikhonov (NETT) in [28] follows this paradigm to learn a regularizer. Here, a pretrained network $\mathcal{E}_\theta \colon X \to \Xi$ is composed with a regularization functional $g \colon \Xi \to [0, +\infty]$, such that $g \circ \mathcal{E}_\theta \colon X \to [0, +\infty]$ takes small values for desired model parameters and penalizes (by producing larger values for) model parameters with artifacts or other unwanted structures. The deep neural network $\mathcal{E}_\theta$ in this approach is allowed to be a rather general architecture, such as the above-mentioned autoencoder. Once trained, the reconstruction is then given as the minimizer of the variational objective

$$J_\theta(x) := f(y^\delta, Ax) + \lambda g\big(\mathcal{E}_\theta(x)\big). \tag{20}$$

Indeed, the NETT approach also provides a provably convergent regularization method under certain analytic conditions on (20), such as weak lower semi-continuity and coercivity of the regularizer $g(\mathcal{E}_\theta(\cdot))$, which can be achieved as follows. First, the usual ReLU activation function is replaced by leaky ReLU defined with a small $\tau > 0$ as

$$\ell\mathrm{ReLU}_\tau(s) := \max(\tau s, s),$$

which tends to $-\infty$ for $s \to -\infty$. In combination with the affine linear maps (weight matrices) in $\mathcal{E}_\theta$, this yields a coercive and weakly lower semi-continuous regularization function $g \circ \mathcal{E}_\theta$ for standard choices of g, such as weighted ℓ_p-norms $g(\xi) = \sum_i v_i |\xi|^p$, with uniformly positive weights v_i and $p \geq 1$. Finally, we note that strong convergence can be achieved by introducing the novel concept of absolute Bregman distances and imposing stronger conditions on the regularizer.

3 Regularization by Plug-and-Play (PnP) Denoising

Denoising is the simplest and arguably the most well-studied inverse problem in imaging, with numerous algorithms developed over the past few decades, particularly for removing additive white Gaussian noise from images. It is, therefore, natural to ask if one can leverage off-the-shelf denoisers for solving more complicated image recovery tasks with a non-trivial forward operator. Venkatakrishnan et al. [49] pioneered the idea of using denoisers within proximal splitting algorithms (e.g., the alternating directions method of multipliers (ADMM) algorithm) in a plug-and-play (PnP) fashion, and the resulting class of algorithms came to be known as the PnP denoising approach. To see the motivation behind using denoisers in place of proximal operators, let us recall the definition of the proximal operator with respect to a (potentially non-smooth) convex

functional $g : X \to \mathbb{R} \cup \{+\infty\}$ and a step-size $\tau > 0$:

$$\text{prox}_{\tau\, g}(x) = \arg\min_u \frac{1}{2}\|x - u\|^2 + \tau\, g(u). \tag{21}$$

As indicated by (21), evaluating the proximal operator amounts to denoising a noisy image x using the Bayesian *maximum a-posteriori probability* (MAP) estimation framework with a Gibbs prior proportional to $\exp(-\tau\, g(u))$. This denoising interpretation of proximal operators underlies the foundation of PnP approaches, which have been shown to produce excellent reconstruction results for a wide range of imaging inverse problems. A classic and widely popular example of PnP denoising would be to consider it in conjunction with forward-backward splitting (FBS), leading to the following iterative reconstruction algorithm:

$$x_{k+1} = D_\sigma\left(x_k - \eta_k \nabla f(x_k)\right). \tag{22}$$

Here, f denotes the data fidelity loss for the underlying inverse problem, $\eta_k > 0$ is the step-size at iteration k, and D_σ is a denoiser that eliminates Gaussian noise of standard deviation σ from its input.

Besides the PnP denoising framework within proximal methods, wherein a denoiser implicitly acts as a regularizer, Romano et al. [38] proposed an alternative approach to explicitly construct a regularizer as

$$g(x) = \frac{1}{2} x^\top \left(x - D_\sigma(x)\right), \tag{23}$$

while utilizing a denoiser $D_\sigma(x)$. One can then seek to minimize the energy functional $f(x) + \lambda\, g(x)$, where g is as defined in (23), leading to fixed-point iterative schemes known as the regularization-by-denoising (RED) algorithms. Nevertheless, it was shown subsequently by Schniter et al. [37] that the *energy minimization* interpretation of the RED algorithms is valid only when (i) the denoiser is *locally homogeneous*, i.e., $D_\sigma\left((1+\epsilon)x\right) = (1+\epsilon) D_\sigma(x)$ holds for all x with sufficiently small ϵ, and (ii) the Jacobian of D_σ is symmetric. These conditions are generally not satisfied by generic denoisers, thereby invalidating the energy minimization-based interpretation of RED. Instead, the authors of [37] developed a new framework called *score-matching* to analyze the convergence of RED algorithms.

In spite of their empirical success, PnP denoising algorithms such as (22) do not immediately inherit the convergence properties of the corresponding optimization scheme (in this specific instance, FBS). Studying the convergence of PnP denoising has received a significant amount of attention in the mathematical imaging community in recent years. Arguably, the most natural form of convergence for PnP algorithms of the form (22) is the stability of the iterations, i.e., to ascertain whether the sequence of iterates x_k generated by a PnP algorithm converges. Such convergence guarantees are typically derived from fixed point theorems, which require showing that the PnP iterations are contractive maps [10, 41]. For instance, [41] established the fixed-point convergence of PnP-ADMM (i.e., PnP with the *alternating direction method of*

multipliers algorithm) under the assumption of Lipschitz continuity of the operator $(D_\sigma - \mathrm{id})$. The specific result is stated in Theorem 2.

Theorem 2 (Fixed-point convergence of PnP-ADMM [41]) *Consider the PnP-ADMM algorithm, given by*

$$\begin{aligned} x_{k+\frac{1}{2}} &= \mathrm{prox}_{\tau f}(z_k), \, x_{k+1} = D_\sigma\left(2x_{k+\frac{1}{2}} - z_k\right), \text{ and} \\ z_{k+1} &= z_k + x_{k+1} - x_{k+\frac{1}{2}}, \end{aligned} \tag{24}$$

where the data-fidelity loss f is assumed to be μ-strongly convex. One can equivalently express (24) as the fixed-point iteration $z_{k+1} = \mathcal{T}(z_k)$, where

$$\mathcal{T} = \frac{1}{2}\mathrm{id} + \frac{1}{2}(2D_\sigma - \mathrm{id})\left(2\,\mathrm{prox}_{\tau f} - \mathrm{id}\right). \tag{25}$$

Suppose, the denoiser satisfies

$$\|(D_\sigma - \mathrm{id})(u) - (D_\sigma - \mathrm{id})(v)\|_2 \leq \epsilon \|u - v\|_2, \tag{26}$$

for all $u, v \in X$ and some $\epsilon > 0$, and the strong convexity parameter μ is such that $\frac{\epsilon}{(1+\epsilon-2\epsilon^2)\mu} < \tau$ holds, the operator $\mathcal{T}$ is contractive and the PnP-ADMM algorithm is fixed-point convergent. That is, $(x_k, z_k) \to (x_\infty, z_\infty)$, where (x_∞, z_∞) satisfy

$$x_\infty = \mathrm{prox}_{\tau f}(z_\infty) \text{ and } x_\infty = D_\sigma(2x_\infty - z_\infty). \tag{27}$$

As noted in [41], fixed-point convergence of PnP-ADMM follows from monotone operator theory if $(2D_\sigma - \mathrm{id})$ is non-expansive, but (26) imposes a less restrictive condition on the denoiser.

While fixed-point convergence ensures that the PnP iterations are stable, the specific fixed point to which they converge does not automatically minimize a variational energy function. To bridge the gap between classical variational approaches and PnP methods, it is important to derive conditions under which the limit point of PnP iterations can be characterized as the minimizer (or, at least a stationary point) of some regularized variational objective (which, of course, depends on the denoiser). This type of convergence is referred to as *objective convergence* and is stronger than fixed-point convergence.

Objective convergence of PnP with classical (pseudo) linear denoisers (e.g., non-local means denoiser) has been established in [34]. Hurault et al. [22] showed that PnP with a denoiser constructed as a gradient field (referred to as gradient-step (GS) denoisers) converges to a stationary point of a (possibly non-convex) variational objective (c.f. Theorem 3). The construction of GS denoisers is motivated by Tweedie's identity: the optimal minimum mean-squared error (MMSE) Gaussian denoiser is given by

$$D_\sigma^*(x) := \mathbb{E}[\mathbf{x}_0 | \mathbf{x} = x] = x + \sigma^2 \nabla \log p_\sigma(x). \tag{28}$$

Here, $\mathbf{x} = \mathbf{x}_0 + \sigma\,\mathbf{z}$, where $\mathbf{z} \sim \mathcal{N}(0, \mathrm{id})$, is the Gaussian noise (with variance σ^2) corrupted version of the clean image $\mathbf{x}_0 \in X \subseteq \mathbb{R}^d$ and

$$p_\sigma(x) = \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} \int \exp\left(-\frac{\|x - x_0\|_2^2}{2\sigma^2}\right) p(x_0)\,\mathrm{d}x_0. \tag{29}$$

Indeed, the optimal Gaussian denoiser is of the form $D_\sigma^*(x) = x - \nabla\, g_\sigma^*(x)$, where g_σ^* is the negative log of the smoothed distribution p_σ defined in (29), which has a structure identical to that of a GS denoiser.

Theorem 3 (Objective convergence of PnP iterations with gradient-step (GS) denoisers [22]) *Suppose, the denoiser is constructed as a gradient-step (GS) denoiser, i.e., $D_\sigma = \mathrm{id} - \nabla g_\sigma$, where g_σ is proper, lower semi-continuous, and differentiable with an L-Lipschitz gradient. The PnP algorithm proposed in [22] is given by*

$$\begin{aligned} x_{k+1} &= \mathrm{prox}_{\tau f}\left(x_k - \tau\,\lambda\,\nabla g_\sigma(x_k)\right) \\ &= \mathrm{prox}_{\tau f} \circ \left(\tau\lambda\, D_\sigma + (1 - \tau\lambda\ \mathrm{id})\right)(x_k), \end{aligned} \tag{30}$$

where $f\colon X \to \mathbb{R} \cup \{+\infty\}$ denotes data-fidelity and is assumed to be convex and lower semi-continuous. Then, the following guarantees hold for $\tau < \frac{1}{\lambda L}$:

1. *The sequence $F(x_k)$, where $F = f + \lambda\, g_\sigma$, is non-increasing and convergent.*
2. *$\|x_{k+1} - x_k\|_2 \to 0$, which indicates that iterations are stable, in the sense that they do not diverge if one iterates indefinitely.*
3. *All limit points of $\{x_k\}$ are stationary points of $F(x)$.*

Notably, the PnP iteration defined by (30) is exactly equivalent to proximal gradient descent on $f + \lambda\, g_\sigma$, with a potentially non-convex g_σ.

While objective convergence ensures a one-to-one connection between PnP iterates with the minimization of a variational objective, it does not provide any guarantees about the regularizing properties of the solution that the iterates converge to. In the same spirit as classical regularization theory, it is therefore desirable to be able to control the implicit regularization effected by the denoiser in PnP algorithms and analyze the asymptotic behavior of the PnP reconstruction as the noise level and the regularization strength tend to vanish. More precisely, assuming that the PnP iterations converge to a solution $\hat{x}\left(y^\delta, \sigma, \lambda\right)$, where σ is a parameter associated with the denoiser and λ is an explicit regularization penalty, one would like to obtain appropriate selection rules for σ and/or λ such that $\hat{x}\left(y^\delta, \sigma, \lambda\right)$ exhibits convergence akin to (9) in the limit as $\delta \to 0$. To the best of our knowledge, some progress in this direction was first made in [15], and the precise convergence result is stated in Theorem 4.

Theorem 4 (Convergent plug-and-play (PnP) regularization [15]) *Consider the PnP-FBS iterates of the form*

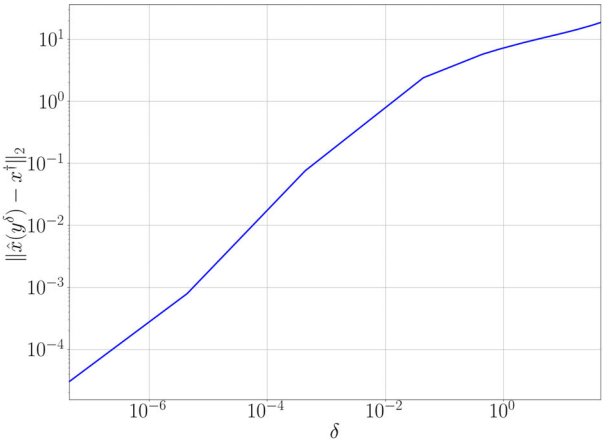
$$x_{\lambda,k+1}^\delta = D_\lambda\left(x_{\lambda,k}^\delta - \eta\, A^*\left(A x_{\lambda,k}^\delta - y^\delta\right)\right), \tag{31}$$

where D_λ is a denoiser with a tuneable regularization parameter λ. Let $\mathrm{PnP}\left(\lambda, y^\delta\right)$ *be the fixed point of the PnP iteration (31). For any $y \in \mathrm{range}(A)$ and any sequence $\delta_k > 0$ of noise levels converging to 0, there exists a sequence λ_k of regularization parameters converging to 0 such that for all y_k with $\|y_k - y^0\|_2 \leq \delta_k$, the following hold under appropriate assumptions on the denoiser (see Definition 3.1 in [15] for details):*

1. $\mathrm{PnP}\left(\lambda, y^\delta\right)$ *is continuous in y^δ for any $\lambda > 0$;*
2. *The sequence* $(\mathrm{PnP}\,(\lambda_k, y_k))_{k\in\mathbb{N}}$ *has a weakly convergent subsequence; and*
3. *The limit of every weakly convergent subsequence of* $(\mathrm{PnP}\,(\lambda_k, y_k))_{k\in\mathbb{N}}$ *is a solution of the operator equation $y^0 = Ax$.*

The result in [15], although the first of its kind, suffers from two main shortcomings: (i) the stability and convergence theory in [15] is based on fairly restrictive assumptions on the denoiser. In particular, the denoiser needs to be *contractive*, which is not satisfied by most practical denoisers, especially denoisers modeled using deep CNNs. (ii) It is not apriori clear how to select the denoiser parameter λ to control the strength of regularization corresponding to different noise levels. In this paper, we address the latter issue by considering linear denoisers, which are simple yet lead to competitive PnP algorithms [34]. Further, we show in Sect. 4 that the well-known idea of *denoiser scaling* [50] does not work for linear denoisers, which also underscores the need for devising an effective strategy to control the regularization of linear denoisers.

However, let us first discuss a more ambitious question here: Are PnP approaches with more general and expressive denoisers also convergent regularization methods, while offering a straightforward approach to control the regularization strength? This question is perhaps more tractable if one can associate the PnP solution (after convergence) with the minimizer of an underlying variational objective. We, therefore, first consider gradient-step denoisers, for which it is possible to establish such a connection (see Theorem 3). Treating λ in (30) as an explicit regularization parameter while using a fixed, pre-trained denoiser, one can interpret the converged PnP solution as a minimizer of $f + \lambda\, g_\sigma$, where λ is varied depending on the noise level δ in the measurement data and σ is kept fixed. The numerical results for image deblurring in Fig. 2 seem to indicate that gradient-step PnP is indeed a convergent regularization scheme, while the classical theory only guarantees stability akin to what is shown in [29] subject to g_σ being coercive and bounded below. In addition, the role of σ as an implicit regularization parameter is not exploited, and it is kept unchanged in our experiments regardless of the noise level in the measurement. This, in part, is due to the fact that the behavior of g_σ w.r.t. σ is non-trivial to characterize in a precise manner, leading to difficulties in tuning σ based on δ. Therefore, in order to rigorously establish convergence, together with developing a principled approach to control the regularization strength arising from the denoiser, we focus our attention on PnP with linear denoisers in the next section.



(a) noise level vs. distance from the g_σ-minimizing solution $x^\dagger$.

(b) $\hat{x}\left(y^{\sigma_0}\right), \sigma_0 = 10\%$ (c) $\hat{x}\left(y^{\sigma_0}\right), \sigma_0 = 8\%$ (d) $\hat{x}\left(y^{\sigma_0}\right), \sigma_0 = 5\%$

(e) $\hat{x}\left(y^{\sigma_0}\right), \sigma_0 = 2\%$ (f) $\hat{x}\left(y^{\sigma_0}\right), \sigma_0 = 1\%$ (g) $\hat{x}\left(y^{\sigma_0}\right), \sigma_0 = 0.5\%$

(h) $\hat{x}\left(y^{\sigma_0}\right), \sigma_0 = 0.1\%$ (i) $x^\dagger, \sigma_0 = 0$ (j) ground-truth

Fig. 2 PnP gradient-step DRUNet denoiser as a convergent regularization method for image deblurring. The PnP scheme for reconstruction minimizes variational energy of the form $f + \lambda\, g_\sigma$, where f is the fidelity and g_σ is the regularizer induced by a pre-trained denoiser. Plot **a** quantitatively demonstrates the convergence of the reconstructions as the noise level decreases. The input blurry image is given by $y^{\sigma_0} = Ax + w$, where A is a Gaussian blur kernel and w is additive Gaussian noise with variance σ_0^2. The images from **b** to **i** are the deblurred images $\hat{x}\left(y^{\sigma_0}\right)$ corresponding to the noise level σ_0 (expressed as the % of maximum pixel value 255.0 in the ground truth), while image **j** is the ground truth image. The regularization parameter is selected as $\lambda = c\,\sigma_0 + \epsilon$, where the constant $c = 0.04$ and $\epsilon = 10^{-4}$

4 Controlling the Regularization Strength in PnP

A fundamental question that arises when applying learned denoisers for solving inverse problems using PnP concerns itself with how to adjust the regularization strength that is applied. Indeed, learned denoisers are typically trained at a fixed noise level, whereas their practical application to inverse problems in a PnP framework and the theoretical notion of convergent regularization both require one to have certain control over the regularization strength.

An approach that has been shown to be beneficial in practice is the *denoiser scaling* approach [50]: given a denoiser D_σ (designed for denoising at a given noise level σ), we introduce an extra scaling parameter $\alpha > 0$, and define the scaled denoisers $\{D_{\sigma,\alpha}\}_{\alpha>0}$ as

$$D_{\sigma,\alpha}(x) = \frac{1}{\alpha} D_\sigma(\alpha x). \tag{32}$$

This choice of scaling is motivated by the fact that if $J : X \to \mathbb{R} \cup \{\infty\}$ is 1-homogeneous (i.e., $J(\tau u) = \tau J(u)$, for $\tau > 0$) and its proximal operator is well-defined, we have

$$\begin{aligned}
\operatorname{prox}_{\tau J}(x) &= \arg\min_y \frac{1}{2}\|x - y\|^2 + \tau J(y) \\
&= \arg\min_y \frac{\tau^2}{2}\left\|\frac{x}{\tau} - \frac{y}{\tau}\right\|^2 + \tau J(y) \\
&= \tau \arg\min_u \frac{1}{2}\left\|\frac{x}{\tau} - u\right\|^2 + J(u) \\
&= \tau \operatorname{prox}_J\left(\frac{x}{\tau}\right).
\end{aligned}$$

In other words, if $D_\sigma = \operatorname{prox}_J$, then $D_{\sigma,\alpha} = \operatorname{prox}_{J/\alpha}$. Let us note that the choice of this particular scaling, while natural (norms and seminorms are 1-homogeneous, for example), is somewhat arbitrary. Indeed, suppose that J is instead c-homogeneous for some $c > 0$, i.e., $J(\delta u) = \delta^c J(u)$ for any u and $\delta > 0$. We have, with $\delta > 0$ arbitrary,

$$\begin{aligned}
\operatorname{prox}_{\tau J}(x) &= \arg\min_y \frac{1}{2}\|x - y\|^2 + \tau J(y) \\
&= \arg\min_y \frac{\delta^2}{2}\left\|\frac{x}{\delta} - \frac{y}{\delta}\right\|^2 + \tau J(y) \\
&= \delta \arg\min_u \frac{1}{2}\left\|\frac{x}{\delta} - u\right\|^2 + \frac{\tau}{\delta^2} J(\delta u) \\
&= \delta \arg\min_u \frac{1}{2}\left\|\frac{x}{\delta} - u\right\|^2 + \frac{\tau}{\delta^{2-c}} J(u).
\end{aligned}$$

Choosing $\delta = \tau^{\frac{1}{2-c}}$, we find that

$$\operatorname{prox}_{\tau J}(x) = \tau^{\frac{1}{2-c}} \operatorname{prox}_J(\tau^{\frac{1}{c-2}} x), \tag{33}$$

which agrees with the result for 1-homogeneous functionals and generalizes it, except for 2-homogeneous functionals where the above derivation does not work. In fact, this leads nicely into a setting where no form of denoiser scaling as in Eq. (33) can possibly be used to control the regularization strength to give a convergent regularization: for linear denoisers, the multiplicative factor inside the denoiser can be pulled out and canceled against the factor outside of it.

4.1 Controlling the Regularization Strength of a Linear Denoiser

Let us consider the setting in which we have a linear denoiser $D_\sigma : X \to X$. If we are to interpret it as a proximal operator of some underlying functional, we must assume that it is a symmetric, positive semi-definite (p.s.d.) operator, and if we assume that the underlying functional is convex as well, then D_σ must be non-expansive in addition. These properties are direct consequences of the characterization of proximal operators given in [31] and generalized (to potentially non-convex functionals) in [19]. Let us restrict to the case where D_σ is non-expansive, bypassing the potential difficulties of non-convexity of the underlying variational problem. In fact, we will assume that D_σ is contractive, i.e. $\|D_\sigma\| < 1$, which as we will see later corresponds to assuming that the underlying regularization functional is coercive. Furthermore, we will assume that D_σ is bounded from below, i.e. $\|D_\sigma(x)\| \geq c\|x\|$ for some $c > 0$, so that D_σ^{-1} exists and is a bounded operator.

Remark 1 In practice, the assumption of symmetry can be relaxed somewhat by taking a different perspective: in [18] it is shown in finite dimensions that any denoiser which is similar to a symmetric p.s.d. matrix is admissible in PnP applications. Indeed, in this case we can find a modified inner product, with respect to which the denoiser is a proximal operator.

Note that the assumptions that we make are ideally suited to the application of the spectral theory of bounded linear operators on Hilbert spaces. Recall that the spectrum of a bounded linear operator $A : X \to X$, $\operatorname{spec}(A)$, is defined as the set of $\lambda \in \mathbb{C}$ such that $A - \lambda \operatorname{id}$ is not boundedly invertible. For bounded and symmetric operators A (such as the ones we consider) $\operatorname{spec}(A)$ is a compact subset of $\mathbb{R}$, and we have a spectral theorem that enables the use of the continuous functional calculus. This is crucial in what follows as it allows us to apply a spectral filtering operation to denoisers to control their regularization strength, by adjusting the weights of the spectral components of the denoiser. We recommend that readers who are interested in studying these topics in more detail consult the textbook [46, Chapter 5].

Let us study the characterization of proximal operators in more detail for the linear denoiser D_σ. The goal is to understand the underlying functional $J : X \to \mathbb{R}$ such that $D_\sigma = \operatorname{prox}_J$. Note first that it is immediate from the definition (Eq. 21) that we

can only hope to recover J up to an additive constant. We have

$$D_\sigma = \text{prox}_J = (\text{id} + \partial J)^{-1},$$

with ∂ being the subdifferential. On the other hand, since D_σ is linear, D_σ^{-1} is linear, and by the equation above $\partial J =: W$ is also linear. As a result of this we have the following, up to an additive constant: $J(x) = \frac{1}{2}\langle x, Wx\rangle$. Furthermore, inverting the equation above and rearranging, we find that $W = D_\sigma^{-1} - \text{id}$. Hence, up to an irrelevant additive constant, we find that the underlying regularization functional J corresponding to D_σ is given by

$$J(x) = \frac{1}{2}\langle x, (D_\sigma^{-1} - \text{id})x\rangle. \tag{34}$$

The most common way of controlling the regularization strength, when we have access to the underlying regularization functional J, is to simply scale it: introduce a parameter $\tau > 0$ and consider $\text{prox}_{\tau J}$. If we apply this to Eq. (34), we obtain

$$\tau J(x) = \frac{1}{2}\langle x, ([\tau D_\sigma^{-1} - (\tau - 1)\,\text{id}] - \text{id})x\rangle,$$

which suggests, by following the above reasoning in reverse, that

$$\text{prox}_{\tau J} = (\tau D_\sigma^{-1} - (\tau - 1)\,\text{id})^{-1} = h_\tau(D_\sigma). \tag{35}$$

Here $h_\tau : \mathbb{R} \to \mathbb{R}$, given by $h_\tau(\lambda) = \lambda/(\tau - \lambda(\tau - 1))$ is applied to D_σ using the functional calculus. The takeaway message of the preceding derivation is that we can perform a *spectral filtering* operation on the linear denoiser D_σ to control its regularization strength. In fact, more general filter functions h_τ than the one seen here can be used, as we will see in what follows.

Remark 2 It is worth contrasting the spectral filtering approach proposed here with well-established spectral filtering approaches to regularization of linear, ill-posed, inverse problems [16]: whereas the traditional approaches operate on the forward operator to enact a regularization effect, we operate on the denoiser (agnostic about the forward operator to which the denoiser will be applied) to control its regularization strength.

To get a better understanding of what the spectral filtering operation does to a denoiser, consider Fig. 3. This will help us get an idea of what we should ask of generalized filter functions, i.e. filter functions that do not just implement a scaling of the underlying regularization functional: as $\tau \to 0$, the effect of the denoiser should vanish at an appropriate rate.

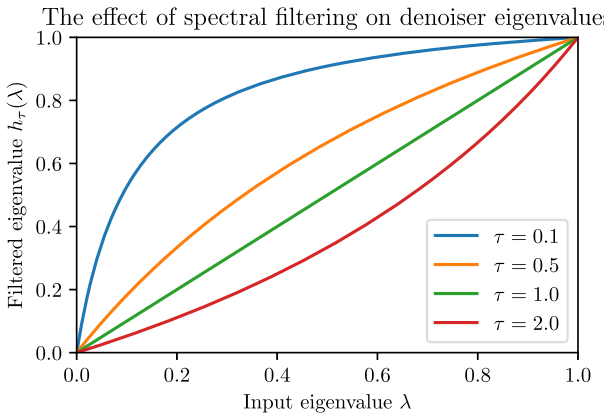


Fig. 3 The effect of filtering the denoiser as in (35). In accordance with intuition, the spectrum is flattened as $\tau \to 0$: as the regularization strength vanishes, the effect of the denoiser should vanish too

4.2 Convergent Regularization Through Generalized Spectral Filtering of Linear Denoisers

In the previous section, we saw that there is a way in which we can spectrally filter a linear denoiser to effectively scale the underlying regularization functional. Now, we will generalize the conditions on the spectral filter and show that this spectral filtering of linear denoisers allows us to obtain a convergent regularization of linear, ill-posed, inverse problems.

Equation (34) tells us that a linear denoiser is related to an underlying regularization functional J as follows: we have $D_\sigma = \operatorname{prox}_J$, where

$$J(x) = \frac{1}{2}\langle x, (D_\sigma^{-1} - \mathrm{id})x\rangle.$$

Furthermore, we have seen the effect of scaling the regularization functional on the corresponding proximal operator. We can generalize this idea and look at

$$J_\tau(x) := \frac{1}{2}\langle x, (h_\tau(D_\sigma)^{-1} - \mathrm{id})x\rangle,$$

where $\{h_\tau : \mathbb{R} \to \mathbb{R}\}_{\tau\in(0,\infty)}$ is a family of spectral filters that we can apply to D_σ using the continuous functional calculus. Let us now derive conditions on the spectral filters h_τ such that this gives a convergent regularization. For one, since we are assuming that D_σ is bounded from below and contractive, we have that $\operatorname{spec}(D_\sigma) \subset (0, 1)$, and the same considerations that led to these assumptions then lead to us asking that $h_\tau(\operatorname{spec}(D_\sigma)) \subset (0, 1)$ for each $\tau > 0$.

Since we would like to think of J_τ as somewhat similar to τJ, we ask the question whether the limit

$$J^*(x) := \lim_{\tau\to 0}\frac{1}{\tau}J_\tau(x) = \lim_{\tau\to 0}\frac{1}{2\tau}\langle x, h_\tau(D_\sigma)^{-1}x\rangle - \frac{1}{2\tau}\|x\|^2$$

exists and is sufficiently well-behaved. Indeed, if this limit is well-defined, a natural result to aim for would be that we have convergence to a J^*-minimizing least-squares solution to the inverse problem with appropriate choices of $\tau \to 0$ as $\delta \to 0$.

Remark 3 In Theorem 5, as above, we will assume that the denoiser is contractive, which by (34) implies that the corresponding regularization functional is coercive. We may be able to relax this assumption, by requiring that the kernel of the forward operator is compatible with the denoiser in the sense that the objective function in (36) is coercive.

Theorem 5 *Suppose that $D_\sigma : X \to X$ is a bounded, linear, self-adjoint operator, which is interpreted as a denoiser. Furthermore, assume that D_σ is positive definite, bounded from below, and contractive (so that* $\mathrm{spec}(D_\sigma) \subset (0,1)$*). Suppose in addition that we have a bounded, linear forward operator $A : X \to Y$ (assuming w.l.o.g. that $\|A\| = 1$), and that $\{h_\tau : \mathbb{R} \to \mathbb{R}\}_{\tau \in (0,\infty)}$ is a collection of continuous scalar functions satisfying*

A.1

$$h_\tau(\mathrm{spec}(D_\sigma)) \subset (0,1) \quad \textit{for any } \tau > 0,$$

A.2

$$r_\tau(\lambda) := \frac{1 - h_\tau(\lambda)}{\tau h_\tau(\lambda)} \quad \textit{converges uniformly for } \lambda \in \textit{ spec }(D_\sigma) \textit{ as } \tau \to 0,$$

with limit r^* and rate $\|r_\tau - r^*\|_{L^\infty(\mathrm{spec}(D_\sigma))} = o(\tau)$,

A.3

$$\underline{c} := \inf_{\tau > 0, \lambda \in \mathrm{spec}(D_\sigma)} r_\tau(\lambda) > 0, \quad \overline{c} := \sup_{\tau > 0, \lambda \in \mathrm{spec}(D_\sigma)} r_\tau(\lambda) < \infty.$$

In this setting, let us define (using the continuous functional calculus to apply scalar functions to D_σ)

$$J_\tau(x) := \frac{1}{2}\langle x, (h_\tau(D_\sigma)^{-1} - \mathrm{id})x\rangle = \frac{\tau}{2}\langle x, r_\tau(D_\sigma)x\rangle.$$

We can compute the solution to the variational problem

$$\hat{x} = \arg\min_{x \in X} \frac{1}{2}\|Ax - y\|^2 + J_\tau(x) \tag{36}$$

using PnP-FBS:

$$\hat{x} = \lim_{k \to \infty} x_k, \qquad \textit{where} \qquad x_{k+1} = h_\tau(D_\sigma)(x_k - A^\top(Ax_k - y)). \tag{37}$$

*By **A.2** we can define $J^*(x) = \lim_{\tau \to 0} J_\tau(x)/\tau$. Now, we obtain a convergent regularization when the regularization parameter τ^δ is chosen appropriately: suppose*

that $\tau^\delta \sim \delta$. Assume that we have an underlying image $x^* \in X$, clean measurements $y = Ax^*$, $\{y^\delta\}_{\delta>0}$ is a sequence in Y satisfying $\|y^\delta - y\| \leq \delta$, and

$$\hat{x}(y^\delta, \tau^\delta) = \operatorname*{arg\,min}_{x \in X} \frac{1}{2}\|Ax - y^\delta\|^2 + J_{\tau^\delta}(x).$$

Then $\hat{x}(y^\delta, \tau^\delta) \to x^\dagger$, where

$$x^\dagger = \operatorname*{arg\,min}_{x \in X \text{ s.t. } A^\top(Ax-y)=0} J^*(x)$$

is the J^*-minimizing least squares solution to the inverse problem $Ax = y$.

Proof First note that under the assumptions of the theorem, PnP-FBS as described in (37) is a contractive fixed-point iteration (so that it has a unique fixed point to which it converges) for any τ and y, with fixed points satisfying the optimality condition of the variational problem in (36).

By ***A.3***, we can define $\underline{J}(x) = \inf_\tau J_\tau(x)/\tau$ and $\overline{J}(x) = \sup_\tau J_\tau(x)/\tau$, so that

$$\frac{\underline{c}}{2}\|x\|^2 \leq \underline{J}(x) \leq \frac{J_\tau(x)}{\tau} \leq \overline{J}(x) \leq \frac{\overline{c}}{2}\|x\|^2. \tag{38}$$

Taking limits, this also gives us that

$$\frac{\underline{c}}{2}\|x\|^2 \leq J^*(x) \leq \frac{\overline{c}}{2}\|x\|^2.$$

The above bounds tell us that the J^*-minimizing least squares solution to the inverse problem is unique, since it is defined by the minimization of a strongly convex functional on a closed linear subspace of X.

We have clean measurements $y = Ax^*$, a set of y^δ such that $\|y - y^\delta\| \leq \delta$ and a parameter choice rule $\delta \mapsto \tau^\delta$ satisfying $\tau^\delta \sim \delta$ as $\tau \to 0$. We are considering the corresponding set of reconstructions

$$\hat{x}(y^\delta, \tau^\delta) = \operatorname*{arg\,min}_{x \in X} \frac{1}{2}\|Ax - y^\delta\|^2 + J_{\tau^\delta}(x).$$

By the remarks above, we can compute these reconstructions using (37). For the sake of the proof, let us also define the variational reconstruction operators with a static regularization functional J^*, as follows

$$\hat{x}_{\text{static}}(y, \tau) := \operatorname*{arg\,min}_{x \in X} \frac{1}{2}\|Ax - y\|^2 + \tau J^*(x). \tag{39}$$

This static regularization approach, with the parameter choice that we are using, is a convergent regularization by the existing theory (this is guaranteed, for example,

by the general result in [42, Proposition 3.32]): $\hat{x}_{\text{static}}(y^\delta, \tau^\delta) \to x^\dagger$ with $x^\dagger$ the J^*-minimizing least squares solution to the inverse problem. Furthermore, the triangle inequality gives us that

$$\|\hat{x}(y^\delta, \tau^\delta) - x^\dagger\| \leq \|\hat{x}(y^\delta, \tau^\delta) - \hat{x}_{\text{static}}(y^\delta, \tau^\delta)\| + \|x_{\text{static}}(y^\delta, \tau^\delta) - x^\dagger\|, \quad (40)$$

so it suffices to show that

$$\|\hat{x}(y^\delta, \tau^\delta) - \hat{x}_{\text{static}}(y^\delta, \tau^\delta)\| \to 0 \quad \text{as } \delta \to 0.$$

We can write

$$\hat{x}(y, \tau) = [A^\top A + \tau r_\tau(D_\sigma)]^{-1} A^\top y$$

and

$$\hat{x}_{\text{static}}(y, \tau) = [A^\top A + \tau r^*(D_\sigma)]^{-1} A^\top y,$$

so we just need to show that $\|M_\tau^{-1} - M_{\tau,\text{static}}^{-1}\| \to 0$ as $\tau \to 0$ (since $\tau^\delta \sim \delta$), where $M_\tau = A^\top A + \tau r_\tau(D)$ and $M_{\tau,\text{static}} = A^\top A + \tau r^*(D)$. We have

$$\begin{aligned} M_\tau^{-1} - M_{\tau,\text{static}}^{-1} &= [M_{\tau,\text{static}} + \tau(r_\tau(D_\sigma) - r^*(D_\sigma))]^{-1} - M_{\tau,\text{static}}^{-1} \\ &= \Big[[\text{id} + \tau M_{\tau,\text{static}}^{-1}(r_\tau(D_\sigma) - r^*(D_\sigma))]^{-1} - \text{id}\Big] M_{\tau,\text{static}}^{-1}. \end{aligned} \quad (41)$$

We will expand the inner matrix inversion using a Neumann series. Note first (by ***A.3***) that $M_{\tau,\text{static}}$ is bounded from below: $\|M_{\tau,\text{static}} x\| \geq \underline{c}\tau\|x\|$. As a result, $\|M_{\tau,\text{static}}^{-1}\| \leq 1/(\underline{c}\tau)$ and we can estimate

$$\begin{aligned} \|\tau M_{\tau,\text{static}}^{-1}(r_\tau(D_\sigma) - r^*(D_\sigma))\| &\leq \tau \frac{1}{\underline{c}\tau} \|r_\tau(D_\sigma) - r^*(D_\sigma)\| \\ &= \frac{\|r_\tau - r^*\|_{L^\infty(\text{spec}(D_\sigma))}}{\underline{c}}. \end{aligned} \quad (42)$$

Since ***A.2*** tells us that $\|r_\tau - r^*\|_{L^\infty(\text{spec}(D_\sigma))} \to 0$ as $\tau \to 0$, this must be smaller than 1 for sufficiently small τ, which is a sufficient condition for absolute convergence of the Neumann series. Using this and (41), we see that

$$\begin{aligned} M_\tau^{-1} - M_{\tau,\text{static}}^{-1} &= \Big[\sum_{n=0}^{\infty} [-\tau M_{\tau,\text{static}}^{-1}(r_\tau(D_\sigma) - r^*(D_\sigma)]^n - \text{id}\Big] M_{\tau,\text{static}}^{-1} \\ &= \Big[\sum_{n=1}^{\infty} [-\tau M_{\tau,\text{static}}^{-1}(r_\tau(D_\sigma) - r^*(D_\sigma))]^n\Big] M_{\tau,\text{static}}^{-1}. \end{aligned}$$

Finally, we can simply estimate its norm from this as follows, using (42) and the fact $\|M_{\tau,\text{static}}^{-1}\| \leq 1/(\underline{c}\tau)$:

$$\begin{aligned}\|M_\tau^{-1} - M_{\tau,\text{static}}^{-1}\| &\leq \sum_{n=1}^{\infty} \left(\frac{\|r_\tau - r^*\|_{L^\infty(\text{spec}(D_\sigma))}}{\underline{c}}\right)^n \frac{1}{\underline{c}\tau} \\ &= \frac{\|r_\tau - r^*\|_{L^\infty(\text{spec}(D_\sigma))}}{\underline{c}} \frac{1}{1 - \frac{\|r_\tau - r^*\|_{L^\infty(\text{spec}(D_\sigma))}}{\underline{c}}} \frac{1}{\underline{c}\tau} \\ &= \frac{1}{\tau} \frac{\|r_\tau - r^*\|_{L^\infty(\text{spec}(D_\sigma))}}{\underline{c}^2 - \underline{c}\|r_\tau - r^*\|_{L^\infty(\text{spec}(D_\sigma))}}.\end{aligned}$$

Since we have assumed that $\|r_\tau - r^*\|_{L^\infty(\text{spec}(D_\sigma))} = o(\tau)$ as $\tau \to 0$, we find by the above reasoning that $\|\hat{x}(y^\delta, \tau^\delta) - \hat{x}_{\text{static}}(y^\delta, \tau^\delta)\| \to 0$. Recalling the inequality in (40) lets us conclude that the spectral filtering approach is a convergent regularization. □

Example 1 Consider the case previously considered in (34) and (35), corresponding to $h_\tau(\lambda) = \lambda/(\tau(1-\lambda)+\lambda)$. We have

$$r_\tau(\lambda) = \frac{1 - h_\tau(\lambda)}{\tau h_\tau(\lambda)} = \frac{1-\lambda}{\lambda}.$$

In particular, the assumptions ***A.3***, ***A.2*** and ***A.1*** are trivially satisfied: we have $h_\tau(\lambda) < \lambda$ for $\lambda > 0$, $r^* = r_\tau$ for all $\tau > 0$ and

$$\inf_{\tau>0, \lambda \in \text{spec}(D_\sigma)} r_\tau(\lambda) = \frac{1 - \lambda_{\max}(D_\sigma)}{\lambda_{\max}(D_\sigma)} > 0$$

and

$$\sup_{\tau>0, \lambda \in \text{spec}(D_\sigma)} r_\tau(\lambda) = \frac{1 - \lambda_{\min}(D_\sigma)}{\lambda_{\min}(D_\sigma)} < \infty.$$

This should come as no surprise, since by the previous discussion, this choice of spectral filtering simply corresponds to the static regularization approach used in the proof of Theorem 5, for which classical theory establishes its convergence properties.

4.3 Experiments

In this section, we will demonstrate the use of the spectral filtering approach to control the regularization strength of a learned denoiser, when applied to an inverse problem using PnP-FBS, showing that it in fact gives rise to a practically convergent regularization method. Since the spectral filtering approach was developed for linear denoisers, we first need to decide on a reasonable design for a linear learnable denoiser.

In this work, we will modify the U-net architecture [39], which continues to be used with great success in image-to-image tasks, and combines a downscaling and upscaling

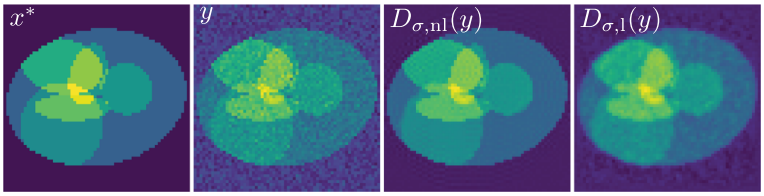


Fig. 4 Comparing the denoising performance of a non-linear U-net ($D_{\sigma,\mathrm{nl}}$) with a linear, symmetric U-net ($D_{\sigma,\mathrm{l}}$) based on the same architecture. Here x^* is a ground truth image and y is the same image, corrupted by Gaussian noise. These images are generated in the same way as the training data was generated. In contrast to $D_{\sigma,\mathrm{nl}}$, $D_{\sigma,\mathrm{l}}$ struggles to reconstruct sharp edges as it does not contain any non-linearity. On the other hand, both denoisers significantly improve the signal-to-noise ratio: y has a PSNR of 24.3 dB, $D_{\sigma,\mathrm{nl}}(y)$ has a PSNR of 34.0 dB and $D_{\sigma,\mathrm{l}}(y)$ has a PSNR of 27.8 dB

path (as in an autoencoder) with skip connections that connect the corresponding scales before and after the bottleneck. The key insight for us is that the U-net architecture is symmetric, in the following sense: if the downscaling and upscaling operations are linear and each other's transposes, and the activation functions and biases are omitted, the U-net is linear and its transpose is a U-net of the same shape (which can be thought of as running the original U-net in reverse). In particular, it is straightforward to see that we can obtain a symmetric linear U-net in this way by tying weights between the downscaling and upscaling paths. Alternatively, and perhaps more simply, we can take the average of a linear U-net and its transpose to get a symmetric linear denoiser. This is the approach that we will take in the experiments considered in this section, since we can leverage the power of JAX [8] to do so: given a linear U-net, we can efficiently compute its vector-Jacobian products to get its transpose. Figure 4 shows a comparison of the denoising performance (in the same setting as the one we will consider for the application to inverse problems below) of such a linear U-net with a comparable non-linear U-net. By this, we mean that the networks have the same sizes and the same number of trainable parameters. While the non-linear U-net allows for better reconstructions, most notably in terms of sharpness, both denoisers remove a significant part of the noise in the noisy images. In what follows, we will use the linear U-net $D_{\sigma,\mathrm{l}}$ and simply call it D_σ.

The experiment that we will consider is concerned with the inverse problem of image reconstruction in computed tomography (CT). We will consider images of size 64×64, consisting of randomly generated ellipse phantoms as in Fig. 6, and simulate CT measurements (sinograms) using the ASTRA toolbox [47, 48] with a parallel beam geometry with 150 equispaced views. A linear U-net is trained as a denoiser on ellipse phantoms corrupted with Gaussian white noise, after which we apply the denoiser in a PnP-FBS manner: denoting the forward operator, which maps images u to (clean) sinograms y by A, the noisy measurements y^δ and the trained denoiser by D_σ, we iterate

$$x_{k+1} = h_\tau(D_\sigma)(x_k - \eta A^\top(Ax_k - y^\delta)), \qquad \hat{x}(y^\delta, \tau) = \lim_{k\to\infty} x_k$$

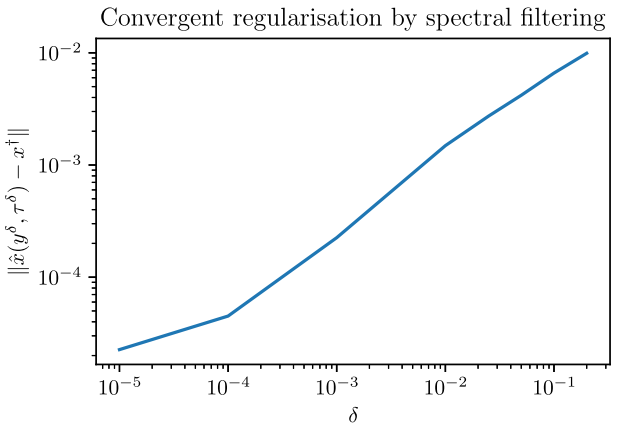


Fig. 5 Applying the spectral filtering approach, we observe convergent regularization in practice. Here $x^\dagger$ is the J^*-minimizing solution to the noiseless least-squares problem as in Theorem 5

where η is a step size, satisfying $\eta \leq 2/\|A\|^2$, so that the limit is well-defined. Here, we simply use the spectral filters h_τ corresponding to scaling the underlying regularization functional as seen in Eq. (35). Considering the size of the images used in this experiment, it is still feasible to implement Eq. (35) by computing a full eigendecomposition of the trained denoiser and applying the filter to the found eigenvalues. An alternative approach that could be used since h_τ is analytic, is to compute a power series expansion. This has the advantage that it requires only repeated applications of the denoiser, i.e. it is not necessary to compute and store an eigenbasis, and the power series can be truncated to give an approximate result.

In order to verify that the proposed spectral filtering provides a convergent regularization we consider a sequence of noisy measurements $\{y^\delta\}_{\delta>0}$ such that $\|y^\delta - y\| \leq \delta$ and a corresponding step size $\tau^\delta \propto \delta$, satisfying the conditions of Theorem 5. The reference $x^\dagger$ is the J^*-minimizing solution that was computed from noiseless measurement data with high accuracy. In Figs. 5 and 6 we show that the spectral filtering approach indeed leads to a numerically verifiable convergent regularization, as predicted by Theorem 5. In comparison to reconstructions obtained with filtered backprojection (FBP) we can observe a successful noise suppression for high noise cases, while the J^*-minimizing solution lacks some sharpness. This suggests the limitations of the linear denoiser in comparison to non-linear networks.

5 Conclusions

The question if a reconstruction algorithm provides a convergent regularization has been long studied in inverse problems, as it provides more than just the knowledge that a solution can be computed at a certain noise level. It tells us that stable solutions exist for all noise realizations and even more importantly that in the limit case, when noise vanishes, we obtain a solution of the underlying operator equation. In other words,

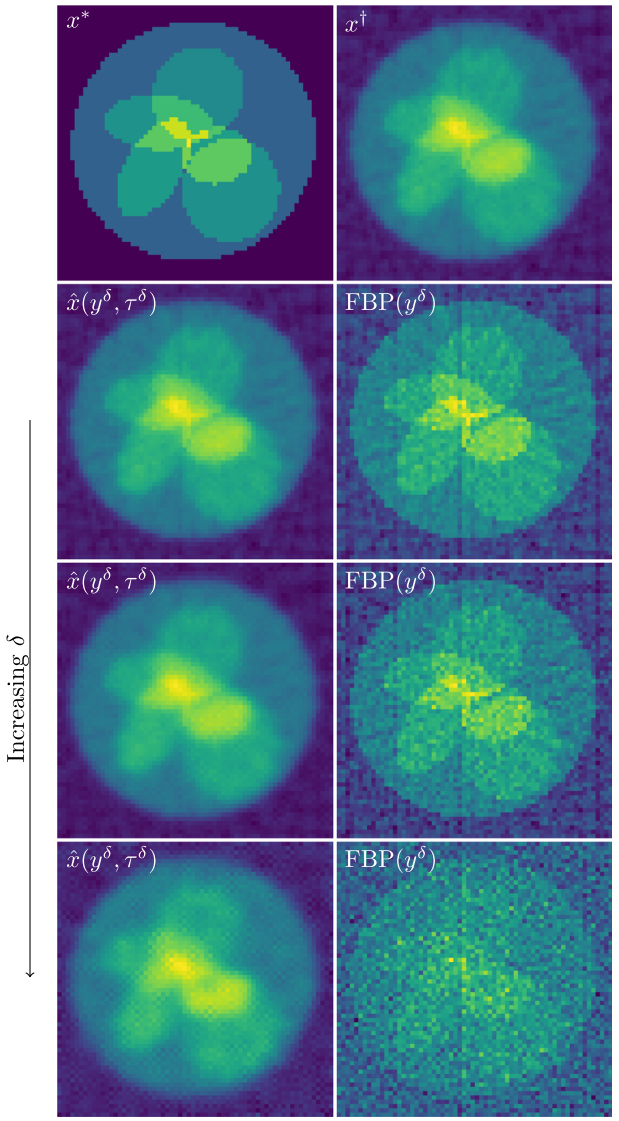


Fig. 6 Applying the spectral filtering approach, we observe convergent regularization in practice. We show a selection of snapshots corresponding to the plot in Fig. 5. Note that x^* (the underlying ground truth) is distinct from $x^\dagger$ since the forward operator has a non-trivial kernel

we can guarantee mathematically that obtained solutions are indeed solutions to the inverse problem.

This is in contrast to some novel data-driven approaches where we may only guarantee that obtained solutions are minimizers of the empirical loss, given suitable training data. Consequently, the concept of convergent data-driven reconstructions has gained considerable interest very recently, see for instance [33]. Here, PnP approaches take a

special role due to their straightforward connection to convex optimization [25] and the possibility to incorporate learned denoisers given by non-linear neural networks. But despite considerable advances in establishing convergence notions, i.e., fixed-point and objective convergence, the question of convergent regularization is still open for general non-linear denoisers.

In this work, we presented a step forward for learned linear denoisers using the novel concept of spectral filtering of the denoiser. The presented approach allows to establish a provably convergent regularization in the PnP framework. Additionally, this convergence is demonstrated numerically on the inverse problem of CT image reconstruction. As established in Theorem 5, there is some freedom in the choice of filters to apply to the denoiser. In future work, this choice could be studied in more detail. In this direction, it is of particular interest to choose spectral filters that are not too computationally costly to evaluate but still give a way to tune the regularization strength of the denoiser. Indeed, in the present implementation of the method, after training, the denoiser is instantiated as a matrix, the eigen-decomposition of which is computed to apply the spectral filtering. By considering spectral filters given by polynomials, for example, we would circumvent the need to instantiate the denoiser as a matrix and compute a full eigen-decomposition. Besides this, it would be of great interest to study whether there is any reasonable generalization of the denoiser filtering approach to the setting in which the denoiser is non-linear.

In fact, we have observed similar convergence behavior numerically even when using a non-linear denoiser in the PnP gradient-step framework (see Fig. 2), suggesting a promising direction for proving that PnP with realistic assumptions on the denoiser can give rise to convergent regularization. The gradient-step framework is, however, just one way of controlling the regularization strength of the learned denoiser. In particular, it relies on flipping the usual splitting of the variational objective and, as a result, requires repeated evaluation of the proximal operator of the data term. This may be very computationally costly if the forward operator is expensive to evaluate. As a result, it is still of great interest to study other ways of controlling the regularization strength of a realistic learned denoiser in PnP that will result in provably convergent regularization.

Acknowledgements CBS acknowledges support from the Philip Leverhulme Prize, the Royal Society Wolfson Fellowship, the EPSRC advanced career fellowship EP/V029428/1, EPSRC Grants EP/S026045/1 and EP/T003553/1, EP/N014588/1, EP/T017961/1, the Wellcome Innovator Awards 215733/Z/19/Z and 221633/Z/20/Z, the European Union Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant agreement No. 777826 NoMADS, the Cantab Capital Institute for the Mathematics of Information and the Alan Turing Institute. FS acknowledges support from the EPSRC advanced career fellowship EP/V029428/1. AH acknowledges support from the Research Council of Finland with the Flagship of Advanced Mathematics for Sensing Imaging and Modelling Proj. 359186. Centre of Excellence of Inverse Modelling and Imaging Proj. 353093, and the Academy Research Fellow Proj. 338408. The authors would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for its support and hospitality during the program on *Mathematics of Deep Learning*, where work on this paper was initiated. This work was supported by EPSRC Grant No. EP/R014604/1. Finally, the authors would like to thank the anonymous reviewers, whose comments have helped improve and clarify the paper.

References

1. Adler, J., Öktem, O.: Learned primal-dual reconstruction. IEEE transactions on medical imaging **37**(6), 1322–1332 (2018)
2. Aharon, M., Elad, M., Bruckstein, A.M.: K-SVD: An algorithm for designing of over-complete dictionaries for sparse representation. IEEE Transactions on Signal Processing **54**(11), 4311–4322 (2006)
3. Allard, W.K., Chen, G., Maggioni, M.: Multi-scale geometric methods for data sets II: Geometric multi-resolution analysis. Applied Computational and Harmonic Analysis **32**(3), 435–462 (2012)
4. Amos, B., Xu, L., Kolter, J.Z.: Input convex neural networks. In: International Conference on Machine Learning, pp. 146–155 (2017)
5. Aspri, A., Banert, S., Öktem, O., Scherzer, O.: A data-driven iteratively regularized landweber iteration. Numerical Functional Analysis and Optimization **41**(10), 1190-1227 (2020)
6. Benning, M., Burger, M.: Modern regularization methods for inverse problems. Acta Numerica **27**, 1–111 (2018)
7. Boink, Y.E., Haltmeier, M., Holman, S., Schwab, J.: Data-consistent neural networks for solving nonlinear inverse problems. Inverse Problems and Imaging **17**(1), 203–229 (2023)
8. Bradbury, J., Frostig, R., Hawkins, P., Johnson, M.J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., Zhang, Q.: JAX: composable transformations of Python+NumPy programs (2018). http://github.com/google/jax
9. Chambolle, A., Holler, M., Pock, T.: A convex variational model for learning convolutional image atoms from incomplete data. Journal of Mathematical Imaging and Vision **62**, 417–444 (2020)
10. Chan, S.H., Wang, X., Elgendy, O.A.: Plug-and-play ADMM for image restoration: Fixed-point convergence and applications. IEEE Transactions on Computational Imaging **3**(1), 84–98 (2017)
11. Chun, I.Y., Zheng, X., Long, Y., Fessler, J.A.: Sparse-view x-ray CT reconstruction using ℓ_1 regularization with learned sparsifying transform. In: 14th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine (Fully3D 2017) (2017)
12. De los Reyes, J.C.: Optimal control of a class of variational inequalities of the second kind. SIAM Journal on Control and Optimization **49**(4), 1629–1658 (2011)
13. De los Reyes, J.C., Schönlieb, C.B., Valkonen, T.: Bilevel parameter learning for higher-order total variation regularisation models. Journal of Mathematical Imaging and Vision **57**(1), 1–25 (2017)
14. De los Reyes, J.C., Villacís, D.: Bilevel optimization methods in imaging. In: Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging: Mathematical Imaging and Vision, pp. 1–34. Springer (2022)
15. Ebner, A., Haltmeier, M.: Plug-and-play image reconstruction is a convergent regularization method. IEEE Transactions on Image Processing **33**, 1476–1486 (2024)
16. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of inverse problems, vol. 375. Springer Science & Business Media (1996)
17. Garcia-Cardona, C., Wohlberg, B.: Convolutional dictionary learning: a comparative review and new algorithms. IEEE Transactions on Computational Imaging **4**(3), 366–381, (2018)
18. Gavaskar, R.G., Athalye, C.D., Chaudhury, K.N.: On Plug-and-Play Regularization Using Linear Denoisers. IEEE Transactions on Image Processing **30**, 4802–4813 (2021)
19. Gribonval, R., Nikolova, M.: A Characterization of Proximity Operators. Journal of Mathematical Imaging and Vision **62**(6), 773–789 (2020)
20. Hintermüller, M., Laurain, A., Löbhard, C., Rautenberg, C.N., Surowiec, T.M.: Elliptic mathematical programs with equilibrium constraints in function space: Optimality conditions and numerical realization. In: Trends in PDE Constrained Optimization, pp. 133–153. Springer International Publishing (2014)
21. Hintermüller, M., Wu, T.: Bilevel optimization for calibrating point spread functions in blind deconvolution. Inverse Problems & Imaging **9**(4), 1139–1169 (2015)

22. Hurault, S., Leclaire, A., Papadakis, N.: Gradient step denoiser for convergent plug-and-play. In: International Conference on Learning Representations (2022)
23. Jin, K.H., McCann, M.T., Froustey, E., Unser, M.: Deep convolutional neural network for inverse problems in imaging. IEEE Transactions on Image Processing **26**(9), 4509–4522 (2017)
24. Kaltenbacher, B., Neubauer, A., Scherzer, O.: Iterative regularization methods for nonlinear ill-posed problems. Radon Series on Computational and Applied Mathematics **6** (2008)
25. Kamilov, U.S., Bouman, C.A., Buzzard, G.T., Wohlberg, B.: Plug-and-play methods for integrating physical and learned models in computational imaging: Theory, algorithms, and applications. IEEE Signal Processing Magazine **40**(1), 85–97 (2023)
26. Kang, E., Min, J., Ye, J.C.: A deep convolutional neural network using directional wavelets for low-dose x-ray CT reconstruction. Medical Physics **44**(10), e360–e375 (2017)
27. Lehmann, E.L., Casella, G.: Theory of Point Estimation, 2nd ed edn. Springer Texts in Statistics. Springer, New York (1998)
28. Li, H., Schwab, J., Antholzer, S., Haltmeier, M.: NETT: solving inverse problems with deep neural networks. Inverse Problems **36**(6), 065005 (2020)
29. Lunz, S., Öktem, O., Schönlieb, C.B.: Adversarial regularizers in inverse problems. In: Advances in Neural Information Processing Systems, pp. 8507–8516 (2018)
30. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. Journal of Machine Learning Research **11**, 19–60 (2010)
31. Moreau, J.J.: Proximité et dualité dans un espace hilbertien. Bulletin de la Société mathématique de France **93**, 273–299 (1965)
32. Mukherjee, S., Dittmer, S., Shumaylov, Z., Lunz, S., Öktem, O., Schönlieb, C.B.: Learned convex regularizers for inverse problems. arXiv preprint arXiv:2008.02839v2 (2021)
33. Mukherjee, S., Hauptmann, A., Öktem, O., Pereyra, M., Schönlieb, C.B.: Learned reconstruction methods with convergence guarantees: A survey of concepts and applications. IEEE Signal Processing Magazine **40**(1), 164–182 (2023)
34. Nair, P., Gavaskar, R.G., Chaudhury, K.N.: Fixed-point and objective convergence of plug-and-play algorithms. IEEE Transactions on Computational Imaging **7**, 337–348 (2021)
35. Natterer, F., Wübbeling, F.: Mathematical methods in image reconstruction. SIAM (2001)
36. Natterer, F.: The mathematics of computerized tomography. Wiley (1986)
37. Reehorst, E.T., Schniter, P.: Regularization by denoising: clarifications and new interpretations. IEEE Transactions on Computational Imaging **5**(1), 52–67 (2019)
38. Romano, Y., Elad, M., Milanfar, P.: The little engine that could: Regularization by denoising (RED). SIAM Journal on Imaging Sciences **10**(4), 1804–1844 (2017)
39. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5–9, 2015, Proceedings, Part III, *Lecture notes in computer science*, vol. 9351, pp. 234–241. Springer (2015)
40. Rubinstein, R., Bruckstein, A.M., Elad., M.: Dictionaries for sparse representation modeling. Proceedings of the IEEE **98**(6), 1045–1057 (2010)
41. Ryu, E., Liu, J., Wang, S., Chen, X., Wang, Z., Yin, W.: Plug-and-play methods provably converge with properly trained denoisers. In: Proceedings of the 36th International Conference on Machine Learning, vol. 97, pp. 5546–5557. PMLR (2019)
42. Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., Lenzen, F.: Variational methods in imaging. Springer (2009)
43. Scherzer, O.: A modified Landweber iteration for solving parameter estimation problems. Applied Mathematics and Optimization **38**(1), 45–68 (1998)
44. Schuster, T.: The method of approximate inverse: theory and applications, vol. 1906. Springer (2007)
45. Schwab, J., Antholzer, S., Haltmeier, M.: Deep null space learning for inverse problems: convergence analysis and rates. Inverse Problems **35**(2), 025008 (2019)
46. Simon, B.: A Comprehensive Course in Analysis. Part 4: Operator Theory. AMS, American Mathematical Society, Providence, Rhode Island (2015)
47. van Aarle, W., Palenstijn, W.J., Cant, J., Janssens, E., Bleichrodt, F., Dabravolski, A., de Beenhouwer, J., Batenburg, K.J., Sijbers, J.: Fast and flexible X-ray tomography using the ASTRA toolbox. Optics Express **24**(22), 25129–25147 (2016)

48. van Aarle, W., Palenstijn, W.J., De Beenhouwer, J., Altantzis, T., Bals, S., Batenburg, K.J., Sijbers, J.: The ASTRA Toolbox: A platform for advanced algorithm development in electron tomography. Ultramicroscopy **157**, 35–47 (2015)
49. Venkatakrishnan, S.V., Bouman, C.A., Wohlberg, B.: Plug-and-play priors for model based reconstruction. In: 2013 IEEE Global Conference on Signal and Information Processing, pp. 945–948 (2013)
50. Xu, X., Liu, J., Sun, Y., Wohlberg, B., Kamilov, U.S.: Boosting the Performance of Plug-and-Play Priors via Denoiser Scaling. In: 2020 54th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, pp. 1305–1312 (2020)
51. Xu, Q., Yu, H., Mou, X., Zhang, L., Hsieh, J., Wang, G.: Low-dose x-ray CT reconstruction via dictionary learning. IEEE Transactions on Medical Imaging **31**(9), 1682–1697 (2012)
52. Zhang, C., Zhang, T., Li, M., Peng, C., Liu, Z., Zheng, J.: Low-dose CT reconstruction via L1 dictionary learning regularization using iteratively reweighted least-squares. BioMedical Engineering OnLine **15**(66), 21 pp. (2016)