Learning Spectral Regularizations for Linear Inverse Problems

Hartmut Bauermeister **Martin Burger** **Michael Moeller**

Abstract

One of the main challenges in linear inverse problems is that a majority of such problems are ill-posed in the sense that the solution does not depend on the data continuously. To analyze this effect and reestablish a continuous dependence, classical theory in Hilbert spaces largely relies on the analysis and manipulation of the singular values of the linear operator and its pseudoinverse with the goal of, on the one hand, keeping the singular values of the reconstruction operator bounded, and, on the other hand, approximating the pseudoinverse sufficiently well for a given noise level. While classical regularization methods manipulate the singular values via explicitly defined functions, this paper considers learning such parameter choice rules in such a way, that one obtains higher quality reconstruction results while still remaining in a setting of provably convergent spectral regularization methods. We discuss different ways of parametrizing our spectral regularization methods via neural networks, interpret existing feed forward networks in the setting of spectral regularization which can become provably convergent via an additional projection, and finally demonstrate their superiority in 1d numerical examples.

1 Introduction

The classical theory for linear inverse problems considers a linear operator $A : X \to Y$ between Hilbert spaces X and Y and asks the question how to reconstruct an unknown $x \in X$ from an observation $y^\delta \in Y$ given as $y^\delta = Ax + n^\delta$, where $n^\delta \in Y$ represents noise of magnitude $\|n^\delta\| = \delta$. As soon as A is a compact linear operator with infinite dimensional range, A admits a singular value decomposition (SVD)

$$Ax = \sum_{n=1}^{\infty} \sigma_n \langle x, u_n \rangle v_n \tag{1}$$

with singular vectors $u_n \in X$ and $v_n \in Y$, and zero being an accumulation point of the corresponding singular values σ_n. Thus, the pseudoinverse $A^\dagger$ of A becomes

$$A^\dagger y = \sum_{n=1}^{\infty} \frac{1}{\sigma_n} \langle y, v_n \rangle u_n. \tag{2}$$

Since zero is an accumulation point of the σ_n, the pseudoinverse is unbounded, i.e., discontinuous. The idea of spectral regularization is to replace the pseudoinverse $A^\dagger$ by a family of operators $R_\alpha : Y \to X$ given by

$$R_\alpha y = \sum_{n=1}^{\infty} g_\alpha(\sigma_n) \langle y, v_n \rangle u_n \tag{3}$$

in such a way that R_α is a continuous on the entire space Y and that it converges pointwise to $A^\dagger$ as the *regularization parameter* α goes to zero. A suitable choice of α as a function of δ (a-priori

NeurIPS 2020 Workshop on Deep Learning and Inverse Problems, virtual.

choice) or of (δ, y^δ) (a-posteriori choice) allows to reestablish the continuous dependence of the solution on the data in the sense that

$$\|R_\alpha y^\delta - A^\dagger y\| \to 0 \qquad \text{as} \qquad \delta \to 0. \tag{4}$$

In this paper we investigate learning a function $\mathcal{N}$ parameterized by θ such that

$$g_\alpha(\sigma) = \mathcal{N}(\sigma, \delta, y^\delta; \theta) \tag{5}$$

provably satisfies (4). We discuss a-priori as well as a-posteriori choices of the regularization, demonstrate how learning improves the results over classical choices, and illustrate the performance boost when turning to nonlinear reconstruction operators.

2 Related Work

Classically, the spectral regularization approaches have been defined manually, e.g. via functions g_α of the form

$$g_\alpha(\sigma) = \begin{cases} \frac{1}{\sigma} & \text{if } \sigma \geq \alpha \\ 0 & \text{otherwise} \end{cases}, \tag{Truncated SVD}$$

$$g_\alpha(\sigma) = \frac{1}{\sigma + \alpha}, \tag{Lavrentiev}$$

$$g_\alpha(\sigma) = \frac{\sigma}{\sigma^2 + \alpha} \tag{Tikhonov}$$

along with suitable parameter choice rules α, see e.g. [8]. The fact that Tikhonov regularization is equivalent to solving a minimization problem with a quadratic penalty, subsequently gave rise to nonlinear variational methods for which convergent regularization methods similar to (4) can be guaranteed in different measures of distances, see e.g. [4, 2].

Due to the rise of deep learning techniques many benchmarks are currently dominated by directly learning a mapping from y^δ to a desired solution x via a neural network. Such techniques are, however, largely lacking a theoretical understanding and rarely even take the noise level δ into account explicitly (see e.g. [20] for an exception).

Several works have considered hybrid methods between regularization techniques that allow for detailed analysis and learning based approaches: For instance, [6, 1, 10] use minimization/regularization algorithms as a template for network architectures, [3, 12, 19, 11] optimize over a learned latent space that contains mostly realistic solutions, [20] consider regularization by parametrization via deep neural networks, and algorithmic schemes that replace the proximal operator of a regularizer with a neural network have been studied in [15, 13]. Such techniques do, however, not correspond to minimization problems anymore unless the network possesses very specific properties, see [16]. Beyond this, safeguarding techniques such as [14] or bilevel optimization problems that learn a parameterized variational regularization (e.g. [18, 7, 5, 9]) are the only way to remain in the regime of energy minimization methods. The additional difficulty of defining networks in such a way that they act on continuous functions rather than their fixed discretizations, makes a convergence analysis in the sense of (4) difficult and rare. On a related note, the work [17] considers learning itself as an ill-posed inverse problem to derive convergence properties similar to (4).

3 Learned Spectral Regularizers

3.1 Architectures

For learning spectral regularizations we consider different types of parameterized functions to represent g_α:

- A-priori parameter choices: We parametrize $\mathcal{N}(\sigma, \delta, y^\delta; \theta)$ by two classical approaches, namely a learned Lavrentiev and a learned Tikhonov regularization given via

$$\mathcal{N}_{\text{Lav}}(\sigma, \delta; \theta) = \frac{1}{\sigma + \delta^p \tilde{N}(\sigma, \delta; \theta)}, \qquad \mathcal{N}_{\text{Tik}}(\sigma, \delta; \theta) = \frac{\sigma}{\sigma^2 + \delta^q \tilde{N}(\sigma, \delta; \theta)}, \tag{6}$$

with $p \leq 1$, $q \leq 2$, and a network

$$\tilde{N}(\sigma, \delta; \theta) = \theta_{\text{scale}} \cdot \text{sigmoid}(\text{FCN}(\sigma, \delta; \theta)) \tag{7}$$

with a 2-layer fully connected network *FCN*, and one additional scale parameter θ_{scale}. To also make a comparison to classical (but noise-level optimal) regularization choices, we additionally drop the dependence of $\tilde{\mathcal{N}}$ on σ in (7) and refer to these methods as the classical Lavrentiev and Tikhonov regularizations.

- A-posteriori parameter choice: Many papers have demonstrated great success in directly predicting a solution $\hat{x}$ for given data y^δ by exploiting spatial regularity of x, e.g. through convolutional neural networks in imaging applications. While the straight forward application of such networks lacks any kind of convergence guarantee, such techniques can be converted to a-posteriori spectral regularizations with convergence guarantees via suitable projectors. Let $\hat{x} = \mathcal{G}(y^\delta; \theta)$ be some prediction of a neural network $\mathcal{G}$ designed to solve the underlying problem directly. Then the choice

$$g_\alpha(\sigma_n) = \frac{\langle u_n, \mathcal{G}(y^\delta; \theta)\rangle}{\langle v_n, y^\delta\rangle} \tag{8}$$

results in the spectral regularization yielding $\mathcal{G}(y^\delta; \theta)$ as a reconstruction result (assuming that $\langle v_n, y^\delta\rangle \neq 0$, and that all σ_n are different). To ensure convergence, we modify such a prediction via the following projection

$$\mathcal{N}(\sigma_n, \delta, y^\delta; \theta) = \text{proj}_{\left[(1-\sqrt{\delta}\theta_l)\frac{\sigma}{\sigma^2+\alpha_l\delta}, (1+\sqrt{\delta}\theta_u)\frac{\sigma}{\sigma^2+\alpha_u\delta}\right]}\left(\frac{\langle u_n, \mathcal{G}(y^\delta; \theta)\rangle}{\langle v_n, y^\delta\rangle}\right) \tag{9}$$

with learnable parameters θ_l and θ_u. The projection can be interpreted as remaining in between a strong and a weak Tikhonov regularization ($\alpha_u << \alpha_l$) up to some tolerance with deceases to zero as $\delta \to 0$. For $\mathcal{G}$ we choose an architecture loosely motivated by unrolling a proximal gradient descent algorithm, see A.2 in the appendix for details.

With the above choices, we can state the following convergence result:

Proposition 1 (Convergent spectral regularization methods)**.** *Both learning based approaches, the a-priori choice* (6) *and the a-posteriori choice* (9)*, are convergent regularization methods in the sense of* (4) *provided that* $q < 2$ *and* $p < 1$.

Proof. See appendix A.1. □

4 Numerical Experiments

To compute the solution of inverse problems numerically, we need to discretize any infinite dimensional problem to a finite one, e.g., by considering a suitable subspace. While this step alone reintroduces regularity as finite dimensional linear operators can never be discontinuous, the resulting problem still remains ill-conditioned due to a quick decay of the singular values σ_n.

To investigate the behavior of our regularization strategy, we consider two inverse problems: The differentiation as well as deblurring of a function $y : [0, 1] \to \mathbb{R}$, giving rise to the linear operators

$$A_{\text{int}}x(t) = \int_0^t x(s)\, ds, \qquad A_{\text{blur}}x(t) = \int_0^1 g(s-t)x(s)\, ds, \tag{10}$$

for g being a Gaussian kernel. In both cases we discretize x by evaluating it at positions $(x_n)_{1\leq n\leq N}$ and approximate the integrals of the operators by simple summations. Further details on the training can be found in the appendix.

Figure 1 (left and middle) shows the learned Tikhonov and Lavrentiev regularizations along with the classical Tikhonov regularization and a naive approach, in which g_α is directly parameterized as a fully connected network depending on σ and δ, for two different noise levels δ as a function of σ. As we can see, all learned regularizers choose a Tikhonov-type shape of g_α, but allow much larger $g_\alpha(\sigma)$ for σ on a medium scale. In particular, the learned Tikhonov and direct (naive) learning yield remarkably similar shapes of g_α. Looking at the spectral regularizers for fixed $\sigma = 0.1$ (right

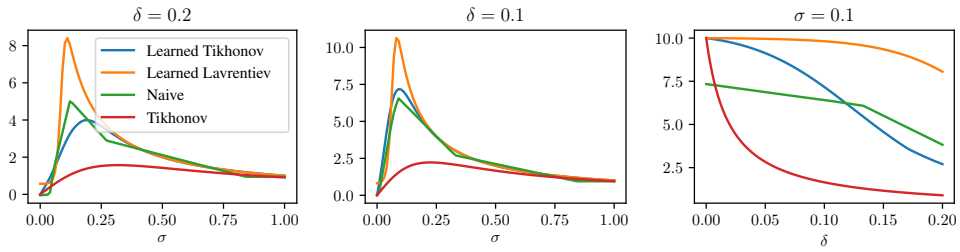


Figure 1: Exemplifying the results of the a-priori parameter choice rules. Left and middle: $\mathcal{N}(\sigma, \delta; \theta)$ as a function of σ for two different noise levels $\delta = 0.2$ and $\delta = 0.1$. Right: $\mathcal{N}(\sigma, \delta; \theta)$ as a function of δ for a fixed $\sigma = 0.1$.

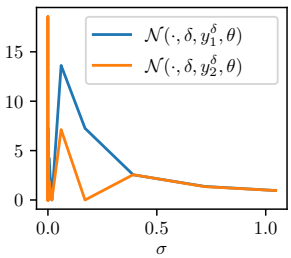
plot in Fig. 1) as a function of δ, one can see that the Tikhonov, learned Tikhonov and Lavrentiev regularizers yield a value of $g_\alpha(\sigma) = 1/\sigma$ for $\delta = 0$. The direct (naive) approach, however, fails to yield such a value, which directly implies that this method is not a convergent regularization in the sense of (4). We conclude that the choice of the architecture with built-in behavior is crucial for obtaining theoretical guarantees.

		Naive	Lav.	Tik.	Learned Lav.	Learned Tik.	A-Post.
Deblur	training	29.58	20.05	27.17	28.59	29.73	31.89
	test	29.56	19.98	27.00	28.55	29.67	31.51
Diff.	training	28.93	21.35	26.44	28.94	29.27	31.63
	test	29.00	21.30	26.45	28.99	29.31	30.74

Table 1: PSNR values during training and testing for deblurring and differentiation for various different regularization strategies.

As for the overall performance of each approach, Table 4 shows the PSNR values for all methods over training and testing in both applications for a fixed discretization of $N = 50$. As we can see, the learning-based methods clearly outperform the classical approach while still remaining provably convergent in the sense of Prop. 1. Moreover, the learned Tikhonov parametrization is superior to its Lavrentiev counterpart. The naive approach does yield good PSNR values (although with a high initialization-depending variance), but does not yield a convergent regularization, i.e., does not yield faithful results for small noise levels. The best results by far are obtained by the learned a-posteriori approach, which converts a direct prediction of the solution x to a spectral regularization.

Shown in the inset figure are two curves of $g_\alpha = \mathcal{N}(\cdot, \delta, y^\delta; \theta)$ for the same y and δ, such that merely the realizations of the noise differ. As we can see, the learned a-posteriori choice is non-monotone, does not yield smooth curves such as in Fig. 1, and differs significantly for different realizations of noisy data. Yet, the PSNR values of this approach is significantly higher, indicating the importance of non-linear regularizations that vary for different y^δ.



5 Conclusions

In this paper we studied spectral regularization methods for linear inverse problems and successfully learned provably convergent regularization methods that outperform their classical counterpart. By considering a-posteriori choice rules in spectral regularizations, we turned to non-linear reconstruction techniques, which yield even better reconstruction results and ultimately raise the quest for establishing provably convergent learned regularization methods in other notions of distance, e.g. resembling the analysis of variational methods using Bregman distances.

A Supplementary Material

A.1 Proof of Proposition 1

The techniques used for proving our proposition are well known and follow the arguments in [8]. For the sake of completeness, we will still give an overview of the proof. The central idea is to show the following results that ensure a convergent spectral regularization method: For a regularization of the form (3) the following criteria ensure (4).

$$g_\alpha(\sigma) \leq C_\alpha \text{ for all } \sigma > 0, \tag{11a}$$

$$g_\alpha(\sigma) \overset{\alpha\to 0}{\to} \frac{1}{\sigma} \text{ for all } \sigma > 0, \tag{11b}$$

$$\sigma g_\alpha(\sigma) \leq \tilde{C} < \infty \text{ for all } \alpha,\ \sigma > 0, \tag{11c}$$

$$\delta C_{\alpha(\delta, y^\delta)} \overset{\delta\to 0}{\to} 0. \tag{11d}$$

Proof. The condition (11a) ensures that R_α is a continuous linear operator for any fixed $\delta > 0$ because $\|R_\alpha\| = C_\alpha$. For y in the domain of $A^\dagger$, $y \in \mathcal{D}(A^\dagger)$, and y^δ with $\|y - y^\delta\| \leq \delta$ we estimate

$$\|A^\dagger y - R_\alpha y^\delta\| = \|A^\dagger y - R_\alpha y + R_\alpha y - R_\alpha y^\delta\| \leq \|A^\dagger y - R_\alpha y\| + \|R_\alpha\|\delta. \tag{12}$$

Now using $\|R_\alpha\| = C_\alpha$ condition (11d) ensures that the second term in the above estimate converges to zero for $\delta \to 0$. As for the first term we find

$$\|R_\alpha y - A^\dagger y\|^2 = \sum_{n=1}^{\infty} \left(g_\alpha(\sigma_n) - \frac{1}{\sigma_n} \right)^2 |\langle v_n, y\rangle|^2, \tag{13}$$

$$= \sum_{n=1}^{\infty} \left(\sigma g_\alpha(\sigma_n) - 1\right)^2 \frac{1}{\sigma_n^2} |\langle v_n, y\rangle|^2. \tag{14}$$

Due to condition (11c) the above sum remains bounded independent of α and therefore is uniformly convergent, such that summation and a limit of $\delta \to 0$ can be exchanged. Finally, condition (11b) ensures that the above term converges to zero.

Left to verify is that our learnable architectures satisfy the conditions (11). For both a-priori choice rules (11a) holds due to the sigmoid function being strictly greater than zero for all inputs. Condition (11b) holds for $p, q > 0$ since the sigmoid function is bounded by 1. Condition (11c) holds with $\tilde{C} = 1$, and (11d) holds if $p < 1$ and $q < 2$. Similarly, the projection operator of the a-posteriori choice (9) ensures the conditions (11) to be met. □

A.2 Details on Network Architectures and Training

The fully-connected networks in the learned Lavrentiev and learned Tikhonov models consist of two fully-connected layers with hidden dimension 10. The naive network consists of 3 fully-connected layers with hidden dimension $2 \to 100 \to 100 \to 1$.

For the a-posteriori prediction $\hat{x} = \mathcal{G}(y^\delta; \theta)$ we utilize an iteration inspired by proximal gradient descent

$$x_i = \mathcal{G}_i(x_{i-1} - \tau A^T(Ax_{i-1} - y); \theta_i) \tag{15}$$

where $x_0 = 0$ and $\tau = 1$. The networks $\mathcal{G}_i$ themselves are 3-layer convolutional networks with hidden layer sizes of 10. For the convoultions we use zero padding and we fix the minimum kernel size such that the receptive field of the output covers the entire input x_{i-1}.

A.2.1 Training

For each resolution $N \in \{10, 11, \ldots, 50\}$ we generate $R = 5000$ many training examples $x_{i,N}$ by sampling random superpositions of sine and cosine functions, applying the discretized operator A and adding zero-mean Gaussian noise of different standard deviations $\delta_{i,N}$ with $\delta_{i,N} \in [0, 0.2]$ to obtain simulated data $y_{i,N}^{\delta_{i,N}}$. Then we train our networks in a supervised way via

$$\min_\theta \sum_{N=10}^{50} \frac{1}{R} \sum_{i=1}^{R} \left\| \sum_{n=1}^{N} \mathcal{N}(\sigma_{n,N}, \delta_{i,N}, y_{i,N}^{\delta_{i,N}}; \theta) \langle y_{i,N}^{\delta_{i,N}}, v_{n,N}\rangle u_{n,N} - x_{i,N} \right\|^2. \tag{16}$$

The superpositions of sine and cosine functions are realized by

$$f(x) = \cos(\omega_1 x) + \gamma \sin(\omega_2 x) \tag{17}$$

where ω_1 and ω_2 are drawn from a standard normal distribution and γ is sampled uniformly from the interval $[-1, 1]$.

References

[1] J. Adler and O. Öktem. Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Problems*, 33(12), December 2017.

[2] M. Benning and M. Burger. Modern regularization methods for inverse problems. *Acta Numerica*, 27:1–111, 2018.

[3] A. Bora, A. Jalal, E. Price, and A.G. Dimakis. Compressed sensing using generative models. In *ICML*, pages 537–546. JMLR. org, 2017.

[4] M. Burger, E. Resmerita, and L. He. Error estimation for bregman iterations and inverse scale space methods in image restoration. *Computing*, 81:109–135, 11 2007.

[5] L. Calatroni, C. Cao, JC. De Los Reyes, C-B. Schönlieb, and T. Valkonen. Bilevel approaches for learning of variational imaging models. *Variational Methods: In Imaging and Geometric Control*, 18:252, 2017.

[6] Y. Chen and T. Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1–1, 08 2016.

[7] Y. Chen, T. Pock, and H. Bischof. Learning ℓ_1-based analysis and synthesis sparsity priors using bi-level optimization. In *NeurIPS*, 2012.

[8] H.W. Engl, M. Hanke, and G. Neubauer. *Regularization of Inverse Problems*. Mathematics and Its Applications. Springer Netherlands, 1996.

[9] J. Geiping and M. Moeller. Parametric majorization for data-driven energy minimization methods. In *ICCV*, pages 10261–10272, 2019.

[10] E. Kobler, T. Klatzer, K. Hammernik, and T. Pock. Variational networks: Connecting variational methods and deep learning. In Volker Roth and Thomas Vetter, editors, *Pattern Recognition*, pages 281–293. Springer International Publishing, 2017.

[11] F. Latorre, A. Eftekhari, and V. Cevher. Fast and provable admm for learning with generative priors. In *NeurIPS*, pages 12004–12016. Curran Associates, Inc., 2019.

[12] Y. Li, S. Liu, J. Yang, and M-H. Yang. Generative face completion. In *CVPR*, pages 3911–3919, 2017.

[13] T. Meinhardt, M. Moeller, C. Hazirbas, and D. Cremers. Learning proximal operators: Using denoising networks for regularizing inverse imaging problems. In *ICCV*, pages 1781–1790, 2017.

[14] M. Moeller, T. Moellenhoff, and D. Cremers. Controlling neural networks via energy dissipation. In *ICCV*, pages 3255–3264, 2019.

[15] J.H. Rick Chang, C-L. Li, B. Poczos, BVK. Vijaya Kumar, and AC. Sankaranarayanan. One network to solve them all–solving linear inverse problems using deep projection models. In *ICCV*, pages 5888–5897, 2017.

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

[17] Lorenzo Rosasco, Andrea Caponnetto, Ernesto De Vito, Umberto De Giovannini, and Francesca Odone. Learning, regularization and ill-posed inverse problems. In *NeurIPS*, page 1145–1152. MIT Press, 2004.

[18] K. G. G. Samuel and M. F. Tappen. Learning optimized MAP estimates in continuously-valued MRF models. In *CVPR*, 2009.

[19] V. Shah and C. Hegde. Solving linear inverse problems using gan priors: An algorithm with provable guarantees. In *ICASSP*, pages 4609–4613, 2018.

[20] K. Zhang, W. Zuo, and L. Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018.