Regularization Theory of the Analytic Deep Prior Approach

Clemens Arndt

March 2022

Abstract

The analytic deep prior (ADP) approach was recently introduced for the theoretical analysis of deep image prior (DIP) methods with special network architectures. In this paper, we prove that ADP is in fact equivalent to classical variational Ivanov methods for solving ill-posed inverse problems. Besides, we propose a new variant which incorporates the strategy of early stopping into the ADP model. For both variants, we show how classical regularization properties (existence, stability, convergence) can be obtained under common assumptions.

1 Introduction

In particular the field of image processing (e.g. denoising, deblurring) is a constant source for challenging inverse problems. The restoration of a corrupted image is typically ill-posed, so regularization techniques are needed to obtain a natural looking result. In other words, the restoration method should incorporate some prior knowledge about the appearance of natural images. However, dependent on the application it can be very difficult to give a mathematically exact definition of what natural looking images are. This makes it hard to encode such prior knowledge in a penalty term for classical variational regularization approaches (e.g. TV regularization [30, 8]).

However, deep learning methods with convolutional neural networks have proven to be quite successful in generating and restoring images [22, 16, 35]. One reason for that is the use of appropriate training data, but [25] shows that just the architecture of an untrained

*Center for Industrial Mathematics, University of Bremen, Germany (carndt@uni-bremen.de)
network can already serve as an image prior. The so-called deep image prior (DIP) approach consists in optimizing the weights of a neural network \( \varphi_\theta \) to minimize the loss function

\[
\frac{1}{2} \| A \varphi_\theta(\mathbf{z}) - y^\delta \|^2
\]

for some forward operator \( A \) and noisy data \( y^\delta \) (the network’s input \( \mathbf{z} \) is randomly chosen and kept fixed). Although no training data and no penalty functional is used, DIP produces remarkable results in different image processing tasks, as can be seen in [25]. Even challenging problems like sparse angle computed tomography [3] or compressive sensing [19] can be solved this way.

Developing regularization theory for deep learning methods is of high interest [2]. A very prosperous approach is to combine classical theory with deep learning (e.g. [26, 29]). The number of papers which analyze DIP from a theoretical point of view is also growing. In [18] a functional is constructed, which measures the ability of the neural network \( \varphi_\theta \) to approximate an arbitrary image. This functional can then be used as a penalty term in a classical variational method. The authors of [32] analyze how fast a DIP network approximates the low-frequency and high-frequency components of the target image. By controlling this so-called spectral bias, overfitting is avoided. In [20] the ability to denoise images is attributed to convolutional layers, which are faster in fitting smooth images than noisy ones. The role and the choice of hyperparameters for DIP approaches is described in [34]. A Bayesian perspective is presented in [9], where DIP is interpreted as a Gaussian process.

The choice of architecture is crucial for applications of DIP. Generative neural networks are a natural choice due to their ability to reproduce natural looking images. But the authors of [13] took a LISTA-like network [17] instead to develop the so-called analytic deep prior (ADP) approach. This may not lead to a better practical performance of DIP, but it’s the foundation for an interesting theory. The main aspect consists in interpreting the training of a neural network as the optimization of a Tikhonov functional. There is an analogy to [1], where the penalty term for a Tikhonov functional is optimized. But in contrast to that, the focus of [13] is on the forward operator inside the functional (see section 2).

This work summarizes deeper investigations of the ADP model. The main result (Theorem 3.3) is an equivalence between the ADP approach and classical Ivanov methods [36]. Out of this follows a complete analysis of the regularization properties of ADP including the existence of solutions, stability of reconstructions and convergence towards the ground truth for vanishing noise.

In practical applications of DIP, gradient descent and early stopping is used to minimize
the loss function (1.1). Thus, a global (or at least a local) minimum is not reached in
general. While this fact was not considered in the theoretical derivation of ADP, we propose
a new variant (called ADP-β) which incorporates the effect of early stopping into the model
(section 3.2). We also analyze the regularization properties of this new approach.

In section 4 we compare different numerical ways to compute ADP and DIP (with a
LISTA-like architecture) solutions of simple inverse problems. We find that numerical solu-
tions of both methods are mostly similar to each other, which is important for using the ADP
theory for interpretations of DIP. But there can also be observed some interesting dispari-
ties between the different numerical ways. This illustrates a crucial difference between the
analytical definition of DIP as a minimization problem and the numerical implementation
as a gradient descent iteration.

2 Preliminaries and methods

We consider an inverse problem based on the operator equation

\[ Ax^\dagger = y^\dagger \]  

(2.1)

where we want to recover the unknown ground truth \( x^\dagger \) as good as possible. The data \( y^\dagger \) is
typically not known exactly, but we have only access to noisy data \( y^\delta \).

Assumption 2.1. We make the following assumptions for the inverse problem (2.1).

- Let \( X, Y \) be Hilbert spaces and \( A \in L(X, Y) \).
- There exists \( x^\dagger \in X \) and for a given \( \delta > 0 \), it holds \( \| y^\delta - y^\dagger \| \leq \delta \) for \( y^\delta \in Y \).
- Let \( R : X \to [0, \infty] \) be a convex, coercive and weakly lower semicontinuous functional
  with \( R \not\equiv \infty \).

We recall the definition of Bregman distances, which we will use in Theorem 3.12 for a
convergence result, similar to the ones in \([6, 21]\).

Definition 2.2 (Bregman distance). For a convex functional \( R : X \to [0, \infty] \) with subdif-
ferential \( \partial R \) and \( \tilde{x}, x \in X \), the Bregman distance is defined as the set

\[ D_R(\tilde{x}, x) = \{ R(\tilde{x}) - R(x) - \langle p, \tilde{x} - x \rangle \mid p \in \partial R(x) \}. \]  

(2.2)

\footnote{Code available at \url{https://gitlab.informatik.uni-bremen.de/carndt/analytic_deep_prior}}
The DIP approach (introduced by [25]) for the inverse problem (2.1) consists in solving
\[
\min_\theta \frac{1}{2} \| A\varphi_\theta(z) - y^\delta \|^2
\]  
(2.3)
via a gradient descent w.r.t. the parameters \( \theta \) of a neural network \( \varphi_\theta \), as already described in the introduction. Despite the use of a neural network, DIP is a model-based approach and not data-based. To derive the ADP approach, we have to make two assumptions (see [13] for details).

The first one is choosing \( \varphi_\theta \) to be a LISTA-like network [17], which consists of several layers of the form
\[
x^{l+1} = S_{\alpha\lambda}(x^l - \lambda B^*(Bx^l - y^\delta))
\]  
(2.4)
where \( B = \theta \) is the trainable parameter. Originally, this architecture is inspired by ISTA [12], an algorithm for finding sparse solutions of the inverse problem (2.1), and \( S_{\alpha\lambda} \) is the shrinkage function. More general, we can choose \( S_{\alpha\lambda} \) to be the proximal mapping of a penalty functional \( R \). Then, (2.4) equals a proximal forward-backward splitting algorithm [11, Theorem 3.4] which converges to the solution of the minimization problem
\[
\min_{x \in X} \frac{1}{2} \| Bx - y^\delta \|^2 + \alpha R(x).
\]  
(2.5)

The second assumption is letting the number of layers tend to infinity. This might be difficult in practice (see section 4), but it causes the output \( \varphi_\theta(z) \) of the network to be a solution of (2.5). Therefore the ADP model (introduced by [13]) is defined as
\[
\min_{B \in L(X,Y)} \frac{1}{2} \| Ax(B) - y^\delta \|^2 \quad \text{s.t.} \quad x(B) = \arg \min_{x \in X} \frac{1}{2} \| Bx - y^\delta \|^2 + \alpha R(x).
\]  
(2.6)

While DIP is about optimizing the weights of a neural network, ADP is about optimizing the forward operator in a Tikhonov functional. If we add an additional regularization term for the operator \( B \), we get the (new) ADP-\( \beta \) model
\[
\min_{B \in L(X,Y)} \frac{1}{2} \| Ax(B) - y^\delta \|^2 + \beta \| B - A \|^2 \quad \text{s.t.} \quad x(B) = \arg \min_{x \in X} \frac{1}{2} \| Bx - y^\delta \|^2 + \alpha R(x).
\]  
(2.7)

The reason for this modification will be explained in section 3.2.

To guarantee uniqueness of \( x(B) \), the functional \( R \) should be strictly convex, but this is
not always required. If we assume \( R \) even to be strongly convex, \( x(B) \) depends continuously on \( B \) as the following theorem states. It will be useful for proving existence and stability results for ADP-\( \beta \).

**Theorem 2.3.** Let \( R: X \to [0, \infty] \) be a strongly convex, coercive and weakly lower semicontinuous functional. Then

\[
x(B) = \arg\min_{x \in X} \frac{1}{2} \| Bx - y^\delta \|^2 + \alpha R(x) \quad (2.8)
\]

depends continuously on \( B \in L(X,Y) \).

The proof can be found in the appendix [A.1](#).

---

**3 Theoretical results**

**3.1 Equivalence to classical methods**

DIP solutions of inverse problems are naturally restricted to be the output of a neural network. Analogously, only elements of the set

\[
U_{\alpha R} = \left\{ \hat{x} \in X \mid \exists B \in L(X,Y) : \hat{x} = \arg\min_{x \in X} \frac{1}{2} \| Bx - y^\delta \|^2 + \alpha R(x) \right\} \quad (3.1)
\]

can be solutions of the ADP approach. By definition

\[
\min_{x \in U_{\alpha R}} \frac{1}{2} \| Ax - y^\delta \|^2 \quad (3.2)
\]

is equivalent to the original ADP problem (2.6). To get a better understanding of this minimization problem we investigate \( U_{\alpha R} \). It will turn out that the set \( U_{\alpha R} \) can be characterized in a much easier way, even without using an operator \( B \in L(X,Y) \). For this purpose, we formulate the following lemmas.

**Lemma 3.1.** Let \( R: X \to [0, \infty] \) be a convex, coercive and weakly lower semicontinuous functional and \( \hat{x} \in X, y^\delta \in Y, y^\delta \neq 0 \) and \( \alpha > 0 \) be arbitrary. If there exists \( v \in \partial R(\hat{x}) \) such that

\[
\alpha \langle v, \hat{x} \rangle \leq \frac{\| y^\delta \|^2}{4} \quad (3.3)
\]

holds, then there exists a linear operator \( B \in L(X,Y) \) which fulfills

\[
\hat{x} = \arg\min_{x \in X} \frac{1}{2} \| Bx - y^\delta \|^2 + \alpha R(x) \quad (3.4)
\]
The proof can be found in the appendix A.2.

Lemma 3.2. Let $R: X \to [0, \infty]$ be a convex, coercive and weakly lower semicontinuous functional and $\hat{x} \in X$, $y^\delta \in Y$, $\alpha > 0$ be arbitrary. If for every $v \in \partial R(\hat{x})$

$$\alpha \langle v, \hat{x} \rangle > \frac{{\| y^\delta \|^2}}{4} \quad (3.5)$$

holds, then there exists no linear operator $B \in L(X, Y)$ which fulfills

$$\hat{x} = \arg \min_{x \in X} \frac{1}{2} \| Bx - y^\delta \|^2 + \alpha R(x). \quad (3.6)$$

The proof can be found in the appendix A.3. For given $y^\delta \in Y$, $\hat{x} \in X$ and a penalty term $R$, these lemmas state whether there exists a linear forward operator $B: X \to Y$ such that $\hat{x}$ is the Tikhonov solution w.r.t. $R$ of the inverse problem w.r.t. $y^\delta$. As a consequence, we can write the ADP minimization problem with a much simpler side constraint.

Theorem 3.3. Let Assumption 2.1 hold. Then, for all $y^\delta \in Y$, $y^\delta \neq 0$, $\alpha > 0$ the formulation

$$\min_{x \in X} \frac{1}{2} \| Ax - y^\delta \|^2 \quad (3.7)$$

is equivalent to the ADP-Problem (2.6).

Proof. According to Lemma 3.1 and Lemma 3.2 there exists a linear operator $B \in L(X, Y)$ such that

$$\hat{x} = \arg \min_{x \in X} \frac{1}{2} \| Bx - y^\delta \|^2 + \alpha R(x) \quad (3.8)$$

if and only if $\hat{x}$ fulfills the side constraint of (3.7).

Remark 3.4. For the standard Tikhonov penalty term $R(x) = \frac{1}{2} \| x \|^2$, it holds $\partial R(x) = x$. In this case we get

$$\min_{x \in X} \frac{1}{2} \| Ax - y^\delta \|^2 \quad (3.9)$$

as an equivalent formulation of the ADP problem (2.6). For $r = \| y^\delta \|^2/(4\alpha)$, this equals the
Ivanov regularization method

\[
\min_{x \in X} \frac{1}{2} \|Ax - y^\delta\|^2 \\
\text{s.t. } \|x\|^2 \leq r.
\]

(3.10)

As [36] shows, this method is in fact equivalent to the Tikhonov method

\[
\min_{x \in X} \frac{1}{2} \|Ax - y^\delta\|^2 + \frac{\tilde{\alpha}}{2} \|x\|^2
\]

(3.11)

for some \(\tilde{\alpha}\) dependent on \(y^\delta\) and \(r\). We note that the Tikhonov parameter \(\tilde{\alpha}\) may be equal to zero and in particular it differs from the parameter \(\alpha\) of the ADP problem (see section 3.2).

**Remark 3.5.** There are also cases in which the side constraint of (3.7) defines a non-convex feasible set. Then, the ADP problem is more difficult to solve. We give a simple two-dimensional example with the penalty term \(R: \mathbb{R}^2 \to [0, \infty)\),

\[
R(x_1, x_2) = \begin{cases} 
3 \cdot |x_1 - 5| & \text{for } 3 \cdot |x_1 - 5| \geq |x_2|, \\
|x_2| & \text{for } |x_2| > 3 \cdot |x_1 - 5|.
\end{cases}
\]

(3.12)

This functional has a non-centered minimum at \((5, 0)^T\) and the absolute value of its gradient \(|\partial R(x)|\) is strongly dependent on the direction. Because of these properties, it’s easy to show that the term \(\langle v, x \rangle, v \in \partial R(x)\) in the side constraint of (3.7) is non-convex w.r.t. \(x \in \mathbb{R}^2\).

### 3.2 Parameter choice and early stopping

By construction of the ADP model, we expect it in application to act like DIP. But in the previous section it turned out that ADP behaves in fact equivalent to classical methods like Tikhonov’s. When we apply ADP to an inverse problem, the question arises whether ADP can also deliver something that is “new” and not equivalent to a Tikhonov solution. This section presents, how the model has to be changed to produce ADP solutions that are more similar to DIP solutions. In the same time, we derive a strategy for choosing the parameter \(\alpha\) of the ADP model.

When we compare the ADP method

\[
\min_{B \in L(X,Y)} \frac{1}{2} \|Ax(B) - y^\delta\|^2 \\
\text{s.t. } x(B) = \arg \min_{x \in X} \frac{1}{2} \|Bx - y^\delta\|^2 + \frac{\alpha_{\text{ADP}}}{2} \|x\|^2
\]

(3.13)
to the equivalent (see Remark 3.4) Tikhonov method

$$\min_{x \in X} \frac{1}{2}\|Ax - y^\delta\|^2 + \frac{\tilde{\alpha}}{2}\|x\|^2,$$

(3.14)

we have to make sure not to confuse the parameters \(\alpha_{\text{ADP}}\) and \(\tilde{\alpha}\) of both models with each other. At first we state the following relation between these parameters.

**Lemma 3.6.** If the solutions of (3.13) and (3.14) coincide, \(\tilde{\alpha} \leq \alpha_{\text{ADP}}\) holds. Equality of the parameters could only occur if \(y^\delta\) was in the kernel of \(A^*\) or a singular vector of \(A\).

The proof can be found in the appendix A.4. In general, we can assume that \(\tilde{\alpha} < \alpha_{\text{ADP}}\) holds. So in any application it makes sense to choose the ADP parameter greater than one would choose the parameter of a Tikhonov model. But independent of the parameter choice, the ADP solution will always be equivalent to a Tikhonov solution (Remark 3.4). To make ADP more similar to DIP, we apply early stopping [15, section 7.8]. This strategy is often used in the application of DIP but wasn’t considered for the ADP model yet.

For a given inverse problem, we could solve the ADP problem (3.13) with a gradient descent algorithm w.r.t. the operator \(B\) (see section 4 for details) and terminate this iteration early. Taking \(B^0 = A\) as initial value leads by definition to \(x(B^0)\) being equal to the Tikhonov solution w.r.t. the parameter \(\alpha_{\text{ADP}}\). We assume the iteration to converge successfully towards the minimizer \(\hat{x}\) of (3.13). Since \(\hat{x}\) is also the minimizer of (3.14), the limit of the iteration is also a Tikhonov solution but w.r.t. the parameter \(\tilde{\alpha}\). Because of \(\tilde{\alpha} < \alpha_{\text{ADP}}\), the starting solution \(x(B^0)\) is a stronger regularized Tikhonov solution than the limit \(\hat{x}\) of the iteration (see figure 1).

If we apply early stopping, we take some \(x(B^k)\) in between, which in general does not equal a Tikhonov solution w.r.t. \(A\) (see figure 1). This strategy makes sense if we expect \(x(B^k)\) to be a better solution than (the Tikhonov solutions) \(x(B^0)\) and \(\hat{x}\). That could be the case if \(x(B^0)\) is slightly over-regularized and \(\hat{x}\) is slightly under-regularized. Because then, the optimal regularization would lay in between.

Now, we come back to the parameter choice. If we have a criterion for estimating a suitable Tikhonov parameter \(\alpha_{\text{Tik}}\) for a given inverse problem, we should try to choose \(\alpha_{\text{ADP}}\) in a way that

$$\tilde{\alpha} < \alpha_{\text{Tik}} < \alpha_{\text{ADP}}$$

(3.15)

holds. Because then, \(x(B^0)\) will be slightly over-regularized and \(\hat{x}\) slightly under-regularized, as proposed.

In the example of figure 1, we see that the ADP solution \(x(B^k)\), obtained with early stopping, is a better approximation for the ground truth \(x^\dagger\) than the most accurate Tikhonov
Figure 1: Comparison of ADP solutions during gradient descent to the Tikhonov method (orange: at the start of the gradient descent, red: by early stopping, gray: limit of the gradient descent). The forward operator is an integration like in \[4.1\] and the data \(y^\delta\) is contaminated with gaussian noise (PSNR=40). The Tikhonov parameter and the stopping criterion for ADP were chosen a posteriori to achieve the most accurate reconstructions.

solution, which corresponds to \(\alpha_{\text{Tik}}\). But this result is strongly dependent on the particular inverse problem. The Tikhonov method is optimal for data that is normally distributed. If the given distribution differs from that, it is theoretically possible that ADP with early stopping produces a better solution than the Tikhonov method.

Finally, we want to include the early stopping strategy directly into the ADP model to be able to investigate its effect on the regularization of inverse problems. Early stopping enforces the iterated variable to stay close to the initial value. Because of \(B^0 = A\), we can expect \(\|B^k - A\|\) to be small for small \(k\). This leads to using \(\|B^k - A\|\) as an additional penalty term in the ADP problem, which has a similar effect as early stopping \cite[section 2.3]{1}, \cite[section 4]{33}. What comes out is the ADP-\(\beta\) model

\[
\min_{B \in L(\mathcal{X}, \mathcal{Y})} \frac{1}{2}\|Ax(B) - y^\delta\|^2 + \beta\|B - A\|^2
\]

s.t. \(x(B) = \arg\min_{x \in \mathcal{X}} \frac{1}{2}\|Bx - y^\delta\|^2 + \alpha R(x)\).  

\[3.16\]
3.3 Properties of ADP

The equivalence between ADP and the Ivanov method (with general convex penalty term $R$), shown in section 3.1, allows to obtain some regularization properties (existence, stability, convergence) for ADP. We suppose that Assumption 2.1 holds. Besides the functional

$$\tilde{R}(x) = \min_{v \in \partial R(x)} \langle v, x \rangle$$

(3.17)

is assumed to be well-defined. Because then, the side constraint of (3.7) can be formulated as

$$\tilde{R}(x) \leq \frac{\|y^\delta\|^2}{4\alpha}.$$  

(3.18)

Due to $\langle v, x \rangle = R(x) + R^*(v)$ for $v \in \partial R(x)$, where $R^*$ denotes the convex conjugated functional, coercivity of $R$ implies coercivity of $\tilde{R}$.

**Remark 3.7** (Existence). There exists a solution of the ADP problem (2.6) if the functional $\tilde{R}$, defined in (3.17), is weakly lower semicontinuous. This follows from the equivalence theorem 3.3 and [37, Theorem 2.1] about the existence of Ivanov solutions.

Uniqueness of solutions and stability w.r.t. the data $y^\delta$ is less trivial. First, the right hand side of the side constraint (3.18) is dependent on $y^\delta$, which isn’t the case for ordinary Ivanov problems. Secondly, we know from Remark 3.5 that the constraint (3.18) does not always define a convex feasible set. Nevertheless, for the special case $R(x) = \frac{1}{2}\|x\|^2$ we can obtain a convenient stability result. In this case, $\tilde{R}(x) = \|x\|^2$ is a strictly convex functional. Additionally, if the given inverse problem is ill-posed, we can assume the ADP solutions to fulfill the constraint (3.18) with equality. Under these conditions, the following theorem provides stability of ADP.

**Theorem 3.8** (Stability). For $R(x) = \frac{1}{2}\|x\|^2$, let $(y_k) \subset Y$ be a sequence with $y_k \rightarrow \hat{y} \in Y$ and assume that the corresponding ADP solutions $x_k, \hat{x}$ are unique and fulfill the side constraint in (3.9) with equality. Then ADP is stable, which means $x_k \rightarrow \hat{x}$.

The proof can be found in the appendix A.5.

To obtain a convergence result for ADP, it makes sense to use standard convergence theorems, either of the Tikhonov method [21, Theorem 4.4] or of the Ivanov method [23, Theorem 2.5], [31, Theorem 3]. They differ especially in the source conditions they require for the ground truth $x^\dagger$ and in the parameter choice rules. If we assume $\tilde{R}$ to be convex, by [36, Theorem 2] and the equivalence theorem 3.3 the Tikhonov problem

$$\min_{x \in X} \frac{1}{2}\|Ax - y^\delta\|^2 + \tilde{\alpha} \tilde{R}(x)$$

(3.19)
is equivalent to the ADP formulation (3.7) for suitable chosen $\tilde{\alpha} \geq 0$.

**Remark 3.9 (Convergence).** Because of the equivalence between (3.7) and (3.19), the convergence of ADP solutions $x^\alpha_\delta$ to $x^\dagger$ for vanishing $\delta$ w.r.t. the Bregman distance can be directly derived from Tikhonov convergence theorems. But the ADP parameter $\alpha$ does not coincide with the Tikhonov parameter $\tilde{\alpha}$. That’s why, for ADP we do not get an explicit parameter choice rule like $\alpha \sim \delta$. Besides, a source condition for $x^\dagger$ has to be fulfilled by the functional $\tilde{R}$ (defined in (3.17)) and not by the penalty term $R$.

### 3.4 Properties of ADP-β

For proving the existence of solutions of variational regularization schemes, [21, Theorem 3.1] provides a useful framework. If we want to apply this for ADP-β, it has to be ensured that $B \mapsto x(B)$ is weak-weak continuous [21, Assumptions 2.1]. But unfortunately, in general this is not the case.

To obtain convenient regularization properties anyway, we restrict to $X = Y = L^2(\Omega)$ with $\Omega \subset \mathbb{R}^n$. In this setting, we consider a forward operator $A: X \rightarrow Y$ that can be parametrized by a function $f \in L^p(\Omega)$, $p \in [1, \infty)$. More precisely, we take a continuous, bilinear operator $T: L^p(\Omega) \times X \rightarrow Y$ and define

$$Ax = T(f, x).$$

The same parametrization of operators by functions is used in [3]. One typical example would be a convolutional operator $T(f, x) = f \ast x$.

The crucial idea is the additional restriction $f \in W^{1,p}(\Omega)$ to take advantage of the compact embedding of Sobolev spaces $W^{1,p}(\Omega) \subset L^p(\Omega)$. A similar strategy is used in [24] for achieving weak-weak continuity of the forward operator.

We define the parametrized ADP-β approach as

$$\min_{g \in W^{1,p}} \frac{1}{2} \|T(f, x_g) - y^\delta\|^2_{L^2} + \beta \|f - g\|^2_{W^{1,p}}$$

s.t. $x_g = \arg \min_{x \in L^2} \frac{1}{2} \|T(g, x) - y^\delta\|^2_{L^2} + \alpha R(x).$  

(3.21)

In particular, this can be interpreted as a Tikhonov method for solving the nonlinear inverse problem $F(g^\dagger) = y^\dagger$ with the forward operator $F: W^{1,p}(\Omega) \rightarrow Y$, $F(g) = T(f, x_g)$.

**Remark 3.10 (Existence).** The forward operator $F$ is weak-strong continuous if the penalty term $R$ is strongly convex. This holds, because weak convergence $g_k \rightarrow g$ w.r.t. $W^{1,p}(\Omega)$
implies convergence by norm in $L^p(\Omega)$, by Theorem 2.3 the convergence of $x_{g_k} \to x_g$ follows, and the bilinear operator $T$ is continuous. This is more than enough to fulfill the assumptions of [21, Theorem 3.1], which provides the existence of a solution of (3.21).

A weak stability result for the parametrized ADP-\(\beta\) method could be directly obtained from [21, Theorem 3.2]. But this particular framework even allows to prove strong stability.

**Theorem 3.11** (Stability). For $p = 2$, let $R$ be a strongly convex penalty term, $(y_k) \subset Y$ a convergent sequence with $y_k \to \hat{y}$ and $(g_k) \subset W^{1,p}(\Omega)$ the corresponding solutions of the ADP-\(\beta\) problem (3.21). Then, $(g_k)$ has a convergent subsequence and the limit of each subsequence is an ADP-\(\beta\) solution corresponding to $\hat{y}$.

The proof can be found in the appendix A.6.

While proving existence and stability of ADP-\(\beta\)-solutions required a smart parametrization and the use of compact embeddings, a convergence theorem (w.r.t. the Bregman distance) can be proven for the general formulation (2.7). Similar to classical results like [21, Theorem 4.4] or [6, Theorem 2], we need to assume a source condition

$$\exists w \in Y : \quad A^*w \in \partial R(x^\dagger). \quad (3.22)$$

The parameter $\beta$ turns out to be really helpful for obtaining a convergence result.

**Theorem 3.12** (Convergence). Let Assumption 2.1 hold, $x^\dagger$ be an $R$-minimizing solution of (2.1) which fulfills the source condition (3.22) and assume there exist ADP-\(\beta\) solutions $\hat{x}_\alpha^\delta$ of (2.7). If $\alpha$ is chosen proportional to $\delta$, there exists $d \in D_R(\hat{x}_\alpha^\delta, x^\dagger)$ which fulfills $d = O(\delta)$.

The proof can be found in the appendix A.7.

## 4 Numerical Computations

**Aim.** We want to see whether there is a similarity between ADP and DIP also on the numerical side. The ADP approach is based on the idea of using a LISTA network in a DIP method. Usually LISTA architectures contain round about ten layers, but ADP is motivated with a network of infinite depth (see section 2). To derive the ADP model, the output of this infinite network is then replaced by the solution of a minimization problem. So the question arises, whether numerically computed ADP solutions of an inverse problem are yet similar to solutions obtained via a DIP with LISTA architecture.

In this section, we present algorithms for the computation of ADP solutions and we compare them with DIP solutions. In doing so, the focus is not on the performance of
the methods (in comparison to other state-of-the-art reconstruction algorithms) but on the similarity of the different solutions.

**Methods.** From Theorem 3.3, we know that the ADP problem is equivalent to an Ivanov problem. This creates a possibility to compute ADP solutions easily, fast and almost exactly (we call this method ADP Ivanov). In contrast to that, it is more difficult to realize a LISTA architecture with infinite depth. But there are at least two possibilities to simulate such a network.

The first idea (Algorithm 1: DIP LISTA $L=\infty$) is to begin with a network $\varphi_B$ of ten layers and to increase the network depth during the training process of the DIP. This is done implicitly with a simple trick. In each training step, the network’s input is set to be the network’s output of the previous step [13, Appendix 3]. So the original input will pass through more and more layers and in each step the last ten layers are optimized (via backpropagation).

**Algorithm 1: DIP LISTA $L=\infty$**

- **initialize** $B_0$, $z_0$ (e.g. $B_0 = A$, $z_0 = \text{random noise}$);
- **for** $k = 0, 1, \ldots$ **do**
  - $z_{k+1} = \varphi_{B_k}(z_k)$;
  - $\text{loss}_k = \frac{1}{2}\|A\varphi_{B_k}(z_k) - y_\delta\|^2$;
  - $B_{k+1} = \text{update}(\nabla_{B_k}\text{loss}_k)$;
- **end**
- **return** $z_k$

**Algorithm 2: ADP IFT**

- **initialize** $B_0$ (e.g. $B_0 = A$);
- **for** $k = 0, 1, \ldots$ **do**
  - 1. Calculate $x(B_k)$ with fixed point iteration
  - 2. IFT provides: $\nabla_{B_k}x(B_k)$;
  - 3. Update:
    - $\text{loss}_k = \frac{1}{2}\|Ax(B_k) - y_\delta\|^2$;
    - $B_{k+1} = \text{update}(\nabla_{B_k}\text{loss}_k)$;
- **end**
- **return** $x(B_k)$

The second idea (Algorithm 2: ADP IFT) is to compute $x(B)$ from (2.6) with a classical algorithm like ISTA. After that, one can compute the gradient of $x(B)$ w.r.t. $B$ (see the proof of [13, Lemma 4.1]) via the implicit function theorem (IFT). Thus, backpropagation through a big amount of layers is avoided.
For the standard DIP approach, we use a LISTA-like architecture of ten Layers (DIP LISTA $L = 10$) and optimize the weights via backpropagation. So, in total we compare four different methods (ADP Ivanov, ADP IFT, DIP LISTA $L = \infty$, DIP LISTA, $L = 10$). Since solving the Ivanov problem results in the exact ADP solution, we use this as a reference for the other three methods (for which we don’t have convergence guarantees).

In all methods we use the elastic net functional $R(x) = \alpha_1 \|x\|_1 + \alpha_2 \|x\|_2^2$ as a penalty term. So there is one parameter for $\ell^1$-regularization (leads to sparsity) and one parameter for $\ell^2$-regularization (leads to stability and smoothness). In the LISTA-architecture, this is realized by subtracting the gradient of the $\ell^2$-term before applying the activation function.

Setting. We consider two different artificial inverse problems (inversion of the integration operator and a deconvolution) on $L^2(I)$ for an interval $I \subset \mathbb{R}$. The forward operators are

$$\begin{align*}
(A_1 x)(t) &= \int_0^t x(s) \, ds \quad \text{and} \quad A_2 x = g * x,
\end{align*}$$

$g$ being a Gaussian function. Both of them lead to ill-posed inverse problems. We chose three different ground truth functions and created data by applying the forward operators and adding normally distributed random noise. This leads to six examples in total, which is enough for some basic observations. Figure 2 shows the reconstructions corresponding to the integration operator $A_1$. The three rows contain the three different ground truth functions and each column contains a different method. For comparison, the actual ADP solution (ADP Ivanov) and the ground truth is displayed in every plot. Since we are only interested in finding similarities and disparities between the solutions of the different methods, the choice of the regularization parameters plays a minor role. So, we took the same values $\alpha_1$, $\alpha_2$ for each method and simply chose them a posteriori for each example to minimize the $L^2$-error between reconstructions and ground truth. Figure 3 shows the analogous results for the deconvolution problem (forward operator $A_2$).

Observations. From these experiments, we can make the following observations. There is a significant difference between using $L = 10$ or $L = \infty$ layers in a LISTA network. With an infinite number of layers, the reconstructions are looking more realistic. The results of the DIP LISTA $L=\infty$ (Algorithm 1) method and of the IFT method (Algorithm 2) are always looking quite similar. This was expected because both of them simulate an infinitely deep LISTA network. Differences are probably due to the different ways the gradients are computed or due to to slow convergence of the methods.

In most of the cases, the reconstructions of these both methods are looking quite similar to the actual ADP solution. But sometimes they contain artifacts (e.g. the peaks in figure 3, third row). It seems that there are some spots which are hard to reconstruct for the
Figure 2: Computation of ADP and DIP reconstructions via the IFT, via a DIP with LISTA network ($L = \infty$ and $L = 10$) and via the equivalent Ivanov problem. The forward operator is $A_1$ (integration). The given data has PSNR=40 due to additive Gaussian noise. The regularization parameters $\alpha_1$, $\alpha_2$ are chosen for each example (row) separately but are the same for each method (column).

DIP methods and others are rather simple. Besides, the ADP problem (2.6) is not a convex minimization problem w.r.t. $B$. So there is no guarantee for the methods which do gradient descent (DIP LISTA $L = \infty$ and the IFT method) to converge towards the global minimizer. Figure 4 shows that the reconstructions of these methods are indeed dependent on the initial value $B_0$ of the algorithms. In contrast to that, the Ivanov problem from Theorem 3.3 is convex (with the elastic net penalty term $R$). That’s probably why the actual ADP solutions are the only ones which never contain strange artifacts and the only ones that are always quite good reconstructions of the ground truth.

The easiest possibility to slightly improve the reconstruction quality is to apply early...
Stopping. In doing so, the most severe artifacts in the reconstructions can be diminished. This case corresponds to the ADP-\(\beta\) approach (see section 3.2), whose additional convex term \(\beta\|B - A\|^2\) is a numerical advantage because it stabilizes the gradient descent for finding the minimizer. Indeed, adding the gradient of the \(\beta\)-term to the update in Algorithm 2 (ADP IFT) can also diminish the severe artifacts. But we do not include experimental results about this, since the most interesting part is the comparison with the equivalent Ivanov problem, which doesn’t exist for ADP-\(\beta\).

The main conclusion is the numerical verification of the derivation of the ADP problem from the DIP approach. It is possible to use the theoretical analysis of the ADP problem for...
interpretations of the DIP approach because of the similarity between the reconstructions from the different numerical methods. However, the examples from figure 4 illustrate that DIP can be formulated as a minimization problem but a numerical computed solution is not automatically a global minimizer of this problem. If early stopping is used, it is probably not even a local minimizer. Hence, there is a significant difference between the theoretical definition and the practical implementation of DIP.

5 Conclusion

ADP and ADP-β were introduced as methods for solving ill-posed inverse problems in a typical Hilbert space setting (Assumption 2.1). Both of them are motivated by considering DIP with a LISTA-like architecture. The main result is an equivalence of ADP to the classical method of Ivanov regularization.

We have proven existence, stability and convergence results for both ADP and ADP-β. The obtained regularization properties are comparable to the ones of classical methods like Tikhonov’s. In principal, these results can be transferred to DIP with LISTA-like networks. But due to non-convexity of the DIP minimization problem, numerically computed DIP solutions can differ significantly from exact ADP solutions, although they are similar in many cases. We conclude that theoretical analyses of the DIP approach should consider the whole optimization process and not only the properties of the minimizer.

One very important part is the early stopping of the DIP optimization process. In the ADP setting, we incorporated this strategy with an additional penalty term, which resulted in the ADP-β model. The effect of this regularization can be seen by comparing the convergence theorems of ADP and ADP-β. Theorem 3.12 provides a parameter choice rule (α ∼ δ) for ADP-β, which is a big advantage over ADP.

A generalization of the ADP regularization results to DIP with general convolutional
neural networks (CNNs) would be very desirable. The LISTA architecture was suitable because of its similarity to proximal splitting algorithms and the possibility to interpret the output as a solution of a variational problem. Finding similar connections for general CNNs is harder. However in [7], CNNs are used to model proximal mappings and in [28], CNNs are interpreted as algorithms for sparse coding. Besides, [10] asserts that most common activation functions are in fact proximal mappings and they establish a theory for characterizing the fixed point sets of neural networks as solutions of variational inequalities. These directions could provide ideas for possible future extensions.

Acknowledgements

I want to thank Dr. Daniel Otero Baguer, Prof. Peter Maaß, Dr. Tobias Kluth and many more colleagues from the University of Bremen and Dr. Yury Korolev from the University of Cambridge for helpful advice and feedback.

A Proofs of theoretical results

A.1 Theorem 2.3

Continuity of $B \mapsto x(B)$.

Proof. Let $(B_k) \subset L(X,Y)$ be a sequence of operators with $B_k \to B \in L(X,Y)$. At first, we mention that the sequence $(x(B_k))$ is bounded because

$$\alpha R(x(B_k)) \leq \frac{1}{2} \|B_k x(B_k) - y^\delta\|^2 + \alpha R(x(B_k)) \leq \frac{1}{2} \|B_k x(B) - y^\delta\|^2 + \alpha R(x(B)) \quad (A.1)$$

holds. Further, we can estimate

$$\frac{1}{2} \|B x(B_k) - y^\delta\|^2 + \alpha R(x(B_k)) = \frac{1}{2} \|B_k x(B_k) - y^\delta + (B - B_k) x(B_k)\|^2 + \alpha R(x(B_k)) \leq \frac{1}{2} (\|B_k x(B_k) - y^\delta\| + \|(B - B_k) x(B_k)\|)^2 + \alpha R(x(B_k)) \quad (A.2)$$

$$= \frac{1}{2} \|B_k x(B_k) - y^\delta\|^2 + \alpha R(x(B_k)) + \|(B - B_k) x(B_k)\| \left(\|B_k x(B_k) - y^\delta\| + \frac{1}{2} \|(B - B_k) x(B_k)\|\right).$$
Because of the boundedness of \( (x(B_k)) \) and the convergence \( B_k \to B \), the term

\[
\| (B - B_k) x(B_k) \| \cdot \left( \| B_k x(B_k) - y^\delta \| + \frac{1}{2} \| (B - B_k) x(B_k) \| \right)
\]  

(A.3)

converges to zero. For the remaining terms, we can estimate

\[
\frac{1}{2} \| B_k x(B_k) - y^\delta \|^2 + \alpha R(x(B_k)) \leq \frac{1}{2} \| B_k x(B) - y^\delta \|^2 + \alpha R(x(B)),
\]  

(A.4)

and \( \| B_k x(B) - y^\delta \|^2 \) converges to \( \| B x(B) - y^\delta \|^2 \). So \( (x(B_k)) \) is a minimizing sequence of the strongly convex functional \( \frac{1}{2} \| B x - y^\delta \|^2 + \alpha R(x) \) because \( x(B) \) is the minimizer. By \[27\] Theorem 1, the minimizing sequence converges to the minimizer \( x(B) \).

\[\square\]

**A.2 Lemma 3.1**

First part of the equivalence theorem for ADP to Ivanov problems.

**Proof.** Let \( \hat{x}, v, y^\delta, \alpha \) and \( R \) be given according to the assumptions. We have to find a linear operator \( B \) such that

\[- B^*(B \hat{x} - y^\delta) \in \alpha \partial R(\hat{x}). \]  

(A.5)

holds. Because of \( v \in \partial R(\hat{x}) \), we just try to solve the equation

\[- B^*(B \hat{x} - y^\delta) = \alpha v \]  

(A.6)

for \( B \). If \( \hat{x} = 0 \), solving would be trivial. Otherwise, we can decompose \( v \) into

\[ v = \mu \hat{x} + v_\perp \quad \text{s.t.} \quad \langle v_\perp, \hat{x} \rangle = 0. \]  

(A.7)

Accordingly it is \( \mu = \langle v, \hat{x} \rangle / \| \hat{x} \|^2 \). With that, we can write the equation from above as

\[ B^* B \hat{x} + \alpha \mu \hat{x} + \alpha v_\perp = B^* y^\delta. \]  

(A.8)

We consider a linear operator \( B : X \to Y \) of the form

\[ Bx = (\sigma_1 \langle x, \hat{x} \rangle + \sigma_2 \langle x, v_\perp \rangle) \cdot y^\delta \]  

(A.9)

with two coefficients \( \sigma_1 \) and \( \sigma_2 \) to be determined later. Then, the adjoint operator is given by

\[ B^* y = \langle y, y^\delta \rangle (\sigma_1 \hat{x} + \sigma_2 v_\perp) \]  

(A.10)
and it holds

\[ B^* B \hat{x} = B^*( (\sigma_1 \|\hat{x}\|^2, y^\delta) = \sigma_1^2 \|\hat{x}\|^2 \|y^\delta\|^2 \hat{x} + \sigma_1 \sigma_2 \|\hat{x}\|^2 \|y^\delta\|^2 v_\perp, \]  

(A.11)

\[ B^* y^\delta = \sigma_1 \|y^\delta\|^2 \hat{x} + \sigma_2 \|y^\delta\|^2 v_\perp. \]  

(A.12)

To fulfill (A.8), we have to solve

\[ \sigma_1 \sigma_2 \|\hat{x}\|^2 \|y^\delta\|^2 \hat{x} + \alpha \mu \hat{x} + \alpha v_\perp = \sigma_1 \|y^\delta\|^2 \hat{x} + \sigma_2 \|y^\delta\|^2 v_\perp. \]  

(A.13)

Because \( \hat{x} \) and \( v_\perp \) are orthogonal to each other, we get the two equations

\[ \sigma_1 \sigma_2 \|\hat{x}\|^2 \|y^\delta\|^2 \hat{x} + \alpha \mu \hat{x} + \alpha v_\perp = \sigma_1 \|y^\delta\|^2 \hat{x} + \sigma_2 \|y^\delta\|^2 v_\perp. \]  

\[ \sigma_1 \|\hat{x}\|^2 \|y^\delta\|^2 \hat{x} + \alpha \mu \hat{x} + \alpha v_\perp = \sigma_1 \|y^\delta\|^2 \hat{x} + \sigma_2 \|y^\delta\|^2 v_\perp. \]  

(A.14)

(A.15)

Notice that (A.15) and the coefficient \( \sigma_2 \) could be ignored if \( v_\perp = 0 \) held.

Equation (A.14) can be solved for \( \sigma_1 \) with a quadratic formula, which leads to

\[ \sigma_1 = \frac{1}{2 \|\hat{x}\|^2} \pm \sqrt{\frac{\alpha \mu}{4 \|\hat{x}\|^4} - \frac{\alpha \mu}{\|\hat{x}\|^2 \|y^\delta\|^2}}. \]  

(A.16)

Accordingly,

\[ \frac{\alpha \mu}{\|y^\delta\|^2} \lesssim \frac{1}{4 \|\hat{x}\|^2} \]  

(A.17)

must hold to get real solutions. We know from above that \( \mu = \langle v, \hat{x} \rangle / \|\hat{x}\|^2 \). If we insert this, we will see that this matches exactly the assumptions of the lemma.

Now, equation (A.15) has to be solved for \( \sigma_2 \). Excluding \( \sigma_2 \) leads to

\[ \sigma_2 (\sigma_1 \|\hat{x}\|^2 \|y^\delta\|^2 - \|y^\delta\|^2) + \alpha = 0. \]  

(A.18)

If the term inside of the parenthesis doesn’t equal zero, there will exist a solution \( \sigma_2 \). If the term equaled zero, equation (A.14) would lead to \( \mu = 0 \). But in this case, we could choose \( \sigma_1 = 0 \) (the quadratic formula allows two solutions), and then it is no problem to find a solution for \( \sigma_2 \), too.

By finding solutions for \( \sigma_1 \) and \( \sigma_2 \), we showed that the operator \( B \) defined in (A.9) solves equation (A.6). So the lemma is proved.
A.3 Lemma 3.2

Second part of equivalence theorem for ADP to Ivanov problems.

**Proof.** Let $\hat{x}, y^\delta, \alpha$ and $R$ be given according to the assumptions. Assume there exists a linear operator $B$ such that

$$0 \in B^*(B\hat{x} - y^\delta) + \alpha \partial R(\hat{x}) \quad (A.19)$$

holds. It follows

$$v := -\frac{1}{\alpha} B^*(B\hat{x} - y^\delta) \in \partial R(\hat{x}). \quad (A.20)$$

We can calculate $\alpha \langle v, \hat{x} \rangle = -\|B\hat{x}\|^2 + \langle y^\delta, B\hat{x} \rangle$. So according to the assumptions,

$$-\|B\hat{x}\|^2 + \langle y^\delta, B\hat{x} \rangle > \frac{\|y^\delta\|^2}{4} \quad (A.21)$$

must hold. But with Young’s inequality, we get

$$-\|B\hat{x}\|^2 + \langle y^\delta, B\hat{x} \rangle \leq -\|B\hat{x}\|^2 + \frac{1}{4}\|y^\delta\|^2 + \|B\hat{x}\|^2 = \frac{\|y^\delta\|^2}{4}. \quad (A.22)$$

Obviously, this is a contradiction. That’s why such an operator $B$ can’t exist. \qed

A.4 Lemma 3.6

Relation between the ADP parameter and the Tikhonov parameter of the equivalent problem.

**Proof.** Let $\hat{x}$ be the solution of the ADP problem (3.13). Because of the equivalence to the Tikhonov method, $\hat{x}$ is the solution of (3.14) in the same time. Besides, $x(A)$ is the Tikhonov solution w.r.t. the parameter $\alpha_{ADP}$ of the inverse problem. Because of the minimizing properties of $\hat{x}$ and $x(A)$,

$$\frac{1}{2}\|A\hat{x} - y^\delta\|^2 \leq \frac{1}{2}\|Ax(A) - y^\delta\|^2, \quad (A.23)$$

$$\frac{1}{2}\|Ax(A) - y^\delta\|^2 + \frac{\alpha_{ADP}}{2}\|x(A)\|^2 \leq \frac{1}{2}\|A\hat{x} - y^\delta\|^2 + \frac{\alpha_{ADP}}{2}\|\hat{x}\|^2 \quad (A.24)$$

holds. It follows $\|x(A)\|^2 \leq \|\hat{x}\|^2$. Both $x(A)$ and $\hat{x}$ are Tikhonov solutions of the same problem (only with different parameters). So $\tilde{\alpha} \leq \alpha_{ADP}$ must hold because the norm of $\hat{x}$ is greater (or equal) than the norm of $x(A)$.
Now, we assume $\tilde{\alpha} = \alpha_{\text{ADP}} > 0$. By Remark 3.4, the problems
\[
\begin{align*}
\min_{x \in X} & \frac{1}{2} \|Ax - y^\delta\|^2 + \frac{\alpha_{\text{ADP}}}{2} \|x\|^2, \\
\min_{x \in X} & \frac{1}{2} \|Ax - y^\delta\|^2 \text{ s.t. } \|x\|^2 \leq \frac{\|y^\delta\|^2}{4\alpha_{\text{ADP}}}
\end{align*}
\] (A.25)
(A.26)
are equivalent. The solution $\hat{x}$ fulfills
\[
(A^*A + \alpha_{\text{ADP}} \cdot \text{Id})\hat{x} = A^*y^\delta
\] (A.27)
and we assume $A^*y^\delta \neq 0$. Then, $\hat{x}$ must fulfill the side constraint of (A.26) with equality, otherwise $A^*A\hat{x} = A^*y^\delta$ would hold, which is a contradiction. Accordingly we get $\alpha_{\text{ADP}}\|\hat{x}\|^2 = \|y^\delta\|^2/4$ and by computing the inner product of (A.27) with $\hat{x}$, it follows
\[
\|A\hat{x}\|^2 + \frac{\|y^\delta\|^2}{4} = \|A\hat{x}\|^2 + \alpha_{\text{ADP}}\|\hat{x}\|^2 = \langle A\hat{x}, y^\delta \rangle.
\] (A.28)
If we then apply the Cauchy-Schwarz and Young’s inequality, we get
\[
\langle A\hat{x}, y^\delta \rangle \leq \|A\hat{x}\| \cdot \|y^\delta\| \leq \|A\hat{x}\|^2 + \frac{\|y^\delta\|^2}{4},
\] (A.29)
which means these inequalities must in fact hold as equalities. Therefore, $A\hat{x}$ and $y^\delta$ must be linear dependent (Cauchy-Schwarz) and $2\|A\hat{x}\| = \|y^\delta\|$ must hold (Young). It follows
\[
A\hat{x} = \frac{1}{2} y^\delta.
\] (A.30)
We can plug this into (A.27) and get
\[
\alpha_{\text{ADP}}\hat{x} = \frac{1}{2} A^* y^\delta.
\] (A.31)
Accordingly $AA^*y^\delta = \alpha_{\text{ADP}}y^\delta$ holds, so $y^\delta$ is a singular vector of $A$. \hfill \Box

### A.5 Theorem 3.8

Stability of the ADP approach.

**Proof.** We follow some of the ideas of the proofs of [14, Theorem 2.1] and [31, Theorem 2].

Let $x_k$ and $\hat{x}$ be unique solutions of (3.9) for $y^\delta = y_k, \hat{y}$ with $y_k \rightarrow \hat{y}$. The sequence $(x_k)$ is bounded, so there exists a weakly convergent subsequence $(x_{k_i})$, $x_{k_i} \rightarrow x_\infty$. For arbitrary
\[ \varepsilon > 0 \text{ and } x \in X \text{ with } \|x\|^2 \leq \|\hat{y}\|^2 \cdot (4\alpha)^{-1} - \varepsilon, \text{ it holds} \]

\[
\|Ax - \hat{y}\| \leq \lim\inf_{l \to \infty} \|Ax_{k_l} - y_{k_l}\| \leq \lim_{l \to \infty} \|Ax - y_{k_l}\| = \|Ax - \hat{y}\| \quad (A.32)
\]

because \(x_{k_l}\) minimizes the ADP problem w.r.t. \(y_{k_l}\) and \(x\) fulfills the side constraint for \(l\) big enough. With \(\varepsilon \to 0\) and because of the uniqueness of the solutions, we obtain \(x_\infty = \hat{x}\).

Arguing with a subsequence of a subsequence leads to the weak convergence \(x_{k_l} \rightharpoonup \hat{x}\) of the whole sequence.

According to the assumptions, it holds \(\|x_{k_l}\|^2 = \|\hat{y}_{k_l}\|^2 \cdot (4\alpha)^{-1}\). So \(y_{k_l} \to \hat{y}\) implies \(\|x_{k_l}\| \to \|\hat{x}\|\) and together with the weak convergence, we finally obtain \(x_{k_l} \to \hat{x}\). \(\square\)

### A.6 Theorem 3.11

Stability of the ADP-\(\beta\) approach.

**Proof.** First, we note that the sequence \((g_k)\) is bounded in \(W^{1,2}(\Omega)\). Hence, there exists at least one weakly convergent subsequence. For any subsequence with \(g_{k_l} \rightharpoonup \hat{g}\), it holds

\[
T(f, x_{g_{k_l}}) - y_{k_l} \to T(f, x_{\hat{g}}) - \hat{y} \quad (A.33)
\]

because of the arguments from Remark 3.10. For arbitrary \(g \in W^{1,2}(\Omega)\),

\[
\frac{1}{2}\|T(f, x_{\hat{g}}) - \hat{y}\|_{L^2}^2 + \beta\|\hat{g} - f\|_{W^{1,2}}^2 \\
\leq \lim\inf_{k \to \infty} \frac{1}{2}\|T(f, x_{g_{k_l}}) - y_{k_l}\|_{L^2}^2 + \beta\|g_{k_l} - f\|_{W^{1,2}}^2 \\
\leq \lim_{k \to \infty} \frac{1}{2}\|T(f, x_{\hat{g}}) - y_{k_l}\|_{L^2}^2 + \beta\|g_{k_l} - f\|_{W^{1,2}}^2 \\
= \frac{1}{2}\|T(f, x_{\hat{g}}) - \hat{y}\|_{L^2}^2 + \beta\|\hat{g} - f\|_{W^{1,2}}^2
\]

(A.34)

holds because of the minimizing property of \(g_{k_l}\) w.r.t. \(y_{k_l}\). Hence, \(\hat{g}\) is a minimizer of (3.21) w.r.t. \(\hat{y}\). If we choose \(g = \hat{g}\), the first and the last line in (A.34) coincide, and we get

\[
\lim_{k \to \infty} \frac{1}{2}\|T(f, x_{g_{k_l}}) - y_{k_l}\|_{L^2}^2 + \beta\|g_{k_l} - f\|_{W^{1,2}}^2 = \frac{1}{2}\|T(f, x_{\hat{g}}) - \hat{y}\|_{L^2}^2 + \beta\|\hat{g} - f\|_{W^{1,2}}^2. \quad (A.35)
\]

If follows \(\lim_{k \to \infty} \|g_{k_l} - f\|_{W^{1,2}}^2 = \|\hat{g} - f\|_{W^{1,2}}^2\). Hence, \((g_{k_l})\) converges by norm to \(\hat{g}\). \(\square\)

### A.7 Theorem 3.12

Convergence of the ADP-\(\beta\) approach.
Proof. According to (3.22), we can choose $d = R(\hat{x}_\delta) - R(x^\dagger) - \langle A^* w, \hat{x}_\delta - x^\dagger \rangle$ and there exists an operator $\hat{B} \in L(X, Y)$ that fulfills $\hat{x}_\delta = x(\hat{B})$.

Because of the minimizing property of $\hat{x}_\delta,$

$$\alpha R(\hat{x}_\delta) \leq \frac{1}{2}\|\hat{B}\hat{x}_\delta - y^\delta\|^2 + \alpha R(\hat{x}_\delta) \leq \frac{1}{2}\|\hat{B}x^\dagger - y^\delta\|^2 + \alpha R(x^\dagger). \quad (A.36)$$

holds. If follows

$$d = R(\hat{x}_\delta) - R(x^\dagger) - \langle A^* w, \hat{x}_\delta - x^\dagger \rangle \leq \frac{1}{2\alpha}\|\hat{B}x^\dagger - y^\delta\|^2 - \langle w, A\hat{x}_\delta - y^\dagger \rangle$$

$$\leq \frac{1}{2\alpha}\left(\|\hat{B}x^\dagger - Ax\| + \|y^\dagger - y^\delta\|\right)^2 + \|w\|\|A\hat{x}_\delta - y^\dagger\| \quad (A.37)$$

$$\leq \frac{1}{2\alpha}\left(\|x^\dagger\|\|\hat{B} - A\| + \delta\right)^2 + \|w\|\|Ax^\delta - y^\dagger\|.$$

We will show $\|\hat{B} - A\| = O(\delta)$ and $\|A\hat{x}_\delta - y^\dagger\| = O(\delta)$ to deduce $d = O(\delta)$ for $\alpha$ chosen proportional to $\delta$. Because of the minimizing property of $\hat{B}$, we get

$$\beta \cdot \|\hat{B} - A\|^2 \leq \frac{1}{2}\|Ax(\hat{B}) - y^\delta\|^2 + \beta \cdot \|\hat{B} - A\|^2 \leq \frac{1}{2}\|Ax(A) - y^\delta\|^2. \quad (A.38)$$

From standard convergence results of the Tikhonov method [21, Theorem 4.4] or [6, Theorem 2], we get $\|Ax(A) - y^\delta\| = O(\delta)$. So $\|\hat{B} - A\| = O(\delta)$ holds.

Besides,

$$\|A\hat{x}_\delta - y^\dagger\| \leq \|Ax(\hat{B}) - y^\delta\| + \|y^\delta - y^\dagger\| \leq \|Ax(A) - y^\delta\| + \delta \quad (A.39)$$

holds and we can use $\|Ax(A) - y^\delta\| = O(\delta)$ again. So $d = O(\delta)$ follows. \hfill \Box

References

[1] G. S. Alberti, E. De Vito, M. Lassas, L. Ratti, and M. Santacesaria. Learning the optimal Tikhonov regularizer for inverse problems. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, Advances in Neural Information Processing Systems, 2021.

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

[3] D. O. Baguer, J. Leuschner, and M. Schmidt. Computed tomography reconstruction using deep image prior and learned reconstruction methods. Inverse Problems, 36(9):094004, 2020.
[4] C. Bishop. *Regularization and complexity control in feed-forward networks*, pages 141–148. EC2 et Cie, 1995. International Conference on Artificial Neural Networks ICANN’95.

[5] I. Bleyer and R. Ramlau. A double regularization approach for inverse problems with noisy data and inexact operator. *Inverse Problems*, 29:025004, 2013.

[6] M. Burger and S. Osher. Convergence rates of convex variational regularization. *Inverse Problems*, 20(5):1411–1421, 2004.

[7] E. Celledoni, M. J. Ehrhardt, C. Etmann, B. Owren, C.-B. Schönlieb, and F. Sherry. Equivariant neural networks for inverse problems. *Inverse Problems*, 37(8):085006, 2021.

[8] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40:120–145, 2011.

[9] Z. Cheng, M. Gadelha, S. Maji, and D. Sheldon. A Bayesian Perspective on the Deep Image Prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[10] P. L. Combettes and J.-C. Pesquet. Deep neural network structures solving variational inequalities. *Set-Valued and Variational Analysis*, 28:491–518, 2020.

[11] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

[12] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.

[13] S. Dittmer, T. Kluth, P. Maass, and D. Otero Baguer. Regularization by Architecture: A Deep Prior Approach for Inverse Problems. *Journal of Mathematical Imaging and Vision*, 62:456–470, 2020.

[14] H. W. Engl, K. Kunisch, and A. Neubauer. Convergence rates for Tikhonov regularisation of non-linear ill-posed problems. *Inverse Problems*, 5(4):523–540, 1989.

[15] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. [http://www.deeplearningbook.org](http://www.deeplearningbook.org)
[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[17] K. Gregor and Y. LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML’10, page 399–406, Madison, WI, USA, 2010. Omnipress.

[18] A. Habring and M. Holler. A generative variational model for inverse problems in imaging. *SIAM Journal on Mathematics of Data Science*, 4(1):306–335, 2022.

[19] R. Heckel and M. Soltanolkotabi. Compressive sensing with un-trained neural networks: Gradient descent finds a smooth approximation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4149–4158. PMLR, 2020.

[20] R. Heckel and M. Soltanolkotabi. Denoising and regularization via exploiting the structural bias of convolutional generators. In *International Conference on Learning Representations*, 2020.

[21] B. Hofmann, B. Kaltenbacher, C. Pöschl, and O. Scherzer. A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. *Inverse Problems*, 23(3):987–1010, 2007.

[22] V. Jain and S. Seung. Natural image denoising with convolutional networks. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2009.

[23] B. Kaltenbacher and A. Klassen. On convergence and convergence rates for Ivanov and Morozov regularization and application to some parameter identification problems in elliptic PDEs. *Inverse Problems*, 34(5):055008, 2018.

[24] T. Kluth, C. Bathke, M. Jiang, and P. Maass. Joint super-resolution image reconstruction and parameter identification in imaging operator: analysis of bilinear operator equations, numerical solution, and application to magnetic particle imaging. *Inverse Problems*, 36(12):124006, 2020.

[25] V. Lempitsky, A. Vedaldi, and D. Ulyanov. Deep Image Prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018.
[26] H. Li, J. Schwab, S. Antholzer, and M. Haltmeier. NETT: solving inverse problems with deep neural networks. Inverse Problems, 36(6):065005, jun 2020.

[27] C. G. Looney. Convergence of minimizing sequences. Journal of Mathematical Analysis and Applications, 61(3):835–840, 1977.

[28] V. Papyan, Y. Romano, J. Sulam, and M. Elad. Theoretical foundations of deep learning via sparse representations: A multilayer sparse model and its connection to convolutional neural networks. IEEE Signal Processing Magazine, 35(4):72–89, 2018.

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

[30] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena, 60(1):259–268, 1992.

[31] T. I. Seidman and C. R. Vogel. Well posedness and convergence of some regularisation methods for non-linear ill posed problems. Inverse Problems, 5(2):227–238, 1989.

[32] Z. Shi, P. Mettes, S. Maji, and C. G. M. Snoek. On Measuring and Controlling the Spectral Bias of the Deep Image Prior. International Journal of Computer Vision, 2022.

[33] J. Sjoberg and L. Ljung. Overtraining, regularization, and searching for minimum with application to neural networks. International Journal of Control, 62, 1994.

[34] Y. Sun, H. Zhao, and J. Scarlett. On architecture selection for linear inverse problems with untrained neural networks. Entropy, 23:1481, 2021.

[35] Y. Tai, J. Yang, and X. Liu. Image super-resolution via deep recursive residual network. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2790–2798, 2017.

[36] V. Vasin. Relationship of several variational methods for the approximate solution of ill-posed problems. Mathematical notes of the Academy of Sciences of the USSR, 7:161–165, 1970.

[37] C. R. Vogel. A constrained least squares regularization method for nonlinear ill-posed problems. SIAM Journal on Control and Optimization, 28(1):34–49, 1990.

[38] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 67(2):301–320, 2005.