OPEN ACCESS
IOP Publishing
Inverse Problems

Inverse Problems **40** (2024) 045021 (49pp) https://doi.org/10.1088/1361-6420/ad2aaa

Bayesian view on the training of invertible residual networks for solving linear inverse problems*

Clemens Arndt[1,**], **Sören Dittmer**[1,2], **Nick Heilenkötter**[1], **Meira Iske**[1], **Tobias Kluth**[1] and **Judith Nickel**[1,**]

[1] Center for Industrial Mathematics, University of Bremen, 28359 Bremen, Germany
[2] Cambridge Image Analysis Group, University of Cambridge, Cambridge CB3 0WA, United Kingdom

E-mail: carndt@uni-bremen.de and judith.nickel@uni-bremen.de

Received 1 August 2023; revised 26 January 2024
Accepted for publication 19 February 2024
Published 6 March 2024

Abstract

Learning-based methods for inverse problems, adapting to the data's inherent structure, have become ubiquitous in the last decade. Besides empirical investigations of their often remarkable performance, an increasing number of works address the issue of theoretical guarantees. Recently, Arndt *et al* (2023 *Inverse Problems* **39** 125018) exploited invertible residual networks (iResNets) to learn provably convergent regularizations given reasonable assumptions. They enforced these guarantees by approximating the linear forward operator with an iResNet. Supervised training on relevant samples introduces data dependency into the approach. An open question in this context is to which extent the data's inherent structure influences the training outcome, i.e. the learned reconstruction scheme. Here, we address this delicate interplay of training design and data dependency from a Bayesian perspective and shed light on opportunities and limitations. We resolve these limitations by analyzing reconstruction-based training of the inverses of iResNets, where we show that this optimization strategy introduces a level of data-dependency that cannot be achieved by

* The authors are listed in alphabetical order.
** Authors to whom any correspondence should be addressed.

approximation training. We further provide and discuss a series of numerical experiments underpinning and extending the theoretical findings.

Keywords: iResNet, learned regularization, linear inverse problems, Bayesian inverse problems

1. Introduction

The mathematical field of inverse problems has many applications, e.g. imaging, image processing, and several more. Inverse problems come with characteristic difficulties summarized under the term 'ill-posedness.' Typically, one wants to recover causes x, which discontinuously depend on some observed measurements z. However, a good reconstruction algorithm needs to be stable; otherwise, it cannot handle noisy measurement data. Still, one naturally wants the reconstructor to be as accurate as possible. This results in a compromise called regularization (see [6] for a recent survey on regularization theory). The more stable the reconstructor becomes, the more the set of causes for which it provides accurate results is restricted.

Hence, this set of accurate performance is critical, and one typically chooses it using *prior knowledge* about the application-specific data. In imaging problems, this knowledge often amounts to solutions x looking somehow 'natural.' However, the mathematical characterization of natural images is challenging. Thus, learned methods often outperform in this area, learning stable and accurate reconstructions from given training data (see, e.g. the early survey [4]).

While many experimental studies confirm the impressive performance of learned methods, the theoretical understanding remains limited. In particular, learned methods often lack stability guarantees. However, the topic is gaining in importance [21]. In the present work, we address this issue by studying invertible residual networks (iResNets) [5]. As proposed in [3], their invertibility makes them readily applicable to linear inverse problems. Arndt *et al* [3] approximates the forward operation ($x \mapsto z$) using the iResNet, the iResNet's inverse, then solves the inverse problem ($z \mapsto x$). Here, one can control the regularization strength by choosing a hyperparameter of the iResNet that directly controls its stability.

Arndt *et al* [3] also develop a regularization theory for these iResNets. For this purpose, they considered particular architectures and uncovered equivalences to filter functions from classical regularization theory. In the present article, we now analyze what iResNets actually learn in practice from the given training data. For this purpose, the Bayesian view is suitable, as it encodes prior knowledge on x and the measurement noise in z as probability distributions. We consider two different ways of training, via the forward and via the inverse mapping, and investigate to which extent the iResNet uses the given information about the data to regularize inverse problems.

The manuscript is structured as follows: Section 2 introduces the problem setting, basic assumptions, and the reconstruction approach using iResNets. The subsequent two sections contain the theoretical analysis of the iResNet's training in a Bayesian setting. First, section 3 considers the so-called approximation training, where the network is trained supervisedly to approximate the forward operator. In particular, we investigate what information the network learns from the training data distribution (i.e. the effect of prior distribution and noise on the network). Second, section 4 considers the so-called reconstruction training, where the iResNet's inverse is trained to map from noisy measurements to a reconstruction. Section 5 complements the theoretical analyses with extensive numerical experiments. We implement the two training types and underpin the theoretical findings of the previous sections.

1.1. Related literature

The Bayesian theory for inverse problems differs from the functional analytic regularization theory. While the functional analytic theory focuses mainly on the stability and convergence of regularization operators, the Bayesian perspective considers the full posterior distribution and uncertainty estimation for reconstructions. Detailed introductions are given in [9, 12, 25]. An overview of the basic theory and Bayesian methods for solving inverse problems is also contained in [4]. Learning-based methods are particularly powerful for solving Bayesian inverse problems, e.g. Adler and Öktem [1] describes two different general concepts for an efficient application of neural networks in this framework.

Laumont *et al* [16] proposes a method that demonstrates the Bayesian theory's advantages for inverse problems using a trained denoiser in a plug-and-play Langevin algorithm. The denoiser is assumed to fulfill a Lipschitz condition (similar to the iResNet, see section 2), implying guaranteed convergence of the algorithm to the posterior distribution. Sherry *et al* [24] leverages convex analysis to create nonexpansive residual networks and uses them to solve inverse problems. This is particularly desirable for denoising and plug-and-play schemes. Furthermore, invertible neural networks are also of interest to generative modeling. In [8], iResNets act as normalizing flows, i.e. learn to map from a base distribution to a target distribution and perform competitive or even superior to alternative architectures. Similar to iResNets, [22] also makes use of invertibility and Lipschitz constraints to get a suitable architecture for the use in convergent Gauss-Newton methods. Arndt *et al* [3] provides a more detailed discussion of the literature concerning learned convergent regularization schemes. Similar to our work, [19] also addresses Bayesian analysis of learning forward and inverse problems. But there, the focus is on a certain 2×2-example and a trivial linear network architecture to illustrate some general properties.

2. Problem setting and basic properties

We consider linear inverse problems based on the operator equation

$$Ax = z \tag{2.1}$$

where $A \in L(X,X)$ is a self-adjoint and positive semidefinite operator and X is a finite-dimensional inner product space, here $X = \mathbb{R}^n$. For simplification, we assume $\|A\| = 1$, which a scaling of the operator can easily obtain. In practice, neural network domains tend to be finite-dimensional; this justifies the restriction to the finite-dimensional case. Also, the Bayesian perspective becomes less standard if the underlying probability theory uses infinite-dimensional probability spaces, and the presented theory would require the extension to Bochner integrals. We, however, expect our observations to generalize to the infinite-dimensional case and aim to treat this in future research.

Due to the properties of A, there exist eigenvalues $\sigma_j^2 \in (0,1]$ and corresponding eigenvectors v_j, such that $\mathcal{N}(A)^\perp = \operatorname{span}\{v_j \,|\, j = 1,\dots,\tilde{n}\}$, $\tilde{n} \leqslant n$. We use this eigendecomposition in some of our theoretical analyses.

The aim is to recover the unknown ground truth vector $x^\dagger$ as well as possible by only having access to a noisy observation $z^\delta = Ax^\dagger + \eta$. The noise η is assumed to be distributed according to a probability density function (pdf) $p_H \colon X \to \mathbb{R}_{\geqslant 0}$. Since there may exist arbitrarily many solutions x which could explain the data z^δ, it is important to incorporate prior knowledge about the unknown solutions. The pdf $p_X \colon X \to \mathbb{R}_{\geqslant 0}$ encodes this knowledge. In practice, p_X may describe the distribution of natural-looking images or the typical structure of a cross-section of the human body (e.g. in CT problems).

To solve the inverse problem (2.1), we use the approach of [3], i.e. we approximate the forward operator A with a (single-layer) invertible residual network (iResNet)

$$\varphi_\theta = \mathrm{Id} - f_\theta, \tag{2.2}$$

where $f_\theta : X \to X$ is some residual function modeled as a (small) neural network. This is done by a supervised training of φ_θ for which a paired dataset $\{x^{(i)}, z^{\delta,(i)}\}_{i=1,\dots,N}$ of i.i.d. (independent and identically distributed) samples $x^{(i)} \sim p_X$, $z^{\delta,(i)} - Ax^{(i)} \sim p_H$ is needed. One can then use the trained network to compute a regularized solution of (2.1) by $\varphi_\theta^{-1}(z^\delta)$. Invertibility of φ_θ is guaranteed using the constraint

$$\mathrm{Lip}\,(f_\theta) \leqslant L \tag{2.3}$$

for some $L < 1$, where the inverse is stable and fulfills $\mathrm{Lip}(\varphi_\theta^{-1}) \leqslant {}^1\!/\!_{(1-L)}$ (see [3, lemma 2.1], [5]).

Remark 2.1. The assumption of a positive semidefinite forward operator A is due to the fact that the invertibility condition (2.3) can also be interpreted as some kind of monotonicity condition for φ_θ. Thus, φ_θ cannot approximate arbitrary linear operators but in particular positive (semi-)definite ones.

A more general linear inverse problem

$$\tilde{A}x = y \tag{2.4}$$

with an arbitrary linear operator, $\tilde{A} \in L(X, Y)$ (X and Y being different spaces), can be translated into the above setting by considering $A = \tilde{A}^*\tilde{A}$ and $z = \tilde{A}^*y$.

In this case the noise η on z may arise from noise $\tilde{\eta}$ on y via $\eta = \tilde{A}^*\tilde{\eta}$. To illustrate this, let us consider the example of Gaussian noise $\tilde{\eta} \sim \mathcal{N}(0, \Sigma)$. Then, it holds $\eta = \tilde{A}^*\tilde{\eta} \sim \mathcal{N}(0, \tilde{A}^*\Sigma\tilde{A})$, which means that $\tilde{A}^*$ transforms the covariance matrix Σ. If $\tilde{A}$ has a nontrivial nullspace, the distribution of $\tilde{A}^*\tilde{\eta}$ is singular, and there exists no pdf. Nevertheless, it is possible to approximate the distribution, e.g. by adding $\varepsilon\,\mathrm{Id}$ to the covariance matrix or restricting the problem to $\mathcal{N}(\tilde{A})^\perp$.

While implicit knowledge about p_X and p_H via the given dataset is sufficient for training φ_θ, we derive some theoretical results using these pdfs explicitly. For this purpose, we need to make the following assumptions.

Assumption 2.1. Let

- $p_X \colon X \to \mathbb{R}_{\geqslant 0}$ be a pdf $\left(\text{i.e. } \int_X p_X(x)\,\mathrm{d}x = 1\right)$ with existing first and second moments (i.e. $p_X(x)\|x\|$ and $p_X(x)\|x\|^2$ are Lebesgue-integrable) and expected value

$$\mu_X = \int_X p_X(x)\,x\,\mathrm{d}x, \tag{2.5}$$

- $p_H \colon X \to \mathbb{R}_{\geqslant 0}$ be a pdf $\left(\text{i.e. } \int_X p_H(\eta)\,\mathrm{d}\eta = 1\right)$ with existing first and second moments (i.e. $p_H(\eta)\|\eta\|$ and $p_H(\eta)\|\eta\|^2$ Lebesgue-integrable) and zero expectation $\left(\text{i.e. } \int_X p_H(\eta)\eta\,\mathrm{d}\eta = 0\right)$, and
- the random variables $x \sim p_X$, $\eta \sim p_H$ be stochastically independent.

The crucial condition to guarantee the invertibility of φ_θ is $\mathrm{Lip}(f_\theta) \leqslant L < 1$. Consequently, the inverse $\psi_\theta = \varphi_\theta^{-1}$ fulfills a property describable as a combination of coercivity and Lipschitz-continuity, which, in turn, trivially implies strong monotonicity. We formulate this equivalence in the following lemma.

Lemma 2.1 (inverse of iResNet). *For $\varphi : X \to X$ and $0 \leqslant L < 1$, the following two conditions are equivalent:*

(1) $\exists f\colon X \to X$ *with* $\mathrm{Lip}(f) \leqslant L$ *such that* $\varphi = \mathrm{Id} - f$
(2) $\exists \psi : X \to X$ *with*

$$\forall z_1, z_2 \in X\colon \qquad \left(1 - L^2\right) \|\psi(z_1) - \psi(z_2)\|^2 + \|z_1 - z_2\|^2 \leqslant 2\langle z_1 - z_2, \psi(z_1) - \psi(z_2)\rangle \tag{2.6}$$

such that $\varphi = \psi^{-1}$.

In particular, (2.6) guarantees the invertibility of ψ.

Proof. We begin with (1) $\Rightarrow$ (2). For arbitrary $x_1, x_2 \in X$, the condition $\mathrm{Lip}(\mathrm{Id} - \varphi) \leqslant L$ implies

$$\begin{aligned} &\|(x_1 - \varphi(x_1)) - (x_2 - \varphi(x_2))\|^2 \leqslant L^2 \|x_1 - x_2\|^2 \\ \Leftrightarrow\ &\|x_1 - x_2\|^2 - 2\langle x_1 - x_2, \varphi(x_1) - \varphi(x_2)\rangle + \|\varphi(x_1) - \varphi(x_2)\|^2 \leqslant L^2 \|x_1 - x_2\|^2 \\ \Leftrightarrow\ &\left(1 - L^2\right)\|x_1 - x_2\|^2 + \|\varphi(x_1) - \varphi(x_2)\|^2 \leqslant 2\langle x_1 - x_2, \varphi(x_1) - \varphi(x_2)\rangle. \end{aligned} \tag{2.7}$$

Since $\mathrm{Lip}(f) \leqslant L$ implies invertibility of φ (see [3, lemma 2.1], [5]), we can define $\psi = \varphi^{-1}$ and $z_i = \varphi(x_i)$. This yields

$$\left(1 - L^2\right)\|\psi(z_1) - \psi(z_2)\|^2 + \|z_1 - z_2\|^2 \leqslant 2\langle z_1 - z_2, \psi(z_1) - \psi(z_2)\rangle \tag{2.8}$$

for arbitrary $z_1, z_2 \in X$.

For the converse implication, we now prove that (2.6) guarantees the invertibility of ψ. Injectivity and Lipschitz continuity follow directly by applying the Cauchy–Schwarz inequality to (2.6), which yields

$$\|z_1 - z_2\|^2 \leqslant 2\|z_1 - z_2\|\|\psi(z_1) - \psi(z_2)\|, \tag{2.9}$$

$$\left(1 - L^2\right)\|\psi(z_1) - \psi(z_2)\|^2 \leqslant 2\|z_1 - z_2\|\|\psi(z_1) - \psi(z_2)\|. \tag{2.10}$$

To prove surjectivity, we construct a convergent sequence (z_k) such that $\psi(z_k)$ converges to an arbitrary $x \in X$. We recursively define

$$z_{k+1} = z_k + \left(1 - L^2\right)(x - \psi(z_k)), \qquad z_0 \in X. \tag{2.11}$$

It can be observed that

$$\begin{aligned} 2\langle x - \psi(z_k), \psi(z_{k+1}) - \psi(z_k)\rangle &= 2\langle z_{k+1} - z_k, \psi(z_{k+1}) - \psi(z_k)\rangle \frac{1}{1 - L^2} \\ &\geqslant \|\psi(z_{k+1}) - \psi(z_k)\|^2 + \frac{1}{1 - L^2}\|z_{k+1} - z_k\|^2 \\ &= \|\psi(z_{k+1}) - \psi(z_k)\|^2 + \frac{1}{1 - L^2}\|\left(1 - L^2\right)(x - \psi(z_k))\|^2 \\ &= \|\psi(z_{k+1}) - \psi(z_k)\|^2 + \left(1 - L^2\right)\|x - \psi(z_k)\|^2 \end{aligned} \tag{2.12}$$

holds. Using this, it follows

$$\begin{aligned}\|x-\psi(z_{k+1})\|^2 &= \|(x-\psi(z_k))-(\psi(z_{k+1})-\psi(z_k))\|^2\\ &= \|x-\psi(z_k)\|^2 - 2\langle x-\psi(z_k),\psi(z_{k+1})-\psi(z_k)\rangle + \|\psi(z_{k+1})-\psi(z_k)\|^2\\ &\leqslant \|x-\psi(z_k)\|^2 - \|\psi(z_{k+1})-\psi(z_k)\|^2 - \left(1-L^2\right)\|x-\psi(z_k)\|^2 + \|\psi(z_{k+1})-\psi(z_k)\|^2\\ &= L^2\|x-\psi(z_k)\|^2. \end{aligned} \tag{2.13}$$

Thus, we have $\|x-\psi(z_{k+1})\| \leqslant L\|x-\psi(z_k)\|$, which implies $\|x-\psi(z_k)\| \leqslant L^k\|x-\psi(z_0)\|$. Hence, it holds $\psi(z_k)\to x$ and (2.6) guarantees convergence of (z_k). Since x was arbitrary, ψ is surjective and therefore invertible. With the argumentation from the beginning in reversed order, we obtain the implication $(2)\Rightarrow(1)$. □

The following remark simplifies condition (2.6) for $X=\mathbb{R}$.

Remark 2.2. In case of $X=\mathbb{R}$ (one-dimensional space), condition (2.6) becomes

$$\forall z_1,z_2\in\mathbb{R}: \qquad \frac{1}{1+L} \leqslant \frac{\psi(z_1)-\psi(z_2)}{z_1-z_2} \leqslant \frac{1}{1-L}, \tag{2.14}$$

which is a constraint on the slope of ψ from above and from below.

This motivates us to think of the condition on ψ as a Lipschitz constraint similar to the one that applies to an iResNet. The following remark shows a direct connection between the iResNet and its inverse.

Remark 2.3 (inverses of iResNets are iResNets). From lemma 2.1, we can deduce that one can write the inverse of an iResNet as a scaled iResNet. The constraint (2.6) is equivalent to

$$\begin{aligned}\left(1-L^2\right)^2\|\psi(z_1)-\psi(z_2)\|^2 - 2\left(1-L^2\right)\langle z_1-z_2,\psi(z_1)-\psi(z_2)\rangle + \|z_1-z_2\|^2 &\leqslant L^2\|z_1-z_2\|^2\\ \Leftrightarrow \left\|\left(\mathrm{Id}-\left(1-L^2\right)\psi\right)(z_1)-\left(\mathrm{Id}-\left(1-L^2\right)\psi\right)(z_2)\right\| &\leqslant L\|z_1-z_2\|\\ \Leftrightarrow \mathrm{Lip}\left(\mathrm{Id}-\left(1-L^2\right)\psi\right) &\leqslant L. \end{aligned} \tag{2.15}$$

By defining $g := \mathrm{Id}-(1-L^2)\psi$ we obtain

$$\psi = \frac{1}{1-L^2}(\mathrm{Id}-g) \quad \text{where } \mathrm{Lip}(g)\leqslant L, \tag{2.16}$$

which is a scaled iResNet $\mathrm{Id}-g$ where g satisfies the same Lipschitz constraint as f in the forward mapping.

3. Approximation training

In [3], the *approximation training* is introduced, in which the iResNet φ_θ is trained to approximate A, i.e. to solve

$$\min_{\theta\in\Theta_L}\frac{1}{N}\sum_{i=1}^{N}\left\|\varphi_\theta\left(x^{(i)}\right)-z^{\delta,(i)}\right\|^2 \tag{3.1}$$

for a given dataset of N pairs $(x^{(i)}, z^{\delta,(i)}) \in X \times X$, $z^{\delta,(i)} = Ax^{(i)} + \eta^{(i)}$. The parameter space Θ_L encodes the architecture choice, and the Lipschitz constraint $\mathrm{Lip}(f_\theta) \leqslant L$. This setting was partly motivated by the so-called local approximation property ([3, theorem 3.1]) characterizing convergence guarantees for the regularized solution $\varphi_\theta^{-1}(z^\delta)$ as $\delta \to 0$. In [3], specific network architectures were trained according to the approximation training and analyzed under which conditions they satisfy the properties of a convergent regularization scheme. This revealed a connection to the classical linear filter-based regularization theory.

In contrast, we now aim to derive more general results without making restrictions on the architecture of the iResNet apart from the constraint on the Lipschitz constant of f. This enables us to analyze the influence of the noise and prior distribution on the trained network and, especially, the regularized solution. To this end, we consider the case of an infinite amount of training data, allowing us to interpret equation (3.1) from a Bayesian point of view. To be more precise, taking the limit $N \to \infty$ in equation (3.1) and exploiting the independence of x and η (assumption 2.1) results in

$$\min_{\theta \in \Theta_L} \mathbb{E}_{x \sim p_X} \mathbb{E}_{\eta \sim p_H} \left(\|\varphi_\theta(x) - Ax - \eta\|^2 \right). \tag{3.2}$$

The Euclidian norm can be decomposed into $\|\varphi_\theta(x) - Ax - \eta\|^2 = \|\varphi_\theta(x) - Ax\|^2 - 2\langle \varphi_\theta(x) - Ax, \eta \rangle + \|\eta\|^2$. Again, because of the independence of x and η and due to $\mathbb{E}_{p_H}(\eta) = 0$, the mixed term vanishes in expectation. Therefore, we obtain

$$\begin{aligned} & \min_{\theta \in \Theta_L} \mathbb{E}_{x \sim p_X} \left(\|\varphi_\theta(x) - Ax\|^2 + \mathbb{E}_{\eta \sim p_H} \left(\|\eta\|^2 \right) \right) \\ \Leftrightarrow \quad & \min_{\theta \in \Theta_L} \mathbb{E}_{x \sim p_X} \left(\|\varphi_\theta(x) - Ax\|^2 \right). \end{aligned} \tag{3.3}$$

Consequently, the noise does not influence the training. We could interpret this positively since the noise cannot lead to approximation errors of φ_θ. However, a big drawback is that φ_θ^{-1}, which shall regularize the inverse problem, neither depends on the noise level. Accordingly, the amount of regularization has to be set manually by choice of L for the noise level δ (see [3]) and is not data-dependent.

What remains is the influence of the prior distribution p_X on the training of φ_θ. We are especially interested in how φ_θ acts on the different eigenspaces of A to analyze the dependence on the size of the eigenvalues. Therefore, we make the rather strong assumption of stochastic independence of the components $x_j = \langle x, v_j \rangle$:

Assumption 3.1. Let $x_j \sim p_{X,j}$ with $p_X(x) = \prod_j p_{X,j}(x_j)$.

Observe that this assumption is implicitly made, for example, when using Tikhonov regularization with $\|\cdot\|^2$-penalty term. Furthermore, assumption 2.1 implies that $p_{X,j}$ has existing first and second moments with

$$\mu_{X,j} = \int_{\mathbb{R}} p_{X,j}(x_j)\, x_j \, \mathrm{d}x, \tag{3.4}$$

which follows from Fubini's theorem and the independence of the components. In this setting, a diagonal structure of the network

$$f_\theta(x) = \sum_j f_{j,\theta}(\langle x, v_j \rangle)\, v_j \quad \text{with} \quad f_{j,\theta} \colon \mathbb{R} \to \mathbb{R}, \tag{3.5}$$

with respect to eigenvectors v_j of A, which was also used in [3], is sufficient to account for the structure of the distribution according to assumption 3.1. Hence, the above minimization

problem can be analyzed for each component separately due to properties of the eigendecomposition, and we get

$$\min_{\theta\in\Theta_L} \mathbb{E}_{x_j\sim p_{X,j}} \left(\left| \left(1-\sigma_j^2\right) x_j - f_{j,\theta}\left(x_j\right) \right|^2 \right). \tag{3.6}$$

This is equivalent to a 1d-setting with $A\colon \mathbb{R}\to\mathbb{R}$, $x_j \mapsto \sigma_j^2 x_j$.

In the following, instead of minimizing over a parameter space Θ_L, we directly consider a function space $\mathcal{F}$ encoding the Lipschitz constraint ($\mathrm{Lip}(f)\leqslant L$) and the architecture choice. For simplicity, in what follows, we omit the index j and consider

$$\min_{f\in\mathcal{F}} \int_{\mathbb{R}} p_X(x) \left| \left(1-\sigma^2\right) x - f(x) \right|^2 \mathrm{d}x. \tag{3.7}$$

If $\mathcal{F}$ allows for (affine) linear functions and in case of $1-\sigma^2\leqslant L$, we can indicate the trivial solution $f=(1-\sigma^2)\,\mathrm{Id}$. Obviously, this solution is unique on $\mathrm{supp}(p_X)$. Thus, for eigenvalues σ^2, which are not too small, the training leads to a perfect approximation of the forward operator and no regularization of the inverse problem. For $1-\sigma^2>L$, the minimization problem gets more interesting due to the Lipschitz constraint. First, we derive the following result, which builds the basis for a subsequent generalization.

Lemma 3.1. *Let* $\mathcal{F}=\{f\in C(\mathbb{R}) \,|\, \exists m\in[-L,L], b\in\mathbb{R}\colon\ f(x)=mx+b\}$ *and* $L<1-\sigma^2$. *Then,*

$$f^*(x) = Lx + \left(1-\sigma^2-L\right)\mu_X \tag{3.8}$$

is the unique solution of the minimization problem (3.7).

Proof. The minimizer can be calculated by using the necessary KKT conditions. A detailed proof can be found in appendix A.1. □

The previous lemma provides the prerequisite for the following theorem, where $\mathcal{F}$ contains arbitrary Lipschitz continuous functions with constrained Lipschitz constant.

Theorem 3.1. *Let* $\mathcal{F}=\{f\in C^{0,1}(\mathbb{R}) \,|\, \mathrm{Lip}(f)\leqslant L\}$, *where* $C^{0,1}$ *denotes the Hölder space of Lipschitz continuous functions. Then,*

$$f^*(x) = \begin{cases} \left(1-\sigma^2\right)x & \text{if } 1-\sigma^2\leqslant L, \\ Lx+\left(1-\sigma^2-L\right)\mu_X & \text{if } 1-\sigma^2>L \end{cases} \tag{3.9}$$

is the solution of the minimization problem (3.7). This solution is unique on $\mathrm{supp}(p_X)$ *and for* $1-\sigma^2>L$ *even on the convex hull of* $\mathrm{supp}(p_X)$.

Proof. We define $F\colon \mathcal{F} \to \mathbb{R}$,

$$F(f) = \int_{\mathbb{R}} p_X(x) \,|\, \left(1-\sigma^2\right) x - f(x) \,|^2 \,\mathrm{d}x \tag{3.10}$$

and start with the case $1-\sigma^2 \leqslant L$. Obviously, it holds $F(f^*) = 0$, so f^* is a minimizer. Using the fundamental lemma of the calculus of variations, one can deduce the uniqueness on $\mathrm{supp}(p_X)$.

Now, consider $1-\sigma^2 > L$ and let $g \in C^{0,1}(\mathbb{R})$, $\mathrm{Lip}(g) \leqslant L$ be an arbitrary function. We will show that $F(g) > F(f^*)$ holds, if $g \neq f^*$ on the convex hull of $\mathrm{supp}(p_X)$.

First, we verify that F is well-defined, i.e, for $f \in \mathcal{F}$

$$\begin{aligned}
F(f) &\leqslant 2 \int_{\mathbb{R}} p_X(x) \left(\left(1-\sigma^2\right)^2 x^2 + |f(x)|^2 \right) \mathrm{d}x \\
&= 2 \int_{\mathbb{R}} p_X(x) \left(\left(1-\sigma^2\right)^2 x^2 + |f(x) - f(0) + f(0)|^2 \right) \mathrm{d}x \\
&\leqslant 2 \int_{\mathbb{R}} p_X(x) \left(1-\sigma^2\right)^2 x^2 \,\mathrm{d}x + 4 \int_{\mathbb{R}} p_X(x) \,|f(x) - f(0)|^2 \,\mathrm{d}x + 4 \int_{\mathbb{R}} p_X(x) \,|f(0)|^2 \,\mathrm{d}x \\
&\leqslant 2 \left(1-\sigma^2\right)^2 \int_{\mathbb{R}} p_X(x) x^2 \,\mathrm{d}x + 4L^2 \int_{\mathbb{R}} p_X(x) x^2 \,\mathrm{d}x + 4f(0)^2 < \infty
\end{aligned} \tag{3.11}$$

holds, as the second moment of p_X exists.

Due to $L < 1-\sigma^2$, the function g has always a smaller slope than $(1-\sigma^2)\,\mathrm{Id}$, which implies that there exists an intersection point x_0 such that $g(x_0) = (1-\sigma^2)\,x_0$. The affine linear function $\tilde{f}(x) = L(x-x_0) + (1-\sigma^2)\,x_0$ possesses the same intersection point.

In case of $g = \tilde{f}$ on the convex hull of $\mathrm{supp}(p_X)$, we simply apply lemma 3.1. This shows that g can be the minimizer only if $\tilde{f} = f^*$.

In the case of $g \neq \tilde{f}$, let us examine the integrand of $F(g)$. For any $x \in \mathbb{R}$, it holds

$$\begin{aligned}
|\left(1-\sigma^2\right) x - g(x)|^2 &= |\left(1-\sigma^2\right) x - \left(g(x) - \tilde{f}(x)\right) - \tilde{f}(x)|^2 \\
&= |\left(1-\sigma^2\right) x - \tilde{f}(x)|^2 - 2\left(\left(1-\sigma^2\right) x - \tilde{f}(x)\right)\left(g(x) - \tilde{f}(x)\right) \\
&\quad + |g(x) - \tilde{f}(x)|^2.
\end{aligned} \tag{3.12}$$

For $x \leqslant x_0$, we have $(1-\sigma^2)x - \tilde{f}(x) \leqslant 0$ and $\mathrm{Lip}(g) \leqslant L$ implies $g(x) - \tilde{f}(x) \geqslant 0$. Thus, we obtain

$$-2\left(\left(1-\sigma^2\right) x - \tilde{f}(x)\right)\left(g(x) - \tilde{f}(x)\right) \geqslant 0, \tag{3.13}$$

which implies $|(1-\sigma^2)x - g(x)|^2 \geqslant |(1-\sigma^2)x - \tilde{f}(x)|^2$. Analogously, for $x \geqslant x_0$, we observe that $(1-\sigma^2)x - \tilde{f}(x) \geqslant 0$ and $g(x) - \tilde{f}(x) \leqslant 0$, which also implies $|(1-\sigma^2)x - g(x)|^2 \geqslant |(1-\sigma^2)x - \tilde{f}(x)|^2$. Therefore, it holds $F(g) \geqslant F(\tilde{f})$.

Finally, we show that $F(g) = F(\tilde{f})$ implies $\tilde{f} = g$ on the convex hull of $\mathrm{supp}(p_X)$. If $F(g) = F(\tilde{f})$, it holds

$$\int_{\Omega} p_X(x) \,|\, \left(1-\sigma^2\right) x - g(x) \,|^2 - p_X(x) \,|\, \left(1-\sigma^2\right) x - \tilde{f}(x) \,|^2 \,\mathrm{d}x = 0. \tag{3.14}$$

for any measurable $\Omega \subset \mathbb{R}$, since the term under the integral is always greater than or equal to zero. The fundamental lemma of the calculus of variations then implies

$$p_X(x) \,|\, \left(1-\sigma^2\right) x - g(x) \,|^2 = p_X(x) \,|\, \left(1-\sigma^2\right) x - \tilde{f}(x) \,|^2 \tag{3.15}$$

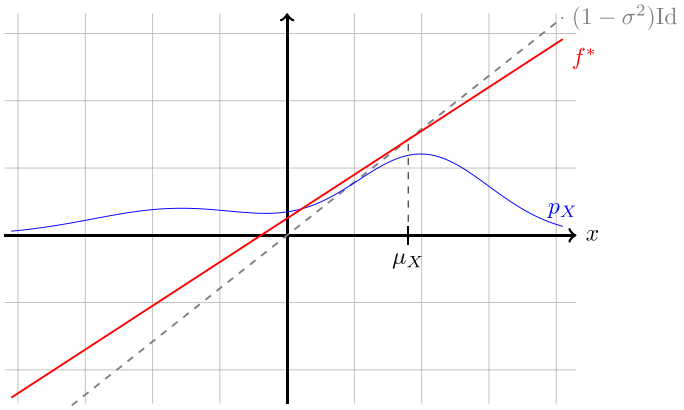


Figure 1. The residual function f^* which results from the approximation training (theorem 3.1) is affine linear and only depends on σ^2, L and μ_X. In case of $\sigma^2 < 1-L$, f^* exhibits the maximum possible slope of L and intersects $(1-\sigma)^2\,\mathrm{Id}$ at the mean μ_X of the prior distribution.

for all $x \in \mathbb{R}$ (since g and $\tilde{f}$ are continuous). Thus, for any $x \in \mathrm{supp}(p_X)$, it holds $g(x) = \tilde{f}(x)$ and for $x_1, x_2 \in \mathrm{supp}(p_X)$, we obtain $g(x_1) - g(x_2) = L(x_1 - x_2)$. Consequently, for any x in between of x_1 and x_2, $g(x) = \tilde{f}(x)$ must also hold, otherwise $\mathrm{Lip}(g) \leqslant L$ would be violated. Hence, g and $\tilde{f}$ coincide on the convex hull of $\mathrm{supp}(p_X)$. □

Figure 1 exemplifies the solution f^* for a Gaussian mixture prior p_X. The inverse φ_θ^{-1} corresponding to the minimizer of (3.7) derived in the previous theorem provides a convergent regularization scheme, which we discuss in the following remark.

Remark 3.1. Due to the affine linear structure of f^*, one can express φ_θ^{-1} as an affine filter-based regularization scheme. The affine linear diagonal architecture was already analyzed in [3, lemma 4.2], i.e. for

$$f_j(x_j) = \min\left\{1-\sigma_j^2, L\right\} x_j + \max\left\{0, 1-\sigma_j^2 - L\right\} \mu_{X,j} \tag{3.16}$$

(which coincides with the solution f^* in (3.9) in theorem 3.1), it holds

$$\varphi_\theta^{-1}(z) = \hat{b}_L + \sum_j \hat{r}_L\left(\sigma_j^2\right) \langle z, v_j \rangle v_j, \tag{3.17}$$

$$\hat{r}_L\left(\sigma_j^2\right) = \frac{1}{\max\left\{\sigma_j^2, 1-L\right\}}, \quad \hat{b}_L = \sum_{\sigma_j^2 < 1-L} \frac{1-\sigma_j^2-L}{1-L} \mu_{X,j} v_j. \tag{3.18}$$

By [3, lemma 3.3], this filter scheme with bias defines a convergent regularization method for $L \to 1$ in case of vanishing noise and a suitable parameter choice $L(\delta)$.

The previous results show that approximation training of a diagonal architecture always leads to an affine linear φ_θ, independent of prior and noise distribution (p_X, p_H). Hence, an affine linear residual layer is the best architecture choice for this task. This implies that φ_θ^{-1} is a reconstruction scheme with minimal data dependency since only the mean μ_X of the prior distribution has an influence. Furthermore, φ_θ^{-1} is equivalent to a classical regularization scheme, where one predefines the amount of regularization by choosing the parameter L depending on the noise level.

For the general approximation training problem

$$\min_{f\in C(\mathbb{R}^n,\mathbb{R}^n),\ \mathrm{Lip}(f)\leqslant L} \mathbb{E}_{x\sim p_X}\left(\|\left(\mathrm{Id}-f\right)(x)-Ax\|^2\right) \tag{3.19}$$

the previous investigations suggest that the solution depends on the second moments of the prior distribution p_X at most. A detailed consideration of the general setting for the approximation training is beyond the scope of the present work.

4. Reconstruction training

The results in the last section show that the approximation training of iResNets is capable to provide a convergent regularization but it turns out that it is insufficient for learning a noise- and more data-dependent regularization. To address this, we instead consider the training

$$\min_{\theta\in\Theta}\frac{1}{N}\sum_{i=1}^{N}\left\|\varphi_\theta^{-1}\left(Ax^{(i)}+\eta^{(i)}\right)-x^{(i)}\right\|^2 \quad \text{s.t. } \mathrm{Lip}\left(f_\theta\right)\leqslant L \tag{4.1}$$

for given training data $\{x^{(i)}\}_i \subset X$, noise realizations $\{\eta^{(i)}\}_i \in X$ and $\varphi_\theta = \mathrm{Id}-f_\theta$. This is also motivated by sufficient conditions for the convergence analysis in [3, remark 4.1]. We refer to this training scheme as the *reconstruction training*. One can also interpret this reconstruction training as a supervised training on data pairs $(x^{(i)},z^{\delta,(i)})$ for $\varphi_\theta^{-1}(z^{\delta,(i)})\approx x^{(i)}$ with $z^{\delta,(i)} = Ax^{(i)}+\eta^{(i)}$.

Using lemma 2.1, we know that

$$\begin{aligned}&\min_{\theta\in\Theta}\frac{1}{N}\sum_{i=1}^{N}\left\|\psi_\theta\left(Ax^{(i)}+\eta^{(i)}\right)-x^{(i)}\right\|^2\\ &\text{s.t. } \left(1-L^2\right)\|\psi_\theta\left(z_1\right)-\psi_\theta\left(z_2\right)\|^2+\|z_1-z_2\|^2\leqslant 2\langle\psi_\theta\left(z_1\right)-\psi_\theta\left(z_2\right),z_1-z_2\rangle \quad \forall z_1,z_2\in X\end{aligned} \tag{4.2}$$

is an equivalent problem, assuming that the architectures of φ_θ and ψ_θ can approximate any continuous function.

Similar to the approximation training, we analyze the case of an unlimited amount of training data with $x\sim p_X$ and $\eta\sim p_H$ fulfilling assumption 2.1. Thus, we obtain the minimization problem

$$\min_{\psi\in\Psi}\int_X\int_X p_X(x)\,p_H(\eta)\,\|\psi\left(Ax+\eta\right)-x\|^2\,\mathrm{d}\eta\,\mathrm{d}x, \tag{4.3}$$

where the set of functions Ψ represents the choice of the architecture and the constraints on the parameters. In this context, we also make use of the density function of $z^\delta = Ax+\eta$, which is given by

$$p_Z\left(z^\delta\right) = \int_X p_X(x)\,p_H\left(z^\delta-Ax\right)\,\mathrm{d}x. \tag{4.4}$$

With this, we can define the space of p_Z-weighted L^2-functions as

$$L^2_{p_Z}(X,X) = \{\psi : X \to X \,|\, \|\psi\|_{p_Z,2} < \infty\}, \tag{4.5}$$

$$\|\psi\|^2_{p_Z,2} = \int_X p_Z(z)\, \|\psi(z)\|^2_X \,\mathrm{d}z, \tag{4.6}$$

which is a Hilbert space. Note that functions from $L^2_{p_Z}(X,X)$ are (only) well-defined on $\mathrm{supp}(p_Z) \subset X$, which is sufficient for our purposes.

At first, we consider the unconstrained case of $\Psi = L^2_{p_Z}(X,X)$. In this setting, the conditional mean[3] $\hat{\psi}(z^\delta) = \mathbb{E}(x|z^\delta)$ is the solution of (4.3) which is in line with the established theory in statistical inverse problems (see, e.g. conditional mean estimator in the discussion of Bayes cost estimators in [12] or [1, proposition 2]).

Lemma 4.1. *Let assumption 2.1 hold and* $\Psi = L^2_{p_Z}(X,X)$. *Then,*

$$\hat{\psi} = \left(z^\delta \mapsto \mathbb{E}\left(x|z^\delta\right)\right) = \left(z^\delta \mapsto \int_X p\left(x|z^\delta\right) x \,\mathrm{d}x\right) \tag{4.7}$$

is the solution of (4.3), which is unique w.r.t. the $L^2_{p_Z}$*-norm.*

Proof. The minimization problem (4.3) can be solved via the first-order optimality condition. A detailed proof can be found in appendix A.2. □

Next, we consider the constrained reconstruction training, where we encode an arbitrary constraint, e.g. (4.2), by choosing Ψ to be a suitable subset of $L^2_{p_Z}(X,X)$.

Lemma 4.2. *Let assumption 2.1 hold,* Ψ *be an arbitrary subset of* $L^2_{p_Z}(X,X)$ *and let* $\hat{\psi} : X \to X$, $z^\delta \mapsto \mathbb{E}(x|z^\delta)$ *be the conditional mean estimator. Then, the minimization problem (4.3) is equivalent to*

$$\min_{\psi \in \Psi} \int_X p_Z\left(z^\delta\right) \|\psi\left(z^\delta\right) - \hat{\psi}\left(z^\delta\right)\|^2 \,\mathrm{d}z^\delta. \tag{4.8}$$

Note that the existence of an actual minimizer is only guaranteed for closed Ψ.

Proof. The minimization problem (4.3) is equivalent to

$$\min_{\psi \in \Psi} \int_X \int_X p_X(x)\, p_H\left(z^\delta - Ax\right) \left(\|\psi\left(z^\delta\right) - x\|^2 - \|\hat{\psi}\left(z^\delta\right) - x\|^2\right) \mathrm{d}z^\delta \,\mathrm{d}x. \tag{4.9}$$

In the proof of lemma 4.1, we have already established that the integrals are finite. To split the integral term into two parts, we use

$$\|\psi\left(z^\delta\right) - x\|^2 - \|\hat{\psi}\left(z^\delta\right) - x\|^2 = \|\psi\left(z^\delta\right)\|^2 - \|\hat{\psi}\left(z^\delta\right)\|^2 - \langle 2x, \psi\left(z^\delta\right) - \hat{\psi}\left(z^\delta\right)\rangle. \tag{4.10}$$

Fubini's theorem and the definition of p_Z (4.4) implies

$$\begin{aligned} &\int_X \int_X p_X(x)\, p_H\left(z^\delta - Ax\right) \left(\|\psi\left(z^\delta\right)\|^2 - \|\hat{\psi}\left(z^\delta\right)\|^2\right) \mathrm{d}z^\delta \,\mathrm{d}x \\ &\quad = \int_X p_Z\left(z^\delta\right) \left(\|\psi\left(z^\delta\right)\|^2 - \|\hat{\psi}\left(z^\delta\right)\|^2\right) \mathrm{d}z^\delta. \end{aligned} \tag{4.11}$$

[3] expected value corresponding to $p(x|z^\delta) = \frac{p_H(z^\delta - Ax)p_X(x)}{p_Z(z^\delta)}$.

Again using Fubini's theorem and the definition of $\hat{\psi}(z^\delta) = \mathbb{E}(x|z^\delta)$ (see proof of lemma 4.1), we obtain

$$\begin{aligned}
&\int_X \int_X p_X(x)\, p_H\left(z^\delta - Ax\right) \langle 2x, \psi\left(z^\delta\right) - \hat{\psi}\left(z^\delta\right)\rangle \,\mathrm{d}z^\delta \,\mathrm{d}x \\
&= \int_X \left\langle \int_X p_X(x)\, p_H\left(z^\delta - Ax\right) 2x \,\mathrm{d}x, \psi\left(z^\delta\right) - \hat{\psi}\left(z^\delta\right) \right\rangle \mathrm{d}z^\delta \\
&= \int_X p_Z\left(z^\delta\right) \left\langle 2\int_X \frac{p_X(x)\, p_H\left(z^\delta - Ax\right)}{p_Z\left(z^\delta\right)} x \,\mathrm{d}x, \psi\left(z^\delta\right) - \hat{\psi}\left(z^\delta\right) \right\rangle \mathrm{d}z^\delta \\
&= \int_X p_Z\left(z^\delta\right) \left\langle 2\hat{\psi}\left(z^\delta\right), \psi\left(z^\delta\right) - \hat{\psi}\left(z^\delta\right) \right\rangle \mathrm{d}z^\delta. \qquad (4.12)
\end{aligned}$$

Thus, (4.9) is equivalent to

$$\min_{\psi \in \Psi} \int_X p_Z\left(z^\delta\right) \left(\|\psi\left(z^\delta\right)\|^2 - \|\hat{\psi}\left(z^\delta\right)\|^2 - \left\langle 2\hat{\psi}\left(z^\delta\right), \psi\left(z^\delta\right) - \hat{\psi}\left(z^\delta\right) \right\rangle \right) \mathrm{d}z^\delta. \qquad (4.13)$$

Now, the assertion follows from $\|\psi(z^\delta)\|^2 - \|\hat{\psi}(z^\delta)\|^2 - \langle 2\hat{\psi}(z^\delta), \psi(z^\delta) - \hat{\psi}(z^\delta)\rangle$ $= \|\psi(z^\delta) - \hat{\psi}(z^\delta)\|^2$. □

Thus, in the constraint case, the function ψ^*, obtained by reconstruction training, aims to approximate the conditional mean estimator for the p_Z-weighted L^2-norm. In other words, reconstruction training with a constraint corresponds to a projection of the conditional mean estimator onto the constraint set with respect to the p_Z-weighted L^2-norm.

Remark 4.1. Since we know from remark 2.3 that the inverse network ψ can be interpreted as a scaled iResNet, we can further compare the minimization problem to the case of approximation training. In the notation of an iResNet, the problem formulated in (4.8) is equivalent to

$$\min_{g \in C(\mathbb{R}^n, \mathbb{R}^n),\ \mathrm{Lip}(g) \leqslant L} \int_X p_Z\left(z^\delta\right) \| \left(\mathrm{Id} - g\right)\left(z^\delta\right) - \left(1 - L^2\right) \mathbb{E}\left(x|z^\delta\right) \|^2 \,\mathrm{d}z^\delta. \qquad (4.14)$$

Thus, the reconstruction training is equivalent to training an iResNet with residual function g ($\mathrm{Lip}(g) \leqslant L$) to fit a scaled version of the posterior expectation estimator. In contrast, approximation training aims to fit the same architecture type to the linear operator A. Overall, this indicates that theoretical and numerical properties (such as data-dependence) for the two strategies are the sole consequences of the training approach, and there is no additional bias due to the architecture choice of an iResNet as the forward mapping when assuming sufficient approximation capability.

So far, the distribution of the noise p_H was fixed. Now, we want to consider a variable noise level $\delta > 0$ by introducing the pdf $p_{H,\delta} \colon X \to \mathbb{R}_{\geqslant 0}$. We do not specify the exact relation of $p_{H,\delta}$ on δ but make the rather informal assumption that $\eta^\delta \sim p_{H,\delta}$ implies $\|\eta^\delta\| \sim \delta$ with high probability. So, $\delta \to 0$ corresponds to the case of vanishing noise. Analogously, let $p_{Z,\delta}$ be defined according to (4.4) s.t. $z^\delta \sim p_{Z,\delta}$ holds for $z^\delta = Ax + \eta^\delta$. The posterior mean now also depends on δ and may therefore be defined as

$$\hat{\psi}_\delta(z) = \int_X \frac{p_{H,\delta}(z - Ax)\, p_X(x)}{p_{Z,\delta}(z)} x \,\mathrm{d}x. \qquad (4.15)$$

Further, we want to specify the set $\Psi \subset L^2_{p_Z}(X, X)$, which represents the inverses of possible iResNet architectures depending on δ and L. To encode the side constraint of (4.2), we define

$$\Psi_L^\delta = \left\{ \psi \in L^2_{p_{Z,\delta}}(X,X) \cap C(\operatorname{supp}(p_{Z,\delta}), X) \mid (4.17) \text{ holds } \forall z_1, z_2 \in \operatorname{supp}(p_{Z,\delta}) \right\}, \tag{4.16}$$

$$\left(1 - L^2\right) \|\psi(z_1) - \psi(z_2)\|^2 + \|z_1 - z_2\|^2 \leqslant 2\langle z_1 - z_2, \psi(z_1) - \psi(z_2)\rangle. \tag{4.17}$$

The set Ψ_L^δ is closed and convex in $L^2_{p_{Z,\delta}}(X,X)$ for any $L < 1$ and $\delta > 0$. Consequently, the minimization problem in lemma 4.2 is well-defined and admits a unique solution. This solution, w.r.t. Ψ_L^δ, $p_{Z,\delta}$ and $\hat{\psi}_\delta$, is denoted by $\psi^*_{L,\delta}$.

The parameter L controls the stability of the elements from Ψ_L^δ and can therefore be interpreted as a regularization parameter. The question remains whether we can also obtain a convergence result, i.e. $\psi^*_{L,\delta}(Ax^\dagger + \eta^\delta) \to x^\dagger$ for $\|\eta^\delta\| \leqslant \delta$, $\delta \to 0$ (vanishing noise) and $L \to 1$, analogous to the convergence theorems of classical methods like Tikhonov's. In the following remark, we discuss difficulties and supporting facts regarding a convergence result.

Remark 4.2. There are different results in the literature for the posterior distribution to converge to one single point $x^\dagger$ (or its respective delta distribution) [7, theorems 1 and 2], [26]. Thus, it might be realistic to expect that the conditional mean estimator $\hat{\psi}_\delta(Ax^\dagger)$ also converges to $x^\dagger$. In classical convergence theorems [10], the ground truth $x^\dagger$ is mostly assumed to be a minimum-norm-solution of (2.1). However, in a Bayesian setting with a learned reconstruction scheme, a criterion based on p_X for characterizing $x^\dagger$ is more appropriate, e.g.

$$x^\dagger = \arg\max_{Ax=z} p_X(x) \qquad \text{for } z \in \mathcal{R}(A) \tag{4.18}$$

as in [7] or the one provided in the subsequent lemma 4.3.

Since $\psi^*_{L,\delta}$ is the projection of $\hat{\psi}_\delta$ onto the set Ψ_L^δ (lemma 4.2), it is not unlikely that $\psi^*_{L,\delta}(Ax^\dagger)$ also converges to $x^\dagger$. However, the projection is w.r.t. the $L^2_{p_{z,\delta}}$-norm, and we are actually interested in pointwise convergence. Besides, all functions ψ from the sets Ψ_L^δ are Lipschitz continuous and they fulfill the monotonicity condition

$$\|z_1 - z_2\|^2 \leqslant 2\langle z_1 - z_2, \psi(z_1) - \psi(z_2)\rangle. \tag{4.19}$$

Thus, one cannot approximate arbitrary $L^2_{p_{z,\delta}}$-functions.

There are two facts that partly address this difficulty. First, the posterior mean $\hat{\psi}_\delta$ is also a Lipschitz continuous function under certain assumptions [9, theorem 4.5, remark 4.6]. Second, a generalized inverse on $\mathcal{R}(A)$ of the form $A^\dagger : z \mapsto x^\dagger$ would indeed fulfill (4.19). This follows from A being self-adjoint and positive semidefinite (we can write $A = \tilde{A}^*\tilde{A}$) and $\|A\| = 1$ by

$$\begin{aligned} \|z_1 - z_2\|^2 = \|Ax_1^\dagger - Ax_2^\dagger\|^2 \leqslant \|\tilde{A}^*\|^2 \|\tilde{A}(x_1 - x_2)\|^2 &= \left\langle \tilde{A}\left(x_1^\dagger - x_2^\dagger\right), \tilde{A}\left(x_1^\dagger - x_2^\dagger\right)\right\rangle \\ &\leqslant \left\langle A\left(x_1^\dagger - x_2^\dagger\right), x_1^\dagger - x_2^\dagger\right\rangle = \langle z_1 - z_2, A^\dagger z_1 - A^\dagger z_2\rangle. \end{aligned} \tag{4.20}$$

Assuming that $\psi^*_{L,\delta}$ converges pointwise to $A^\dagger$ on $\mathcal{R}(A)$ would be sufficient to obtain a convergence result for noisy data as well. If $\|z^\delta - z\| \leqslant \delta$ holds and one chooses L in a way that $L \to 1$ and $\frac{\delta}{1-L} \to 0$ for $\delta \to 0$, the desired convergence

$$\begin{aligned} \|\psi^*_{L,\delta}(z^\delta) - A^\dagger(z)\| &\leqslant \|\psi^*_{L,\delta}(z^\delta) - \psi^*_{L,\delta}(z)\| + \|\psi^*_{L,\delta}(z) - A^\dagger(z)\| \\ &\leqslant \frac{1}{1-L}\|z^\delta - z\| + \|\psi^*_{L,\delta}(z) - A^\dagger(z)\| \to 0 \end{aligned} \tag{4.21}$$

would follow, since $\operatorname{Lip}(\psi^*_{L,\delta}) \leqslant \frac{1}{1-L}$ [3, lemma 2.1].

In the following, we provide a result for a potential candidate for $x^\dagger$ for the convergence analysis.

Lemma 4.3. *Let assumption 2.1 hold with $p_{H,\delta} = p_H$ indicating the dependence on δ. In addition, let $\hat{\psi}_\delta : X \to X$, $z^\delta \mapsto \mathbb{E}(x|z^\delta)$ be the conditional mean estimator with*

$$p\left(x|z^\delta\right) = \frac{p_{H,\delta}\left(z^\delta - Ax\right) p_X(x)}{p_{Z,\delta}\left(z^\delta\right)} \tag{4.22}$$

and $p_{Z,\delta}(z^\delta) = \int_X p_X(x)\, p_{H,\delta}(z^\delta - Ax)\,\mathrm{d}x$. We further make the following assumptions:

(i) *Noise on $\mathcal{R}(A)^\perp = \mathcal{N}(A)$ and $\mathcal{R}(A) = \mathcal{N}(A)^\perp$ (as $A = A^*$ and X is finite-dimensional) is stochastically independent, i.e. there exist pdfs $p^\dagger_{H,\delta} : \mathcal{N}(A)^\perp \to \mathbb{R}_{\geqslant 0}$ and $p^0_{H,\delta} : \mathcal{N}(A) \to \mathbb{R}_{\geqslant 0}$ such that $p_{H,\delta}(\eta) = p^0_{H,\delta}(P_{\mathcal{N}(A)}\eta) \cdot p^\dagger_{H,\delta}(P_{\mathcal{N}(A)^\perp}\eta)$,*
(ii) *$p^0_{H,\delta}$ and $p^\dagger_{H,\delta}$ define Dirac sequences with respect to $\delta \to 0$,*
(iii) *p_X is compactly supported and continuous. We define $\mathcal{R}_{p_X}(A) := \{Ax \,|\, x \in \mathrm{supp}(p_X)\} \subset \mathcal{R}(A)$,*
(iv) *For any $z \in \mathcal{R}_{p_X}$ there exists a $\bar{\delta}$ such that for all $\delta \in (0, \bar{\delta}]$ it holds $z \in \mathrm{supp}(p_{Z,\delta})$.*

We then have pointwise convergence of $\hat{\psi}_\delta$ for $z \in \mathcal{R}_{p_X}(A)$ such that it holds

$$\begin{aligned} &\hat{\psi}_\delta(z) \underset{\delta\to 0}{\longrightarrow} A^\dagger z + \int_{\mathcal{N}(A)} p\left(x_0|A^\dagger z\right) x_0 \,\mathrm{d}x_0 \\ &\text{with } p\left(x_0|A^\dagger z\right) = \frac{p_X\left(x_0 + A^\dagger z\right)}{\int_{\mathcal{N}(A)} p_X\left(x_0' + A^\dagger z\right)\,\mathrm{d}x_0'} \end{aligned} \tag{4.23}$$

i.e. it converges to the minimum-norm solution $A^\dagger z$ plus the conditional expectation $\mathbb{E}(x_0|A^\dagger z) \in \mathcal{N}(A)$ in the nullspace.

Proof. The proof can be found in appendix A.3. □

4.1. Reconstruction training for diagonal architecture

In order to derive more specific results for the minimizer of (4.3), we make the assumption of stochastic independence of the components $x_j = \langle x, v_j\rangle \sim p_{X,j}$ and $\eta_j = \langle \eta, v_j\rangle \sim p_{H,j}$ similar to the setting in the last section on the approximation training:

Assumption 4.1. Let $p_X(x) = \prod_j p_{X,j}(x_j)$ and $p_H(\eta) = \prod_j p_{H,j}(\eta_j)$.

Observe that the first and second moments of $p_{X,j}$ and $p_{H,j}$ exist due to assumption 2.1 with

$$\mu_{X,j} = \int_{\mathbb{R}} p_{X,j}(x_j)\, x_j \,\mathrm{d}x_j \qquad \text{and} \qquad \mu_{H,j} = \int_{\mathbb{R}} p_{H,j}(\eta_j)\, \eta_j \,\mathrm{d}\eta_j = 0 \tag{4.24}$$

for all $j \in \mathbb{N}$. In addition, the density of $z_j = \sigma_j^2 x_j + \eta_j$ is given by

$$p_{Z,j}(z_j) = \int_{\mathbb{R}} p_{X,j}(x_j)\, p_{H,j}\left(z_j - \sigma_j^2 x_j\right)\,\mathrm{d}x_j \tag{4.25}$$

with

$$\mu_{Z,j} = \int_{\mathbb{R}} p_{Z,j}(z_j)\, z_j \,\mathrm{d}z_j = \int_{\mathbb{R}}\int_{\mathbb{R}} z_j\, p_{X,j}(x_j)\, p_{H,j}\left(z_j - \sigma_j^2 x_j\right)\,\mathrm{d}x_j\,\mathrm{d}z_j = \sigma_j^2\, \mu_{X,j}. \tag{4.26}$$

Analogously to the approximation training, in this setting, it is sufficient to consider a diagonal structure of φ, which then implies the same structure for ψ, i.e.

$$\psi(z) = \sum_j \psi_j(\langle z, v_j\rangle) v_j \tag{4.27}$$

and the resulting optimization problem to train each ψ_j now reads

$$\min_{\psi_j \in \Psi_j} \int_{\mathbb{R}} \int_{\mathbb{R}} p_{X,j}(x_j)\, p_{H,j}(\eta_j)\, \|\psi_j(\sigma_j^2 x_j + \eta_j) - x_j\|^2 \,\mathrm{d}\eta_j \,\mathrm{d}x_j \tag{4.28}$$

with a suitable set of functions Ψ_j. Recall that $\psi_j : \mathbb{R} \to \mathbb{R}$ represents the inverse of an iResNet if ψ_j satisfies

$$\frac{1}{1+L} \leqslant \frac{\psi_j(z_1) - \psi_j(z_2)}{z_1 - z_2} \leqslant \frac{1}{1-L} \qquad \forall z_1, z_2 \in \mathbb{R} \tag{4.29}$$

for some $0 \leqslant L < 1$, cf remark 2.2. Therefore, we consider the set

$$\Psi_j = \left\{ \psi_j : \mathbb{R} \to \mathbb{R} \,\middle|\, \psi_j(z_j) = m z_j + b \text{ for } z_j \in \mathbb{R} \text{ with } m \in \mathbb{R}, \frac{1}{1+L} \leqslant m \leqslant \frac{1}{1-L}, b \in \mathbb{R} \right\} \tag{4.30}$$

for some $0 \leqslant L < 1$. The following lemma provides a formula for the minimizer of problem (4.28). For better readability, we leave out the index j in the subsequent derivations.

Lemma 4.4. *The unique solution to the minimization problem (4.28) with Ψ_j as in (4.30) is given by*

$$\psi^*(z) = mz + \left(1 - \sigma^2 m\right) \mu_X \quad \text{for } z \in \mathbb{R} \tag{4.31}$$

with

$$m = \begin{cases} \frac{1}{1+L} & \text{if } \frac{\sigma^2 Var_{p_X}(x)}{\sigma^4 Var_{p_X}(x) + Var_{p_H}(\eta)} < \frac{1}{1+L}, \\ \frac{\sigma^2 Var_{p_X}(x)}{\sigma^4 Var_{p_X}(x) + Var_{p_H}(\eta)} & \text{if } \frac{1}{1+L} \leqslant \frac{\sigma^2 Var_{p_X}(x)}{\sigma^4 Var_{p_X}(x) + Var_{p_H}(\eta)} \leqslant \frac{1}{1-L}, \\ \frac{1}{1-L} & \text{if } \frac{\sigma^2 Var_{p_X}(x)}{\sigma^4 Var_{p_X}(x) + Var_{p_H}(\eta)} > \frac{1}{1-L}. \end{cases} \tag{4.32}$$

Proof. The minimization problem (4.28) can be solved using the necessary KKT conditions. A detailed proof can be found in appendix A.4. □

Note that the inverse of the function $\psi^* : \mathbb{R} \to \mathbb{R}$ is given by $\varphi^*(x) = x - f^*(x)$ with

$$f^*(x) = \begin{cases} -Lx + \left(1 + L - \sigma^2\right) \mu_X & \text{if } \frac{\sigma^2 \mathrm{Var}_{p_X}(x)}{\sigma^4 \mathrm{Var}_{p_X}(x) + \mathrm{Var}_{p_H}(\eta)} < \frac{1}{1+L}, \\ \left(1 - \frac{\sigma^4 \mathrm{Var}_{p_X}(x) + \mathrm{Var}_{p_H}(\eta)}{\sigma^2 \mathrm{Var}_{p_X}(x)}\right) x + \frac{\mathrm{Var}_{p_H}(\eta)}{\sigma^2 \mathrm{Var}_{p_X}(x)} \mu_X & \text{if } \frac{1}{1+L} \leqslant \frac{\sigma^2 \mathrm{Var}_{p_X}(x)}{\sigma^4 \mathrm{Var}_{p_X}(x) + \mathrm{Var}_{p_H}(\eta)} \leqslant \frac{1}{1-L}, \\ Lx + \left(1 - L - \sigma^2\right) \mu_X & \text{if } \frac{\sigma^2 \mathrm{Var}_{p_X}(x)}{\sigma^4 \mathrm{Var}_{p_X}(x) + \mathrm{Var}_{p_H}(\eta)} > \frac{1}{1-L} \end{cases} \tag{4.33}$$

for $x \in \mathbb{R}$. In the case of noiseless data, i.e. $\mathrm{Var}_{p_H}(\eta) = 0$, f^* corresponds to the function f^* derived in lemma 3.1 and theorem 3.1.

The function ψ^* plays an important role in the case of Gaussian prior and noise distributions, which we will deal with in the following corollary.

Corollary 4.1. *Assume that $p_X : \mathbb{R} \to \mathbb{R}$ and $p_H : \mathbb{R} \to \mathbb{R}$ are Gaussian probability density functions. Then, the function ψ^* of lemma 4.4 is a solution to the minimization problem (4.28) with*

$$\Psi = \left\{ \psi \in C^1(\mathbb{R}) \cap L^2_{p_Z}(\mathbb{R}) \;\middle|\; \frac{1}{1+L} \leqslant \psi'(z) \leqslant \frac{1}{1-L} \text{ for all } z \in \mathbb{R} \right\}. \quad (4.34)$$

Proof. In lemma 4.1 we have seen that the unconstrained solution of problem (4.28) is given by $\hat{\psi}(z) = \mathbb{E}(x|z)$ for all $z \in \mathbb{R}$. In the case of Gaussian noise and prior distributions, $\mathbb{E}(x|z)$ can be expressed as

$$\mathbb{E}(x|z) = \frac{\sigma^2 \mathrm{Var}_{p_X}(x)}{\sigma^4 \mathrm{Var}_{p_X}(x) + \mathrm{Var}_{p_H}(\eta)} z + \left(1 - \frac{\sigma^4 \mathrm{Var}_{p_X}(x)}{\sigma^4 \mathrm{Var}_{p_X}(x) + \mathrm{Var}_{p_H}(\eta)}\right) \mu_X \quad (4.35)$$

for all $z \in \mathbb{R}$ [25, theorem 6.20 and equation (2.16a)], which is an element of $C^1(\mathbb{R}) \cap L^2_{p_Z}(\mathbb{R})$. In combination with lemma (4.2), minimization problem (4.3) can be rewritten as

$$\min_{\psi \in \Psi} \int_{\mathbb{R}} p_Z(z) \left(\psi(z) - \frac{\sigma^2 \mathrm{Var}_{p_X}(x)}{\sigma^4 \mathrm{Var}_{p_X}(x) + \mathrm{Var}_{p_H}(\eta)} z - \left(1 - \frac{\sigma^4 \mathrm{Var}_{p_X}(x)}{\sigma^4 \mathrm{Var}_{p_X}(x) + \mathrm{Var}_{p_H}(\eta)}\right) \mu_X \right)^2 \mathrm{d}z. \quad (4.36)$$

The same reasoning as in the proof of theorem 3.1 now shows that ψ^* of lemma 4.4 is a solution to the minimization problem. □

Figure 2 illustrates the behavior of the unconstrained solution $\hat{\psi}$ and the constrained solution ψ^* in case of Gaussian probability density functions for varying noise and small singular values ($1 - \sigma^2 > L$). Note that both solutions can be rewritten to depend on μ_Z instead of on μ_X using $\mu_Z = \sigma^2 \mu_X$. It can be observed that the noise level affects the slope of the unconstrained solution, with decreasing values at higher noise levels. Thus, $\hat{\psi}$ violates the invertibility condition (4.29) for very small and very large values of $\mathrm{Var}_{p_H}(\eta)$ leading to $\psi^* \neq \hat{\psi}$ in these cases (left and right image of figure 2).

4.1.1. General behavior of ψ^*. The previous results deal with special cases where either the architecture or the probability density functions are known. In order to derive more general results, we make use of the theory of optimal control. For this, we need to restrict ourselves to piecewise continuously differentiable functions ψ with bounded domain, i.e. we consider the set

$$\Psi = \left\{ \psi \in C^0([z_0, z_1]) \;\middle|\; \psi \text{ piecewise continuously differentiable with } \frac{1}{1+L} \leqslant \psi'(z) \leqslant \frac{1}{1-L} \right\} \quad (4.37)$$

with fixed $z_0, z_1 \in \mathbb{R}$ and $\Pr(z \leqslant z_0)$[4] $\leqslant \varepsilon$, $\Pr(z \geqslant z_1) \leqslant \varepsilon$ for some small $\varepsilon > 0$ to stay close to the previous setting. Furthermore, to apply the optimal control theory, we need to split the optimization problem into two successive minimization problems. First, we minimize over all

[4] Pr denotes the probability w.r.t. $z \sim p_Z$.

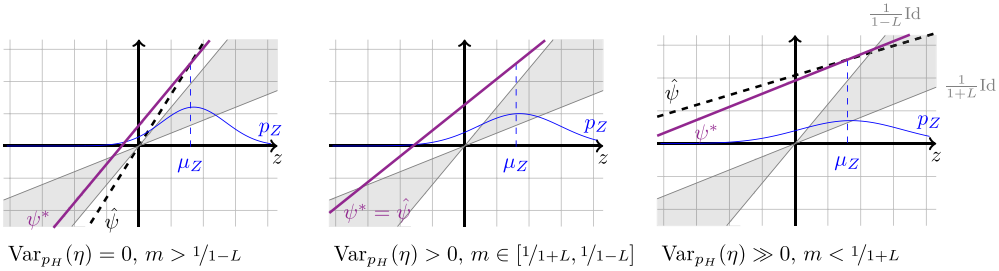


Figure 2. Illustration of the constrained solution ψ^* and the unconstrained solution $\hat{\psi}$ in the case of Gaussian probability density functions p_X and p_H, cf corollary 4.1. The slope of the unconstrained solution $\hat{\psi}$ is denoted by m, i.e. $m = \frac{\sigma^2 \mathrm{Var}_{p_X}(x)}{\sigma^4 \mathrm{Var}_{p_X}(x) + \mathrm{Var}_{p_H}(\eta)}$. The plots exemplify the behavior of ψ^* and $\hat{\psi}$ for small singular values ($1 - \sigma^2 > L$) assuming a fixed prior distribution p_X but increasing variance of p_H. In the case that $\mathrm{Var}_{p_H}(\eta) = 0$ (left), the slope of the unconstrained solution exceeds $\frac{1}{1-L}$. If the noise increases, the slope of the unconstrained solution decreases, resulting in $m \in [\frac{1}{1+L}, \frac{1}{1-L}]$ (middle). For very noisy data, the slope of the unconstrained solution is smaller than $\frac{1}{1+L}$ (right), again resulting in $\psi^* \neq \hat{\psi}$. Observe that ψ^* and $\hat{\psi}$ are equal to $\frac{1}{\sigma^2}\mu_Z = \mu_X$ for $z = \mu_Z = \sigma^2 \mu_X$ in all cases.

functions $\psi \in \Psi$ with fixed starting point $\psi(z_0) = \psi^0 \in \mathbb{R}$. Then, the starting point minimizing the objective function is determined. In combination with lemma 4.2, the minimization problem thus reads

$$\min_{\psi^0 \in \mathbb{R}} \left(\min_{\psi \in \Psi \cap \{\psi \,|\, \psi(z_0) = \psi^0\}} \frac{1}{2} \int_{z_0}^{z_1} p_Z(z) \, |\psi(z) - \hat{\psi}(z)|^2 \, \mathrm{d}z \right). \tag{4.38}$$

We would like to stress that the minimization problem defined in lemma 4.2 is not equivalent to the initial one of equation (4.3) due to the bounded domain of ψ. However, this error is negligible for small ε and the two minimization problems coincide if $\mathrm{supp}(p_Z) \subset [z_0, z_1]$.

Remark 4.3. The restriction to a bounded domain of ψ might seem artificial at first. Nevertheless, in applications, the dataset rarely contains samples belonging to low-density regions of p_Z, and thus, these cases are covered in our setting.

The inner minimization problem can be solved with the help of Pontryagin's maximum principle, resulting in the following necessary and sufficient conditions for the derivative of ψ.

Lemma 4.5. *Let $\psi_0 \in \Psi \cap \{\psi \,|\, \psi(z_0) = \psi^0\}$ be a solution of the minimization problem*

$$\min_{\psi \in \Psi \cap \{\psi \,|\, \psi(z_0) = \psi^0\}} \frac{1}{2} \int_{z_0}^{z_1} p_Z(z) \, |\psi(z) - \hat{\psi}(z)|^2 \, \mathrm{d}z. \tag{4.39}$$

Then, in all points of differentiability, the derivative ψ_0' must satisfy the necessary and sufficient conditions

$$\psi_0'(z) = \begin{cases} \frac{1}{1+L} & \text{if } \lambda(z) > 0 \\ f_0(z) & \text{if } \lambda(z) = 0 \\ \frac{1}{1-L} & \text{if } \lambda(z) < 0 \end{cases} \quad \text{with } z \in [z_0, z_1] \tag{4.40}$$

for some $f_0 : [z_0, z_1] \to \mathbb{R}$ *satisfying*

$$\frac{1}{1+L} \leqslant f_0(z) \leqslant \frac{1}{1-L} \quad \forall z \in [z_0, z_1] \tag{4.41}$$

and $\lambda : [z_0, z_1] \to \mathbb{R}$ *with*

$$\lambda'(z) = -p_Z(z)\left(\psi_0(z) - \hat{\psi}(z)\right) \text{ and } \lambda(z_1) = 0. \tag{4.42}$$

Proof. Let us denote the set of all points $z \in [z_0, z_1]$ where ψ_0 is differentiable by D. For problem (4.39), Pontryagin's maximum principle, see [18, *9.6 theorem 1] and [23, theorem 1], provides the necessary conditions

$$\psi_0'(z) = u_0(z) \quad \text{for all } z \in D \text{ and some piecewise continuous function } u_0 : [z_0, z_1] \to \mathbb{R} \tag{4.43}$$

$$\frac{1}{1+L} \leqslant u_0(z) \leqslant \frac{1}{1-L} \quad \forall z \in [z_0, z_1] \tag{4.44}$$

$$\lambda'(z) = -p_Z(z)\left(\psi_0(z) - \hat{\psi}(z)\right) \text{ with } \lambda(z_1) = 0 \tag{4.45}$$

$$H(\psi_0, u_0, \lambda, z) \leqslant H(\psi_0, u, \lambda, z) \quad \forall u \text{ satisfying } (4.44) \tag{4.46}$$

with the Hamiltonian function

$$H(\psi, u, \lambda, z) = \lambda(z)\, u(z) + \frac{1}{2} p_Z(z)\, |\psi(z) - \hat{\psi}(z)|^2. \tag{4.47}$$

Condition (4.46) is equivalent to setting

$$u_0(z) = \begin{cases} \frac{1}{1+L} & \text{if } \lambda(z) > 0 \\ f_0(z) & \text{if } \lambda(z) = 0 \\ \frac{1}{1-L} & \text{if } \lambda(z) < 0 \end{cases} \qquad \text{for some function } f_0 \text{ satisfying } (4.41). \tag{4.48}$$

Furthermore, For a function ψ_0 satisfying the conditions of Pontryagin's maximum principle to be a solution of problem (4.39), the Hamiltonian needs to be jointly convex in ψ and u and the constrained set defined by equation (4.44) needs to be convex, see [23, theorem 2]. Both of these conditions are satisfied in our setting, and thus, the proof is complete. □

We would like to remark that λ can be expressed as

$$\lambda(z) = \int_z^{z_1} p_Z(\tilde{z}) \left(\psi_0(\tilde{z}) - \hat{\psi}(\tilde{z})\right) \mathrm{d}\tilde{z}. \tag{4.49}$$

whenever p_Z and $\hat{\psi}$ are continuous. To illustrate the previous lemma, let us look at a very simple example. Assume that $\hat{\psi}'(z) > {}^1\!/_{1-L}$ for all $z \in [z_0, z_1]$. Then, lemma 4.5 in combination with a minimization over the starting points ψ^0 shows that a solution to problem (4.38) is given by

$$\psi^*(z) = \frac{1}{1-L} z + \frac{\int_{z_0}^{z_1} p_Z(\tilde{z})\, \hat{\psi}(\tilde{z})\, \mathrm{d}\tilde{z} - \frac{1}{1-L} \int_{z_0}^{z_1} p_Z(\tilde{z})\, \tilde{z}\, \mathrm{d}\tilde{z}}{\int_{z_0}^{z_1} p_Z(\tilde{z})\, \mathrm{d}\tilde{z}} \quad \text{for } z \in [z_0, z_1]. \tag{4.50}$$

In addition, the general behavior of the solution to problem (4.38) is illustrated in figure 3. This exemplifies that the best possible architecture choice, when considering the reconstruction training loss, is not necessarily an affine linear one, unlike in the approximation training studied in the last section. This is partly because of the influence of the noise distribution p_H, which

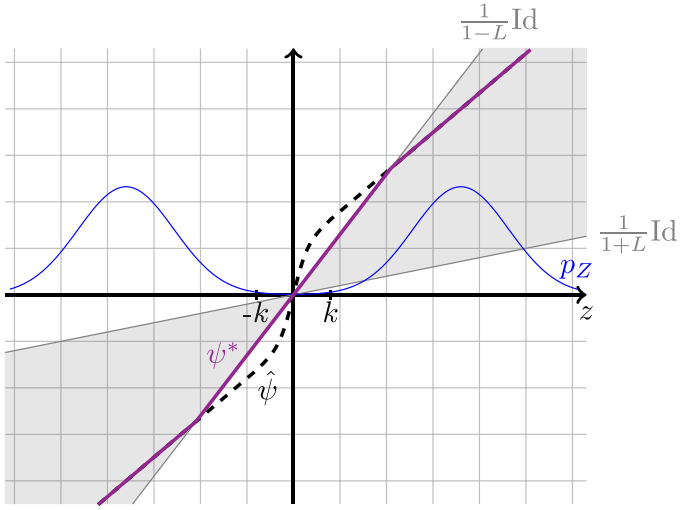


Figure 3. Behavior of the solution ψ^* of (4.38) in the case that p_z can be represented as a Gaussian mixture, cf remark 5.1. Observe that the slope of the unconstrained solution $\hat{\psi}$ exceeds $1/1-L$ in the interval $[-k,k]$ resulting in $\lambda(z)<0$ for $z\in[-k-\varepsilon,k+\varepsilon]$ with $\varepsilon>0$. As a result, ψ^* is equal to $z\mapsto 1/1-Lz$ for $z\in[-k-\varepsilon,k+\varepsilon]$ and to $\hat{\psi}$ outside this interval.

cancels out when using the approximation training loss. Moreover, the variance of the prior and noise distribution influences the best architecture and parameter choice, which is not the case in the approximation training setting, where only the expectation of the prior distribution influences the solution.

5. Numerical experiments

To study the implications of the previously developed theory for the practical application of iResNets for solving inverse problems, we perform experiments on two forward operators, where we compare approximation training (3.1) to reconstruction training (4.1). In all experiments, we train single-layer iResNets with diagonal (3.5) structure where the residual functions f_θ comprise multiple layers.

In the setting presented in the following sections, we consider a discrete convolution with a smoothing kernel $a\in\mathbb{R}^{9\times 9}$ that is depicted in figure 4 and zero padding to preserve dimensionality. Since the resulting Toeplitz matrix M_a that performs the convolution with a is symmetric and positive definite, this serves immediately as a self-adjoint operator $A=M_a$. The second inverse problem we aim to solve is given by a discrete Radon operator $\tilde{A}:\mathbb{R}^{28\times 28}\to\mathbb{R}^{30\times 41}$, such that $A=\tilde{A}^*\tilde{A}$, which is in line with the setting used in the prior work [3]. We restrict the discussion of the numerical results to the convolution operator, and for the sake of completeness, the results for the Radon operator are provided in appendix D in figures 12–16 and table 2.

In both cases, we train our models on the MNIST handwritten digits dataset [17], where we treat images as flattened vectors in $\mathbb{R}^{28\cdot 28}$. In addition, we study an artificially generated bimodal Gaussian dataset for which we sample the prior distribution in every singular vector independently from a (bimodal) Gaussian mixture distribution. The pdf in every j therefore reads

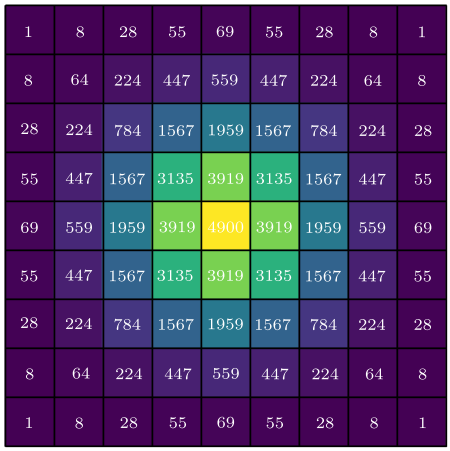


Figure 4. The filter kernel a used in the convolution operator M_a.

$$p_{X,j}(x_j) = \rho_1\, g_{v,t_1}(x_j) + \rho_2\, g_{v,t_2}(x_j) \tag{5.1}$$

where $\rho_1 = 0.35$, $\rho_2 = 0.65$, $v = 0.15$, $t_1 = -0.6$, $t_2 = 0.6$ and $g_{s,t}$ is the pdf of a Gaussian with standard deviation s and expected value t. This bimodal structure enables us to further explore the data dependency of the optimized models.

The architecture of the subnetworks included in the diagonal architecture and their training is realized identical to [3]: Each subnetwork consists of a three-layer fully connected network, equipped with 35 hidden neurons in each of the first two layers and one output neuron. In addition, we apply the ReLU activation function to the first and second layers of each subnetwork. Figure 7 of [3] depicts the architecture. The network weights are optimized using Adam [13] with a learning rate of 0.001, reduced by the factor 0.96 after every epoch. Extending the approach in [20], we parameterize the network weights in the k-th linear layer of all subnetworks to fulfill the Lipschitz constraint $L^{(k)}$. We do so by choosing the weights $W_{j,k}$ as $W_{j,k} = \min(1, L^{(k)}/\mathrm{Lip}(\tilde{W}_{j,k}))\tilde{W}_{j,k}$, where $\tilde{W}_{j,k}$ are unconstrained matrices. We choose the individual layer Lipschitz constants as $L^{(1)} = L^{(2)} = 1$, $L^{(3)} = L$.

Reconstruction training is accomplished by computing the inverse of the iResNets through the usual fixed point iteration and backpropagating through the unrolled iteration to optimize the network weights. To ensure sufficient accuracy of the inverse, especially for the models with a high Lipschitz bound, we need to make a suitable choice for the number of fixed point iterations. For this purpose, we utilize the models that were trained via approximation training. This is motivated by our observation that each of these models reaches its Lipschitz bound and is therefore as unstable as permitted within the constraint. For each L, we select a number of fixed point iterations that results in approximately 2% error of $\varphi^{-1}_{\theta(L)}$ evaluated on the test dataset. As a result, one has to run the iterative inversion in every training step, resulting in a much greater computational effort for the reconstruction training than approximation training. The Lipschitz bound, realized by a proper parameterization of the f_θ, has to be computed only once per iteration.

Of course, a numerically more efficient training approach would be to extend on remark 2.3 and construct ψ, the inverse of an iResNet, as a scaled iResNet that is trained on reversed data points but similar to the approximation approach. However, we currently do not have guarantees on the approximation capability of the involved (fully-connected) iResNets architectures

and their inverses. To provide a fair comparison, we aim to enforce the same inductive bias in both training methods by choosing the same forward mapping architecture. The source code corresponding to the experiments in this section is openly available [27].

In the described settings, we perform experiments for varying noise levels δ_ℓ and Lipschitz constants L_m, where we choose

$$\hat{\delta}_\ell = \begin{cases} \left(\frac{1}{3}\right)^{7-\ell} & \text{for } 0 < \ell < 7 \\ 0 & \text{for } \ell = 0. \end{cases} \qquad \ell = 0, \dots, 6 \tag{5.2}$$

$$L_m = 1 - \left(\frac{1}{3}\right)^m, \qquad m = 1, \dots, 5 \tag{5.3}$$

and the resulting noise η is Gaussian noise with standard deviation $\delta_\ell = \hat{\delta}_\ell \cdot \mathrm{std}_{\mathrm{dataset}}$, where $\mathrm{std}_{\mathrm{dataset}}$ denotes the averaged standard deviation of the coefficients $\mathrm{std}_j := \mathrm{std}(\langle x^{(i)}, v_j \rangle_{i=1,\dots,N})$ of the current dataset (i.e. standard deviation with respect to i, mean with respect to j). The corresponding numbers of fixed point iterations for the chosen error bound are $i_{1,2} = 30, i_3 = 100$, $i_4 = 300$ and $i_5 = 800$. In the following, we discuss the results in terms of the learned solutions, the resulting data-dependent filter functions, and the regularization and approximation properties of the models.

5.1. Learned inverse mappings

To compare the characteristics of the approximation and reconstruction training to our theoretical findings, we plot the learned one-dimensional inverse mappings ψ_j in the different components j (corresponding to the singular values σ_j) for the bimodal dataset. We visualize the results in figure 5 for a large and a small eigenvalue of A.

In the case of approximation training, we observe the predicted affine linear behavior in the support of the data distribution, limited by the Lipschitz constraints and aligned to the expected value of the training data. This result is independent of the noise and optimal only for small noise levels. At the boundaries of the data distribution seen during training, the proper behavior of the optimal solution from theorem 3.1 is not learned properly.

For the case of reconstruction training, we can again corroborate our theoretical findings numerically. For this purpose, we compute the posterior expectation $\mathbb{E}(x_j | z_j^\delta)$ for our setting.

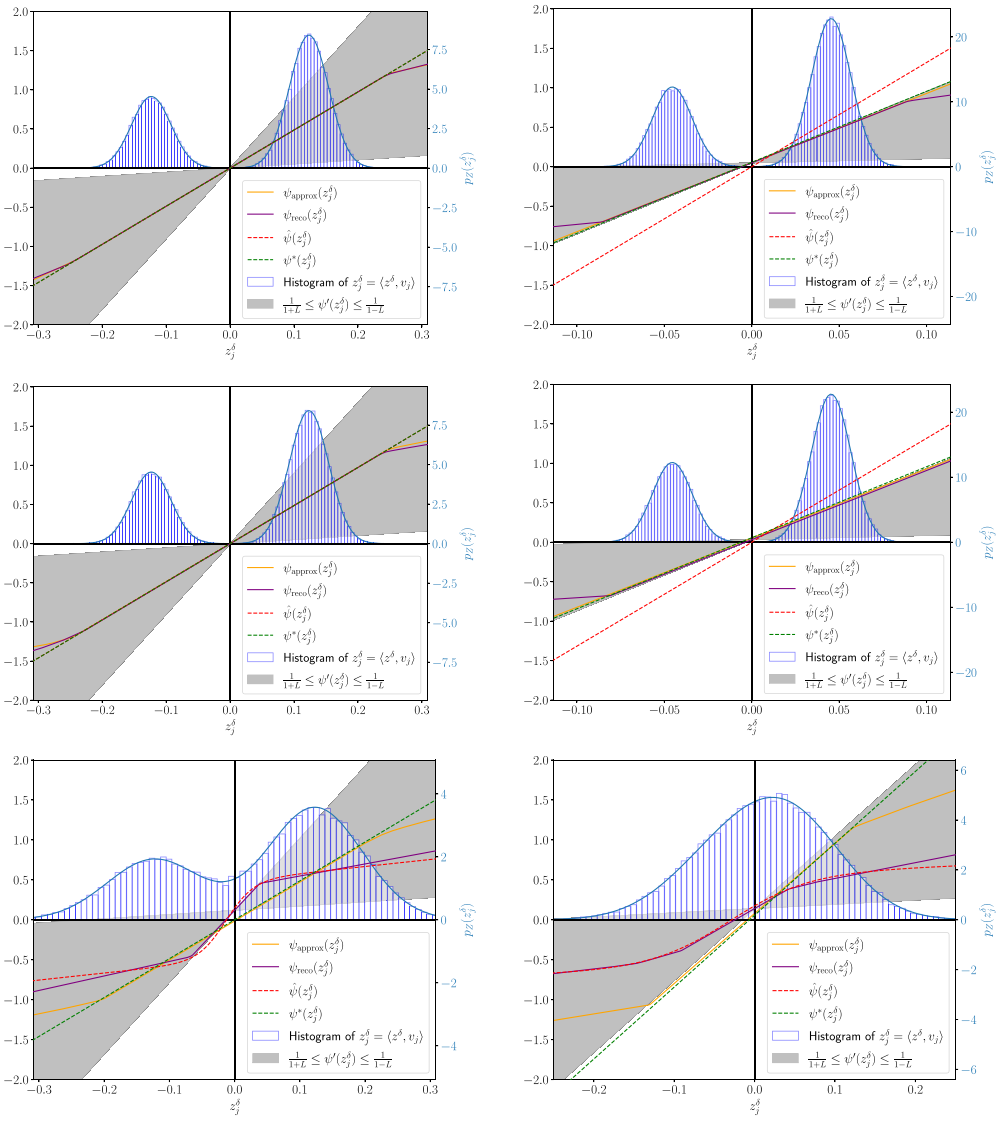
Remark 5.1. For the multimodal Gaussian distribution with pdf p_X and Gaussian noise p_H,

$$p_H(z) = g_{w,0}(z) \tag{5.4}$$

$$p_X(x) = \sum_{k=1}^{K} \rho_k \, g_{v_k, t_k}(x), \tag{5.5}$$

we have

$$p_z(z) = \sum_{k=1}^{K} \rho_k \, g_{u_k, t_k \sigma^2}(z) \tag{5.6}$$



(a) $\sigma_j^2 = 0.205,\ j = 95$ (b) $\sigma_j^2 = 0.075,\ j = 150$

Figure 5. Reconstructions $\psi^*_{\text{approx}}(z_j^\delta)$ trained via approximation training and $\psi^*_{\text{reco}}(z_j^\delta)$ trained via reconstruction training at Lipschitz bound L_2 for different singular values and for noise levels 'zero' (δ_0, *top row*), 'small' (δ_1, *middle row*) and 'large' (δ_5, *bottom row*) for $A = M_a$.

and the posterior expectation $\hat{\psi}$ reads

$$\begin{aligned}\mathbb{E}(x|z) &= \frac{\sum_{k=1}^{K} \frac{\rho_k}{u_k^2} \left(\sigma^2 v_k^2 z + t_k w^2\right) g_{u_k, t_k\sigma^2}(z)}{\sum_{k=1}^{K} \rho_k g_{u_k, t_k\sigma^2}(z)} \\ &= \sigma^2 z \frac{\sum_{k=1}^{K} \frac{v_k^2 \rho_k}{u_k^2} g_{u_k, t_k\sigma^2}(z)}{\sum_{k=1}^{K} \rho_k g_{u_k, t_k\sigma^2}(z)} + w^2 \frac{\sum_{k=1}^{K} \frac{t_k \rho_k}{u_k^2} g_{u_k, t_k\sigma^2}(z)}{\sum_{k=1}^{K} \rho_k g_{u_k, t_k\sigma^2}(z)},\end{aligned} \tag{5.7}$$

where $u_k = \sqrt{w^2 + (\sigma^2 v_k)^2}$. We note that this recovers the linear behavior $\mathbb{E}(x|z) = z/\sigma^2$ in the noise-free case while it adds a correction term in the noisy case that pulls and pushes data points towards more likely results.

For regions within the support of the data distribution, where the constraint (2.6) permits the model to approximate $\mathbb{E}(x_j|z_j^\delta)$, we observe in figure 5 that the learned solutions match well with the posterior expectation. If the model reaches the limiting constraint, it exhausts the possible slope to be as close to the posterior expectation as possible. This results in a much more data-dependent inversion scheme, where reconstructions that were more likely to appear during training are favored. Consequently, the model can compensate for larger noise levels based on additional learned knowledge about the data. The behavior of the learned solution thus coincides with the theoretically founded one in figure 3.

In the case of large noise, the reconstruction-based model regularizes and does not necessarily exhaust the Lipschitz constraint, while the approximation model always tries to fit the operator as well as possible. If noise is absent, the learned mappings coincide for both training strategies.

5.2. Learned filter functions

As a link to classical regularization theory, we visualize the data-dependent filter functions that correspond to the learned models. For this purpose, we evaluate the filter r_L, where

$$(\mathrm{Id} - f_{\theta,j})^{-1}(s(q)) - \hat{b}_{L,j} = r_L\left(\sigma_j^2, s(q)\right) s(q) \text{ for } s \in \mathbb{R}, \tag{5.8}$$

for each singular value σ_j at data points $s(q) := \sigma_j^2(\mu_{X,j} + q \cdot \mathrm{std}_j)$, where we subtract the axis intercept $\hat{b}_{L,j} = (\mathrm{Id} - f_{\theta,j})^{-1}(0)$. The variable $q \in \mathbb{R}$ determines the number of standard deviations std_j away from the mean value $\mu_{X,j} = \frac{1}{N}\sum_{i=1}^{N} \langle x^{(i)}, v_j \rangle$ in the image of the dataset with respect to the fixed singular value of the operator. For simplicity, we define

$$R_L(\sigma_j, q) := r_L\left(\sigma_j^2, s(q)\right) s(q). \tag{5.9}$$

The results for approximation and reconstruction training are visualized as surface plots for both datasets in figures 6 and 11. In all cases, the r_L show a sensible behavior, damping small singular values and roughly satisfying $r_L \to 1$ as $\sigma^2 \to 1$.

On the bimodal dataset in figure 6, data dependency occurs in all filter functions. In the case of approximation training, this is visible only in the sense that proper regularizing behavior is learned exclusively within the support of the data distribution. Reconstruction training shows a more complex dependency since the posterior mean aims to push points that lie close to 0 towards the two peaks of the bimodal data distribution if noise is present in the data, as can also be seen in figure 5. As a result, the medium singular values, in which an amplified slope in the solution is, on the one hand, permitted by the Lipschitz constraint and, on the other hand, also necessary due to the present influence of noise, are elevated near the center of

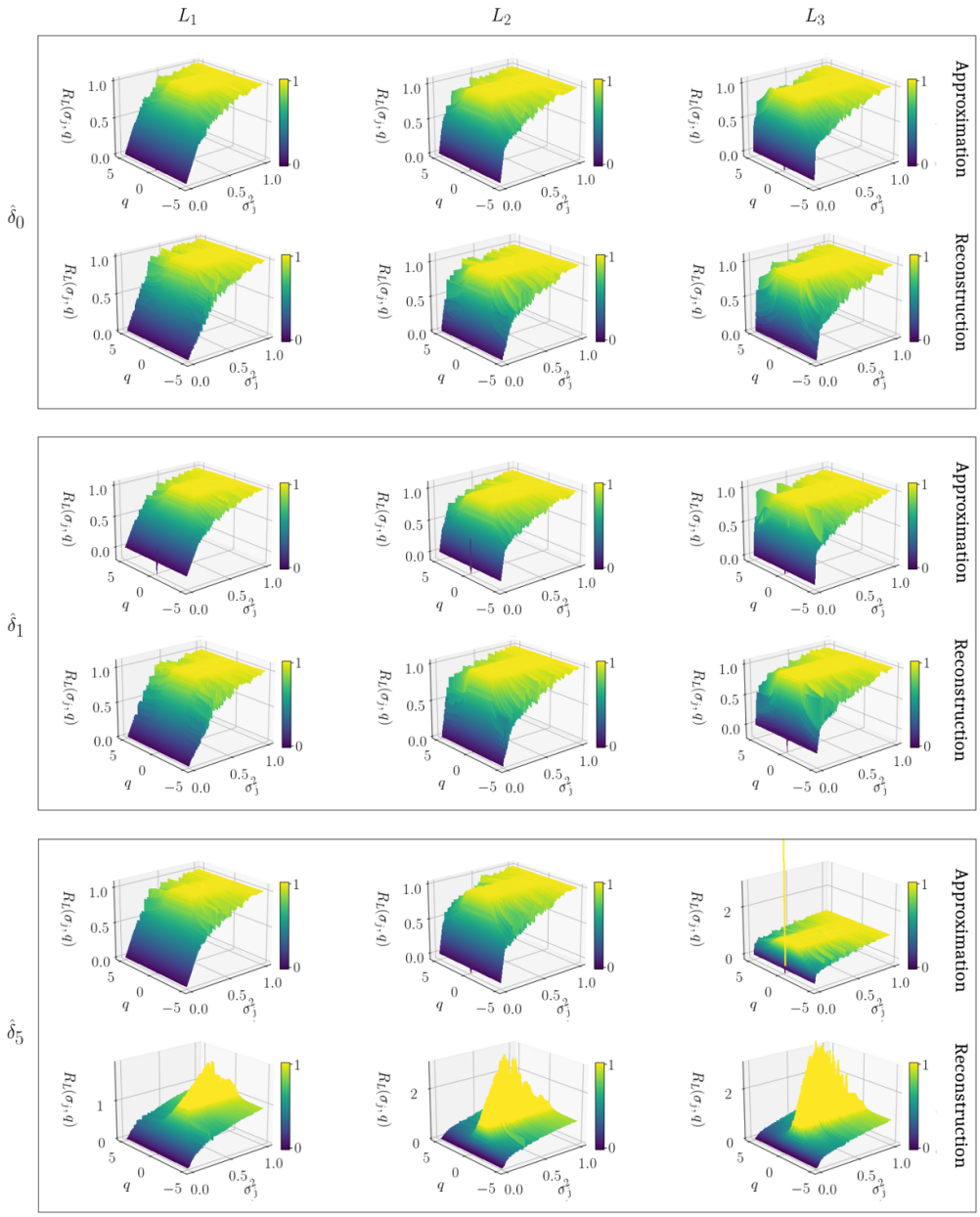


Figure 6. Filter functions $R_L(\sigma_j, q)$ as defined in (5.9) corresponding to trained networks $\varphi_{\theta(L_m),\delta_\ell}$ for $m = 1, 2, 3$ *(columns)* and $\ell = 0, 1, 5$ *(rows)*, trained via approximation training *(top)* and via reconstruction training *(bottom)* on the bimodal dataset for $A = M_a$.

the distribution. This results in filter functions that may not lead to convergent regularization schemes but include data-driven corrections for the observed noise.

On MNIST in figure 11, the stronger data dependence of the reconstruction training is not directly visible; all filters appear to be approximately constant in the range of 5 standard deviations around the mean. This is likely to be due to the fact that the distributions in the singular values are all approximately unimodal. Therefore, the correction could be similar to the simple regularizing behavior of the neural networks in the approximation training approach. In return, the action of L as a regularization parameter becomes visible. Especially for small L, the filter functions show similarities to the expected filter function; see remark 3.1. The learned filter functions in approximation training are very similar for every noise level, which is in line with the developed theory. In contrast, the filters learned in reconstruction training show stronger regularization (i.e. more dampening of small singular values) for larger noise levels, adapting to the data seen during training. This again fits well within the theoretical results developed in section 4 and is especially visible for large $L = L_3$, where the model is, in principle, able to fit the operator well. However, the filter functions for the largest noise $\hat{\delta}_5$ and L_2, L_3 look very similar, indicating that the model learns strong regularization from data at the cost of a worse operator approximation.

5.3. Reconstruction quality and convergence

To compare the performance of the different models in terms of reconstruction quality, we show images and filter functions for a single MNIST digit in figure 7.

Especially for large noise, the visual quality clearly benefits from the reconstruction training compared to the approximation approach. This supports the argument that additional data dependence may be desirable for large-noise applications. In the same case, the approximation training shows its regularization properties and indicates proper parameter choice rules: The image quality improves for smaller Lipschitz constraints, which imply a strong regularization. This behavior is not visible in the reconstruction approach, where a large L generally seems to improve the quality. The filter plots we provide for the given sample in figure 7(right) also underlined this. In this case, they show a similar graph for L_2 and L_3. The filter plots also reveal that the models optimized via approximation training are independent of the noise seen during training and, therefore, learn identical filter functions for all noise levels. Table 1 also reveals an advantage of this method for small noise levels.

Another way to study the convergence in L and δ of the trained models is to evaluate the approximation properties in terms of the overall errors on the dataset with respect to different error measures. To evaluate the localized approximation property that has been introduced in [3, theorem 3.1], we define

$$\mathcal{E}_{\text{mean}}\left(\varphi_{\theta(L)}, A\right) = \frac{1}{N}\sum_{i=1}^{N}\left\|\varphi_{\theta(L)}\left(x^{(i)}\right) - Ax^{(i)}\right\|, \tag{5.10}$$

$$\mathcal{E}_{x^{(m)}}\left(\varphi_{\theta(L)}, A\right) = \left\|\varphi_{\theta(L)}\left(x^{(m)}\right) - Ax^{(m)}\right\|, \tag{5.11}$$

estimating the approximation error of the trained model for the whole dataset or a single sample. In figure 8, we plot this error for the models trained without noise and varying L. In the case of approximation training, we observe slightly superlinear convergence in the dataset on average, indicating that the property is satisfied for many samples. As proven in prior work [3], this implies that the training constructs a convergent regularization scheme for these samples. For reconstruction-based training, this is not fulfilled on average. To preserve some

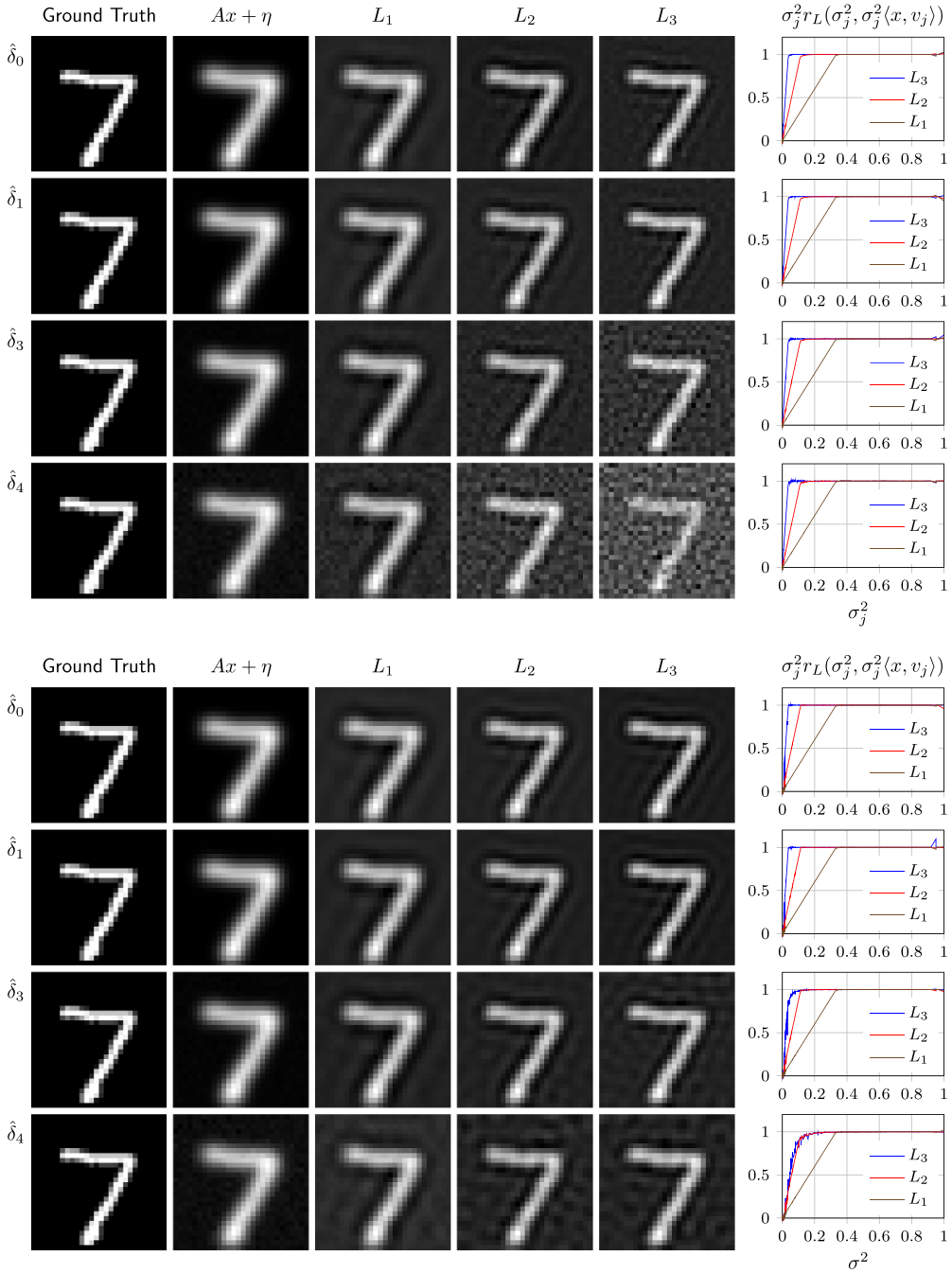


Figure 7. Reconstructions of an MNIST sample $x = x^{(1)}$ from the test dataset by computing $\varphi^{-1}_{\theta(L_m,\delta_\ell)}(Ax+\tilde{\eta})$ with $\tilde{\eta} \sim \mathcal{N}(0, \delta_\ell \mathrm{Id})$ for Lipschitz bounds L_m with $m = 1,2,3$ *(columns)* and noise levels $\delta_\ell = \hat{\delta}_\ell \cdot \mathrm{std}_{\mathrm{MNIST}}$ with $\ell = 0,1,3,4$ *(rows)* together with corresponding filter functions for $A = M_a$. The top subfigure depicts the reconstructions from networks trained via approximation training, and the bottom subfigure corresponds to the networks optimized via reconstruction training.

Table 1. SSIM and MSE measures corresponding to reconstructions of $x^{(1)}$ in figure 7. Bold values indicate the best reconstruction quality with respect to the corresponding error measure for a given noise level.

	SSIM			MSE		
Approximation training	L_1	L_2	L_3	L_1	L_2	L_3
δ_0	0.7401	0.7984	**0.8106**	0.0101	0.0058	**0.0037**
δ_1	0.7377	0.7930	**0.8219**	0.0103	0.0056	**0.0036**
δ_3	0.7261	**0.7455**	0.6575	0.0105	**0.0063**	0.0094
δ_4	**0.6791**	0.6236	0.5085	**0.0112**	0.0132	0.0469
	SSIM			MSE		
Reconstruction training	L_1	L_2	L_3	L_1	L_2	L_3
δ_0	0.7417	0.8046	**0.8255**	0.0102	0.0057	**0.0040**
δ_1	0.7403	0.8051	**0.8240**	0.0103	0.0056	**0.0039**
δ_3	0.7336	0.7774	**0.7961**	0.0105	0.0061	**0.0046**
δ_4	0.7208	**0.7242**	0.7096	0.0106	0.0070	**0.0066**

insights on the convergence of the method, we extend on the weaker but sufficient condition in [3, remark 3.2] and define

$$\widetilde{\mathcal{E}}_{\text{mean}}\left(\varphi_{\theta(L)},A\right)=\frac{1}{N}\sum_{i=1}^{N}\left\|x^{(i)}-\varphi_{\theta(L)}^{-1}\left(Ax^{(i)}\right)\right\| \tag{5.12}$$

$$\widetilde{\mathcal{E}}_{x^{(m)}}\left(\varphi_{\theta(L)},A\right)=\left\|x^{(m)}-\varphi_{\theta(L)}^{-1}\left(Ax^{(m)}\right)\right\|, \tag{5.13}$$

which is closer to the target in reconstruction training. In this case, $\widetilde{\mathcal{E}}_{x^{(m)}}(\varphi_{\theta(L)},A)\xrightarrow{L\to 1}0$ would be sufficient for local convergence. Figure 8 indicates that this property can still be satisfied, however, with slow convergence rates.

In addition, we evaluate the reconstruction error of the training approaches for varying noise levels. Figure 9 depicts the results for the mean squared error

$$\text{MSE}_{\text{reco}}^{\delta_\ell}\left(\varphi_{\theta(L,\delta_\ell)},A\right)=\frac{1}{N}\sum_{i=1}^{N}\left\|x^{(i)}-\varphi_{\theta(L,\delta_\ell)}^{-1}\left(Ax^{(i)}+\eta^{(i)}\right)\right\|^2 \tag{5.14}$$

and averaged structural similarity index measure (SSIM) as defined in [28], computed for the dataset by

$$\text{SSIM}^{\delta_\ell}\left(\varphi_{\theta(L,\delta_\ell)},A\right)=\frac{1}{N}\sum_{i=1}^{N}\text{SSIM}\left(x^{(i)},\varphi_{\theta(L,\delta_\ell)}^{-1}\left(Ax^{(i)}+\eta^{(i)}\right)\right). \tag{5.15}$$

Here, one can again see that the approximation training comes with a typical parameter choice rule known from regularization theory, where one has to choose $L\to 1$ while $\delta\to 0$. In contrast, the reconstruction training performs best with the largest L among all noise levels since the regularization emanates from the data. Overall, the reconstruction training method outperforms the approximation training for all large δ but stays slightly behind for very small noise.

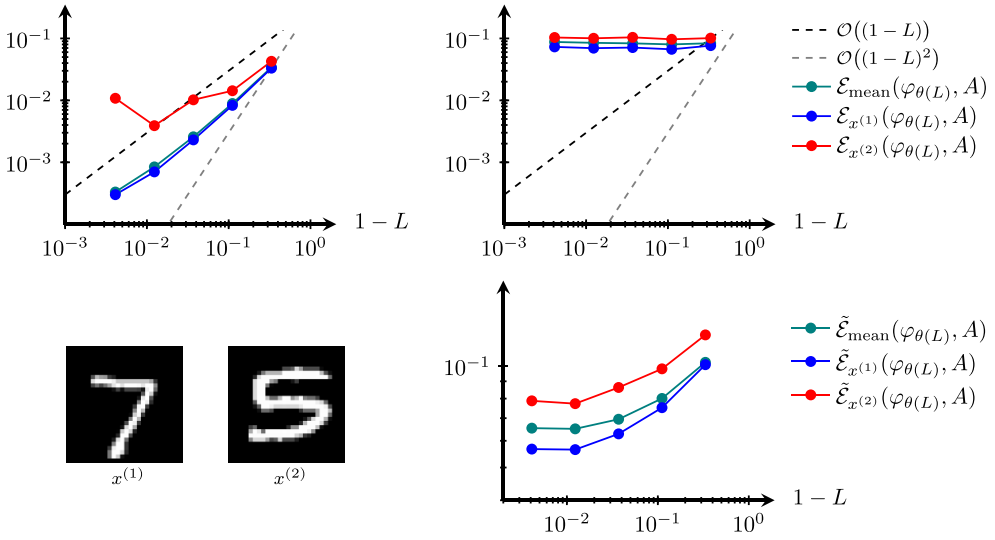


Figure 8. Test samples $x^{(1)}$ and $x^{(2)}$ *(bottom left)*. Evaluations of the local approximation property via $\mathcal{E}_{\text{mean}}(\varphi_{\theta(L_m)},A)$, $\mathcal{E}_{x^{(1)}}(\varphi_{\theta(L_m)},A)$ and $\mathcal{E}_{x^{(2)}}(\varphi_{\theta(L_m)},A)$ for the approximation training *(top left)* and the reconstruction training *(top right)*, and evaluations of the generalized approximation property via $\tilde{\mathcal{E}}_{\text{mean}}(\varphi_{\theta(L_m)},A)$, $\tilde{\mathcal{E}}_{x^{(1)}}(\varphi_{\theta(L_m)},A)$ and $\tilde{\mathcal{E}}_{x^{(2)}}(\varphi_{\theta(L_m)},A)$ for the reconstruction training *(bottom right)* for $L_m = 1 - 1/3^m$ with $m = 1,\dots,5$, $A = M_a$ on the MNIST test dataset.

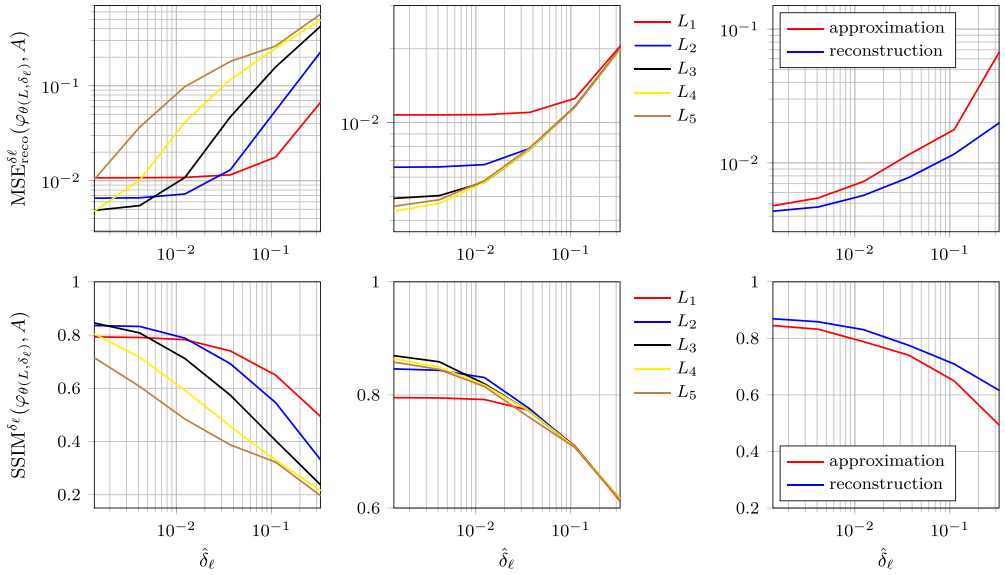


Figure 9. Reconstruction errors $\text{MSE}^{\delta_\ell}_{\text{reco}}(\varphi_{\theta(L,\delta_\ell)},A)$ *(top row)* and $\text{SSIM}^{\delta_\ell}(\varphi_{\theta(L,\delta_\ell)},A)$ *(bottom row)* for networks trained on noisy samples with noise levels $\hat{\delta}_\ell$ for $\ell = 0,\dots,6$ and reconstructions from noisy samples of the same noise level for the approximation training *(left)* and for the reconstruction training *(middle)* with Lipschitz bounds L_m on the MNIST dataset for $A = M_a$. Outcomes of optimal parameter choices for both training strategies over different noise levels can be seen on the right-hand side.

6. Discussion and outlook

The present work can be seen as a continuation of [3]. There, the authors investigated regularization properties of the proposed iResNet reconstruction approach for specific network architectures trained according to the approximation training on samples to impart data dependency. Here, we have extended the theory by focusing on the question of to what extent the training data distribution influences the optimal parameters of the iResNet and, in turn, the resulting reconstruction scheme. To this end, we considered the training of the iResNets from a Bayesian perspective and focused on two different loss functions, namely the approximation training and the reconstruction training.

In the approximation training, our results for the diagonal architecture show that for all possible prior and noise distributions, the best-suited residual function f is an affine linear one whose optimal parameters depend on the mean of the prior distribution, the eigenvalue σ_j^2 and the Lipschitz constant $L < 1$. Thus, in this setting, the data dependency on the training outcome is minimal, with no influence from the noise distribution and, especially, the regularization properties of the resulting reconstruction scheme. Instead, the amount of regularization of φ_θ^{-1} is solely controlled by the Lipschitz constant L, which also becomes apparent by the observed equivalence of φ_θ^{-1} to a convergent filter-based regularization scheme with bias.

In contrast, the prior and noise distributions significantly impact the optimal architecture and parameters when considering the reconstruction training. Here, we realize the network by an iResNet's inverse and train it to approximate A^{-1}, resulting in a stable reconstruction scheme. We showed that we can interpret the optimal network as an approximation of the conditional mean estimator $z^\delta \mapsto \mathbb{E}(x|z^\delta)$ w.r.t. the p_Z-weighted L^2-norm, where p_Z is the density function of $Ax + \eta$. Hence, the optimal architecture choice and the corresponding optimal parameters depend on the prior and the noise distribution. Consequently, this indicates that the amount of regularization of the trained network is controlled by the Lipschitz constant L and possibly by the amount of noise in the training data.

The theoretical findings are validated and further corroborated by a series of numerical investigations on the MNIST dataset and an artificially generated dataset following a bimodal Gaussian distribution for two different forward operators. In particular, the results highlight that in reconstruction training, the noise distribution influences the regularization properties of the network. In the approximation training, the noise does not influence the regularization properties; they are solely controlled by the Lipschitz constant L. As a result, the reconstruction training leads to superior regularizations in high noise regimes, whereas the approximation training is more suitable in low noise regimes due to better convergence properties.

These investigations of the approximation and reconstruction training illustrate how the loss function determines the influence of the prior and noise distribution on the reconstruction scheme and shed light on which architectures are suitable. Investigating a link to MAP estimation and how it could be represented in terms of an iResNet might allow for revealing further links to regularization theory in future works (see also a more detailed discussion in appendix B).

The presented results allow further investigations and can serve as a foundation for future research directions. In the approximation training case, it might be desirable to relax the naive Bayes assumption and to consider the general non-diagonal network case (3.19). In line with the found limited data dependency in the diagonal case, we conjecture a dependency on second moments of the prior distribution p_X at most. In the general network case, we showed that reconstruction training leads to a data-dependent and stable reconstruction scheme that approximates the mean of the posterior distribution, where the degree of stability can be controlled by the Lipschitz constant L. What is left to prove is a convergence property as discussed in

remark 4.2, which could, in principle, further manifest the superiority of the reconstruction training approach and provide additional guarantees. Here, potential generalizations also could incorporate alternative loss functions in the integrand of the reconstruction training loss, which might result in approximations of alternative estimators induced by Bayes costs [12]. In this context, as a starting point, one may limit oneself to linear estimators to further investigate the relation to learned MAP estimators, respectively Tikhonov regularization, as in [2, 11]. In order to obtain convergence guarantees as well as data dependence, when desirable, one can explore a noise-controlled convex combination of both training losses. Theoretical as well as numerical investigations in this direction remain future research.

In addition, remark 2.3 serves as a basis to improve the numerical implementation of the reconstruction training by constructing the network as a scaled iResNet, resulting in a more efficient training approach as mentioned in section 5. This is a reasonable approach when one is not interested in directly comparing the approximation and reconstruction training. Numerical investigations, including the general non-diagonal network architecture, remain immediate future research.

Besides these improvements, extending our results to deeper network architectures, e.g. a concatenation of iResNets, could be beneficial. This would allow for more expressive network architectures and further improve the reconstruction quality of the networks. Finally, it might be worthwhile to generalize the results to nonlinear inverse problems, allowing for an application to a larger number of operators.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://gitlab.informatik.uni-bremen.de/inn4ip/iresnet-regularization.

Acknowledgments

N Heilenkötter, M Iske, and J Nickel acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project number 281474342/GRK2224/2. T Kluth acknowledges support from the DELETO project funded by the Federal Ministry of Education and Research (BMBF, Project Number 05M20LBB).

Appendix A. Proofs

A.1. Proof of lemma 3.1

Proof. For a function f of the form $f(x) = mx + b$ with the constraint $m^2 \leqslant L^2$, we can solve (3.7) by using the Lagrangian

$$K(m,b,\lambda) = \int_{\mathbb{R}} p_X(x) \left| \left(1 - \sigma^2 - m\right) x - b \right|^2 \mathrm{d}x + \lambda \left(m^2 - L^2\right), \tag{A.1}$$

where the integral is well-defined due to the existence of the first and second moments of p_X. The convexity, coercivity, and continuity of the integral term w.r.t. (m, b) imply that a minimizer exists. Therefore, we can calculate the minimizer using the necessary conditions (KKT conditions)

$$\frac{\partial K}{\partial m}(m,b,\lambda) = -\int_{\mathbb{R}} 2p_X(x)\left(\left(1-\sigma^2-m\right)x-b\right)x\,\mathrm{d}x + 2\lambda m \overset{!}{=} 0, \tag{A.2}$$

$$\frac{\partial K}{\partial b}(m,b,\lambda) = -\int_{\mathbb{R}} 2p_X(x)\left(\left(1-\sigma^2-m\right)x-b\right)\mathrm{d}x \overset{!}{=} 0, \tag{A.3}$$

$$\lambda\left(m^2-L^2\right) \overset{!}{=} 0, \tag{A.4}$$

$$\lambda \geqslant 0. \tag{A.5}$$

Rearranging (A.2) for m and (A.3) for b leads to

$$m = \frac{\int_{\mathbb{R}} 2p_X(x)\left(\left(1-\sigma^2\right)x-b\right)x\,\mathrm{d}x}{\int_{\mathbb{R}} 2p_X(x)x^2\,\mathrm{d}x + 2\lambda} = \frac{\left(1-\sigma^2\right)\mathbb{E}_{p_X}\left(x^2\right)-b\mu_X}{\mathbb{E}_{p_X}\left(x^2\right)+\lambda}, \tag{A.6}$$

$$b = \frac{\int_{\mathbb{R}} 2p_X(x)\left(1-\sigma^2-m\right)x\,\mathrm{d}x}{\int_{\mathbb{R}} 2p_X(x)\,\mathrm{d}x} = \left(1-\sigma^2-m\right)\mu_X, \tag{A.7}$$

where we use the abbreviated notation $\mathbb{E}_{p_X}$ for $\mathbb{E}_{x\sim p_X}$. Now, plugging $b=(1-\sigma^2-m)\mu_X$ into the equation for m implies

$$\begin{aligned}
& m = \frac{\left(1-\sigma^2\right)\mathbb{E}_{p_X}\left(x^2\right)-\left(1-\sigma^2-m\right)\mu_X^2}{\mathbb{E}_{p_X}\left(x^2\right)+\lambda} \\
\Leftrightarrow \quad & \left(1-\frac{\mu_X^2}{\mathbb{E}_{p_X}\left(x^2\right)+\lambda}\right)m = \frac{\left(1-\sigma^2\right)\left(\mathbb{E}_{p_X}\left(x^2\right)-\mu_X^2\right)}{\mathbb{E}_{p_X}\left(x^2\right)+\lambda} \\
\Leftrightarrow \quad & \frac{\mathrm{Var}_{p_X}(x)+\lambda}{\mathbb{E}_{p_X}\left(x^2\right)+\lambda}m = \frac{\left(1-\sigma^2\right)\mathrm{Var}_{p_X}(x)}{\mathbb{E}_{p_X}\left(x^2\right)+\lambda} \\
\Leftrightarrow \quad & m = \left(1-\sigma^2\right)\frac{\mathrm{Var}_{p_X}(x)}{\mathrm{Var}_{p_X}(x)+\lambda}.
\end{aligned} \tag{A.8}$$

Since $1-\sigma^2 > L$ holds by assumption, we need $\lambda > 0$ to ensure $m \leqslant L$. Then, (A.4) directly implies $m = L$. As m is uniquely determined we also know that b is unique with $b = (1-\sigma^2-L)\,\mu_X$. □

A.2. Proof of lemma 4.1

Proof. We denote the objective function by $F\colon L^2_{p_Z}(X,X) \to [0,\infty)$,

$$F(\psi) = \int_X \int_X p_X(x)\,p_H\left(z^\delta - Ax\right)\,\|\psi\left(z^\delta\right)-x\|^2\,\mathrm{d}z^\delta\,\mathrm{d}x, \tag{A.9}$$

which coincides with (4.3) with the substitution $z^\delta = Ax+\eta$. Note that for all $\psi \in L^2_{p_Z}(X,X)$, it holds $F(\psi) < \infty$ since

$$\|\psi\left(z^\delta\right)-x\|^2 = \|\psi\left(z^\delta\right)\|^2 - 2\langle\psi\left(z^\delta\right),x\rangle + \|x\|^2 \leqslant 2\|\psi\left(z^\delta\right)\|^2 + 2\|x\|^2 \tag{A.10}$$

and the integrals

$$\int_X \int_X p_X(x)\, p_H\left(z^\delta - Ax\right) \|\psi(z)\|^2 \,\mathrm{d}z^\delta \,\mathrm{d}x = \int_X p_Z\left(z^\delta\right) \|\psi\left(z^\delta\right)\|^2 = \|\psi\|^2_{p_Z,2},$$

$$\begin{aligned}\int_X \int_X p_X(x)\, p_H\left(z^\delta - Ax\right) \|x\|^2 \,\mathrm{d}z^\delta &= \int_X p_X(x)\, \|x\|^2 \int_X p_H\left(z^\delta - Ax\right) \mathrm{d}z^\delta \,\mathrm{d}x \\ &= \int_X p_X(x)\, \|x\|^2 \,\mathrm{d}x \end{aligned} \tag{A.11}$$

are both finite.

Besides, note that F is convex w.r.t. ψ since $\psi \mapsto \|\psi(z^\delta) - x\|^2$ is convex for any $x, z^\delta \in X$. Thus, we can find the minimizer of F by setting its derivative to zero.

To compute the derivative of F, we consider

$$\begin{aligned} F(\psi + h) &= \int_X \int_X p_X(x)\, p_H\left(z^\delta - Ax\right) \|\psi\left(z^\delta\right) + h\left(z^\delta\right) - x\|^2 \,\mathrm{d}z^\delta \,\mathrm{d}x \\ &= \int_X \int_X p_X(x)\, p_H\left(z^\delta - Ax\right) \|\psi\left(z^\delta\right) - x\|^2 \,\mathrm{d}z^\delta \,\mathrm{d}x \\ &\quad + 2\int_X \int_X p_X(x)\, p_H\left(z^\delta - Ax\right) \langle \psi\left(z^\delta\right) - x, h\left(z^\delta\right)\rangle \,\mathrm{d}z^\delta \,\mathrm{d}x \\ &\quad + \int_X \int_X p_X(x)\, p_H\left(z^\delta - Ax\right) \|h\left(z^\delta\right)\|^2 \,\mathrm{d}z^\delta \,\mathrm{d}x. \end{aligned} \tag{A.12}$$

The first of the three summands is $F(\psi)$, the last one equals $\|h\|^2_{p_Z,2}$ and the second summand equals $\partial F(\psi)h$. Note that the second summand is finite since $F(\psi)$ and $\|h\|^2_{p_Z,2}$ are both finite. Using Fubini's theorem, we obtain

$$\begin{aligned} \partial F(\psi)\, h &= 2\int_X \int_X p_X(x)\, p_H\left(z^\delta - Ax\right) \langle \psi\left(z^\delta\right) - x, h\left(z^\delta\right)\rangle \,\mathrm{d}z^\delta \,\mathrm{d}x \\ &= 2\int_X \left\langle \int_X p_X(x)\, p_H\left(z^\delta - Ax\right) \left(\psi\left(z^\delta\right) - x\right) \mathrm{d}x, h\left(z^\delta\right) \right\rangle \mathrm{d}z^\delta. \end{aligned} \tag{A.13}$$

We are looking for $\hat{\psi}$ such that $\partial F(\hat{\psi})h = 0$ for any $h \in L^2_{p_Z}(X,X)$. Hence, the fundamental lemma of the calculus of variations implies

$$\begin{aligned} & \int_X p_X(x)\, p_H\left(z^\delta - Ax\right) \left(\hat{\psi}\left(z^\delta\right) - x\right) \mathrm{d}x = 0 \\ \Leftrightarrow \quad & p_Z\left(z^\delta\right) \cdot \hat{\psi}\left(z^\delta\right) = \int_X p_X(x)\, p_H\left(z^\delta - Ax\right) x \,\mathrm{d}x \\ \Leftrightarrow \quad & \hat{\psi}\left(z^\delta\right) = \frac{\int_X p_X(x)\, p_H\left(z^\delta - Ax\right) x \,\mathrm{d}x}{p_Z\left(z^\delta\right)} \end{aligned} \tag{A.14}$$

for almost all $z^\delta \in \mathrm{supp}(p_Z)$. According to Bayes' formula

$$p\left(x|z^\delta\right) = \frac{p\left(z^\delta|x\right) p_X(x)}{p_Z\left(z^\delta\right)} = \frac{p_H\left(z^\delta - Ax\right) p_X(x)}{p_Z\left(z^\delta\right)}, \tag{A.15}$$

$\hat{\psi}(z^\delta)$ is the expected value of the posterior density function $p(x|z^\delta)$.

Finally, we need to make sure that $\hat{\psi}(z^\delta) = \mathbb{E}(x|z^\delta)$ is an $L^2_{p_Z}$-function. For this, we use Jensen's inequality for conditional expectations [14, theorem 8.20, corollary 8.21] and obtain

$$\|\mathbb{E}\left(x|z^\delta\right)\|^2 = \sum_{j=1}^{n} |\mathbb{E}\left(x_j|z^\delta\right)|^2 \leqslant \sum_{j=1}^{n} \mathbb{E}\left(|x_j|^2 \,|\, z^\delta\right) = \mathbb{E}\left(\|x\|^2 \,|\, z^\delta\right). \tag{A.16}$$

Then, by the definition of conditional expectations, we get

$$\int_X p_Z\left(z^\delta\right) \mathbb{E}\left(\|x\|^2 \,|\, z^\delta\right) \mathrm{d}z^\delta = \mathbb{E}_{p_Z}\left(\mathbb{E}\left(\|x\|^2 \,|\, z^\delta\right)\right) = \mathbb{E}\left(\|x\|^2\right), \tag{A.17}$$

which is finite, and the proof is complete. □

A.3. Proof of lemma 4.3

Proof. Due to (iii), it immediately follows that p_X is uniformly continuous. From (iii), we also deduce that the marginal of p_X on $\mathcal{N}(A)^\perp$, $p_{X,\mathcal{N}(A)^\perp} : \mathcal{N}(A)^\perp \to \mathbb{R}_{\geqslant 0}$ with

$$p_{X,\mathcal{N}(A)^\perp}(x) = \int_{\mathcal{N}(A)} p_X(x_0 + x)\ \mathrm{d}x_0 \tag{A.18}$$

is compactly supported and bounded and thus also uniformly continuous. Analogously, we can deduce uniform continuity of the mappings $g_0 : \mathcal{N}(A)^\perp \to \mathcal{N}(A)$ and $g_\dagger : \mathcal{N}(A)^\perp \to \mathcal{N}(A)^\perp$ with

$$g_0(x) = \int_{\mathcal{N}(A)} p_X(x_0 + x)\, x_0\ \mathrm{d}x_0 \tag{A.19}$$

and

$$g_\dagger(x) = \int_{\mathcal{N}(A)} p_X(x_0 + x)\ \mathrm{d}x_0\, x. \tag{A.20}$$

Next, by using $X = \mathcal{N}(A) \oplus \mathcal{N}(A)^\perp$ we now observe that for arbitrary $z \in X$ due to (i) it holds

$$\begin{aligned} p_{Z,\delta}(z) &= \int_X p_{H,\delta}(z - Ax)\, p_X(x)\ \mathrm{d}x \\ &= \int_{\mathcal{N}(A)^\perp} p_{H,\delta}(z - Ax_1) \int_{\mathcal{N}(A)} p_X(x_0 + x_1)\ \mathrm{d}x_0\, \mathrm{d}x_1 \\ &= p^0_{H,\delta}\left(P_{\mathcal{N}(A)} z\right) \int_{\mathcal{R}(A)} p^\dagger_{H,\delta}\left(P_{\mathcal{N}(A)^\perp} z - y_1\right) \int_{\mathcal{N}(A)} p_X\left(x_0 + A^\dagger y_1\right)\ \mathrm{d}x_0\, \mathrm{d}y_1 |\det A^\dagger|, \end{aligned} \tag{A.21}$$

where the transformation $x_1 = A^\dagger y_1$, with generalized inverse $A^\dagger : \mathcal{R}(A) \to \mathcal{N}(A)^\perp$, is used in the last equality. Analogously, we further obtain for arbitrary $z \in \mathrm{supp}(p_{Z,\delta}) \subset X$

$$\begin{aligned}\hat{\psi}_\delta(z) &= \frac{1}{p_{Z,\delta}(z)} \int_X p_{H,\delta}(z - Ax)\, p_X(x)\, x \,\mathrm{d}x \\ &= \frac{1}{p_{Z,\delta}(z)} \int_{\mathcal{N}(A)^\perp} p_{H,\delta}(z - Ax_1) \int_{\mathcal{N}(A)} p_X(x_0 + x_1)\,(x_0 + x_1)\ \mathrm{d}x_0\, \mathrm{d}x_1 \\ &= \frac{p^0_{H,\delta}\left(P_{\mathcal{N}(A)} z\right)}{p_{Z,\delta}(z)} \int_{\mathcal{R}(A)} p^\dagger_{H,\delta}\left(P_{\mathcal{N}(A)^\perp} z - y_1\right) \\ &\qquad \int_{\mathcal{N}(A)} p_X\left(x_0 + A^\dagger y_1\right)\left(x_0 + A^\dagger y_1\right)\ \mathrm{d}x_0\, \mathrm{d}y_1 |\det A^\dagger|. \end{aligned} \tag{A.22}$$

Exploiting the previously derived representation of $p_{Z,\delta}$ and (i) yields

$$\begin{aligned}\hat{\psi}_\delta(z) &= \frac{\int_{\mathcal{R}(A)} p^\dagger_{H,\delta}\left(P_{\mathcal{N}(A)^\perp} z - y_1\right)\left(\int_{\mathcal{N}(A)} p_X\left(x_0 + A^\dagger y_1\right)\left(x_0 + A^\dagger y_1\right)\ \mathrm{d}x_0\right)\ \mathrm{d}y_1}{\int_{\mathcal{R}(A)} p^\dagger_{H,\delta}\left(P_{\mathcal{N}(A)^\perp} z - y_1\right) \int_{\mathcal{N}(A)} p_X\left(x_0 + A^\dagger y_1\right)\ \mathrm{d}x_0\, \mathrm{d}y_1} \\ &= \frac{\int_{\mathcal{R}(A)} p^\dagger_{H,\delta}\left(P_{\mathcal{N}(A)^\perp} z - y_1\right)\left(g_0\left(A^\dagger y_1\right) + g_\dagger\left(A^\dagger y_1\right)\right)\ \mathrm{d}y_1}{\int_{\mathcal{R}(A)} p^\dagger_{H,\delta}\left(P_{\mathcal{N}(A)^\perp} z - y_1\right) p_{X,\mathcal{N}(A)^\perp}\left(A^\dagger y_1\right)\ \mathrm{d}y_1}. \end{aligned} \tag{A.23}$$

We now consider the denominator and nominator separately. In the denominator we first observe that the mapping $p_{X,\mathcal{N}(A)^\perp}(A^\dagger y_1)$ is also uniformly continuous due to the continuity of $A^\dagger$. Exploiting $\mathcal{R}(A) = \mathcal{R}(A^*) = \mathcal{N}(A)^\perp$, the approximation property [15, sec II] of the Dirac sequence $p^\dagger_{H,\delta}$ delivers uniform convergence of the denominator, i.e. this implies pointwise convergence such that for any $z \in X$ it holds

$$\int_{\mathcal{R}(A)} p^\dagger_{H,\delta}\left(P_{\mathcal{N}(A)^\perp} z - y_1\right) p_{X,\mathcal{N}(A)^\perp}\left(A^\dagger y_1\right)\ \mathrm{d}y_1 \underset{\delta\to 0}{\longrightarrow} p_{X,\mathcal{N}(A)^\perp}\left(A^\dagger P_{\mathcal{N}(A)^\perp} z\right). \tag{A.24}$$

Analogous arguments apply to the nominator of (A.23) by exploiting the approximation property of the Dirac sequence implying uniform convergence and thus pointwise convergence such that for any $z \in X$, it holds

$$\begin{aligned}&\int_{\mathcal{R}(A)} p^\dagger_{H,\delta}\left(P_{\mathcal{N}(A)^\perp} z - y_1\right)\left(g_0\left(A^\dagger y_1\right) + g_\dagger\left(A^\dagger y_1\right)\right)\ \mathrm{d}y_1 \\ &\quad \underset{\delta\to 0}{\longrightarrow} g_0\left(A^\dagger P_{\mathcal{N}(A)^\perp} z\right) + g_\dagger\left(A^\dagger P_{\mathcal{N}(A)^\perp} z\right). \end{aligned} \tag{A.25}$$

Due to (iv) for any fixed $z \in \mathcal{R}_{p_X}(A)$ the representation of $\hat{\psi}_\delta$ in (A.23) is well-defined for sufficiently small δ. Consequently, we have convergent sequences in the nominator and in the denominator such that the quotient converges by standard sequence arguments. As $\mathcal{R}_{p_X}(A) \subset \mathcal{R}(A) = \mathcal{N}(A)^\perp$ we thus obtain the desired pointwise convergence

$$\hat{\psi}_\delta(z) \underset{\delta\to 0}{\longrightarrow} \frac{g_\dagger\left(A^\dagger z\right) + g_0\left(A^\dagger z\right)}{p_{X,\mathcal{N}(A)^\perp}\left(A^\dagger z\right)} = A^\dagger z + \int_{\mathcal{N}(A)} \frac{p_X\left(x_0 + A^\dagger z\right)}{\int_{\mathcal{N}(A)} p_X\left(x_0' + A^\dagger z\right)\ \mathrm{d}x_0'} x_0 \,\mathrm{d}x_0. \tag{A.26}$$

□

A.4. Proof of lemma 4.4

Proof. We solve the minimization problem (4.28) by applying the KKT conditions. To this end, consider the Lagrange functional

$$K(m,b,\lambda_1,\lambda_2) = \int_{\mathbb{R}}\int_{\mathbb{R}} p_X(x)\, p_H(\eta)\, \left(m\left(\sigma^2 x+\eta\right)+b-x\right)^2 \mathrm{d}\eta\,\mathrm{d}x \\ + \lambda_1\left(m-\frac{1}{1-L}\right)+\lambda_2\left(\frac{1}{1+L}-m\right) \tag{A.27}$$

for $m,b,\lambda_1,\lambda_2\in\mathbb{R}$. Observe that the integral is well-defined as the first and second moments of p_X and p_H exist by assumption. Moreover, K is convex and coercive w.r.t. (m,b) and the integral is continuous w.r.t. (m,b). Consequently, there exists a minimizer of problem (4.28) which satisfies the necessary KKT conditions

$$\frac{\partial K}{\partial m}(m,b,\lambda_1,\lambda_2) = \int_{\mathbb{R}}\int_{\mathbb{R}} 2p_X(x)\, p_H(\eta)\, \left(\sigma^2 x+\eta\right)\left(m\left(\sigma^2 x+\eta\right)+b-x\right)\mathrm{d}\eta\,\mathrm{d}x \\ + \lambda_1 m-\lambda_2 m \overset{!}{=} 0, \tag{A.28}$$

$$\frac{\partial K}{\partial b}(m,b,\lambda_1,\lambda_2) = \int_{\mathbb{R}}\int_{\mathbb{R}} 2p_X(x)\, p_H(\eta)\, \left(m\left(\sigma^2 x+\eta\right)+b-x\right)\mathrm{d}\eta\,\mathrm{d}x \overset{!}{=} 0, \tag{A.29}$$

$$\lambda_1\left(m-\frac{1}{1-L}\right)\overset{!}{=}0, \tag{A.30}$$

$$\lambda_2\left(\frac{1}{1+L}-m\right)\overset{!}{=}0, \tag{A.31}$$

$$\lambda_1,\lambda_2\geqslant 0. \tag{A.32}$$

Case $\lambda_1=\lambda_2=0$: equation (A.28) implies

$$\begin{aligned}
&\int_{\mathbb{R}}\int_{\mathbb{R}} p_X(x)\, p_H(\eta)\, \left(\sigma^2 x+\eta\right)\left(m\left(\sigma^2 x+\eta\right)+b-x\right)\mathrm{d}\eta\,\mathrm{d}x \\
&\quad= \int_{\mathbb{R}}\int_{\mathbb{R}} p_X(x)\, p_H(\eta)\, \left(m\sigma^4x^2+2m\sigma^2x\eta+m\eta^2+b\sigma^2x+b\eta-\sigma^2x^2-\eta x\right)\mathrm{d}\eta\,\mathrm{d}x \\
&\quad= m\sigma^4\mathbb{E}_{p_X}\left(x^2\right)+2m\sigma^2\mu_X\mu_H+m\mathbb{E}_{p_H}\left(\eta^2\right)+b\sigma^2\mu_X+b\mu_H-\sigma^2\mathbb{E}_{p_X}\left(x^2\right)-\mu_H\mu_X \\
&\quad= m\sigma^4\mathbb{E}_{p_X}\left(x^2\right)+m\mathbb{E}_{p_H}\left(\eta^2\right)+b\sigma^2\mu_X-\sigma^2\mathbb{E}_{p_X}\left(x^2\right) \\
&\quad\overset{!}{=} 0
\end{aligned} \tag{A.33}$$

and equation (A.29) gives

$$\int_{\mathbb{R}}\int_{\mathbb{R}} 2p_X(x)p_H(\eta)\left(m\left(\sigma^2x+\eta\right)+b-x\right)\mathrm{d}\eta\,\mathrm{d}x = m\sigma^2\mu_X+m\mu_H+b-\mu_X\overset{!}{=}0 \tag{A.34}$$

resulting in

$$b = \mu_X - m\sigma^2\mu_X. \tag{A.35}$$

Inserting the last equation into equation (A.33) yields

$$\begin{aligned} & m\sigma^4\mathbb{E}_{p_X}\left(x^2\right) + m\mathbb{E}_{p_H}\left(\eta^2\right) + \sigma^2\mu_X^2 - m\sigma^4\mu_X^2 - \sigma^2\mathbb{E}_{p_X}\left(x^2\right) \overset{!}{=} 0 \\ \Leftrightarrow \quad & m\sigma^4\mathrm{Var}_{p_X}(x) + m\mathrm{Var}_{p_H}(\eta) - \sigma^2\mathrm{Var}_{p_X}(x) = 0 \\ \Leftrightarrow \quad & m = \frac{\sigma^2\mathrm{Var}_{p_X}(x)}{\sigma^4\mathrm{Var}_{p_X}(x) + \mathrm{Var}_{p_H}(\eta)} \end{aligned} \tag{A.36}$$

and

$$b = \left(1 - \frac{\sigma^4\mathrm{Var}_{p_X}(x)}{\sigma^4\mathrm{Var}_{p_X}(x) + \mathrm{Var}_{p_H}(\eta)}\right)\mu_X. \tag{A.37}$$

The formulas for m and b correspond to the unconstrained solution of problem (4.28). In order to satisfy the constraint on m, we need to guarantee that

$$\frac{1}{1+L} \leqslant \frac{\sigma^2\mathrm{Var}_{p_X}(x)}{\sigma^4\mathrm{Var}_{p_X}(x) + \mathrm{Var}_{p_H}(\eta)} \leqslant \frac{1}{1-L}. \tag{A.38}$$

If this is not satisfied, we must require $\lambda_1 > 0$ or $\lambda_2 > 0$, which we will deal with in the following paragraphs.
Case $\lambda_1 > 0$, $\lambda_2 = 0$: If $\lambda_1 > 0$, equation (A.30) directly yields

$$m = \frac{1}{1-L}. \tag{A.39}$$

For b, we again obtain

$$b = \mu_X - m\sigma^2\mu_X = \left(1 - \frac{\sigma^2}{1-L}\right)\mu_X \tag{A.40}$$

as equation (A.29) is independent of λ_1 and λ_2. Furthermore, equation (A.28) implies

$$\begin{aligned} & \int_{\mathbb{R}}\int_{\mathbb{R}} p_X(x)\, p_H(\eta)\left(\sigma^2 x + \eta\right)\left(m\left(\sigma^2 x + \eta\right) + b - x\right)\mathrm{d}\eta\,\mathrm{d}x + \lambda_1 m \overset{!}{=} 0 \\ \Leftrightarrow \quad & \lambda_1 = 2\,\frac{\sigma^2\mathrm{Var}_{p_X}(x) - m\sigma^4\mathrm{Var}_{p_X}(x) - m\mathrm{Var}_{p_H}(\eta)}{m}. \end{aligned} \tag{A.41}$$

In combination with the condition $\lambda_1 > 0$ we now obtain

$$\frac{\sigma^2\mathrm{Var}_{p_X}(x)}{\sigma^4\mathrm{Var}_{p_X}(x) + \mathrm{Var}_{p_H}(\eta)} > \frac{1}{1-L}. \tag{A.42}$$

Case $\lambda_1 = 0$, $\lambda_2 > 0$: The same line of reasoning as in the preceding case yields

$$m = \frac{1}{1+L} \tag{A.43}$$

$$b = \mu_X - m\sigma^2\mu_X = \left(1 - \frac{\sigma^2}{1+L}\right)\mu_X \tag{A.44}$$

$$\frac{\sigma^2 \mathrm{Var}_{p_X}(x)}{\sigma^4 \mathrm{Var}_{p_X}(x) + \mathrm{Var}_{p_H}(\eta)} < \frac{1}{1+L}. \tag{A.45}$$

Observe that the case $\lambda_1 > 0$ and $\lambda_2 > 0$ is not possible as there is no m satisfying equations (A.30) and (A.31) simultaneously for $0 < L < 1$. The uniqueness of the solution directly follows from the uniqueness of m and b, which concludes the proof. □

Appendix B. MAP estimation and its interpretation as an iResNet

Consider a linear inverse problem $\tilde{A}x = y$ with $\tilde{A} \in L(X, Y)$ as in remark 2.1, where noisy data $y^\delta \in Y$ is given. One established approach in Bayesian inverse problems is the so-called MAP (maximum *a posteriori*) estimator

$$x_{\mathrm{MAP}} = \arg\max_{x \in X} p\left(x|y^\delta\right). \tag{B.1}$$

The posterior density $p(x|y^\delta)$ can be derived via Bayes rule from the pdf's of the prior ($x \sim p_X$), the noise ($y^\delta - Ax \sim \tilde{p}_H$) and the data ($y^\delta \sim p_Y$). Using this and the monotonicity of the logarithm, one obtains

$$\begin{aligned} x_{\mathrm{MAP}} &= \arg\max_{x \in X} \frac{\tilde{p}_H\left(y^\delta - \tilde{A}x\right) p_X(x)}{p_Y\left(y^\delta\right)} \\ \Leftrightarrow x_{\mathrm{MAP}} &= \arg\max_{x \in X} \tilde{p}_H\left(y^\delta - \tilde{A}x\right) p_X(x) \\ \Leftrightarrow x_{\mathrm{MAP}} &= \arg\min_{x \in X} -\log\left(\tilde{p}_H\left(y^\delta - \tilde{A}x\right)\right) - \log\left(p_X(x)\right). \end{aligned} \tag{B.2}$$

Here, one can observe a well-known similarity to variational regularization schemes. In the case of Gaussian noise with noise level $\delta > 0$, i.e.

$$\tilde{p}_H(\tilde{\eta}) \propto \exp\left(-\frac{1}{2\delta^2}\|\tilde{\eta}\|^2\right) \tag{B.3}$$

it holds

$$x_{\mathrm{MAP}} = \arg\min_{x \in X} -\frac{1}{2}\|\tilde{A}x - y^\delta\|^2 - \delta^2 \log\left(p_X(x)\right). \tag{B.4}$$

The negative log-likelihood (NLL) $-\log p_X$ can be interpreted as a penalty term, weighted with the squared noise level δ^2. If $-\log p_X$ is differentiable, we can use the first-order optimality condition and derive

$$\begin{aligned} 0 &= \tilde{A}^*\left(\tilde{A}x_{\mathrm{MAP}} - y^\delta\right) - \delta^2 \partial\left(\log p_X\right)\left(x_{\mathrm{MAP}}\right) \\ \Rightarrow \tilde{A}^* y^\delta &= \tilde{A}^*\tilde{A}x_{\mathrm{MAP}} - \delta^2 \partial\left(\log p_X\right)\left(x_{\mathrm{MAP}}\right) \\ \Rightarrow x_{\mathrm{MAP}} &= \left(\mathrm{Id} - \left(\mathrm{Id} - \tilde{A}^*\tilde{A}x_{\mathrm{MAP}} + \delta^2 \partial\left(\log p_X\right)\right)\right)^{-1}\left(\tilde{A}^* y^\delta\right), \end{aligned} \tag{B.5}$$

where the last implication only holds if $\tilde{A}^*\tilde{A} - \delta^2\partial(\log p_X)$ is invertible (which is guaranteed, e.g. in case of a convex NLL).

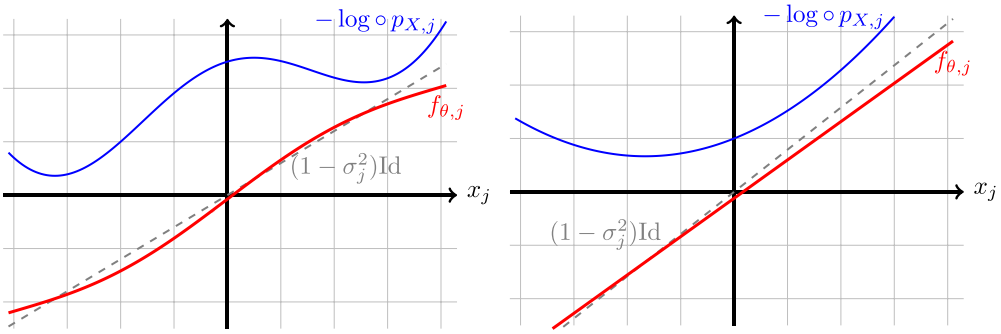


Figure 10. MAP estimation with iResNet. For large singular values $\sigma_j^2 > 1-L$, the NLL is not very restricted (can be nonconvex) to guarantee $\mathrm{Lip}(f_{\theta,j}) \leqslant L < 1$. For small singular values $\sigma_j^2 \leqslant 1-L$, the NLL has to be (strongly) convex to guarantee that the slope of $f_{\theta,j}$ is smaller than $1-\sigma_j^2$ (and smaller than L).

Now, we can interpret (B.5) as an iResNet approach for solving the inverse problem

$$Ax = z \tag{B.6}$$

where $A = \tilde{A}^*\tilde{A}$ and $z = \tilde{A}^* y$. It holds $x_{\mathrm{MAP}} = \varphi_\theta^{-1}(z^\delta) = (\mathrm{Id} - f_\theta)^{-1}(\tilde{A}^* y^\delta)$ if the residual layer is given by

$$f_\theta = \mathrm{Id} - A + \delta^2 \partial(\log p_X). \tag{B.7}$$

Thus, MAP estimation with an iResNet is possible as long as the above-defined f_θ has a Lipschitz constant of at most $L < 1$.

We can derive conditions for the prior p_X and the noise level δ from this Lipschitz constraint by making use of assumption 3.1, i.e. stochastic independence of the components $x_j \sim p_{X,j}$ and the eigendecomposition of A. In this setting, the components can be handled separately and, thus,

$$f_{\theta,j} = \left(1-\sigma_j^2\right)\mathrm{Id} + \delta^2 \partial(\log p_{X,j}). \tag{B.8}$$

To obtain further insights, we distinguish between large eigenvalues (i.e. $\sigma_j^2 > 1-L$) and small ones (i.e. $\sigma_j^2 \leqslant 1-L$).

Remark B.1. The prior $p_{X,j}$ corresponding to a large eigenvalue can have a rather arbitrary character. The only important property is that the derivative of the NLL (i.e. $\partial(\log p_{X,j})$) must be Lipschitz continuous. Because in this case, for small enough δ, it holds

$$\mathrm{Lip}\left(f_{\theta,j}\right) \leqslant \left(1-\sigma^2\right) + \delta^2 \mathrm{Lip}\left(\partial(\log p_{X,j})\right) \leqslant L. \tag{B.9}$$

This is visualized in the left plot of figure 10.

However, for smaller eigenvalues, it holds $1-\sigma^2 \geqslant L$. Hence, the prior must decrease the slope of $f_{\theta,j}$. This holds true if the NLL of the prior is convex (or even strongly convex). Because then, $\partial(\log p_{X,j})$ is monotonously decreasing and δ can again be chosen s.t.

$$\mathrm{Lip}\left(f_{\theta,j}\right) = \mathrm{Lip}\left(\left(1-\sigma^2\right)\mathrm{Id} + \partial(\log p_{X,j})\right) \leqslant L \tag{B.10}$$

holds. An example for such a $p_{X,j}$ is a Gaussian prior, where $\partial(\log p_{X,j})$ is a linear function with a negative slope (see the right plot of figure 10 and the subsequent derivations).

B.1. Example prior distributions

To exemplify the previous observations, let us look at two commonly used prior distributions, namely the Gaussian distribution and the Laplace distribution. In the case of the Gaussian distribution, we have

$$p_{X,j}(x) = \frac{1}{\sqrt{2\pi\, b_j^2}} \exp\left(-\frac{(x-\mu_j)^2}{2b_j^2}\right) \tag{B.11}$$

for all $j \in \mathbb{N}$ with mean $\mu_j \in \mathbb{R}$ and variance $b_j^2 > 0$. Consequently, for the residual layer, we obtain

$$f_j(x) = \left(1 - \sigma_j^2 - \frac{\delta^2}{b_j^2}\right) x + \frac{\delta^2 \mu_j}{b_j^2} \quad \text{for } x \in \mathbb{R} \tag{B.12}$$

with Lipschitz constant $\mathrm{Lip}(f_j) = |1 - \sigma_j^2 - {\delta^2}/{b_j^2}|$. Hence, similar to the observations in the previous remark, for singular values with $1 - \sigma_j^2 < L$, the Lipschitz constraint $\mathrm{Lip}(f_j) \leqslant L$ is fulfilled for all δ and b_j. In the case $1 - \sigma_j^2 \geqslant L$, δ and b_j need to satisfy ${\delta^2}/{b_j^2} \geqslant 1 - \sigma_j^2 - L$ to guarantee $\mathrm{Lip}(f_j) \leqslant L$.

In the case of the prior distribution being a Laplacian, i.e.

$$p_{X,j}(x) = \frac{1}{\sqrt{2b_j}} \exp\left(-\frac{|x-\mu_j|}{b_j}\right) \quad \text{for } x \in \mathbb{R} \tag{B.13}$$

with mean $\mu_j \in \mathbb{R}$ and variance $2b_j^2 > 0$, the subgradient of $\log p_{X,j}$ is given by

$$\partial(\log p_{X,j})(x) = \begin{cases} \frac{1}{b_j} & x < \mu_j \\ \left[-\frac{1}{b_j}, \frac{1}{b_j}\right] & x = \mu_j \\ -\frac{1}{b_j} & x > \mu_j \end{cases} \quad \text{for } x \in \mathbb{R}. \tag{B.14}$$

Consequently, the residual layer for the Laplacian prior distribution is given by

$$f_j(x) = \begin{cases} \left(1 - \sigma_j^2\right) x + \frac{\delta^2}{b_j} & x < \mu_j \\ \left[\left(1 - \sigma_j^2\right) x - \frac{\delta^2}{b_j}, \left(1 - \sigma_j^2\right) x + \frac{\delta^2}{b_j}\right] & x = \mu_j \\ \left(1 - \sigma_j^2\right) x - \frac{\delta^2}{b_j} & x > \mu_j \end{cases} \quad \text{for } x \in \mathbb{R}, \tag{B.15}$$

which is not Lipschitz-continuous. As a result, the Laplace distribution does not satisfy the conditions of remark B.1.

In summary, the previous considerations illustrate that MAP estimation with a Gaussian noise model can also be represented by the proposed iResNet approach for certain prior distributions, guaranteeing invertibility.

Appendix C. Approximation training in diagonal architecture with dependent data and noise distribution

In section 3, we assumed that the random variables $x \sim p_X$ and $\eta \sim p_H$ are independent. However, one can obtain a more general version of lemma 3.1 with less restrictive assumptions on the joint data and noise distribution. To this end, we denote by $p\colon \mathbb{R}^2 \to [0, \infty)$ the joint probability density function with marginal distributions $p_X(x) = \int_{\mathbb{R}} p(x,\eta)\,\mathrm{d}\eta$, $p_H(\eta) =$

$\int_{\mathbb{R}} p(x,\eta)\,\mathrm{d}x$ and assume that the respective first and second moments exist. In this setting, the minimization problem of the approximation training reads

$$\min_{f\in\mathcal{F}} \int_{\mathbb{R}^2} p\,(x,\eta)\;|\left(1-\sigma^2\right)x-\eta-f(x)\,|^2\,\mathrm{d}\,(x,\eta) \tag{C.1}$$

and the subsequent lemma provides a closed-form solution of the minimizer of problem (C.1) in the case $L<1-\sigma^2$.

Lemma C.1. *Let* $\mathcal{F}=\{f\in C(\mathbb{R})\,|\,\exists m\in[-L,L], b\in\mathbb{R}\colon\ f(x)=mx+b\}$ *and* $L<1-\sigma^2$. *Then,*

$$f^*\,(\hat{x}) = \begin{cases} L\hat{x}+\left(1-\sigma^2-L\right)\mu_X-\mu_H & \text{if } \frac{\mathrm{Cov}_p(x,\eta)}{\mathrm{Var}_{p_X}(x)} < 1-\sigma^2-L \\ \left(1-\sigma^2-\frac{\mathrm{Cov}_p(x,\eta)}{\mathrm{Var}_{p_X}(x)}\right)\hat{x}+\frac{\mathrm{Cov}_p(x,\eta)}{\mathrm{Var}_{p_X}(x)}\mu_X-\mu_H & \text{if } \frac{\mathrm{Cov}_p(x,\eta)}{\mathrm{Var}_{p_X}(x)} \in \left[1-\sigma^2-L,1-\sigma^2+L\right] \\ -L\hat{x}+\left(1-\sigma^2+L\right)\mu_X-\mu_H & \text{if } \frac{\mathrm{Cov}_p(x,\eta)}{\mathrm{Var}_{p_X}(x)} > 1-\sigma^2+L \end{cases} \tag{C.2}$$

is the unique solution of the minimization problem (C.1), where μ_X, μ_H *denote the expected values of the marginal distributions,* Cov_p *the covariance w.r.t.* $(x,\eta)\sim p$ *and* Var_{p_X} *the variance w.r.t.* $x\sim p_X$.

Proof. For a function f of the form $f(x)=mx+b$ with the constraint $m^2\leqslant L^2$, we can solve (C.1) by using the Lagrangian

$$K(m,b,\lambda)=\int_{\mathbb{R}^2} p\,(x,\eta)\;|\left(1-\sigma^2-m\right)x-\eta-b|^2\,\mathrm{d}\,(x,\eta)+\lambda\left(m^2-L^2\right). \tag{C.3}$$

Observe that the integral is well-defined due to the existence of the first and second moments of p. In addition, the convexity, coercivity, and continuity of the integral term w.r.t. (m,b) implies that a minimizer exists. The minimizer must satisfy the necessary conditions (KKT conditions)

$$\frac{\partial K}{\partial m}\,(m,b,\lambda) = -2\int_{\mathbb{R}^2} p\,(x,\eta)\left(\left(1-\sigma^2-m\right)x-\eta-b\right)x\,\mathrm{d}\,(x,\eta)+2\lambda m \overset{!}{=} 0, \tag{C.4}$$

$$\frac{\partial K}{\partial b}\,(m,b,\lambda) = -2\int_{\mathbb{R}^2} p\,(x,\eta)\left(\left(1-\sigma^2-m\right)x-\eta-b\right)\mathrm{d}\,(x,\eta) \overset{!}{=} 0, \tag{C.5}$$

$$\lambda\left(m^2-L^2\right) \overset{!}{=} 0, \tag{C.6}$$

$$\lambda \geqslant 0. \tag{C.7}$$

Exploiting the marginal distribution p_X, p_H and rearranging (C.4) for m and (C.5) for b leads to

$$m=\frac{\left(1-\sigma^2\right)\mathbb{E}_{p_X}\left(x^2\right)-b\,\mu_X-\mathbb{E}_p\left(x\cdot\eta\right)}{\mathbb{E}_{p_X}\left(x^2\right)+\lambda}, \tag{C.8}$$

$$b=\left(1-\sigma^2-m\right)\mu_X-\mu_H, \tag{C.9}$$

where we use the abbreviated notation $\mathbb{E}_p$ for $\mathbb{E}_{(x,\eta)\sim p}$, $\mathbb{E}_{p_X}$ for $\mathbb{E}_{x\sim p_X}$ and $\mathbb{E}_{p_H}$ for $\mathbb{E}_{\eta\sim p_H}$. Combining both equations yields

$$\begin{aligned}
&\left(1-\frac{\mu_X^2}{\mathbb{E}_{p_X}(x^2)+\lambda}\right)m = \frac{\left(1-\sigma^2\right)\left(\mathbb{E}_{p_X}\left(x^2\right)-\mu_X^2\right)-\left(\mathbb{E}_p\left(x\cdot\eta\right)-\mu_X\mu_H\right)}{\mathbb{E}_p\left(x^2\right)+\lambda}\\
\Leftrightarrow\quad &\frac{\mathrm{Var}_{p_X}(x)+\lambda}{\mathbb{E}_{p_X}(x^2)+\lambda}m = \frac{\left(1-\sigma^2\right)\mathrm{Var}_{p_X}(x)-\mathrm{Cov}_p(x,\eta)}{\mathbb{E}_{p_X}(x^2)+\lambda}\\
\Leftrightarrow\quad &m = \left(1-\sigma^2\right)\frac{\mathrm{Var}_{p_X}(x)}{\mathrm{Var}_{p_X}(x)+\lambda}-\frac{\mathrm{Cov}_p(x,\eta)}{\mathrm{Var}_{p_X}(x)+\lambda}. \qquad \text{(C.10)}
\end{aligned}$$

In order to determine the value of λ, we need to distinguish two cases.

(I): $\mathrm{Cov}_p(x,\eta)\leqslant 0$:
Since $1-\sigma^2>L$ holds by assumption, the case $\lambda=0$ is not possible, and we need $\lambda>0$ to ensure $m\leqslant L$. Then, (C.6) directly implies $m=L$ as (C.10) cancels out the possibility $m=-L$. Thus we also know $b=(1-\sigma^2-L)\mu_X-\mu_H$.

(II): $\mathrm{Cov}_p(x,\eta)>0$:
In this case, we need to distinguish the cases $\lambda=0$ and $\lambda>0$.
(IIa): $\lambda=0$:
In this case, we have

$$m = \left(1-\sigma^2\right)-\frac{\mathrm{Cov}_p(x,\eta)}{\mathrm{Var}_{p_X}(x)}. \qquad \text{(C.11)}$$

The constraint that $m\leqslant L$ and $m\geqslant -L$ only holds true if

$$\mathrm{Cov}_p(x,\eta)\geqslant\left(1-\sigma^2-L\right)\mathrm{Var}_{p_X}(x) \qquad \text{(C.12)}$$

$$\wedge\ \mathrm{Cov}_p(x,\eta)\leqslant\left(1-\sigma^2+L\right)\mathrm{Var}_{p_X}(x) \qquad \text{(C.13)}$$

taking into account that $1-\sigma^2>L$.
(IIa): $\lambda>0$:
In this case, we can have either $m=L$ or $m=-L$. Rearranging (C.10) yields

$$\lambda = \frac{1}{m}\left(\left(1-\sigma^2-m\right)\mathrm{Var}_{p_X}(x)-\mathrm{Cov}_p(x,\eta)\right). \qquad \text{(C.14)}$$

From this we deduce that $\lambda>0$ holds if either

* $m=L$ and $\mathrm{Cov}_p(x,\eta)<(1-\sigma^2-L)\mathrm{Var}_{p_X}(x)$, or
* $m=-L$ and $\mathrm{Cov}_p(x,\eta)>(1-\sigma^2+L)\mathrm{Var}_{p_X}(x)$.

Exploiting (C.9) yields b in either case, which provides the desired f^*. In combination with the observation that m and b are uniquely determined, the proof is complete. □

Appendix D. Additional numerical experiments

The numerical results in section 5 are illustrated for the convolution operator $A=M_a$. In the following, additional illustrations for the convolution operator and all corresponding results for the Radon operator, described in section 5, are provided.

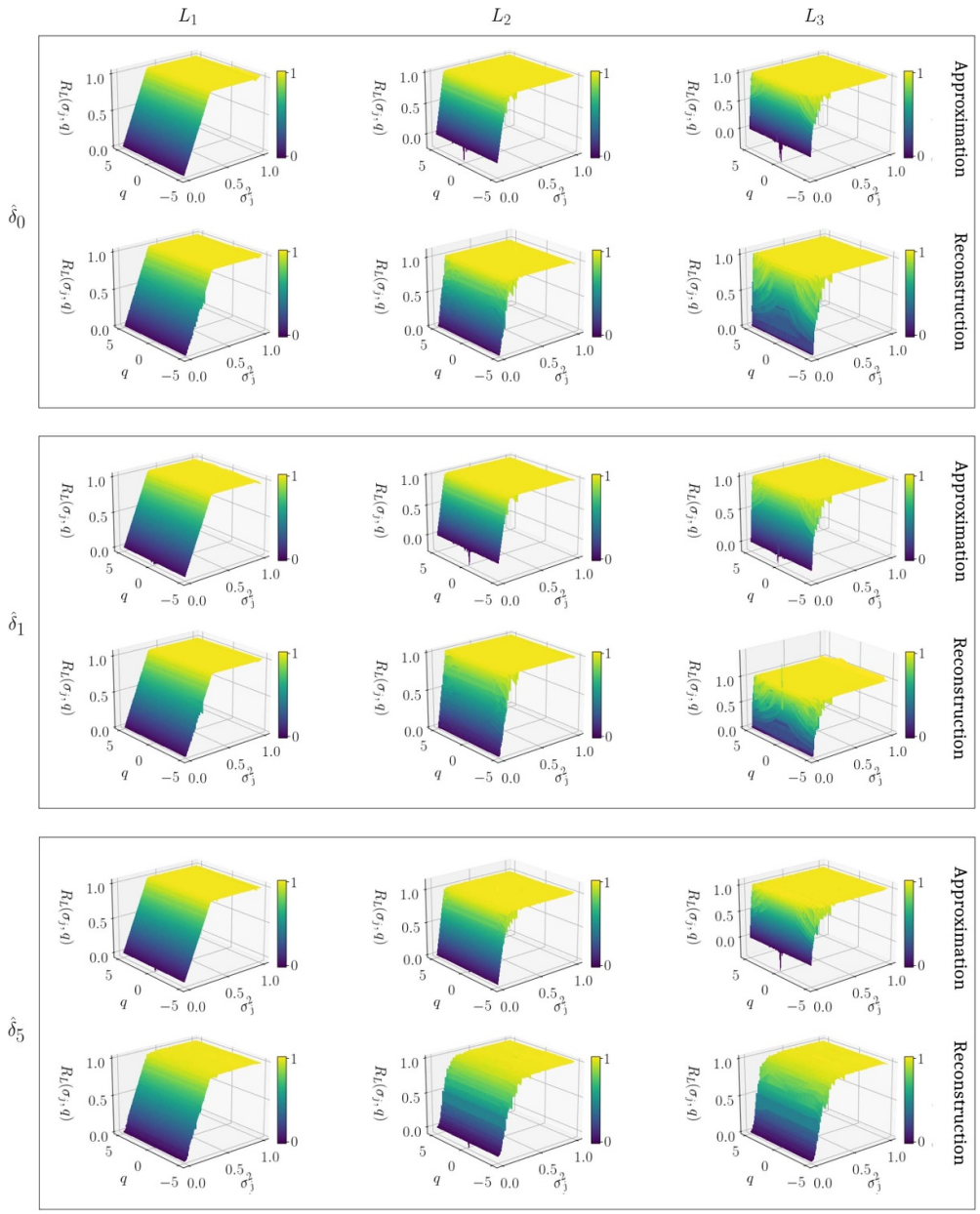
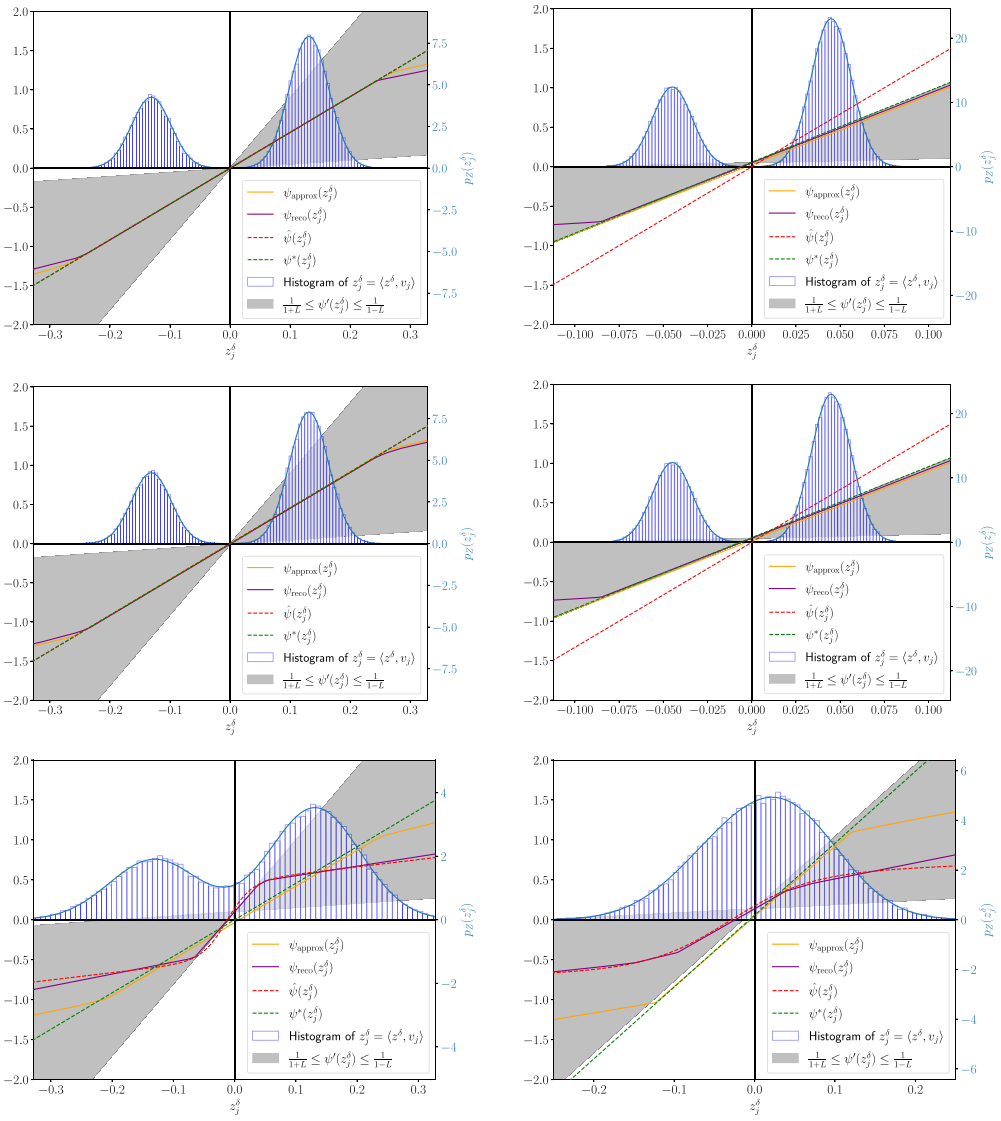


Figure 11. Filter functions $R_L(\sigma_j, q)$ as defined in (5.9) corresponding to trained networks $\varphi_{\theta(L_m,\delta_\ell)}$ for $m = 1,2,3$ *(columns)* and $\ell = 0,1,5$ *(rows)*, trained via approximation training *(top)* and via reconstruction training *(bottom)* on the MNIST dataset for $A = M_a$.



(a) $\sigma_j^2 = 0.219,\ j = 8$ (b) $\sigma_j^2 = 0.075,\ j = 51$

Figure 12. Reconstructions $\psi^*_{\text{approx}}(z_j^\delta)$ trained via approximation training and $\psi^*_{\text{reco}}(z_j^\delta)$ trained via reconstruction training at Lipschitz bound L_2 for different singular values and for noise levels 'zero' (δ_0, *top row*), 'small' (δ_1, *middle row*) and 'large' (δ_5, *bottom row*) for $\tilde{A}$ the Radon operator.

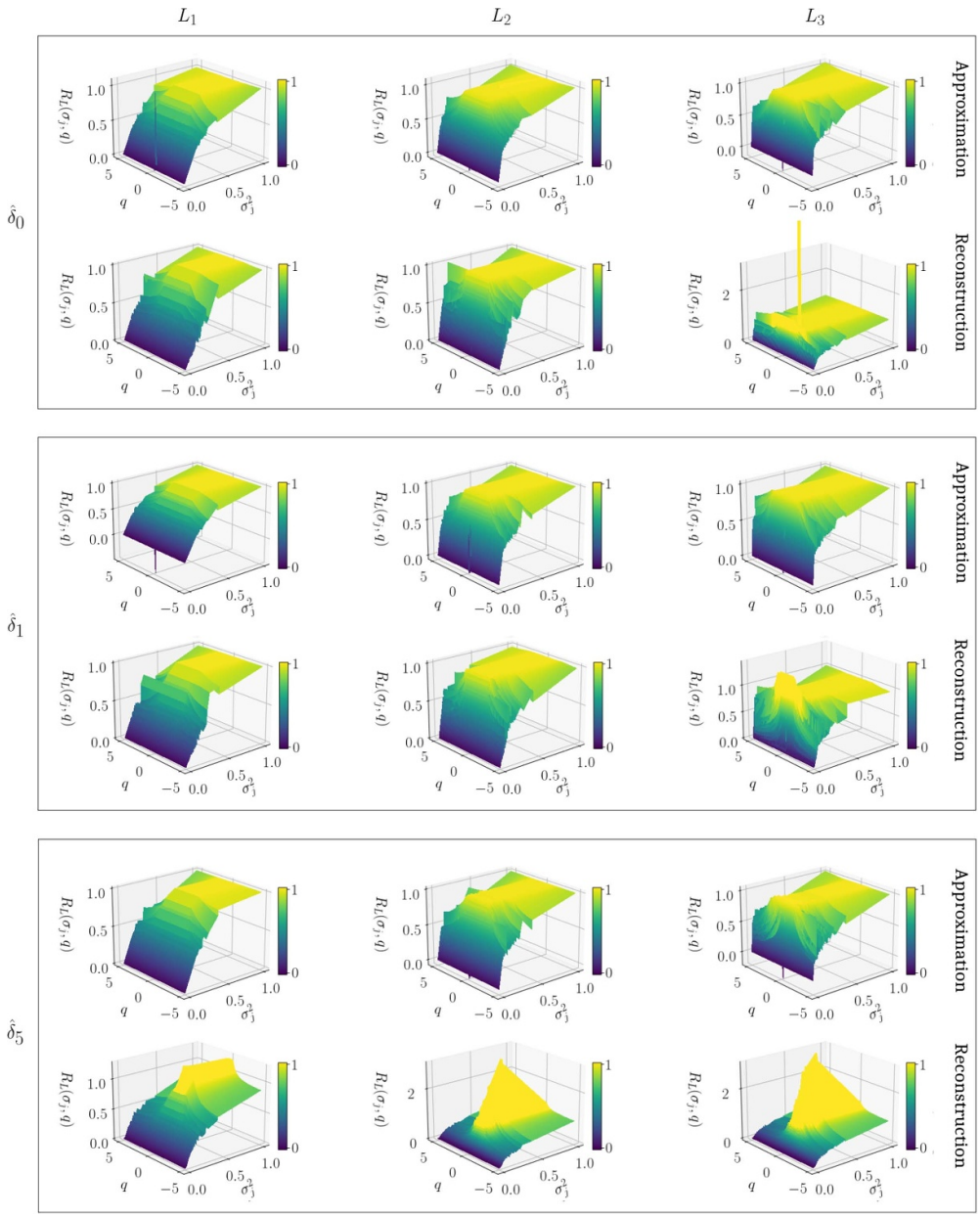


Figure 13. Filter functions $R_L(\sigma_j, q)$ as defined in (5.9) corresponding to trained networks $\varphi_{\theta(L_m,\delta_\ell)}$ for $m = 1, 2, 3$ *(columns)* and $\ell = 0, 1, 5$ *(rows)*, trained via approximation training *(top)* and via reconstruction training *(bottom)* on the bimodal dataset for $\tilde{A}$ the Radon operator.

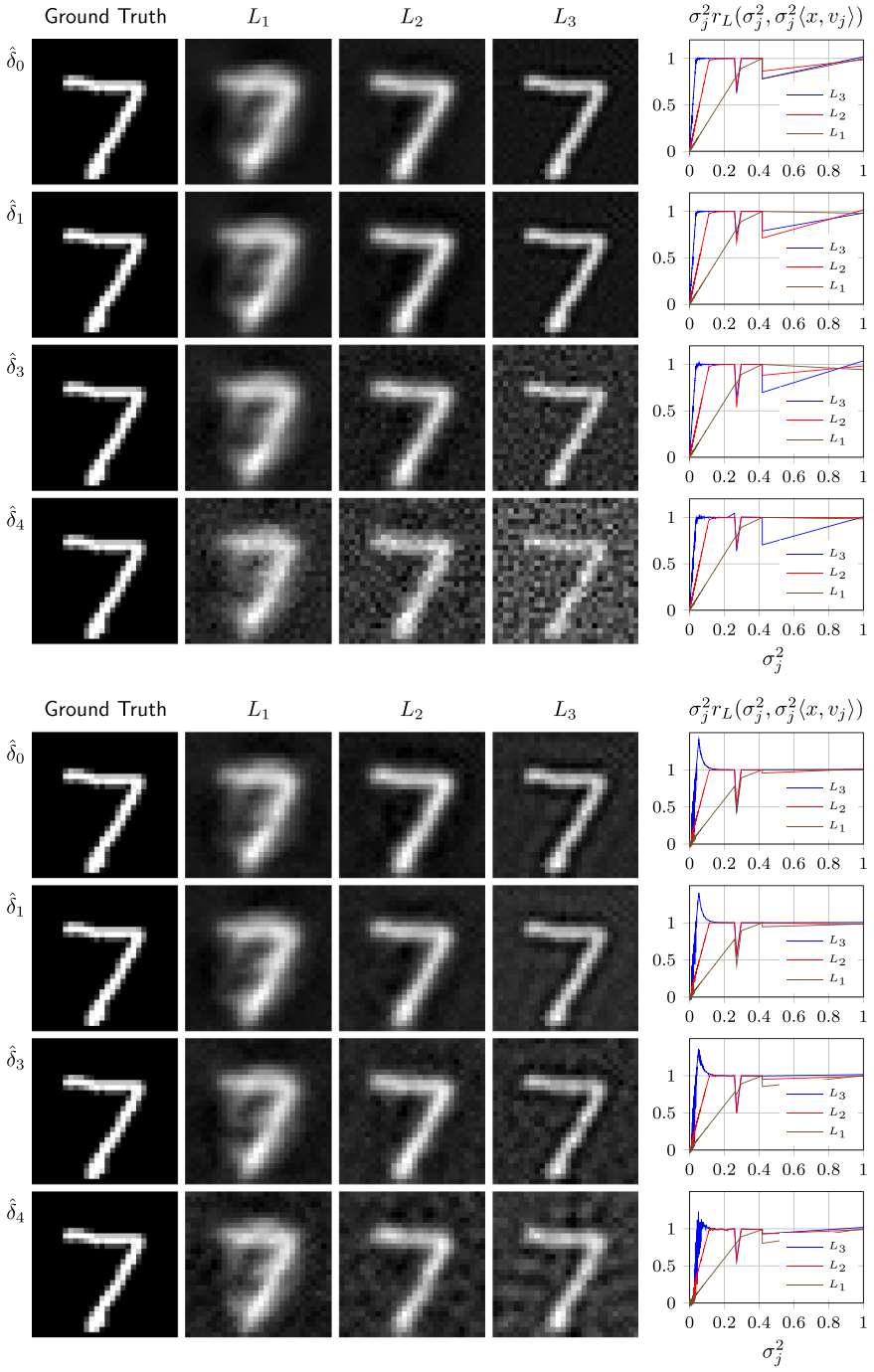


Figure 14. Reconstructions of an MNIST sample $x = x^{(1)}$ from the test dataset by computing $\varphi^{-1}_{\theta(L_m,\delta_\ell)}(Ax + \tilde{\eta})$ with $\tilde{\eta} \sim \mathcal{N}(0, \delta_\ell \mathrm{Id})$ for Lipschitz bounds L_m with $m = 1, 2, 3$ *(columns)* and noise levels $\delta_\ell = \hat{\delta}_\ell \cdot \mathrm{std}_{\mathrm{MNIST}}$ with $\ell = 0, 1, 3, 4$ *(rows)* together with corresponding filter functions for $\tilde{A}$ the Radon operator. The top subfigure depicts the reconstructions from networks trained via approximation training, and the bottom subfigure corresponds to the networks optimized via reconstruction training.

Table 2. SSIM and MSE measures corresponding to reconstructions of $x^{(1)}$ in figure 14. Bold values indicate the best reconstruction quality with respect to the corresponding error measure for a given noise level.

	SSIM			MSE		
Approximation training	L_1	L_2	L_3	L_1	L_2	L_3
δ_0	0.3950	0.7202	**0.8671**	0.0307	0.0117	**0.0023**
δ_1	0.3913	0.7083	**0.8521**	0.0306	0.0120	**0.0025**
δ_3	0.3871	0.6569	**0.6620**	0.0306	0.0129	**0.0087**
δ_4	0.3774	**0.5480**	0.5033	0.0311	**0.0190**	0.0459
	SSIM			MSE		
Reconstruction training	L_1	L_2	L_3	L_1	L_2	L_3
δ_0	0.4161	0.6899	**0.8545**	0.0290	0.0124	**0.0029**
δ_1	0.3989	0.6865	**0.8464**	0.0304	0.0126	**0.0031**
δ_3	0.3913	0.6512	**0.7267**	0.0300	0.0134	**0.0055**
δ_4	0.3934	0.5990	**0.6254**	0.0298	0.0149	**0.0126**

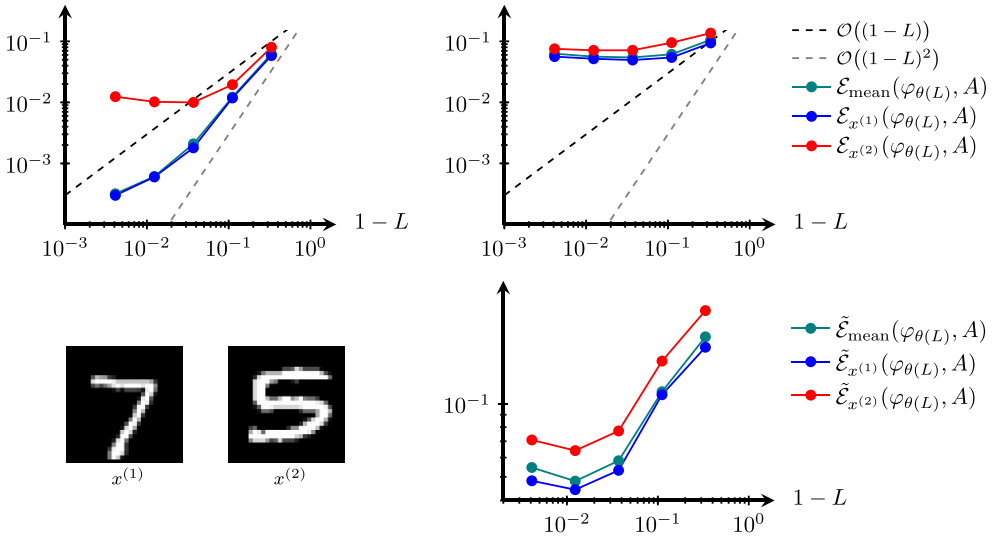


Figure 15. Test samples $x^{(1)}$ and $x^{(2)}$ *(bottom left)*. Evaluations of the local approximation property via $\mathcal{E}_{\text{mean}}(\varphi_{\theta(L_m)},A)$, $\mathcal{E}_{x^{(1)}}(\varphi_{\theta(L_m)},A)$ and $\mathcal{E}_{x^{(2)}}(\varphi_{\theta(L_m)},A)$ for the approximation training *(top left)* and the reconstruction training *(top right)*, and evaluations of the generalized approximation property via $\tilde{\mathcal{E}}_{\text{mean}}(\varphi_{\theta(L_m)},A)$, $\tilde{\mathcal{E}}_{x^{(1)}}(\varphi_{\theta(L_m)},A)$ and $\tilde{\mathcal{E}}_{x^{(2)}}(\varphi_{\theta(L_m)},A)$ for the reconstruction training *(bottom right)* for $L_m = 1 - 1/3^m$ with $m = 1,\dots,5$ and $\tilde{A}$ the Radon operator on the MNIST test dataset.

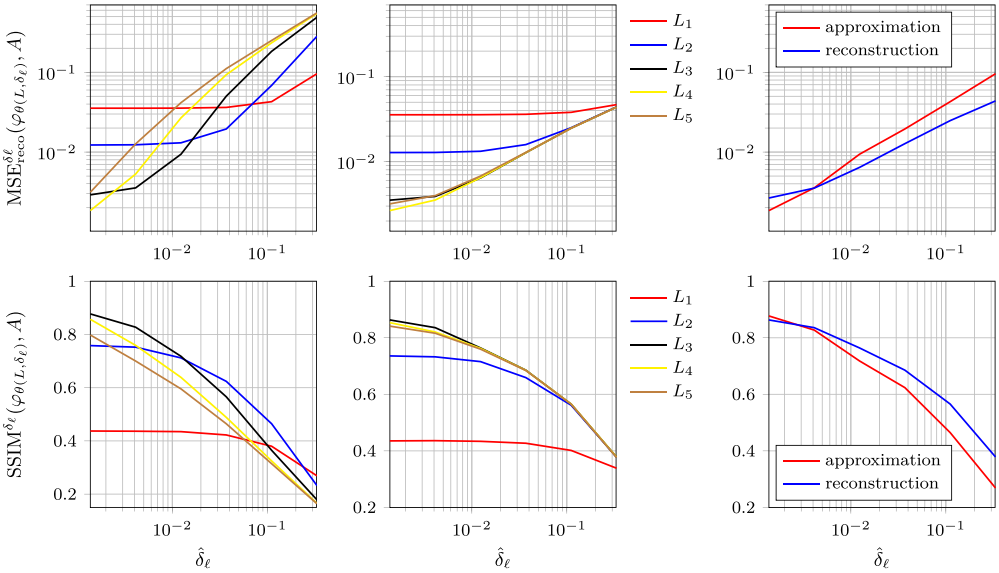


Figure 16. Reconstruction errors $\mathrm{MSE}^{\delta_\ell}_{\mathrm{reco}}(\varphi_{\theta(L,\delta_\ell)},A)$ *(top row)* and $\mathrm{SSIM}^{\delta_\ell}(\varphi_{\theta(L,\delta_\ell)},A)$ *(bottom row)* for networks trained on noisy samples with noise levels δ_ℓ for $\ell = 0,\dots,6$ and reconstructions from noisy samples of the same noise level for the approximation training *(left)* and for the reconstruction training *(middle)* with Lipschitz bounds L_m on the MNIST dataset for the Radon operator $\tilde{A}$. Outcomes of optimal parameter choices for both training strategies over different noise levels can be seen on the right-hand side.

ORCID iDs

Clemens Arndt https://orcid.org/0000-0001-5607-4074
Nick Heilenkötter https://orcid.org/0000-0001-7693-8618
Meira Iske https://orcid.org/0009-0009-1198-6358
Tobias Kluth https://orcid.org/0000-0003-4814-142X
Judith Nickel https://orcid.org/0000-0001-6599-7540

References

[1] Adler J and Öktem O 2018 Deep Bayesian inversion (arXiv:1811.05910)
[2] Alberti G S, De Vito E, Lassas M, Ratti L and Santacesaria M 2021 Learning the optimal tikhonov regularizer for inverse problems *Advances in Neural Information Processing Systems* vol 34 pp 25205–16
[3] Arndt C, Denker A, Dittmer S, Heilenkötter N, Iske M, Kluth T, Maass P and Nickel J 2023 Invertible residual networks in the context of regularization theory for linear inverse problems *Inverse Problems* **39** 125018
[4] Arridge S, Maass P, Öktem O and Schönlieb C-B 2019 Solving inverse problems using data-driven models *Acta Numer.* **28** 1–174
[5] Behrmann J, Grathwohl W, Chen R T, Duvenaud D and Jacobsen J-H 2019 Invertible residual networks *Int. Conf. on Machine Learning* (PMLR) pp 573–82
[6] Benning M and Burger M 2018 Modern regularization methods for inverse problems *Acta Numer.* **27** 1–111

[7] Bochkina N 2013 Consistency of the posterior distribution in generalized linear inverse problems *Inverse Problems* **29** 095010
[8] Chen R T Q, Behrmann J, Duvenaud D K and Jacobsen J-H 2019 Residual flows for invertible generative modeling *Advances in Neural Information Processing Systems* vol 32, ed H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox and R Garnett (Curran Associates, Inc.)
[9] Dashti M and Stuart A 2017 *The Bayesian Approach to Inverse Problems* (Springer) pp 311–428
[10] Engl H W, Hanke M and Neubauer A 1996 *Regularization of Inverse Problems* vol 375 (Springer)
[11] Kabri S, Auras A, Riccio D, Bauermeister H, Benning M, Moeller M and Burger M 2022 Convergent data-driven regularizations for CT reconstruction (arXiv:2212.07786)
[12] Kaipio J and Somersalo E 2006 *Statistical and Computational Inverse Problems* vol 160 (Springer)
[13] Kingma D P and Ba J 2015 Adam: a method for stochastic optimization *3rd Int. Conf. on Learning Representations, ICLR 2015 (Conf. Track Proc.) (San Diego, CA, USA, 7–9 May 2015)* ed Y Bengio and Y LeCun
[14] Klenke A 2020 *Probability Theory: A Comprehensive Course* (*Universitext*) (Springer)
[15] Königsberger K 2013 *Analysis 2* (Springer)
[16] Laumont R, Bortoli V D, Almansa A, Delon J, Durmus A and Pereyra M 2022 Bayesian imaging using plug & play priors: When langevin meets tweedie *SIAM J. Imaging Sci.* **15** 701–37
[17] LeCun Y 1998 The MNIST database of handwritten digits (available at: http://yann.lecun.com/exdb/mnist/)
[18] Luenberger D G 1969 *Optimization by Vector Space Methods* (*Wiley Professional Paperback Series*) (Wiley)
[19] Maass P 2019 Deep learning for trivial inverse problem *Compressed Sensing and Its Applications* (Springer)
[20] Miyato T, Kataoka T, Koyama M and Yoshida Y 2018 Spectral normalization for generative adversarial networks (arXiv:1802.05957)
[21] Mukherjee S, Hauptmann A, Öktem O, Pereyra M and Schönlieb C-B 2023 Learned reconstruction methods with convergence guarantees: a survey of concepts and applications *IEEE Signal Process. Mag.* **40** 164–82
[22] Scherzer O, Hofmann B and Nashed Z 2023 Gauss–Newton method for solving linear inverse problems with neural network coders *Sampling Theory Signal Process. Data Anal.* **21** 25
[23] Seierstad A and Sydsaeter K 1977 Sufficient conditions in optimal control theory *Int. Econ. Rev.* **18** 367–91
[24] Sherry F, Celledoni E, Ehrhardt M J, Murari D, Owren B and Schönlieb C-B 2018 Designing stable neural networks using convex analysis and odes (arXiv:1802.05957)
[25] Stuart A M 2010 Inverse problems: a Bayesian perspective *Acta Numer.* **19** 451–559
[26] Vollmer S J 2013 Posterior consistency for Bayesian inverse problems through stability and regression results *Inverse Problems* **29** 125011
[27] Arndt C, Dittmer S, Heilenkötter N, Iske M, Kluth T and Nickel J 2024 iResNet Regularization (available at: https://gitlab.informatik.uni-bremen.de/inn4ip/iresnet-regularization)
[28] Wang Z, Bovik A, Sheikh H and Simoncelli E 2004 Image quality assessment: from error visibility to structural similarity *IEEE Trans. Image Process.* **13** 600–12