Mathematisches Forschungsinstitut Oberwolfach

Report No. 13/2021

DOI: 10.4171/OWR/2021/13

Deep Learning for Inverse Problems (hybrid meeting)

Organized by
Simon Arridge, London
Peter Maaß, Bremen
Carola-Bibiane Schönlieb, Cambridge UK

7 March – 13 March 2021

Abstract. Machine learning and in particular deep learning offer several data-driven methods to amend the typical shortcomings of purely analytical approaches. The mathematical research on these combined models is presently exploding on the experimental side but still lacking on the theoretical point of view. This workshop addresses the challenge of developing a solid mathematical theory for analyzing deep neural networks for inverse problems.

Mathematics Subject Classification (2010): 65J22, 65Y20, 94C99.

Introduction by the Organizers

The workshop *Deep Learning for Inverse Problems*, organized by Simon Arridge (London), Peter Maaß (Bremen) and Carola-Bibiane Schönlieb (Cambridge) was well attended with 28 participants and aimed at bringing together experts from different scientific directions to contribute mathematically proven results in the theory of deep neural networks for inverse problems. Most participants attended the workshop online, which required a special schedule in order to allow a maximal attendance from our participants from Peru, USA or Korea. We also organized several online meetings in the evening for a panel discussion or for reviving the Oberwolfach spirit in terms of a casual exchange of ideas. Five participants attended in person, which created a particular relaxed yet intense atmosphere for them. This resulted in long blackboard discussions and planning for future research and publications.

The scientific focus of the workshop was on inverse problems, which classically start with an analytical description $F : X \to Y$ of the forward operator in some function spaces X and Y. The main target in inverse problems is to reconstruct an unknown x^* from given noisy data $y^\delta \sim F(x^*)$, where the generalized inverse F^{-1} is unbounded. However, these purely analytic models are typically just an approximation to the real application and their extension are often restricted due to the high degree of complexity or an only partial understanding of the underlying physical processes. Furthermore, the input space of many applications will be just a subspace of the whole function space X and obey an unknown stochastic distribution.

The huge field of machine learning provides several data-driven approaches to tackle these problems by using training datasets to either construct a problem-adapted forward operator and use an established inversion method or to solve the inverse problem directly. In particular deep learning approaches using neural networks with multiple internal layers have become popular over the last decade. However, no consistent mathematical theory on deep neural networks for inverse problems has been developed yet besides the stunning experimental results, which have been published so far for many different types of applications to inverse problems.

The talks focussed on different aspects of deep learning for inverse problems. The opening talk by Jin Keun Seo highlighted several stunning deep learning solutions for applications in medical imaging. This set the stage in terms of open challenges, the need for appropriate mathematical concepts and the specific intricacies of inverse problems. Further survey talks highlighted general concepts for data driven regularization of inverse problems, imaging learning problems or statistical learning theory (Schönlieb, Rosasco, De los Reyes).

New mathematical concepts based on microlocal analysis or reduced order models were introduced for achieving or explaining particular properties of neural networks (Öktem, Hauptmann, de Hoop). Similarly, turning analytic results into deep learning concepts is a major source of inspiration for the design and construction of deep learning schemes, which was addressed in talks on learning penalty terms, incorporating structure preserving architectures or general regularization properties of networks (Etmann, Dittmer, Betcke, Haltmeier).

Complementary to that is the development of a mathematical foundation for existing learning concepts such as explaining the success of U-Net architectures or generalized perceptron learning (Liu, Benning). And, naturally, there were several talks presenting deep learning solution for particularly challenging inverse problems (Siltanen, Boiger, Duff).

Finally we want to highlight the panel discussion chaired jointly by Barbara Kaltenbacher and Simon Arridge, which produces a wealth of ideas for future research for supporting regularization theory to data driven concepts.

Workshop (hybrid meeting): Deep Learning for Inverse Problems

Table of Contents

Abstracts

Deep Learning for ill-posed inverse problems in medical imaging

Jin Keun Seo
(joint work with Chang Min Hyun)

1. Abstract

Recently, with the significant developments in deep learning (DL) techniques, solving underdetermined inverse problems has become one of the major concerns in the medical imaging domain, where underdetermined problems are motivated by the willingness to provide high resolution medical images with as little data as possible, by optimizing data collection in terms of minimal acquisition time, cost-effectiveness, and low invasiveness. DL methods appear to have a strong capability to explore the prior information of the expected images via training data, which allows to deal with the uncertainty of solutions to ill-posed inverse problems. However, there is a tremendous lack of a rigorous mathematical foundation which would allow us to understand the reasons why deep learning methods perform that well. In this talk, we will try to discuss mathematical interpretations of DL-based low-dimensional nonlinear representations of expected solutions to ill-posed inverse problems.

2. Deep learning techniques for Inverse Problems

Inverse problems are to find physical quantities (e.g., electrical impedance in EIT, attenuation in CT, nuclear spin density in MRI) that are observable or measurable and their values change with position and time to form signals. Whether or not an inverse problem is well-posed may be dependent on how the solution is expressed. Many problems are ill-posed because we are overly ambitious or lacking in expressiveness. There are mainly two types of inverse problems. Type 1 is characterized by having data that is much smaller in dimension compared to the input (i.e., undersampled models that violate the Nyquist criteria in the sense that the number of equations is much smaller than the number of unknowns)[4]; Type 2 refers to inaccurate forward models with data contaminated by various noise and artifacts (e.g., inverse problems with forward modeling errors associated with various uncertain factors and with the measured data being insensitive to local perturbation of the input)[3].

The talk starts with the example of lung electrical impedance tomography (EIT), which is known to be a nonlinear and ill-posed inverse problem. As an example of a 16-channel EIT system for respiratory monitoring of sleep apnea, we have to deal with the uncertainty of a number of free parameters (pixel dimension-data dimension = 16384-208) [5]. Deep learning framework may provide a nonlinear regression on training data which acts as learning complex prior knowledge on the output. The first network is a variational autoencoder (VAE) network that allows to achieve compact representation (or low dimensional manifold learning)

for prior information of lung EIT images. For the second step, only the decoder part of this network is used. This decoder part takes only very few latent variables and transforms them back to produce an image on the learned manifold of meaningful reconstructions. The second network now takes a data vector and maps it onto the latent variables, which are then input to the decoder. This approach exploits the potential of neural networks for constructing low dimensional nonlinear representations of approximate solution maps.

For a theoretical analysis, we introduce the M-RIP condition for deep learning-based solvability of ill-posed inverse problems (Type 1) in medicine[1]. Assuming that medical data are on or near a low-dimensional manifold embedded in high-dimensional ambient space, we need to fit a nonlinear solution manifold to training data. We have not yet succeeded in fitting low-dimensional manifolds to real high-dimensional image data using various DL techniques, including autoencoders and GANs. Only in a relatively low-dimensional ambient space, VAEs have achieved somewhat successful manifold learning. Manifold learning as a low-dimensional representation of high-resolution medical images would be an important future research topic. We close with some general remarks for further research directions.

3. Discussion and Challenging Issues

AI algorithms should be explainable and transparent in order that doctors can backtrack AI diagnosis. AI algorithms should be properly configured to reduce black box prediction as much as possible.

Many experiments have shown that well-trained neural networks work only in the immediate vicinity of the regression manifold generated from the training data. Even if two images are almost the same from the viewpoint of radiologists, deep neural networks may produce different results, because they are vulnerable to various noise-like perturbations. Hence, normalizing data is an important part of improving a network's generalization ability (by enhancing out-of-distribution robustness), but it can be very challenging. Data normalization and standardization can reduce diversity in images caused by variation among scanners or imaging protocols [1].

DL techniques have expanded our ability by sophisticated ?disentangled representation learning? though training data, and appear to overcome limitations of existing mathematical methods in handling various ill-posed problems. The DL approach is a completely different paradigm from the classical regularized data-fitting approaches that use a ?single? data-fidelity with regularization, and has excellent ability to learn complex prior knowledge of the output by effectively utilizing prior and additional information as a ?group? data fidelity.

For example, until 2015, achieving automatic segmentation of amniotic fluid from ultrasound images was almost impossible due to the limitations of classical segmentation techniques (e.g. energy-based segmentation methods using active contour or level set). However, this is now achieved by DL technology.

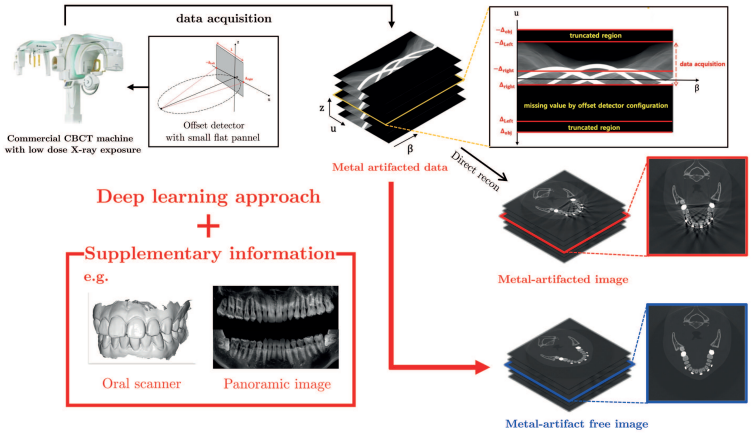


FIGURE 1. Metal artifact reduction in low dose dental cone beam computed tomography.

As another example, metal artifact reduction (MAR) in CT has been a very challenging problem for the last 40 years. In dental cone beam computed tomography (CBCT), metal-induced artifacts are becoming an increasingly common problem, as the number of aged people with artificial prostheses and metallic implants is rapidly increasing with the speedily aging population. Because dental CBCT is designed to use a much lower radiation dose than the conventional multi-detector CT, it tends to produce more artifacts. The field of view size in dental CBCT is usually small as compared to the size of a patient?s head, because a small detector is employed to reduce system costs down [6]. In these reasons, achieving MAR in CBCT is very difficult. Currently, commercially available MAR algorithms include SEMAR (Toshiba Medical Systems), O-MAR (Philips Healthcare), iMAR (Siemens Healthineers), and Smart MAR (GE Healthcare). However, the existing MAR methods do not reduce metal artifacts effectively in low-dose CBCT environments and may introduce new streaking artifacts that did not previously exist. However, by the virtue of the DL technology, this problem can be solved by leveraging information from oral scans that contain accurate 3D images of tooth surface and gingiva in high resolution.

REFERENCES

[1] C. M. Hyun, S. H. Baek, M. Lee, S. M. Lee, J. K. Seo, *Deep Learning-Based Solvability of Underdetermined Inverse Problems in Medical Imaging*, Medical Image Analysis (2021)

[2] H. C. Cho, S. Sun, C. M. Hyuna, J. Kwon, B. Kim, Y. Park, J. K. Seo, *Automated ultrasound assessment of amniotic fluid index using deep learning*, Medical Image Analysis (2021)

[3] H. S. Park, J. Baek, S. K. You, J. K. Choi, and J. K. Seo, *Unpaired image denoising using a generative adversarial network in X-ray CT*, IEEE Access (2019)

[4] C. M. Hyun; H. P. Kim, S. M. Lee, S. Lee; J. K. Seo,, *Deep learning for undersampled MRI reconstruction*, Physics in Medicine and Biology , vol 63, no 13 (2018)

[5] J. K. Seo, K. C. Kim, A. Jargal, K. Lee and B. Harrach, *A learning-based method for solving ill-posed nonlinear inverse problems: a simulation : a simulation study of Lung EIT*, SIAM Journal on Imaging Sciences (2019)
[6] T. Bayaraa, C. M. Hyun, T. J. Jang, S. M. Lee, and J. K. Seo, *A Two-Stage Approach for Beam Hardening Artifact Reduction in Low-Dose Dental CBCT*, IEEE Access (2020)

Microlocal analysis and deep learning for tomographic reconstruction

Ozan Öktem
(joint work with Gitta Kutyniok, Hector Andrade-Loarca, Philip Petersen)

In many applications, the singular part of a function representing a signal carries important information. An example is edges of an image in imaging applications. Understanding how this singular part is transformed by an operator is therefore important in both mathematics and applications. Microlocal analysis is a powerful mathematical theory that offers means for such an analysis. It turns out that the location of singularities (which is already described by the singular support of the function) is not enough for describing how singularities propagate as the function is transformed. For each point in the singular support one also needs to keep track of the directions in frequency space that causes the singularity. Formalising this leads to the definition of the wavefront set of the function.

Since its introduction in the early 1970's by Sato [13] and Hörmander [9], the wavefront set has proven itself useful in both pure and applied mathematical research. One can through the microlocal canonical relation relate the wavefront set of a function to the wavefront set of its transformation for certain operators. In particular, applying a pseudodifferential operator does not introduce new singularities and elliptic operators completely preserve singularities, see [12, section 3.3] for precise formulations. Similar results can be formulated for more general class of Fourier integral operators and many integral operators that are frequently encountered in analysis, scientific computing, and physical sciences [9, 8].

Microlocal analysis in tomography. In applications it is common to have data that represents noisy realisation of a function that has undergone a known transformation (forward operator). A natural task is therefore to recover the function of interest, or a feature thereof, from such data. Such inverse problems frequently arise in applications, like those that involve imaging/sensing technologies where the forward operator is a pseudodifferential or Fourier integral operator. The typical example is tomographic imaging in medicine, which can be phrased as the task of recovering a real-valued function on the plane/space from its ray transform sampled on a known manifold of lines.

In many applications it is sufficient to recover the wavefront set of the signal, like recovering edges of an image from tomographic data. Microlocal analysis has been successfully used to determine when this is possible and it also provides the foundations for reconstruction methods for recovering the wavefront set from noisy indirect observations. As an example, one can show that a point (=singularity) in the wavefront set of a function is recoverable from ray transform data if and only

if the ray transform is sampled on a line that goes through the point and has a co-normal that is in the wavefront set. The recovery is moreover mildly ill-posed in the sense that singularities in data are weaker than those of the function by 1/2 Sobolev order (strong enough to allow stable detection in practice). The reader may consult the surveys [10, 12] for further details.

In contrast to its theoretical successes, microlocal analysis has had more limited impact in computational signal processing. A key reason has been the difficulty in computationally extracting and manipulating the wavefront set of a digitised signal. Certain transforms from applied harmonic analysis, like the curvelet and shearlet transform, offer an alternative possibility to identify the wavefront set. In particular, the connection between the behaviour of these transforms, and the wavefront set has been analysed in [7, 11]. These approaches characterise the wavefront set through the rate of decay of the respective transforms, which in turn becomes computationally unfeasible in large-scale signal processing applications.

Deep learning and microlocal analysis. As shown in [4], a successful wavefront set extractor needs to be tailored to the function class of interest. The relevant function classes in applications are, however, difficult to characterise analytically. In fact, it is shown in [4] that one cannot extract a 'digital' wavefront set of L^2-functions in a consistent manner. An alternative is to formulate wavefront set extraction as a statistical estimation problem and then learn a wavefront set extraction operator from supervised data by deep learning. This led to the approach in [4] of training a deep neural network classifier, DeNSE, to predict the wavefront set from the shearlet coefficients of surpervised training data. DeNSE was successfully used for edge extraction [4] and later in [5] it was also used for tomographic image reconstruction.

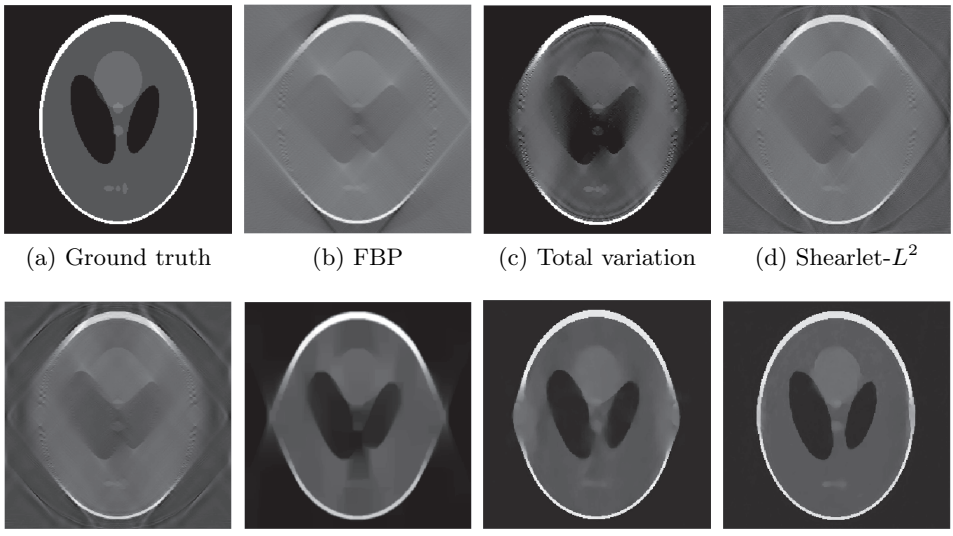
Microlocally consistent deep neural networks for reconstruction. Empirical experience shows that performance of deep learning approaches for reconstruction significantly improves if one accounts for how data is generated (forward model), e.g., as in [1, 3] for tomographic image reconstruction. Such domain adapted deep neural networks can be assembled by unrolling a suitable iterative scheme [3, Sec. 4.9.1].

A natural next step is to aim for further domain adaptation, like encoding a priori knowledge of the singularities that are recoverable (microlocal characterisation is visible singularities) and how these relate to singularities in data (microlocal canonical relation). This ensures the reconstruction is microlocally consistent.

Our approach for ensuring microlocal consistency in tomographic reconstruction builds on the learned-primal dual (LPD) approach in [1]. We use the framework in [2] to encode the above knowledge about singularities into the LPD approach. This results in a microlocally consistent version of LPD (microlocal LPD) that has the following four components: (1) The DeNSE neural network is used to extract a digital wavefront set of data. (2) The canonical relation for the nonlinear reconstruction operator given by LPD can be derived and this allows us to map the digital wavefront set in data, which was provided by DeNSE, to a visible digital wavefront set in the image domain. (3) The microlocal characterisation

of visible singularities can then be used to split a wavefront set in the image domain into visible and invisible parts. (4) The task of mapping the visible part of a (digital) wavefront set into the full wavefront set is phrased as a statistical estimation problem (wavefront set in-painting) that is solved by a trained a deep neural network. The framework in [2] is then used to combine this network with the one in the LPD method.

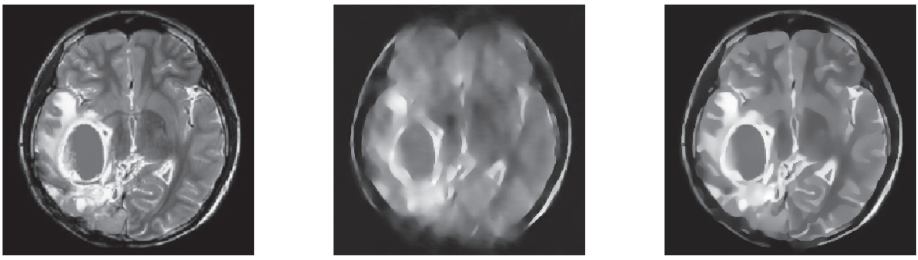
The resulting microlocal LPD is a deep neural network for reconstruction that transforms singularities in a microlocally consistent manner. This has major advantages in limited data problems, like limited angle tomography in the plane as shown in figures 1 and 2.



(a) Ground truth (b) FBP (c) Total variation (d) Shearlet-L^2

(e) Shearlet-L^1 (f) PhantomNet [6] (g) LPD [1] (h) Microlocal LPD

FIGURE 1. Limited angle tomography (40° missing wedge). Reconstruction in (b) is by filtered backprojection (FBP) whereas the total variation (c) and shearlet compressed sensing approaches (d)-(e) are based on sparsity promoting regularisation. Reconstructions in (f)–(h) are obtained by deep learning approaches that are trained against limited angle tomography data. The one in (h) obtained by the microlocal LPD method is clearly the best.



(a) Ground truth (b) LPD (c) Microlocal LPD

FIGURE 2. Same set-up as in Figure 1, but this time with a more complex ground true image (a). The images shows the benefit of enforcing microlocal consistency by comparing LPD (b) against its microlocal consistent variant (c).

REFERENCES

[1] J. Adler and O. Öktem, *Learned primal-dual reconstruction*, IEEE Transactions on Medical Imaging **37**(6) (2018), 1322–1332.

[2] J. Adler, S. Lunz, O. Verdier, C.-B. Schönlieb, and O. Öktem, *Task adapted reconstruction for inverse problems*, ArXiv **cs.CV**:1809.00948 (2018). Submitted to Inverse Problems.

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

[4] H. Andrade-Loarca, G. Kutyniok, O. Öktem, and P. Petersen, *Extraction of digital wavefront sets using applied harmonic analysis and deep neural networks*, SIAM Journal on Imaging Sciences **12**(4) (2019), 1936– 1966.

[5] H. Andrade-Loarca, G. Kutyniok, and O. Öktem, *Shearlets as feature extractor for semantic edge detection: The model-based and data-driven realm*, Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences **476**(2243) (2020), 20190841.

[6] T. A. Bubba, G. Kutyniok, M. Lassas, M. März, W. Samek, S. Siltanen, and V. Srinivasan, *Learning the invisible: A hybrid deep learning-shearlet framework for limited angle computed tomography*, Inverse Problems **35**:064002 (2019).

[7] E. J. Candés and D. L. Donoho, *Continuous curvelet transform: I. resolution of the wavefront set*, Applied and Computational Harmonic Analysis **19**(2) (2005), 162–197.

[8] E. J. Candés, L. Demanet, and L. Ying, *Fast computation of Fourier integral operators*, SIAM Journal on Scientific Computing **29**(6) (2007), 2464–2493

[9] L. Hörmander, *Fourier integral operators. I*, Acta Mathematica **127** (1971), 79–183.

[10] V. P. Krishnan and E. T. Quinto, *Microlocal analysis in tomography*, volume 2 of *Handbook of Mathematical Methods in Imaging*. Springer, 2nd edition, 2015.

[11] G. Kutyniok and D. Labate, *Resolution of the wavefront set using continuous shearlets*, Transactions of the American Mathematical Society **361**(5) (2009), 2719–2754.

[12] M. Salo, *Applications of Microlocal Analysis in Inverse Problems*, Mathematics **8**(7) (2020), 1184.

[13] M. Sato, *Regularity of hyperfunctions solutions of partial differential equations*. In Actes du Congrès international des mathématiciens, volume 2, pages 785?794, Paris, 1971. Gauthier-Villars.

Globally injective ReLU networks, injective flows and uncertainty quantification

Maarten de Hoop

(joint work with Ivan Dokmanić, AmirEhsan Khorashadizadeh, Konik Kothari, Matti Lassas, Michael Puthawala)

Injectivity plays an important role in generative models where it enables inference; in inverse problems and compressed sensing with generative priors it is a precursor to well posedness. We establish sharp characterizations of injectivity of fully-connected and convolutional ReLU layers and networks. First, through a layerwise analysis, we show that an expansivity factor of two is necessary and sufficient for injectivity by constructing appropriate weight matrices. We show that global injectivity with iid Gaussian matrices, a commonly used tractable model, requires larger expansivity between 3.4 and 10.5. We also characterize the stability of inverting an injective network via worst-case Lipschitz constants of the inverse. We then use arguments from differential topology to study injectivity of deep networks and prove that any Lipschitz map can be approximated by an injective ReLU network. Finally, using an argument based on random projections, we show that an end-to-end – rather than layerwise – doubling of the dimension suffices for injectivity. We then use these result to generalize invertible normalizing flows to obtain injective flows. Our results establish a theoretical basis for the study of nonlinear inverse and inference problems using neural networks.

Reduced models in learned image reconstruction

Andreas Hauptmann

(joint work with Simon Arridge, Sebastian Lunz, Carola-Bibiane Schönlieb, Tanja Tarvainen)

1. Model-corrections in inverse problems

In applications where the forward model is given by the solution of a partial differential equation, model reduction techniques are often used to reduce computational cost, which leads to known approximation errors. Here, we will discuss how such model errors can be corrected with data-driven methods and used for image reconstruction. In what follows, we restrict ourselves to linear inverse problems. Let $x \in X$ be the unknown quantity of interest we aim to reconstruct from measurements $y \in Y$, where X, Y are Hilbert spaces and x and y fulfil the relation

$$Ax = y, \tag{1}$$

where $A : X \to Y$ is the accurate forward operator. We assume that the evaluation of the accurate operator A is computationally expensive and we rather want to use an approximate model $\widetilde{A} : X \to Y$, which introduces an inherent approximation

$$\widetilde{A}x = \widetilde{y}. \tag{2}$$

leading to a systematic model error $\delta y = y - \widetilde{y}$. In the following we will discuss two principled approaches how we can use an approximate model in the framework of learned image reconstruction and the possibility to establish reconstruction guarantees.

2. Implicit model correction

The most straight-forward approach is to directly use the approximate model in the framework of learned iterative reconstructions [1, 2]. That is, we aim to formulate a network Λ_Θ, that is designed to mimic a gradient descent scheme. In particular, we train the networks to perform an iterative update, such that

$$x_{k+1} = \Lambda_\Theta \left(\nabla_x \frac{1}{2} \|Ax_k - y\|_Y^2, x_k \right), \tag{3}$$

where $\nabla_x \frac{1}{2}\|Ax_k - y\|_Y^2 = A^*(Ax_k - y)$. Now, one could use an approximate model instead of the accurate model and compute an approximate gradient given by $\widetilde{A}^*(\widetilde{A}x_k - y)$ for the update in (3), as proposed in [3]. The network Λ_Θ then *implicitly* corrects the model error to produce the new iterate. That means, the correction and regularisation are hence trained simultaneously with the update in (3). Such approaches are typically trained by using a loss function, like the L^2-loss, to measure the distance between reconstruction and a ground truth phantom. This way a substantial speed-up, compared to classical variational approaches, can be achieved with improved reconstruction quality. Nevertheless, such implicit corrections offer limited insights into how approximate models are corrected for and hence we consider in the following an *explicit* correction that can then be subsequently used in a variational framework.

3. Explicit model correction and a convergence result

Let us now consider corrections for this approximation error via a parameterisable, possibly nonlinear, mapping $F_\Theta : Y \to Y$, applied as a correction to $\widetilde{A}$. Typically, this mapping would be given by a (convolutional) neural network. This leads to a corrected operator A_Θ of the form

$$A_\Theta = F_\Theta \circ \widetilde{A}. \tag{4}$$

We aim to choose the correction F_Θ such that ideally $A_\Theta(x) \approx Ax$ for some $x \in X$ of interest. The primary question that we aim to answer is, whether such corrected models (4) can be subsequently used in variational regularisation approaches. Thus, it is natural to require that the obtained solutions involving the corrected operator A_Θ and the accurate operator A, are close, that is

$$\operatorname*{arg\,min}_{x \in X} \frac{1}{2}\|A_\Theta(x) - y\|_Y^2 + \lambda R(x) \approx \operatorname*{arg\,min}_{x \in X} \frac{1}{2}\|Ax - y\|_Y^2 + \lambda R(x), \tag{5}$$

with regularisation functional R and associated hyper-parameter λ. Solutions are then usually computed by an iterative algorithm. Here we consider first order methods to draw connections to learned iterative schemes as in (3). In particular,

we consider a classic gradient descent scheme, assuming differentiable R. Then, given an initial guess x_0, we can compute a solution by the iterative process

$$x_{k+1} = x_k - \gamma_k \nabla_x \left(\frac{1}{2} \|A_\Theta x_k - y\|_Y^2 + \lambda R(x_k) \right), \tag{6}$$

with appropriately chosen step size $\gamma_k > 0$. When using (6) for the corrected operator it seems natural to ask for a *gradient consistency* of the approximate gradient $\nabla_x \|A_\Theta(x) - y\|_Y^2 \approx \nabla_x \|Ax - y\|_Y^2$. We recall that the correction F_Θ in (4) is given by a nonlinear neural network and following the chain rule we obtain

$$\frac{1}{2} \nabla_x \|A_\Theta(x) - y\|_2^2 = \widetilde{A}^* \left[DF_\Theta(\widetilde{A}x) \right]^* \left(F_\Theta(\widetilde{A}x) - y \right). \tag{7}$$

Here, we denote by $DF_\Theta(y)$ the Fréchet derivative of F_Θ at y, which is a linear operator $Y \to Y$. That means, to satisfy the gradient consistency condition, we would need

$$\widetilde{A}^* \left[DF_\Theta(\widetilde{A}x) \right]^* \left(F_\Theta(\widetilde{A}x) - y \right) \approx A^*(Ax - y). \tag{8}$$

This solution comes with its own drawback: the range of the corrected fidelity term's gradient (7) is limited by the range of the approximate adjoint, $\mathbf{rng}(\widetilde{A}^*)$. Thus, we identify the key difficulty here in the differences of the range of the accurate and the approximate adjoints rather than the differences in the forward operators themselves. Indeed, a correction of the forward operator via composition with a parametrised model F_Θ in measurement space is not able to yield gradients close to the gradients of the accurate data term if $\mathbf{rng}(\widetilde{A}^*)$ and $\mathbf{rng}(A^*)$ are too different, see also Theorem 3.1 in [4].

3.1. **Obtaining a Forward-Adjoint Correction.** To achieve a gradient consistent model correction we propose to learn two networks instead. That is, we learn a network F_Θ that corrects the forward model and another network G_Φ that corrects the adjoint, such that we have

$$A_\Theta := F_\Theta \circ \widetilde{A}, \quad A_\Phi^* := G_\Phi \circ \widetilde{A}^*$$

These corrections can then be obtained as follows. Given a set of training samples (x^i, Ax^i), we train the forward correction F_Θ acting in measurement space Y, for the adjoint we train the network G_Φ acting on image space X, that yields the two losses

$$\min_\Theta \sum_i \|F_\Theta(\widetilde{A}x^i) - Ax^i\|_Y \text{ and } \min_\phi \sum_i \|G_\Phi(\widetilde{A}^* r^i) - A^* r^i\|_X. \tag{9}$$

Here, we can choose the direction $r^i = F_\Theta(\widetilde{A}x^i) - y^i$ for the adjoint loss. This ensures that the adjoint correction is in fact trained in directions relevant when solving the variational problem. We can then use both corrections to compute approximate gradients of the data fidelity term $\|Ax - y\|_Y^2$ as

$$A^*(Ax - y) \approx \left(G_\Phi \circ \widetilde{A}^* \right) \left(F_\Theta(\widetilde{A}x) - y \right). \tag{10}$$

To establish our convergence results, we can now consider the two functionals

$$\mathcal{L}(x) := \frac{1}{2}\|Ax - y\|_Y^2 + \lambda R(x),\ \mathcal{L}_\Theta(x) := \frac{1}{2}\|A_\Theta(x) - y\|_Y^2 + \lambda R(x)$$

and using the forward-adjoint correction we obtain, under suitable conditions outlined in [4], the main theorem.

Theorem 3.1 (Convergence to a neighbourhood of the accurate solution $\hat{x}$ [4])**.** *Let $\epsilon > 0$ and suitable δ (controlling the subdifferential of $\mathcal{L}_\Theta$). Assume both adjoint and forward operator are fit up to a $\delta/4$-margin, i.e.*

$$\|A\|_{X\to Y}\|(A - A_\Theta)(x_n)\|_Y < \delta/4, \quad \|(A^* - A_\Phi^*)(A_\Theta(x_n) - y)\|_X < \delta/4$$

for all y and x_n obtained during gradient descent over $\mathcal{L}_\Theta$. Then eventually the gradient descent dynamics over $\mathcal{L}_\Theta$ will reach an ϵ neighbourhood of the accurate solution $\hat{x}$.

References

[1] J. Adler, O. Öktem. *Solving ill-posed inverse problems using iterative deep neural networks*, Inverse Problems **33(12)** (2017), 124007.

[2] A. Hauptmann, F. Lucka, M. Betcke, N. Huynh, J. Adler, B. Cox, P. Beard, S. Ourselin, S. Arridge, *Model-based learning for accelerated, limited-view 3-d photoacoustic tomography*, IEEE Transactions on Medical Imaging **37(6)** (2018), 1382–1393.

[3] A. Hauptmann, B. Cox, F. Lucka, N. Huynh, M. Betcke, P. Beard, S. Arridge, *Approximate k-space models and deep learning for fast photoacoustic reconstruction*, In International Workshop on Machine Learning for Medical Image Reconstruction (2018), 103–111. Springer, Cham.

[4] S. Lunz, A. Hauptmann, T. Tarvainen, C.B. Schönlieb, S. Arridge, *On learned operator correction in inverse problems*, SIAM Journal on Imaging Sciences, **14(1)** (2021), 92–127.

Sparsity and NETT regularization for compressed sensing photoacoustic tomography

Markus Haltmeier

Photoacoustic tomography (PAT) is an emerging imaging technique that combines the high resolution of ultrasound imaging with the high contrast of optical tomography [6]. In PAT, a semi-transparent sample is illuminated by short pulses of optical energy, which induces an acoustic pressure wave $p\colon \mathbb{R}^3 \times [0,\infty) \to \mathbb{R}$ depending on spatial position $x \in \mathbb{R}^3$ and time $t \geq 0$. The initial pressure distribution $f\colon \mathbb{R}^3 \to \mathbb{R}$ is proportional to the internal light absorption characteristics of the sample and provides valuable diagnostic information. Detectors located on a measurement surface S that (partially) surrounds the sample measure the acoustic pressure from which the initial pressure distribution is recovered. In the case of complete data, several reconstruction methods for recovering the initial pressure have been developed including the variable and the constant sound speed case. Here, we model the acoustic pressure by the standard wave equation $\partial_t^2 p - \Delta_x p = 0$ and write $\mathrm{W} f := p|_{S\times[0,\infty)}$ for the restriction of the acoustic pressure to the measurement surface.

Data $\mathrm{W}\, f$corresponds to full continuously sampled data in PAT. In practical application, however, the pressure can only be measured at a finite number of sampling points. This means we measure data $p(z_k, \cdot)$ with sensor locations $z_1, z_2, \ldots, z_n$. As shown in [3], using classical Shannon sampling theory, the number of spatial measurements determines the resolution of the reconstructed PA source. In practice, however, collecting a large number of spatial measurements requires either a large number of parallel data acquisition channels or a large number of sequential measurements. This either increases the cost and technical complexity of the system or significantly increases measurement time. In order to reduce the number of detectors while maintaining spatial resolution, CSPAT has been investigated in several works [2, 4]. The basic idea is to use general linear measurements of the form $y = \mathrm{A}\ \mathrm{W}\, f$, where A is a linear measurement operator that, that only acts in the spatial variable. The term compressed sensing refers to the fact that the number of measurements is to be chosen is much smaller than the number of initial sampling points. In such a situation, $y = \mathrm{A}\ \mathrm{W}\, f$ constitutes a highly underdetermined linear system of equations and can only be solved with additional information on the unknown to be recovered.

In this talk we presented two different classes of reconstruction methods for compressed PAT that use different type of prior information accounting for the non-uniqueness of the underlying reconstruction problem. The first class of methods is based on sparsity. In this class we considered a two-step approach where the operators A and W are inverted consecutively. Another sparsity based method that we presented and that has been introduced in [4] is based on an intertwining relation for the solution operator of the wave equation with the Laplacian. Applying the second derivative to the modeling equation results in the linear equation $\partial_t^2 y = \mathrm{A}\ \mathrm{W}(\Delta f)$. Assuming a sparsity prior on Δf in [4] it is proposed to jointly recover f and $h = \delta f$ by minimizing

$$\frac{1}{2}\|\,\mathrm{A}\ \mathrm{W}\, f - y\|_2^2 + \frac{1}{2}\|\,\mathrm{A}\ \mathrm{W}\, h - \partial_t^2 y\|_2^2 + \frac{\alpha}{2}\|\Delta f - h\|_2^2 + \lambda\|\Delta\|_1 + I_C(f)\,,$$

where α is a tuning parameter and λ the regularization parameter. Moreover, I_C implements a positivity constraint, i.e. with $C = [0, \infty)^n$, the function I_C is defined by $I_C(f) = 0$ if $f \in C$ and $I_C(\mathbf{f}) = \infty$ otherwise. Another method that we presented considers minimizers of the optimization problem

$$\mathcal{N}_\theta(f) = \frac{1}{2}\|\,\mathrm{A}\ \mathrm{W}\, f - y\|_2^2 + \frac{\lambda}{2}\,\mathcal{R}_\theta(f)\,, \tag{1}$$

where $\mathcal{R}_\theta$ is a trained regularizer and $\lambda > 0$ the regularization parameter. The resulting reconstruction approach is called NETT (for network Tikhonov regularization), as it is a generalized form of Tikhonov regularization using a NN as trained regularizer. In [5] it has been shown that under reasonable conditions, the NETT approach is well-posed and yields a convergent regularization method. In particular, minimizers of (1) exist, are stable with respect to data perturbations, and minimizers of (1) converge to $\mathcal{R}_\theta$-minimizing solutions of the equation $\mathrm{A}\ \mathrm{W}\, f = y$ as the noise level goes to zero. The regularizer includes prior given in the form of training data $f_1, \ldots, f_N$ and is constructed such that is has a small

value for f_i and large values for reconstructions of f_i with artifacts which may be due to under-sampling or noise. For details we refer to [1].

References

[1] S. Antholzer, J. Schwab, J. Bauer-Marschallinger, Peter B., and Markus H. NETT regularization for compressed sensing photoacoustic tomography. In *Photons Plus Ultrasound: Imaging and Sensing 2019*, volume 10878, page 108783B. International Society for Optics and Photonics, 2019.

[2] S. Arridge, P. Beard, M. Betcke, B. Cox, N. Huynh, F. Lucka, O. Ogunlade, and E. Zhang. Accelerated high-resolution photoacoustic tomography via compressed sensing. *Phys. Med. Biol.*, 61(24):8908, 2016.

[3] M. Haltmeier. Sampling conditions for the circular Radon transform. *IEEE Trans. Image Process.*, 25(6):2910–2919, 2016.

[4] M. Haltmeier, M. Sandbichler, et al. A sparsification and reconstruction strategy for compressed sensing photoacoustic tomography. *J. Acoust. Soc. Am.*, 143(6):3838–3848, 2018.

[5] H. Li, J. Schwab, S. Antholzer, and M. Haltmeier. NETT: solving inverse problems with deep neural networks. *Inverse Problems*, 36(6):065005, 2020.

[6] M. Xu and L. V. Wang. Photoacoustic imaging in biomedicine. *Rev. Sci. Instrum.*, 77(4):041101, 2006.

Learning and interpolation with Matérn kernels

Lorenzo Rosasco

(joint work with Nicolò Pagliana, Alessandro Rudi, Ernesto De Vito)

1. Learning & interpolation

Let $(\mathcal{X} \times \mathcal{Y}, \rho)$ be a probability space with $\mathcal{X} \subseteq \mathbb{R}^d, \mathcal{Y} \subseteq \mathbb{R}$. Supervised learning with least squares problem amounts to estimating the minimizer of the *expected risk*

$$f_\rho = \operatorname*{arg\,min}_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{E}(f) \qquad \mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} (f(x) - y)^2 \, \mathrm{d}\rho(x, y) \; .$$

having access to ρ only through a finite number of data $\mathbf{z} = (\mathbf{x}, \mathbf{y}) = (x_i, y_i)_{i=1}^n$.

A classic idea in statistical learning is that for ood estimation there should be a trade off between fitting the examples $\mathbf{z}$ and controlling the complexity of the solution. Just focusing on fitting the data may lead to over-fitting. This classic point of view is contrasted by recent empirical observations suggesting that with over-parameterized models it is often possible to fit the data arbitrary well without degrading learning accuracy. This phenomenon has been observed in neural networks, but also other models like kernel methods, and begs the question of whether interpolation can be reconciled with classical learning theory [1, 2, 3].

Here, we consider the minimum norm interpolating estimator in the RKHS $\mathcal{H}$ defined by a kernel k,

$$\widehat{f} = \operatorname*{arg\,min}_{f\in\mathcal{H}} \|f\|_{\mathcal{H}}^2 \quad \text{such that} \quad f(x_i) = y_i \quad \forall i \in \{1, \dots, n\}.$$

Such an estimator can be explicitly computed as

$$\widehat{f}(x) = k(x, \mathbf{x})^\top \widehat{K}^{-1}\mathbf{y} \quad \text{where} \quad (k(x, \mathbf{x}))_i = k(x, x_i)\,, \;\; \widehat{K}_{i,j} = k(x_i, x_j)$$

This estimator can be seen as the limit of the ridge regression estimator where $\widehat{K}^{-1}$ is replaced by $(\widehat{K} + \lambda n\,\mathrm{I})^{-1}$ for λ going to 0.

2. Excess Risk for Interpolating Kernel Estimators

For the sake of simplicity, we assume a well specified model, that is $f_\rho \in \mathcal{H}$.

Then, we derive the following high probability bound on the excess risk,

$$\mathcal{E}(\widehat{f}) - \mathcal{E}(f_\rho) \leq \sigma^2 \frac{\tau_n d_{\mathrm{eff}}(\tau_n)}{\sigma_{\min}(\widehat{K})} + \left\| (\widehat{P} - \mathrm{I}) f_\rho \right\|_\rho^2 \quad \text{with} \quad \tau_n \approx \frac{\log n}{n}$$

where σ is the variance of $y - f_\rho(x)$, $\sigma_{\min}(\widehat{K})$ denotes the minimum eigenvalue of the kernel matrix $\widehat{K}$, d_{eff} quantifies the dimension of the hypothesis space $\mathcal{H}$ and is called *effective dimension* [4], $\widehat{P}$ is the orthonormal projection from $\mathcal{H}$ to $\widehat{\mathcal{H}} = \operatorname{span}(k(\cdot, x_i) \colon x_i \in \mathbf{x})$ and $\|\cdot\|_\rho$ denotes the L^2 norm induced by the measure ρ over $\mathcal{X}$.

Comments: the first term in the error bound quantify the variance of the estimator and strongly depend on the noise σ and on the *stability* of the method, where the stability is encoded in the ratio $\frac{d_{\mathrm{eff}}(\tau_n)}{\sigma_{\min}(\widehat{K})}$ which depend on the dimension of the hypothesis space $\mathcal{H}$ and on the numerical stability of the empirical method through $\sigma_{\min}(\widehat{K})$. The second term is the error of approximating a function in $\mathcal{H}$ with its projection into a finite dimensional random subspace $\widehat{\mathcal{H}}$. This term is also know as interpolation error in approximation theory.

3. Learning with Matérn Kernels

We specialize the above bound to a family of kernel functions, the Matérn kernel defined as

$$k(x, x') = Q(x - x') \quad \text{with} \quad Q(z) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\gamma} \|z\| \right)^\nu \mathbf{K}_\nu \left(\frac{\sqrt{2\nu}}{\gamma} \|z\| \right).$$

where γ and ν are respectively the *bandwidth* and *smoothness* of the kernel. This family of kernels allows to consider very smooth kernels for $\nu \gg 1$ as well as not very smooth kernel for $\nu = 0.5$ (corresponding to the Laplace kernel. Different Sobolev spaces $W^{\nu+d/2}$ are associated to each kernel. With this choice of kernels we can estimate the quantities in the bound to obtain

$$\mathcal{E}(\widehat{f}) - \mathcal{E}(f_\rho) \lesssim C_1\; \sigma^2 \gamma^{\frac{4\nu^2}{d+2\nu}} n^{\frac{4\nu}{d}} + C_2 \left(\frac{1}{\gamma^d\; n} \right)^{1+\frac{2\nu}{d}} \|f_\rho\|_{W_2^{\nu+d/2}(\mathcal{X})}\,.$$

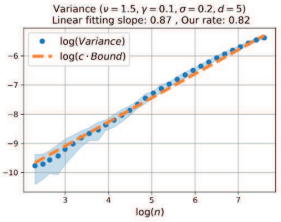
Comments: We observe that the stability term increases with a large number of data and benefits from choosing a small bandwidth γ. The intuition behind this is that the minimum eigenvalue of the kernel matrix depends on the distances between the points and as soon the points start to be close (e.g. when $n \sim w^2$)

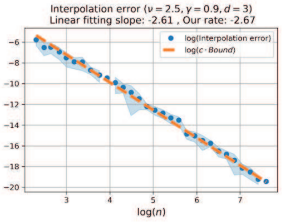
then the method will be forced to find a function which has different values for close points and this will lead to instability. This instability can be controlled by the bandwidth γ which controls the area in which one points influence the others and by the smoothness ν which allows to consider larger space of functions.

The interpolation error decrease with the number of points, the decrease is faster for large ν and γ because in this case because in this case our target f_ρ is in a smaller space and it is easier to reconstruct. We also observe that if decreasing γ will improve the stability term then it will make worse the interpolation error leading to a trade off. From the bound, we can see that even without explicit regularization, the stability of the method can be determined by different factor like the kernel itself, in particular its smoothness and bandwidth. While we cannot expect the method to perform well as the number of point becomes exponential in the input dimension [5, 6], we might still perform well with no need of extra regularization terms.

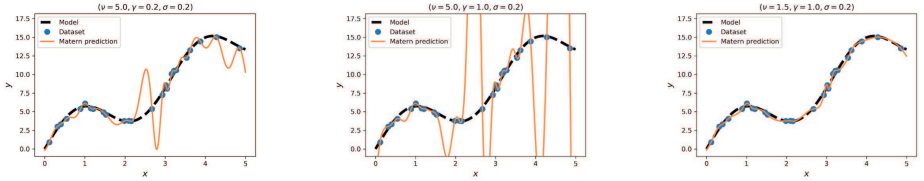
4. Numerical Simulations

To illustrate the result and test the validity of the bounds, we consider simulated data obtained by the model $y_i = f_\rho(x_i) + \epsilon_i$ for a fixed function f_ρ, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and x_i uniform distributed over $[0, 1]^d$. Here we plot the comparison of the variance and interpolation error (blue dots) with our theoretical rate in n (orange) and observe that our bound are tight in this cases.





Here we plot some 1-dimensional simulation of the comparison between the model function f_ρ (black) and our estimator $\widehat{f}$ (orange). In the center we observe unstable behavior of the estimator, while in the left and in the right we show how the parameters γ and ν respectively stabilize the estimator. The numerical results show that indeed the bounds can predict the qualitative behavior of the algorithm.



References

[1] Zhang, Chiyuan and Bengio, Samy and Hardt, Moritz and Recht, Benjamin and Vinyals, Oriol, *Understanding Deep Learning Requires Rethinking Generalization*, International Conference on Learning Representations (ICLR) (2017).

[2] Belkin, Mikhail and Ma, Siyuan and Mandal, Soumik, *To understand deep learning we need to understand kernel learning*, arXiv preprint arXiv:1802.01396 (2018).
[3] Belkin, Mikhail and Hsu, Daniel and Ma, Siyuan and Mandal, Soumik, *Reconciling modern machine-learning practice and the classical bias–variance trade-off*, Proceedings of the National Academy of Sciences (2019).
[4] Caponnetto, A. and De Vito, E., *Optimal rates for the regularized least-squares algorithm*, Foundations of Computational Mathematics (2007).
[5] Rakhlin, Alexander, and Xiyu Zhai, *Consistency of interpolation with laplace kernels is a high-dimensional phenomenon*, In Conference on Learning Theory (PMLR), (2019).
[6] Bartlett, Peter L., Philip M. Long, Gábor Lugosi, and Alexander Tsigler, *Benign overfitting in linear regression*, Proceedings of the National Academy of Sciences 117, no. 48 (2020): 30063-30070.

On the structure of bilevel imaging learning problems

Juan C. De los Reyes
(joint work with D. Villacís)

In recent years, novel optimization ideas have been applied to image restoration tasks in combination with machine learning approaches, to improve the reconstruction of images by optimally choosing different quantities/functions of interest. A fruitful approach in this sense is bilevel optimization, where the imaging problems are considered as lower-level constraints, while on the upper-level a loss function based on a training set is used (see, e.g., [2, 3]).

Let us consider a training dataset of pairs $(u_k^{\text{train}}, f_k)$, for $k = 1, \dots, K$, where each u_k^{train} corresponds to ground-truth data and f_k to the corresponding noisy image. We consider the following class of *bilevel optimization* problems with variational imaging models as lower-level constraints:

$$\text{(1a)} \qquad \underset{(\lambda,\sigma,\alpha,\beta)}{\text{minimize}} \quad \sum_{k=i}^{K} \ell(u_k; u_k^{\text{train}})$$

$$\text{(1b)} \qquad \text{subject to} \quad u_k = \underset{v\in\mathbb{R}^n}{\arg\min}\, \mathcal{E}(v, \lambda, \sigma, \alpha, \beta; f_k),$$

$$\text{(1c)} \qquad P(\lambda), R(\sigma), Q(\alpha), S(\beta) \geq 0,$$

where the variational energy is given by

$$\begin{aligned}\mathcal{E}(v, \lambda, \sigma, \alpha, \beta; f_k) := &\sum_{j=1}^{K}\sum_{i=1}^{k_j} P_j(\lambda_j)_i \mathcal{D}_j(v; f_k)_i + \sum_{j=1}^{L}\sum_{i=1}^{l_j} R_j(\sigma_j)_i |(\mathbb{B}_j v)_i| \\ &+ \sum_{j=1}^{M}\sum_{i=1}^{m_j} Q_j(\alpha_j)_i \|(\mathbb{K}_j v)_i\|_2 + \sum_{j=1}^{N}\sum_{i=1}^{n_j} S_j(\beta_j)_i \|(\mathbb{E}_j v)_i\|_F,\end{aligned}$$

with vector parameters $\lambda_j, \sigma_j, \alpha_j, \beta_j$ and operators $P_j : \mathbb{R}^{|\lambda_j|} \mapsto \mathbb{R}^{k_j}$, $R_j : \mathbb{R}^{|\sigma_j|} \mapsto \mathbb{R}^{l_j}$ $Q_j : \mathbb{R}^{|\alpha_j|} \mapsto \mathbb{R}^{m_j}$ and $S_j : \mathbb{R}^{|\beta_j|} \mapsto \mathbb{R}^{n_j}$, which are assumed to be at least twice continuous differentiable, and encompass the scalar, the scale-dependent and the patch-based regularization cases. The functions $\mathcal{D}_j(v; f_k)_i$ correspond to different data fidelity models that may be considered at once. The notation $|\cdot|$, $\|\cdot\|_2$ and

$\|\cdot\|_F$ stands for the absolute value, the Euclidean norm and the Frobenius norm, respectively.

The bilevel optimization problem structure (1) is involved to be analyzed, as classical nonlinear or bilevel programming results cannot be directly utilized: Standard *constraint qualification conditions* fail for all nontrivial cases [4]. As a remedy, tools from nonsmooth variational analysis (see, e.g., [5, 6]) have to be employed to cope with the difficulties related with the lack of differentiability of the solution mapping or the failure of standard constraint qualifications.

For instance, in the case of a single training pair (u^{train}, f) and a scale-dependent gaussian total variation denoising model

$$u = \arg\min_{v\in\mathbb{R}^n} \sum_{i=1}^{n} \lambda_i |v_i - f_i|^2 + \sum_{i=1}^{n} |(\mathbb{K}v)_i|_2,$$

with $\mathbb{K}$ the discrete gradient operator, an *M-stationay point* may be characterized through the existence of Lagrange multipliers $(\mathbb{K}^\top\mu, p)$ that satisfy:

$$\begin{aligned}
&\lambda \circ p + \mathbb{K}^\top\mu = \nabla_u \ell(u; u^{\text{train}}),\\
&(u-f)\circ p + \vartheta = 0,\\
&\mu_j = \frac{1}{|(\mathbb{K}u)_j|_2}\left[(\mathbb{K}p)_j - \langle (\mathbb{K}p)_j, q_j\rangle\ q_j\right], && \text{if } (\mathbb{K}u)_j \neq 0,\\
&(\mathbb{K}p)_j = 0, && \text{if } (\mathbb{K}u)_j = 0,\ |q_j|_2 < 1,\\
&\left.\begin{array}{l}(\mathbb{K}p)_j = 0 \vee\\ (\mathbb{K}p)_j = cq_j (c\in\mathbb{R}),\ \langle \mu_j, q_j\rangle = 0, \vee\\ (\mathbb{K}p)_j = cq_j (c\geq 0),\ \langle \mu_j, q_j\rangle \geq 0,\end{array}\right\} && \text{if } (\mathbb{K}u)_j = 0,\ |q_j|_2 = 1,\\
&0 \leq \lambda \perp \vartheta \geq 0.
\end{aligned}$$

Moreover, a detailed study of the nonsmooth properties of the lower-level solution operator may lead to the design of novel solution algorithms or neural network architectures for hyperparameter learning. In particular, in the case of variational imaging models such as (1b) it can be proved that the solution operator is locally Lipschitz continuous and directionally differentiable, which is already a valuable property. In addition, a proper chracterization of the Bouligand subdifferential may be obtained, which leads to a generalized adjoint system and convergence properties of nonsmooth trust-region algorithms for the solution of the bilevel instances [1].

References

[1] Constantin Christof and Juan C. De los Reyes and Christian Meyer, *A nonsmooth trust-region method for locally Lipschitz functions with application to optimization problems constrained by variational inequalities*, SIAM J. Optim., **30(3)** (2020), 2163-2196.

[2] Juan C. De los Reyes and Carola B. Schönlieb, *Image denoising: learning the noise model via nonsmooth PDE-constrained optimization*, Inverse Problems & Imaging, **7(4)** (2013), 1183–1214.

[3] Juan C. De los Reyes and Carola B. Schönlieb, and Tuomo Valkonen, *Bilevel parameter learning for higher-order total variation regularisation models*, Journal of Mathematical Imaging and Vision, **57** (2017), 1–25.
[4] Juan C. De los Reyes & David Villacís, *Bilevel Optimization Methods in Imaging*, Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging, (to appear).
[5] Boris S. Mordukhovich, *Variational Analysis and Applications*, Springer Verlag (2018).
[6] Jirí V. Outrata, *A generalized mathematical program with equilibrium constraints*, SIAM Journal on Control and Optimization, **38(5)** (2000), 1623–1638.

Generalised perceptron learning

Martin Benning
(joint work with Xiaoyu Wang)

We have demonstrated that Rosenblatt's perceptron learning algorithm [1] can be cast as an energy minimisation problem. A (generalised) perceptron can be considered as an artificial neuron or one-layer feed forward neural network of the form

$$y = \sigma\left(Wx + b\right) .$$

Here σ denotes the (point-wise) activation function, $W \in \mathbb{R}^{n\times m}$ is the weight-matrix and $b \in \mathbb{R}^n$ is the bias-vector. The vector $x \in \mathbb{R}^m$ and the vector $y \in \mathbb{R}^n$ denote the input, respectively the output, of the perceptron. If we consider activation functions σ that are proximal maps, i.e.

$$\sigma(z) := \operatorname*{arg\,min}_{u\in\mathbb{R}^n} \left\{ \frac{1}{2}\|u - z\|^2 + \Psi(u) \right\} ,$$

for a proper, lower semi-continuous and convex function $\Psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, then we can define a loss function for which the gradient does not require the differentiation of the activation function. This loss function is defined as

$$L(y, \sigma(z)) := \frac{1}{2}\|y - \sigma(z)\|^2 + D_\Psi^{z-\sigma(z)}(y, \sigma(z)) , \tag{1}$$

for the (valid) subgradient $z - \sigma(z) \in \partial\Psi(\sigma(z))$. Here, the generalised Bregman distance with respect to the function Ψ for the subgradient $z - \sigma(z) \in \partial\Psi(\sigma(z))$ is defined as

$$D_\Psi^{z-\sigma(z)}(y, \sigma(z)) := \Psi(y) - \Psi(\sigma(z)) - \langle z - \sigma(z), y - \sigma(z)\rangle .$$

In [2] it has been shown that the gradient of the loss function L as defined in (1) with respect to the argument z is simply

$$\nabla_z L(y, \sigma(z)) = \sigma(z) - y .$$

Hence, for proximal activation functions, the perceptron learning algorithm can be interpreted as a stochastic or incremental gradient descent method applied to the energy $L(y, \sigma(Wx + b))$ for L as defined in (1). A nice consequence of this interpretation is that generalisations of Rosenblatt's algorithm can easily be

constructed. One example discussed in [2] is the Rosenblatt ISTA algorithm, which is the Iterative Soft-Thresholding Algorithm (ISTA) [3] applied to the energy

$$L(y, \sigma(Wx + b)) + \alpha \|W\|_1 .$$

Here $\|\cdot\|_1$ denotes the matrix one-norm, and $\alpha > 0$ is a regularisation parameter. Another example discussed in the talk is a Bregman Alternating Direction Method of Multipliers (BADMM) [4] for the minimisation of $L(y, \sigma(Wx + b))$. Based on the augmented Lagrange function

$$\mathcal{L}_\delta(W, b, \{z_i\}_{i=1}^s; \{\mu_i\}_{i=1}^s) = \frac{1}{s} \sum_{i=1}^{s} \Bigg[L(y_i, \sigma(z_i)) + \langle \mu_i, z_i - Wx_i - b \rangle + \frac{\delta}{2} \|z_i - Wx_i - b\|^2 \Bigg] ,$$

and the BADMM variant

$$\begin{aligned}
W^{k+1} &= \arg\min_W \mathcal{L}_\delta(W, b^k, \{z_i^k\}_{i=1}^s; \{\mu_i^k\}_{i=1}^s) + \frac{\gamma}{2} \|W - W^k\|^2 , \\
b^{k+1} &= \arg\min_b \mathcal{L}_\delta(W^{k+1}, b, \{z_i^k\}_{i=1}^s; \{\mu_i^k\}_{i=1}^s) , \\
z_i^{k+1} &= \arg\min_z \mathcal{L}_\delta(W^{k+1}, b^{k+1}, \{z_i\}_{i=1}^s; \{\mu_i^k\}_{i=1}^s)) + D_{\frac{1}{2\tau}\|\cdot\|^2 - L(y_i, \sigma(\cdot))}(z, z_i^k) , \\
\mu_i^{k+1} &= \mu_i^k + \delta \nabla_{\mu_i} \mathcal{L}_\delta(W^{k+1}, b^{k+1}, \{z_i^{k+1}\}_{i=1}^s; \{\mu_i^k\}_{i=1}^s)) ,
\end{aligned}$$

for all $i \in \{1, \ldots, s\}$, we can derive the following iterative procedure to minimise $L(y, \sigma(Wx + b))$: for suitable initial values, we compute

$$\begin{aligned}
W^{k+1} &= \left(\gamma W^k + \delta \sum_{i=1}^{s} \left(z_i^k - b + \frac{1}{\delta} \mu_i^k \right) x_i^\top \right) \left(\gamma I + \delta \sum_{i=1}^{s} x_i x_i^\top \right)^{-1} , \\
b^{k+1} &= \frac{1}{s} \sum_{i=1}^{s} \left[z_i^k + \frac{1}{\delta} \mu_i^k - W^{k+1} x_i \right] \\
z_i^{k+1} &= \frac{z_i^k - \tau \left(\sigma(z_i^k) - y_i + \delta \left(W^{k+1} x_i + b^{k+1} \right) \right)}{1 + \tau\delta} , \\
\mu_i^{k+1} &= \mu_i^k + \delta \left(z_i^{k+1} - W^{k+1} x_i - b^{k+1} \right) ,
\end{aligned}$$

for $k = 1, 2, \ldots$ and $i \in \{1, \ldots, s\}$. The interpretation of Rosenblatt's perceptron algorithm as an energy minimisation algorithm naturally paves the way for new algorithms. The BADMM framework furthermore has the potential to be used for the training of multi-layer perceptrons, respectively artificial neural networks, without requiring a differential of the activation functions. Another advantage is that BADMM can easily incorporate convex, non-smooth regularisations that act on hidden layers of a network, which can be useful for the regularised training of autoencoders.

References

[1] Frank Rosenblatt, *The perceptron, a perceiving and recognizing automaton*, Project Para. Cornell Aeronautical Laboratory (1957).

[2] Xiaoyu Wang, Martin Benning *Generalised perceptron learning*, 12th Annual Workshop on Optimization for Machine Learning (2020).

[3] Ingrid Daubechies, Michel Defrise, Christine De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Commun. Pure App. Math. 57(11), 1413?1457 (2004)

[4] Huahua Wang, Arindam Banerjee, *Bregman alternating direction method of multipliers*, In Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2, pp. 2816-2824 (2014).

Structure Preserving Deep Learning

Christian Etmann

(joint work with Elena Celledoni, Matthias J Ehrhardt, Robert I McLachlan, Brynjulf Owren, Carola-Bibiane Schönlieb, Ferdia Sherry)

Deep learning with mathematical guarantees has been a research focus of the last few years. Here, we want to look at a subset of mathematical guarantees, which in particular arise from the view of neural networks as discretised continuous systems. A more extensive discussion of this topic is given in [1], where additional guarantees (e.g. in the form of equivariant and invertible networks) are explored.

1. Introduction

At the heart of our analysis lies the realisation, that the widely-used *residual networks* (ResNets) [2] can be viewed as such a discretisation of an ODE. ResNets are given via the iteration

$$\begin{aligned} z^0 &= x \\ z^{k+1} &= z^k + hf(z^k, \theta^k), \quad k = 0, \dots, K-1, , \end{aligned} \tag{1}$$

which defines a mapping $x \mapsto z^K$. Here, z^k is from some (common) feature space $\mathcal{X}$ and θ^k is from a (common) parameter space Θ for all k. Typically, $f : \mathcal{X} \times \Theta \to \mathcal{X}$ is some combination of classical neural network layers (such as convolutional or fully-connected layers with nonlinearities).

As observed in [3], in this case, one can view the ResNet iteration (1) as an Euler discretisation of the initial value problem

$$\dot{z}(t) = f(z(t), \theta(t)), \quad t \in [0, T], \quad z(0) = x, \tag{2}$$

with $z^k = z(kh)$ and $\theta^k = \theta(kh)$, if we consider the parameters respectively activations as functions $\theta : [0, T] \to \Theta$ and $z : [0, T] \to \mathcal{X}$. With this, training a neural network can be phrased as an optimal control problem [5]. Different discretisation schemes then lead to different properties of the assumed underlying ODE being preserved.

2. A discussion on stability

A central consideration in both ODEs and neural networks is that of *stability*, so the well-developed stability theory of ODEs and their discretisation offers valuable insights into the stability of neural networks. However, it is important to qualify in which sense one is talking about stability, and how the different notions of stability of ODEs carry over to neural networks. A classical notion of stability of ODEs considers the stability of equilibrium points, i.e., points $z^0 = z(0)$ with $f(t, z^0) = 0$ for $t \in [0, \infty)$. This type of stability can be studied in terms of Lyapunov functions, that is, functions $V(z)$ that are non-increasing along solution trajectories. Functions that are constant along solutions are called first integrals, and a particular instance is the energy function of autonomous Hamiltonian systems.

In this view, trajectories for perturbations of the initial value are studied for all $t \in [0, \infty)$. However, this notion of stability has limited applicability to neural networks, as one typically only considers finite time horizons $t \in [0, T]$ (as in the previous section). Furthermore, equilibria hold no particular significance in neural networks.

Therefore, a more directly applicable view of stability is whether $z(T)$ does not change 'too much' when the initial value $x = z(0)$ is perturbed. That is, small perturbations in the data should not lead to large deviations in the end result. Such considerations lead to asking whether $z(T)$ is, e.g., Lipschitz continuous (or more generally, uniformly continuous) in $x = z(0)$. This means that for any two solutions $z_1(t)$ and $z_2(t)$ to (2), we have

$$\|z_2(T) - z_1(T)\| \leq C\|z_2(0) - z_1(0)\| \tag{3}$$

for some $C \geq 0$. It is well-known that this type of estimate can be obtained in several different ways, depending on the properties of the underlying vector field in (2). In particular, the Lipschitz constant C now depends on f as well as θ. Thus, in order to guarantee that C is not too large, one may want to impose certain restrictions on f or θ.

3. Structure preserving discretisations

Given certain properties of the ODE, such as (3), a neural network that is created as a discretisation of this ODE does not necessarily have the same property. In the following, we will provide an example, where *non-expansiveness* (meaning (3) with $C = 1$) of the ODE is preserved even after discretisation. Considerations like this lead to a similar analysis as in [4].

If the ODE

$$\dot{z}(t) = f(z(t), \theta(t))$$

has the property that

$$\begin{aligned} &\langle f(z_2(t), \theta(t)) - f(z_1(t), \theta(t)), z_2(t) - z_1(t)\rangle \\ \leq &\nu \|f(z_2(t), \theta(t)) - f(z_1(t), \theta(t))\|^2 \end{aligned} \tag{4}$$

for $t \in [0, T]$ for any two different solutions z_1 and z_2 if $\nu < 0$, then *any* Runge-Kutta scheme with strictly positive weights preserves the non-expansiveness, given a certain additional restriction to the step size. This can be viewed as a specific Lipschitz guarantee, as desired in the previous section.

References

[1] Celledoni, E., Ehrhardt, M. J., Etmann, C., McLachlan, R. I., Owren, B., Schönlieb, C. B., & Sherry, F. (2020). Structure preserving deep learning. arXiv preprint arXiv:2006.03364.

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[3] Haber, E., & Ruthotto, L. (2017). Stable architectures for deep neural networks. Inverse Problems, 34(1), 014004.

[4] Zhang, L., & Schaeffer, H. (2020). Forward stability of resnet and its variants. Journal of Mathematical Imaging and Vision, 62(3), 328-351.

[5] Martin Benning, Elena Celledoni, Matthias J. Ehrhardt, Brynjulf Owren, and Carola- Bibiane Schönlieb. Deep learning as optimal control problems: models and numerical methods. Journal of Computational Dynamics, 6(2):171?198, 2019

Learning a denoiser without ground truth

Sören Dittmer

(joint work with Carola-Bibiane Schönlieb, Peter Maass)

Learned solvers of inverse problems usually require some form of ground truth data. However, while high-dose computed tomography (CT) provides good ground truth approximations for low-dose CT, ground truth samples for other problems are often hard to come by.

We present a reconstruction method that does not require any ground truth data. More specifically, we train a denoising method flexible enough to be applied to noisy measurements. We can therefore use the denoiser as a preprocessing step for arbitrary reconstruction methods.

We denote the probability density function (pdf) of hypothetical clean measurements as p_y, the pdf of noise as p_η and assume the noise is additive and independent. This yields a distribution of noisy measurements

$$p_{y^\delta} = p_y * p_\eta. \tag{1}$$

In this setting an assumed ideal denoiser

$$G^* : Y :\ni y^\delta \mapsto y \in Y \tag{2}$$

would fulfill the following three push forward equations. First, we would have

$$G^*_\# p_{y^\delta} = p_y, \tag{3}$$

i.e., the samples generated by the denoiser G^*, given noisy samples from p_{y^δ}, should follow the law given by p_y. In English, denoised measurements should look like clean measurements.

The second forward equation that G^* is supposed to fulfil is

$$\left(G^*_{\#} p_{y^\delta}\right) * p_\eta = p_{y^\delta}. \tag{4}$$

This follows directly from (1) and (4) and formalizes that if we "renoise" denoised measurements, they should look like noisy measurements again.

The third and last push forward equation is

$$\left(\mathrm{id} - G^*\right)_{\#} p_{y^\delta} = p_\eta. \tag{5}$$

This can be interpreted as the requirement that the part of the noisy measurement that the denoiser removes should look like noise.

We approximate such a denoiser via the optimization problem

$$\arg\min_G \; D\left(p_{y^\delta} \middle\| \left(G_{\#} p_{y^\delta}\right) * p_\eta\right) + D\left(p_\eta \middle\| \left(\mathrm{id} - G\right)_{\#} p_{y^\delta}\right). \tag{6}$$

Here D is a distance measure between probability distributions. One can implement this optimization problem via a Generative Adversarial Network (GAN) [1] setup with one generator and two critics/discriminators. One critic for each of the summands in (6). Depending on the type of GAN loss one chooses, one can realize different choices of D.

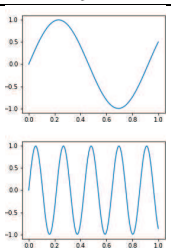
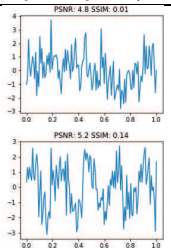
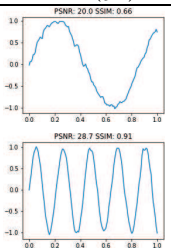
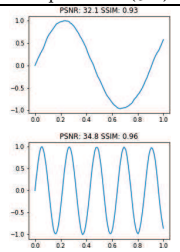
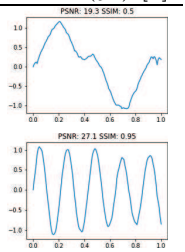
y	$y^\delta = y + \eta$	$G_{\text{ours}}(y^\delta)$	$G_{\text{supervised}}(y^\delta)$	$G_{\text{SURE}}(y^\delta)$ [2]
				
PSNR stats:	mean: 6.1 median: 6.0	mean: 25.3 median: 25.5	mean: 31.0 median: 30.3	mean: 23.4 median: 23.4
SSIM stats:	mean: 0.03 median: 0.03	mean: 0.79 median: 0.81	mean: 0.90 median: 0.93	mean: 0.78 median: 0.80

TABLE 1. Reconstructions of sines. Here, we compare – from left to right – the ground truth, the noisy measurement, our setup, a supervised setup, and a SURE setup. The stats are over 10,000 samples.

The training of such a GAN setup only requires samples of noisy measurements, i.e., samples from p_{y^δ} and samples of noise, i.e., samples from p_η. Critically, these samples do not need to come in pairs. We therefore avoid the need of any samples from p_y during the training.

The numerical results presented in Table 1 are based on a Wasserstein GAN [4] setup of the Gradient-Penalty type [3, 5]. The noise used is Gaussian and sine waves of varying frequencies give p_y.

References

[1] Goodfellow, Ian and Pouget-Abadie, Jean and Mirza, Mehdi and Xu, Bing and Warde-Farley, David and Ozair, Sherjil and Courville, Aaron and Bengio, Yoshua, *Advances in neural information processing systems*, Topology (2014), 2672–2680.

[2] Metzler, Christopher A and Mousavi, Ali and Heckel, Reinhard and Baraniuk, Richard G, *Unsupervised Learning with Stein's Unbiased Risk Estimator*, arXiv preprint arXiv:1805.10531 (2018)

[3] Gulrajani, Ishaan and Ahmed, Faruk and Arjovsky, Martin and Dumoulin, Vincent and Courville, Aaron C, *Improved training of wasserstein gans*, Advances in neural information processing systems (2017), 5767–5777.

[4] Arjovsky, Martin and Chintala, Soumith and Bottou, Léon, *Wasserstein gan*, arXiv preprint arXiv:1701.07875 (2017).

[5] Petzka, Henning and Fischer, Asja and Lukovnicov, Denis, *On the regularization of Wasserstein GANs*, arXiv preprint arXiv:1709.08894 (2017).

Learned Stochastic Primal-Dual Reconstruction

Marta M. Betcke

(joint work with Andreas Hauptmann, Won Tek Hong, Francesc Rul·lan)

1. From variational to learned reconstruction methods

Limited data problems are ubiquitous in real word applications. Be it the geometrical constraints on the scanner design such as one sided access only (limited angle/view problems) or the requirement to limit the dose delivered to the patient or data acquisition time (sparse sampling), such situations result in incomplete data which precludes good quality reconstruction using analytical inversion formulas derived for the complete data scenario.

A common way to approach such incomplete data problems is via variational formulation where the data-model fit is balanced with a prior knowledge about the image properties

$$\min_{u\in\mathbb{U}} \|P(u) - h\|_2^2 + \lambda R(u), \tag{1}$$

where $P : \mathbb{U} \to \mathbb{H}$ is possibly nonlinear forward operator and $R : \mathbb{U} \to \mathbb{R} \cup \{+\infty\}$ an appropriate regularisation functional, and $\lambda \in \mathbb{R}_+$ strikes the balance between the data fidelity and prior.

The variational formulation using non-strictly convex regularisation functionals enabled edge preserving reconstructions which revolutionised imaging applications. Many highly sophisticated regularisation functionals (convex or non-convex) especially tailored to imaging applications have been since developed including total variation and its higher order and directional variants, sparsity in a frame like Wavelets and their directional variants, composite regularisers via infimal convolution, joint regularisers e.g. total nuclear variation, to name a few. While these regularisers perform great in many applications, they have one common limitation: they are an analytical description of the image constructed with a particular property in mind e.g. total variation favours piecewise constant images, directional

Wavelet sparsity favours images with discontinuities along smooth curves. Real application images not only contain a mixture of such traits, which can be to certain extent addressed via composite regularisers, frequently their inherent qualities are hard to encode via analytical models.

The deep learning revolution of the last decade has seen a surge of interest in machine learning methods, in particular deep neural networks, in many fields including inverse problems and image reconstruction. A consensus quickly emerged that hybrid model-data based approaches hold most promise in image reconstruction applications, where generally the non-trivial forward operator admits analytical description with well understood properties and high dimensionally of the images implies that training data is not abundant. The *unrolling* proposed in [1] has been applied to a number of image reconstruction applications across modalities. We particularly mention the learned primal dual (LPD) method [2] which applies the unrolling concept to the popular primal dual hybrid gradient (PDHG) method [3].

2. Learned satochastic primal-dual method (SLPD)

Our proposed stochastic learned primal dual (SLPD) method builds on LPD [2] and stochastic primal-dual hybrid gradient (SPDHG) variant of PDHG proposed in [4].

We consider a general setting

$$\min_{u\in\mathbb{U}} F(P(u)) + G(u),$$

where $P : \mathbb{U} \to \mathbb{H}$ is a bounded operator which admits split into partial operators $P_i : P_i(x) = h_i$ and $h_i, i \in 1, \ldots, n$ is the ith component of $h \in \mathbb{H}$. If P is linear, its adjoint can be written as $P^*h = \sum_{i=1}^n P_i^* h_i$. For nonlinear P we can use similar formulation on Fréchet derivative ∂P. The data fidelity $F : \mathbb{H} \to \mathbb{R} \cup \{+\infty\}$ is assumed separable i.e. $F(h) = \sum_i^n F_i(h_i)$, and the regularisation $G : \mathbb{U} \to \mathbb{R} \cup \{+\infty\}$ is convex.

For proper, convex, lower semicontinuous F the problem can be reformulated as a saddle point

$$\min_{u\in\mathbb{U}} \max_{h\in\mathbb{H}} \langle P(u), h\rangle - F^*(h) + G(u),$$

where F^* is the Fenchel conjugate of F, which under the split becomes

$$\min_{u\in\mathbb{U}} \max_{h\in\mathbb{H}} \sum_{i=1}^{n} \langle P_i(u), h_i\rangle - F_i^*(h_i) + G(u).$$

SPDHG [4] capitalises on the partial operator structure and updates in each iteration only a random subset S_k of dual variables which only requires application of the partial forward P_{S_k} and partial adjoint $P_{S_k}^*$ operators. This has an advantage whenever such partial operators can be computed at a fraction of cost of the full operator, which is frequently the case in tomography e.g. Radon/X-ray transform (projection angles), US tomography (sources), Photoacoustic tomography (when using ray based acoustic solvers [5, 7]).

We propose a following modifications to the SPDHG method (note that practical implementations will make further assumptions). We highlight with colour the influences from LPD and SPDHG and refer the reader to [9, 8] for details.

Learnt Stochastic Primal Dual (SLPD)

Initialisation:

$u^0 = 0 \in \mathbb{U}^{N^p}$, $h^0 = 0 \in \mathbb{H}^{N^d}$

for $k = 0, \dots, K$ **do**

Select $S_{k+1} \subset \{1, \dots, n\}$

$h^{k+1} = \Gamma_{\theta_k^d}(h^k, P_{S_{k+1}}(u^{k,(2)}), g_{S_{k+1}})$

$u^{k+1} = \Gamma_{\theta_k^p}(u^k, [\partial P_{S_{k+1}}(u^{k,(1)})]^*(h^{k+1,(1)}))$,

end for

return $u^{K,(1)}$

We would like to mention two important advantages of the proposed LSPD method: i) it can be potentially used as a one pass method which allows a reconstruction from an unrecorded stream of data (all 1 angle epoch reconstructions), ii) it can be trained using subsampling S_k not fixed but randomly drawn from a chosen distribution for each training pair. The resulting learned method is then independent of a particular subsampling draw (pattern), and only depends on the underlying subsampling distribution (**Fig 2.** random sampling example).

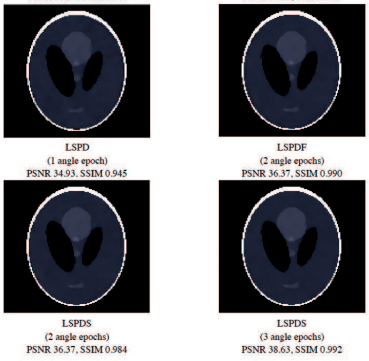


Fig 1. LSPD reconstructions from uniformly subsampled fixed 60 angles (over 180 degrees), grey level [0,1]

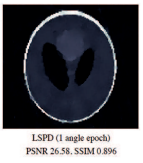


Fig 2. LSPD reconstruction from randomly drawn (for each training pair) 60 angles (over 180 degrees), grey level [0,1]

Method	Angle Epochs	PSNR (dB)	SSIM
Filtered Back-Projection	—	21.24	0.822
TV Regularisation[1]	—	28.06	0.929
Learned Primal	10	36.24	0.988
Learned Primal-Dual	10	39.06	0.989
Learned stochastic Primal-Dual flexible/static	1	34.93	0.945
Learned stochastic Primal-Dual flexible	2	36.37	0.990
Learned stochastic Primal-Dual static	2	37.53	0.984
Learned stochastic Primal-Dual static	3	38.63	0.992

TABLE 1: Comparison of the different reconstruction schemes for the simulated ellipse data in the UA setup

References

[1] K. Gregor, and Y. LeCun, *Learning fast approximations of sparse coding*. In ICML 2010 - Proceedings, 27th International Conference on Machine Learning, (2010) 399-406.

[2] J. Adler, and O. Oktem, *Learned Primal-Dual Reconstruction*, IEEE Trans Med Imaging **37(6)** (2018),1322-1332.

[3] A. Chambolle, and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vision, **40** (2011), 120–145.

[4] A. Chambolle, M.J. Ehrhardt, P. Richtárik, and C.-B. Schönlieb, *Stochastic Primal-Dual Hybrid Gradient Algorithm with Arbitrary Sampling and Imaging Applications*, SIAM Journal on Optimization, **28:4** (2018), 2783-2808
[5] F. Rullan, and M.M. Betcke, *Hamilton-Green solver for the forward and adjoint problems in photoacoustic tomography*, arXiv:1810.13196.
[6] F. Rullan, N. Huynh, B. Cox, P. Beard, and M.M. Betcke, *Stochastic reconstruction methods for PAT (SPD(HG)^2)*, in preparation.
[7] F. Rullan, *Photoacoustic tomography: flexible acoustic solvers based on geometrical optics*, PhD thesis, UCL, 2020.
[8] W.T. Hong, *Learned Stochastic Primal-Dual Reconstruction*, MSc thesis, UCL, 2018.
[9] M.M. Betcke, A. Hauptmann, W.T. Hong, and F. Rullan, *Learned Stochastic Primal-Dual Reconstruction with applications in X-ray and photo-acoustic tomography*, in preparation.

Solving Inverse Imaging Problems with Generative Machine Learning Models

Margaret Duff
(joint work with Neill D.F. Campbell, Matthias J. Ehrhardt)

Solving an inverse problem is the task of computing an unknown physical quantity from indirect measurements found via a forward model. Let $A : X \to Y$ be a forward process that takes an image $x \in X$ to data $y \in Y$. The inverse problem takes data, y, often corrupted by noise, and finds image, x, such that $y = A(x)$.

Generative models learn, from observations, approximations to high-dimensional data distributions. Consider some set of 'feasible' solutions to the inverse problem, for example in an MRI image reconstruction problem where the data is of a knee, example feasible images could be given by a dataset of other knee MRI images. Assume that due to the similarities between the images, such as repeating patterns, textures and backgrounds, that this set of feasible images lie on some lower dimensional manifold in the space, X. A trained generator $G : Z \to X$, takes values in a known lower dimensional latent space, Z, and outputs images on this manifold. If, in addition, a prior probability distribution, p_z, is given on the latent space, Z, we further ask that the generator maps high probability points in the latent space to high probability points in the image space. Mathematically we ask that the pushforward of the prior by the function, G, is close to the unknown distribution of feasible images.

We approach the inverse problem using variational regularisation of the form

$$x^* \in \arg\min_{x \in X} \|Ax - y\|_2^2 + \lambda R_G(x), \tag{1}$$

where $R_G : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ is a regularisation functional that penalises values of $x \in \mathcal{X}$ that are far from the range of the trained generator, G, and the constant λ is a regularisation parameter. The idea is that learned priors could provide more specific information than a hand-crafted regulariser and thus a better reconstruction. Any learning would also not require paired training data and would be done independently of the forward model. This makes the method very flexible in real-world scenarios where noise levels and forward model parameters may change.

We discuss three different choice of R_G.

- Taking $R_G(x) = \min_{z\in Z} \iota_{\{0\}}(G(z)-x) + \|z\|_2^2$, where $\iota_{\mathcal{C}}(t) = \begin{cases} 0 & t \in \mathcal{C} \\ \infty & t \notin \mathcal{C} \end{cases}$. This was introduced by [1] and restricts images x to be in the range of the generator. The additional regularisation on z is chosen to match the prior on the latent space, usually taken to be a standard normal distribution. This gives the solution $x^* = G(z^*)$ where

$$z^* \in \operatorname*{arg\,min}_{z\in Z} \|AG(z) - y\|_2^2 + \lambda\|z\|_2^2. \tag{2}$$

- Taking $R_G(x) = \min_{z\in Z} \|G(z) - x\|_2^2 + \mu\|z\|_2^2$ encourages x to lie close to the range of the generator, leading to the solution of the inverse problem

$$x^*, z^* \in \operatorname*{arg\,min}_{x\in X, z\in Z} \|Ax - y\|_2^2 + \lambda\|G(z) - x\|_2^2 + \lambda\mu\|z\|_2^2. \tag{3}$$

- Taking $R_G(x) = \min_{z\in Z, u\in X} \iota_{\{0\}}(G(z)+u-x) + \|u\|_1 + \mu\|z\|_2^2$ we restrict x to be within a sparse deviation from the range of the generator. This was originally introduced by [2]. The deviation, u, in the image space, is restricted to be sparse using the 1-norm. This gives the solution $x^* = G(z^*) + u^*$ where

$$z^*, u^* \in \operatorname*{arg\,min}_{z\in Z, u\in X} \|A(G(z) + u) - y\|_2^2 + \lambda\|u\|_1 + \lambda\mu\|z\|_2^2. \tag{4}$$

Including a generator in a variational framework has been considered by a variety of authors. From the initial applications in compressed sensing [1, 3] there have been applications in blind inverse problems [4], seismic imaging [5], inpainting [6] and photo up-sampling [7].

The approach raises a wide range of questions. Assuming that solutions of (1) can be found, some theoretical results exist. For example papers such as [8, 9], consider convergence of solutions of variational approaches with learned regularisers as the error in the data goes to zero. However, the addition of a generator in (1) makes the minimisation problem potentially non-linear and non-convex and any theoretical results on the optimisation of (1) very difficult. Careful initialisation of optimisation schemes for (1) is needed, for example with an approximate reconstruction.

The choice of training of the generator is also very important. We also require that the ground truth image lies in, or close to, the range of the generator. Common generative models include variational autoencoders (VAEs) [10] and generative adversarial networks (GANs) [11]. Although both have provided good image generation results, VAEs often suffer from blurred images, generating images that may not be feasible. GANs can also suffer from mode collapse, where the generator just outputs a subset of the feasible images. This subset could just be a few images or large enough that the mode collapse is difficult to detect [12].

Evaluating a generative model's ability to reconstruct an entire distribution, in the context of inverse problems, is an area of future work. In addition, it may be important to have a generator that is smooth with respect to the latent space, Z,

in order to use gradient descent based methods to minimise (1). We could also ask that the generator is somewhat stable with respect to the latent space, so that points close together in the latent space map to similar points in the image space.

References

[1] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," *ICML 2017*, vol. 2, pp. 822–841, 2017.

[2] M. Dhar, A. Grover, and S. Ermon, "Modelling Sparse Deviations for Compressed Sensing using Generative Models," in *ICML 2018*, vol. 3, pp. 1990–2005, 2018.

[3] S. Tripathi, Z. C. Lipton, and T. Q. Nguyen, "Correction by Projection: Denoising Images with Generative Adversarial Networks," *ArXiv Preprint*, 2018.

[4] M. Asim, F. Shamshad, and A. Ahmed, "Blind image deconvolution using pretrained generative priors," *30th British Machine Vision Conference 2019, BMVC 2019*, 2020.

[5] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *CVPR*, pp. 6882–6890, IEEE, 2017.

[6] A. Lahiri, A. K. Jain, D. Nadendla, and P. K. Biswas, "Faster Unsupervised Semantic Inpainting: A GAN Based Approach," *Proceedings - International Conference on Image Processing, ICIP*, vol. 2019-Septe, pp. 2706–2710, 2019.

[7] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, pp. 2437–2445, 2020.

[8] D. Obmann, J. Schwab, and M. Haltmeier, "Deep synthesis regularization of inverse problems," *ArXiv Preprint*, Feb 2020.

[9] H. Li, J. Schwab, S. Antholzer, and M. Haltmeier, "NETT: Solving inverse problems with deep neural networks," *Inverse Problems*, vol. 36, no. 6, 2020.

[10] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *ICLR 2014* , 2014.

[11] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *NeurIPS*, pp. 2672–2680, 2014.

[12] S. Arora, A. Risteski, and Y. Zhang, "Do GANs Learn the Distribution? Some Theory and Empirics," *ICLR 2018*, pp. 1–16, 2018.

Interpreting U-Nets via Task-Driven Multiscale Dictionary Learning

Tianlin Liu

(joint work with Anadi Chaman, David Belius, Ivan Dokmanić)

1. Introduction

U-Nets [1] have been tremendously successful in many imaging inverse problems. In an effort to understand the source of this success, we show that one can reduce a U-Net to a tractable, well-understood sparsity-driven dictionary model while retaining its strong empirical performance. We achieve this by extracting a certain multiscale convolutional dictionary from the standard U-Net. This dictionary imitates the structure of the U-Net in its convolution, scale-separation, and skip connection aspects, while doing away with the nonlinear parts. We show that this model can be trained in a task-driven dictionary learning framework and yield comparable results to standard U-Nets on a number of relevant tasks, including

CT and MRI reconstruction. These results suggest that the success of the U-Net may be explained mainly by its multiscale architecture and the induced sparse representation.

2. Multiscale convolutional dictionaries

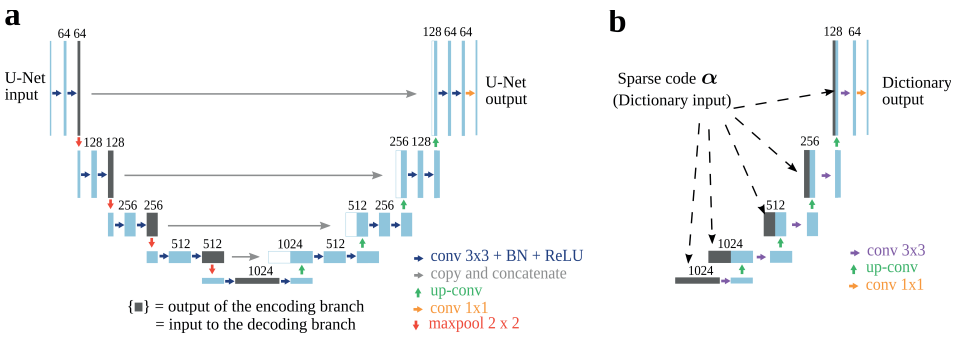


Figure 1. **The U-Net and the multiscale convolutional dictionary considered in this work.** (**a**): The U-Net processes input images using convolution, scale-separation, and skip connections operations in conjunction with ReLU non-linearlities and batch-normalization modules indicated by colored arrows. (**b**): The dictionary considered in this work constitutes the main ingredients of U-Net, but with removed non-linearlities, batch-normalization, and additive biases.

Based on the U-Net architecture (Figure 1**a**), we construct a simple linear model by keeping the U-Net's essential ingredients – convolution and scale separation – but removing non-linearities, batch normalization, and additive biases (Figure 1**b**). The resulting model is thus an overcomplete dictionary written as $\boldsymbol{D}_{\boldsymbol{\gamma}} \in \mathbb{R}^{d\times N}$ with atoms $\boldsymbol{\gamma}$; we refer it to as the synthesis dictionary.

With a given synthesis dictionary $\boldsymbol{D}_{\boldsymbol{\gamma}}$ that describes the image generation process, we next consider how to infer the sparse code $\boldsymbol{\alpha}$, so that the linear transformation $\boldsymbol{D}_{\boldsymbol{\gamma}}\boldsymbol{\alpha}$ well approximates the image $\boldsymbol{y}$ we wish to model, that is, $\boldsymbol{D}_{\boldsymbol{\gamma}}\boldsymbol{\alpha} \approx \boldsymbol{y}$. To that end, we make use of paired input-target data of the form $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^d \times \mathbb{R}^d$ as in the context of supervised learning. The input $\boldsymbol{x}$ here could be a noisy image associated to the clean image $\boldsymbol{y}$. Since $\boldsymbol{x}$ and $\boldsymbol{y}$ are associated, we interpret $\boldsymbol{\alpha}$ as a latent representation of both $\boldsymbol{x}$ and $\boldsymbol{y}$. Specifically, we posit that each input-target data pair $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^d \times \mathbb{R}^d$ admits a shared underlying sparse code $\boldsymbol{\alpha}$ with respect to two dictionaries $\boldsymbol{D}_{\boldsymbol{\theta}}$ and $\boldsymbol{D}_{\boldsymbol{\gamma}}$:

$$\boldsymbol{D}_{\boldsymbol{\theta}}\boldsymbol{\alpha} \approx \boldsymbol{x} \quad \text{and} \quad \boldsymbol{D}_{\boldsymbol{\gamma}}\boldsymbol{\alpha} \approx \boldsymbol{y} \quad \text{for } \boldsymbol{\alpha} \text{ sparse.} \tag{1}$$

Here, the analysis dictionary $\boldsymbol{D}_{\boldsymbol{\theta}} \in \mathbb{R}^{d\times N}$ has the same structure of the synthesis dictionary $\boldsymbol{D}_{\boldsymbol{\gamma}}$, albeit with a different set of atoms $\boldsymbol{\theta}$.

3. Bi-level optimization

In light of the formulation of (1), to turn an input $\boldsymbol{x}$ to approximate the target $\boldsymbol{y}$, we consider the following bi-level minimization problem over a training dataset of M input-target pairs $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^M$:

$$\begin{aligned} &\underset{\{\boldsymbol{\theta},\boldsymbol{\gamma}\},\boldsymbol{\lambda}>0}{\text{minimize}} \quad \frac{1}{2M}\sum_{i=1}^{M} \|\boldsymbol{D}_{\boldsymbol{\gamma}}\boldsymbol{\alpha}_{\boldsymbol{x}_i,\boldsymbol{\theta}} - \boldsymbol{y}_i\|_2^2 \\ &\qquad\text{where} \quad \boldsymbol{\alpha}_{\boldsymbol{x}_i,\boldsymbol{\theta}} = \underset{\boldsymbol{\alpha}\geq 0}{\arg\min}\, \frac{1}{2}\|\boldsymbol{D}_{\boldsymbol{\theta}}\boldsymbol{\alpha} - \boldsymbol{x}_i\|_2^2 + \|\boldsymbol{\lambda}\odot\boldsymbol{\alpha}\|_1, \end{aligned} \tag{2}$$

for $\odot$ being the Hadamard product. In plain words, the objective of Equation (2) aims to minimize the discrepancy between the ground-truth signal $\boldsymbol{y}$ and the model prediction $\boldsymbol{D}_{\boldsymbol{\gamma}}\boldsymbol{\alpha}_{\boldsymbol{x},\boldsymbol{\theta}}$, where the latter is a signal synthesized from a sparse code $\boldsymbol{\alpha}_{\boldsymbol{x},\boldsymbol{\theta}}$ via the synthesis dictionary $\boldsymbol{D}_{\boldsymbol{\gamma}}$; the code $\boldsymbol{\alpha}_{\boldsymbol{x},\boldsymbol{\theta}}$, defined through the argmin function in Equation (2) is a sparse representation of the input image $\boldsymbol{x}$ with respect to the analysis dictionary $\boldsymbol{D}_{\boldsymbol{\theta}}$ under a Lasso-like objective. Here, the sparsity-controlling parameter $\boldsymbol{\lambda}$ is a vector, weighting codes component-wise. We have required the code $\boldsymbol{\alpha}$ to be non-negative to enhance interpretability.

4. Approximately solve bi-level optimization through unrolling

To solve the bi-level optimization task (2), we use a two-step procedure: (i) infer the underlying sparse code $\boldsymbol{\alpha}$ using an unrolled sparse coding algorithm $\mathcal{S}$ respect to the analysis dictionary $\boldsymbol{D}_{\boldsymbol{\theta}}$; and (ii) linearly transform the inferred sparse code using the synthesis dictionary $\boldsymbol{D}_{\boldsymbol{\gamma}}$:

$$\boldsymbol{x} \xmapsto{\text{Sparse coding}} \boldsymbol{\alpha} \coloneqq \mathcal{S}(\boldsymbol{x}, \boldsymbol{D}_{\boldsymbol{\theta}}) \xmapsto{\text{Linear synthesis}} \hat{\boldsymbol{y}} \coloneqq \boldsymbol{D}_{\boldsymbol{\gamma}}\boldsymbol{\alpha}. \tag{3}$$

Here, we choose $\mathcal{S}$ to be a learned variant of the ISTA sparse coding algorithm. The computational graph (3), dubbed ISTA U-Net, can be trained under the standard supervised learning paradigm.

5. Numerical experiments

We use the LoDoPaB-CT dataset [2], which contains more than 40000 pairs of human chest CT images and their simulated low photon count measurements. To train U-Nets and ISTA U-Nets, we first transformed the sinogram measurements to the image space using filtered back projection (FBP), which was the only pre-processing we performed. Table 2 and Figure 2 show the reconstruction results on the test set. The U-Net variants achieved superior results. Importantly, ISTA U-Net replicated U-Net's strong performance in this task. Additional numerical results can be seen in our full paper [3].

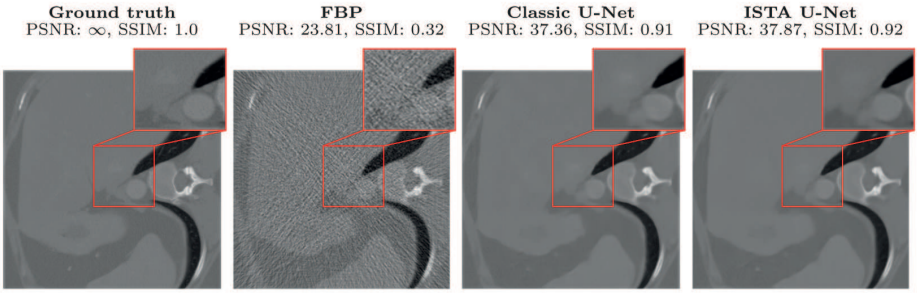


FIGURE 2. **Reconstructions of a test sample from the LoDoPaB-CT dataset.**

	PSNR (dB)	SSIM
FBP	30.37	0.74
Small U-Net	35.48	0.84
Classic U-Net	35.71	0.84
ISTA U-Net	35.83	0.84

TABLE 2. **Performance on the LoDoPaB-CT test data.**

REFERENCES

[1] Ronneberger, O., Fischer, P., and Brox, T. *U-net: Convolutional networks for biomedical image segmentation.* In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F. (eds.), Medical Image Computing and Computer-Assisted Intervention, pp. 234?241. Springer International Publishing, 2015.

[2] Leuschner, J., Schmidt, M., Baguer, D. O., and Maaß, P. *The LoDoPaB-CT dataset: A benchmark dataset for low-dose CT reconstruction methods.* CoRR abs/1910.01113 (2019).

[3] Liu, T., Chaman, A., Belius, D., and Dokmanić, I. *Interpreting U-Nets via Task-Driven Multiscale Dictionary Learning.* CoRR abs/2011.12815 (2020).

Inverse Models for Particle Accelerators

ROMANA BOIGER

(joint work with Adelmann Andreas, Bellotti Renato)

Forward and especially inverse modeling of particle accelerators is of great interest, not only for the initial machine design but also during operation. Particularly for the latter, inverse models could potentially be used to help optimizing the requested beam conditions for accelerators with frequent re-tuning of settings or even offer the means for online beam-optimization. With OPAL, Adelmann et al. [1] provide a parallel open source tool for forward simulating particle accelerators, based on high fidelity nonlinear physical models. OPAL predicts beam parameters for given accelerator settings with generally high accuracy and reliability. This

simulation is computationally very expensive, due to the high model complexity. Hence, determining particle accelerator settings from given beam parameters, i.e. solving the inverse problem, is not yet efficiently possible.

The main focus of this work is the direct solution of the inverse problem, using deep learning algorithms. Therewith the use of the forward model and costly optimization methods or the hand tuning of the accelerator settings could be avoided. In references [2, 3] different deep learning algorithms for solving inverse problems are suggested, such as invertible neural networks, autoencoders, invertible residual networks, or autoregressive flows. Especially the invertible architectures have the advantage, that not only the inverse problem, but also the forward problem can be solved, without any additional effort. For the forward simulation and further optimization it was shown e.g. in [4] that surrogate models based on neural networks can speed up calculation by orders of magnitude. Hence it is worth taking a look at the solution of the inverse problem using neural networks as well. In a first step we use invertible neural networks as suggested by [2]. They have the following specific structure: the main building blocks are affince coupling blocks, followed by permutation layers. The affine coupling blocks, are designed such that the input u is split into two parts u_1 and u_2, and also the output consists of two parts v_1 and v_2. Input and output are connected via functions s and t that are neural networks themselves. The forward pass is given by:

$$v_1 = u_1 \odot \exp\left(s_2(u_2)\right) + t_2(u_2)$$
$$v_2 = u_2 \odot \exp\left(s_1(v_1)\right) + t_1(v_1)$$

Hence the inverse pass is:

$$u_2 = (v_2 - t_1(v_1)) \odot \exp\left(-s_1(v_1)\right)$$
$$u_1 = (v_1 - t_2(u_2)) \odot \exp\left(-s_2(u_2)\right)$$

Additionally, a latent variable following a predefined distribution is added to the output that accounts for the information loss from input to output. The above structure can only be used, if input and output have the same dimension. In order to fulfill this condition on both, the input and output layer, zero or low noise padding can be used. With that, applying the forward after the inverse prediction and vice-versa should give the identity, which holds up to some numerical errors. Due to the invertible structure of the affine coupling blocks and together with a specific loss function, mentioned in [2] it is possible to train forward and inverse prediction simultaneously.

We used this method for a first proof of concept to solve the forward and inverse problem for specific accelerators, modelled with OPAL to get a reliable dataset. First numerical results were quite promising and showed the general applicability of surrogate models to solve the inverse problem. Furthermore, a significant speed-up in the forward simulation was obtained. The usage of the inverse surrogate model for optimization purposes can, depending on the problem of course, speed-up the time-to-solution and reduction of computational cost more than 98 %, compared to using exclusively OPAL for the same task.

In summary, deep learning methods and especially invertible neural networks can solve the forward and inverse problem for particle accelerators in principle and can reduce both prediction time and computational cost tremendously.

References

[1] A. Adelmann, P. Calvo, M. Frey, A. Gsell, U. Locans, C. Metzger-Kraus, N. Neveu, C. Rogers, S. Russell, S. Sheehy, J. Snuverink, and D. Winklehner. *OPAL a versatile tool for charged particle accelerator simulations*. arXiv:1905.06654, 2019.

[2] L. Ardizzone, J. Kruse , C. Rother , and U. Koethe. *Analyzing inverse problems with invertible neural networks*. In International Conference on Learning Representations, 2019.

[3] J. Kruse, L. Ardizzone, C. Rother, and U. Köthe. *Benchmarking Invertible Architectures on Inverse Problems*. First Workshop on Invertible Neural Networks and Normalizing Flows (ICML 2019), Long Beach, USA, June 2019.

[4] A. Edelen, N. Neveu, M.Frey, Y.Huber, C.Mayes, and A.Adelmann. *Machine learning for orders of magnitude speedup in multiobjective optimization of particle accelerator systems*. In Physical Review Accelerators and Beams., 23(4),2020.

Learning from electric X-ray images: the new EIT

Samuli Siltanen

(joint work with Juan Pablo Agnelli, Aynur Cöl, Matti Lassas, Rashmi Murthy, Matteo Santacesaria)

A fundamental connection between Electrical Impedance Tomography (EIT) and classical X-ray tomography was found in [Greenleaf et al 2018]. There it was shown that a one-dimensional Fourier transform applied to the spectral parameter of Complex Geometric Optics (CGO) solutions produces generalised projections, enabling a novel filtered back-projection type nonlinear reconstruction algorithm for EIT. This approach is called Virtual Hybrid Edge Detection (VHED).

One of the medically most promising applications of EIT is stroke imaging. There are two main types of stroke: (1) brain haemorrhage and (2) ischemic stroke caused by a blood clot. The symptoms for those two conditions are the same, but the treatments are completely the opposite. There are two main uses for EIT here: (a) classifying the type of stroke already in the ambulance with a cost-effective portable device, and (b) monitoring the state of recovering stroke patients in the intensive care unit.

The main difficulty in using EIT for head imaging is the resistive skull. Because of that, the relevant signal from the brain is weak and almost buried in noise. Given the extreme ill-posedness of the inverse conductivity problem, it is quite a challenge to design a robust EIT algorithm for either (a) or (b).

VHED offers a way to divide the information in EIT measurements into geometrically understood pieces. One could wish that those pieces are less sensitive to noise than a full reconstructed image of the conductivity. This presentation shows how machine learning can be used for classifying stroke (problem (a)) above based on VHED profiles. Examined are fully connected neural networks (FCNN), convolutional neural networks (CNN) and recurrent neural networks (RNN). Perhaps

surprisingly, CNNs offer the worst performance, while RNNs are slightly better than FCNNs.

Data-Driven Regularization for Inverse Problems

Subhadip Mukherjee and Carola-Bibiane Schönlieb
(joint work with Sebastian Lunz, Sören Dittmer, Zakhar Shumaylov, Ozan Öktem)

1. Introduction

Inverse problems arise in virtually every modern medical imaging modality such as computed tomography (CT), magnetic resonance imaging (MRI), etc., wherein the key objective is to reconstruct some parameter of interest $\boldsymbol{x}^* \in \mathbb{X}$ based on an indirect and possibly noisy measurement (data):

$$\boldsymbol{y}^\delta = \mathcal{A}(\boldsymbol{x}^*) + \boldsymbol{e} \in \mathbb{Y}. \tag{1}$$

Here, $\mathbb{X}$ and $\mathbb{Y}$ are appropriately defined Hilbert spaces, and the measurement noise $\boldsymbol{e}$ satisfies $\|\boldsymbol{e}\|_2 \leq \delta$. An inverse problem is said to be ill-posed if it has no or multiple solutions, or if its solution does not vary continuously in the data. Classical variational approaches attempt to alleviate the issue of ill-posedness by involving hand-crafted prior information on possible reconstructions:

$$\hat{\boldsymbol{x}}_\lambda(\boldsymbol{y}^\delta) \in \underset{\boldsymbol{x}\in\mathbb{X}}{\arg\min} \|\boldsymbol{y}^\delta - \mathcal{A}(\boldsymbol{x})\|_{\mathbb{Y}}^2 + \lambda\mathcal{R}(\boldsymbol{x}). \tag{2}$$

Here, the regularization functional $\mathcal{R} : \mathbb{X} \to \mathbb{R}$ is chosen such that it penalizes undesirable solutions. The penalty parameter $\lambda > 0$ trades-off between data consistency and regularization and should be selected based on the noise-level δ. Several decades of research has gone into hand-crafting appropriate regularizers with provable properties [11], but they fall short in terms of data-adaptability.

With the recent surge of research in deep learning, attempts have been made to learn the regularizer in a data-driven manner. The line of research that this report focuses on directly uses a neural network to parametrize the regularization functional, allowing to reconstruct from the observed data by solving a variational optimization problem [7, 5, 12, 4]. The idea of using a trained neural network as a regularizer was considered in [5] (referred to as network Tikhonov (NETT)) and more recently in [4] (referred to as total deep variation (TDV)). The regularization by denoising (RED) approach [9, 10, 6] also belongs to this class, wherein one constructs an explicit regularizer from an image denoiser by penalizing the inner product of the image with its denoising residual. This report specifically considers the adversarial regularizer (AR) framework introduced in [7] and its convex variant (adversarial convex regularizer (ACR)) proposed in [8].

2. Adversarial regularization

The principle idea in AR is to replace a hand-crafted regularizer with a data-adaptive one, parametrized by a deep neural network. The parametric regularizer $\{\mathcal{R}_\theta\}_{\theta\in\Theta}$ is first trained to discern ground-truth images from images containing artifacts. One does not need paired images to approximate the training objective in AR, which makes the framework unsupervised in theory. Subsequently, the trained AR is deployed in a variational scheme for solving an ill-posed inverse problem.

Given a training dataset containing an ensemble of ground-truth images $\{\boldsymbol{x}_i\}_{i=1}^{N_1}$ and unregularized reconstructions $\{\mathcal{A}^\dagger\,\boldsymbol{y}_j\}_{j=1}^{N_2}$, sampled i.i.d. from the respective marginal distributions $\pi_{\boldsymbol{x}}$ and $\mathcal{A}^\dagger_{\#}\,\pi_{\boldsymbol{y}}$ (the push-forward of the data distribution π_Y by the pseudo-inverse of $\mathcal{A}$), respectively, the training objective reads

$$\min_{\theta\in\Theta}\left(\frac{1}{N_1}\sum_{i=1}^{N_1}\left[\mathcal{R}_\theta\left(\boldsymbol{x}_i\right)\right]-\frac{1}{N_2}\sum_{j=1}^{N_2}\left[\mathcal{R}_\theta(\mathcal{A}^\dagger\,\boldsymbol{y}_j)\right]\right)\ \text{s.t.}\ \mathcal{R}_\theta\in 1-\text{Lipschitz}. \tag{3}$$

The 1-Lipschitz condition in (3) encourages the output of $\mathcal{R}_\theta$ to transition smoothly with respect to the input, thus making the corresponding variational loss stable. The 1-Lipschitz constraint is enforced by adding a gradient-penalty term [3] to the training loss in (3). AR enjoys the following important theoretical properties:

- As a consequence of the Kantorovich-Rubinstein duality [2], a perfectly trained regularizer (which is a 1-Lipschitz functional and achieves the minima in (3), with parameter θ^*) approximates the Wasserstein distance: $\mathbb{W}(\pi_{\boldsymbol{x}},\mathcal{A}^\dagger_{\#}\,\pi_{\boldsymbol{y}}) = \mathbb{E}_{\mathcal{A}^\dagger_{\#}\,\pi_{\boldsymbol{y}}}\left[\mathcal{R}_{\theta^*}(\boldsymbol{x})\right]-\mathbb{E}_{\pi_{\boldsymbol{x}}}\left[\mathcal{R}_{\theta^*}(\boldsymbol{x})\right]$.
- Let $\boldsymbol{u}\sim\pi_{\text{noisy}}:=\mathcal{A}^\dagger_{\#}\,\pi_{\boldsymbol{y}}$, and $g_\eta(\boldsymbol{u})=\boldsymbol{u}-\eta\,\nabla_{\boldsymbol{u}}\mathcal{R}_{\theta^*}(\boldsymbol{u})$ be the gradient-descent update at $\boldsymbol{u}$ with step-size $\eta>0$ in the negative gradient direction of $\mathcal{R}_{\theta^*}$. Then, we have $\mathbb{W}(\pi_{\boldsymbol{x}},(g_\eta)_{\#}\,\pi_{\text{noisy}})\le\mathbb{W}(\pi_{\boldsymbol{x}},\pi_{\text{noisy}})$, where $(g_\eta)_{\#}\,\pi_{\text{noisy}}$ is the distribution of $g_\eta(\boldsymbol{u})$. That is, the regularizer seeks to align the distribution of the reconstruction to that of the ground-truth images.
- Assume that $\pi_{\boldsymbol{x}}$ is supported on the weakly compact set $\mathbb{M}$ such that $\mathbb{M}^{\text{c}}$ has zero measure and $(\mathcal{P}_{\mathbb{M}})_{\#}\,\pi_{\text{noisy}}=\pi_{\boldsymbol{x}}$, where $\mathcal{P}_{\mathbb{M}}(\boldsymbol{x}):=\underset{\boldsymbol{u}\in\mathbb{M}}{\arg\min}\,\|\boldsymbol{x}-\boldsymbol{u}\|$ is the projection of $\boldsymbol{x}$ on to the image manifold $\mathbb{M}$. Then, the regularizer approximates the distance from $\mathbb{M}$: $\mathcal{R}_{\theta^*}(\boldsymbol{x}):=\min_{\boldsymbol{u}\in\mathbb{M}}\|\boldsymbol{x}-\boldsymbol{u}\|$. Therefore, the corresponding variational problem finds a solution that is consistent with the data and is close to the true image manifold.
- Let $\lim_{n\to\infty}\left\|\boldsymbol{y}_n-\boldsymbol{y}^\delta\right\|_{\mathbb{Y}}=0$, $|\mathcal{R}_{\theta^*}(\boldsymbol{x})|\to\infty$ as $\|\boldsymbol{x}\|\to\infty$ (coercive), and

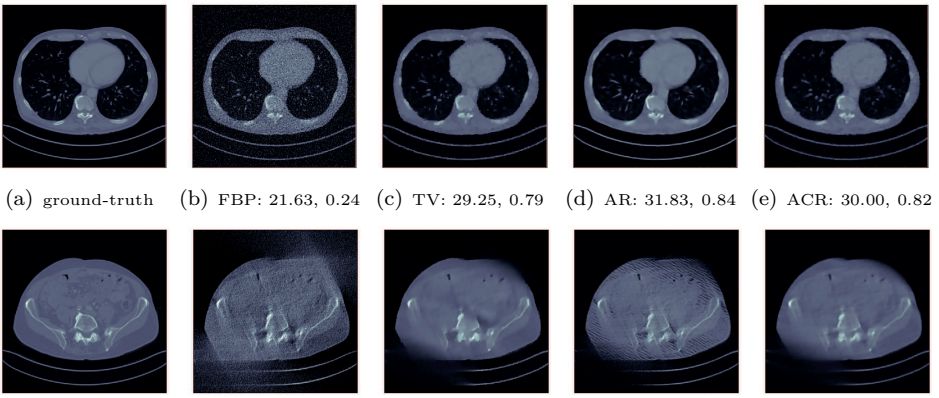
$$\boldsymbol{x}_n\in\underset{\boldsymbol{x}\in\mathbb{X}}{\arg\min}\,\|\boldsymbol{y}_n-\mathcal{A}(\boldsymbol{x})\|_2^2+\lambda\,\mathcal{R}_{\theta^*}(\boldsymbol{x}).$$

  Then, $\boldsymbol{x}_n$ has a weakly convergent sub-sequence and the limit is a minimizer of the variational loss $\left\|\boldsymbol{y}^\delta-\mathcal{A}(\boldsymbol{x})\right\|_2^2+\lambda\,\mathcal{R}_{\theta^*}(\boldsymbol{x})$.

One can derive precise stability estimates and stronger convergence guarantees when $\mathcal{R}_\theta$ is convex, as we will show next.

3. Adversarial convex regularization

In [8], we proposed to parametrize $\mathcal{R}_\theta$ using an input-convex neural network (ICNN) [1]. For simplicity, we explain the construction in the finite-dimensional setting $\mathbb{X} = \mathbb{R}^n$.



(a) ground-truth (b) FBP: 21.63, 0.24 (c) TV: 29.25, 0.79 (d) AR: 31.83, 0.84 (e) ACR: 30.00, 0.82

(f) ground-truth (g) FBP: 21.61, 0.17 (h) TV: 25.74, 0.80 (i) AR: 26.83, 0.71 (j) ACR: 27.98, 0.84

Figure 1. Comparison of different reconstruction methods for sparse-view (first-row) and limited-angle (second-row) CT (in terms of PSNR (dB) and SSIM). For sparse-view CT, ACR outperforms TV, but AR turns out to be better than ACR. However, for limited-angle CT, ACR outperforms both TV and AR in terms of reconstruction quality.

It is straight-forward to show that $\mathcal{R}_\theta(\boldsymbol{x}) := \mathcal{H}_{\text{avg}}(\boldsymbol{z}_{L+1}(\boldsymbol{x})) + \rho_0 \|\boldsymbol{x}\|_2^2$ is strongly-convex in $\boldsymbol{x}$, where $\mathcal{H}_{\text{avg}}$ is an averaging operator and the entries of $\boldsymbol{z}_{L+1}(\boldsymbol{x})$ are convex in $\boldsymbol{x}$. $\boldsymbol{z}_{L+1}(\boldsymbol{x})$ is computed as $\boldsymbol{z}_{i+1}(\boldsymbol{x}) = \varphi\left(\mathcal{B}_i\left(\boldsymbol{z}_i(\boldsymbol{x})\right) + \mathcal{W}_i(\boldsymbol{x}) + \boldsymbol{b}_i\right)$, $i = 0, \cdots, L$, where $\mathcal{B}_i$'s are point-wise non-negative and linear, and the (element-wise) nonlinear activation φ is convex and monotonically non-decreasing (e.g., ReLU/leaky-ReLU). This construction of a convex regularizer uses the facts that a non-negative combination of convex functions is convex and the composition $f_1 \circ f_2$ is convex when both f_1 and f_2 are convex and f_1 is monotone non-decreasing. We showed in [8] that the ACR so constructed has the following properties:

- Stability: $\left\|\hat{\boldsymbol{x}}_\lambda\left(\boldsymbol{y}^\delta\right) - \hat{\boldsymbol{x}}_\lambda(\boldsymbol{y}^0)\right\|_2 \leq \frac{\beta_1 \delta}{\lambda \rho_0}$, where $\beta_1 := \sup_{\boldsymbol{x} \in \mathbb{X}} \frac{\|\mathcal{A}(\boldsymbol{x})\|_{\mathbb{Y}}}{\|\boldsymbol{x}\|_{\mathbb{X}}} < \infty$ is the spectral-norm of (linear) $\mathcal{A}$ and $\boldsymbol{y}^0$ denotes clean data.

- Well-posedness: For $\delta \to 0$, $\lambda(\delta) \to 0$, and $\frac{\delta}{\lambda(\delta)} \to 0$, the solution $\hat{\boldsymbol{x}}_\lambda\left(\boldsymbol{y}^\delta\right)$ converges to the $\mathcal{R}$-minimizing solution below with respect to the norm on $\mathbb{X}$:
$$\boldsymbol{x}^\dagger := \arg\min_{\boldsymbol{x}} \mathcal{R}(\boldsymbol{x}) \text{ subject to } \mathcal{A}(\boldsymbol{x}) = \boldsymbol{y}^0.$$
- Convergence of sub-gradient-descent: There exist step-sizes $\eta_k^* = 2\lambda\rho_0 \frac{\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}\|_{\mathbb{X}}^2}{\|\boldsymbol{z}_k\|_{\mathbb{X}}^2}$ such that the updates $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k^* \boldsymbol{z}_k$, where $\boldsymbol{z}_k \in \partial\left(\lambda\,\mathcal{R}_{\theta^*}(\boldsymbol{x}_k)\right)$, converge to true minimizer $\hat{\boldsymbol{x}}$ with respect to the norm topology.

Some sample reconstructions on the Mayo-clinic low-dose CT data are shown in Figure 1.

References

[1] B. Amos, L. Xu, and J. Z. Kolter, *Input convex neural networks*, in International Conference on Machine Learning, 2017, pp. 146–155.

[2] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein GAN*, arXiv preprint arXiv:1701.07875v3, (Dec. 2017).

[3] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, *Improved Training of Wasserstein GANs*, arXiv preprint arXiv:1704.00028v3, (Dec. 2017).

[4] E. Kobler, A. Effland, K. Kunisch, and T. Pock, *Total Deep Variation for Linear Inverse Problems*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 7549–7558.

[5] H. Li, J. Schwab, S. Antholzer, and M. Haltmeier, *NETT: Solving Inverse Problems with Deep Neural Networks*, arXiv preprint arXiv:1803.00092v3, (Dec. 2019).

[6] J. Liu, Y. Sun, C. Eldeniz, W. Gan, H. An, and U. S. Kamilov, *RARE: image reconstruction Using Deep priors learned without groundtruth*, IEEE J. Selected Topics in Signal Processing, 14 (2020), pp. 1088–1099.

[7] S. Lunz, O. Öktem, and C.-B. Schönlieb, *Adversarial regularizers in inverse problems*, in Advances in Neural Information Processing Systems, 2018, pp. 8507–8516.

[8] S. Mukherjee, S. Dittmer, Z. Shumaylov, S. Lunz, O. Öktem, and C.-B. Schönlieb, *Learned convex regularizers for inverse problem*, arXiv preprint arXiv:2008.02839v2, (Mar. 2021).

[9] E. T. Reehorst, and P. Schniter, *Regularization by denoising: clarifications and new interpretations*, IEEE Transactions on Computational Imaging, 5 (2019), pp. 52–67.

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

[11] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen, *Variational methods in imaging*, Springer, 2009.

[12] D. Ulyanov, A. Vedaldi, and V. Lempitsky, *Deep image prior*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9446–9454.

Reporter: Matthias Beckmann

Participants

Dr. Jonas Adler
Department of Mathematics
Royal Institute of Technology
Lindstedtsvägen 25
100 44 Stockholm
SWEDEN

Prof. Dr. Simon R. Arridge
Department of Computer Science
University College London
Gower Street
London WC1E 6BT
UNITED KINGDOM

Dr. Matthias Beckmann
Fachbereich Mathematik
Universität Hamburg
Bundesstraße 55
20146 Hamburg
GERMANY

Dr. Martin Benning
School of Mathematical Sciences
Queen Mary
University of London
Mile End Road
London E1 4NS
UNITED KINGDOM

Dr. Marta M. Betcke
Department of Computer Science
University College London
90 High Holborn
London WC1V 6LJ
UNITED KINGDOM

Dr. Romana Boiger
Paul Scherrer Institute (PSI)
Forschungsstrasse 111
5232 Villigen
SWITZERLAND

Prof. Dr. Michael Bronstein
Imperial College London
South Kensington Campus
London SW7 2AZ
UNITED KINGDOM

Prof. Dr. Maarten V. de Hoop
Simons Chair in Computational and
Applied Mathematics and Earth Science
Rice University
Houston TX 77005
UNITED STATES

Prof. Dr. Juan Carlos De los Reyes
Centro de Modelización Matemática
(MODEMAT)
Edificio #12, 6to piso
Escuela Politécnica Nacional
Ladrón de Guevara E11-253
170525 Quito
ECUADOR

Dr. Sören Dittmer
Zentrum für Technomathematik
Fachbereich 3, AG Technomathematik
Universität Bremen
Bibliothekstrasse 5
28359 Bremen
GERMANY

Margaret Duff
Dept. of Mathematical Sciences
University of Bath
Claverton Down
Bath BA2 7AY
UNITED KINGDOM

Prof. Dr. Yonina C. Eldar
Department of Mathematics
Technion - Israel Institute of
Technology
Haifa 32000
ISRAEL

Dr. Christian Etmann
Centre for Mathematical Sciences
University of Cambridge
Wilberforce Road
Cambridge CB3 0WB
UNITED KINGDOM

Prof. Dr. Mark Girolami
University of Cambridge
Department of Engineering
Trumpington Street
Cambridge CB2 1PZ
UNITED KINGDOM

Prof. Dr. Markus Haltmeier
Institut für Mathematik
Universität Innsbruck
Technikerstr. 13
6020 Innsbruck
AUSTRIA

Asst. Prof. Dr. Andreas Hauptmann
Research Unit of Mathematical Sciences
University of Oulu
FI-90014 University of Oulu
Pentti Kaiteran katu 1
P.O. Box P.O.Box 8000
90014 Oulu
FINLAND

Prof. Dr. Barbara Kaltenbacher
Institut für Mathematik
Alpen-Adria-Universität Klagenfurt
Universitätsstrasse 65-67
9020 Klagenfurt
AUSTRIA

Dr. Tobias Kluth
Zentrum für Technomathematik
Fachbereich 3
Universität Bremen
Postfach 330 440
28334 Bremen
GERMANY

Tianlin Liu
Departement Mathematik und Informatik
Universität Basel
Spiegelgasse 1
4051 Basel
SWITZERLAND

Dr. Felix Lucka
Centrum Wiskunde & Informatica (CWI)
Science Park 123
1098 XG Amsterdam
NETHERLANDS

Prof. Dr. Peter Maaß
Zentrum für Technomathematik
Fachbereich 3
Universität Bremen
Postfach 330 440
28334 Bremen
GERMANY

Prof. Dr. Ozan Oktem
Department of Mathematics
KTH - Royal Institute of Technology
SE Stockholm 10044
SWEDEN

Prof. Dr. Lorenzo Rosasco
University of Genova and Massachusetts Institute of Technology and Istituto Italiano di Tecnologia
via dodecaneso 35
16146 Genova
ITALY

Prof. Dr. Carola-Bibiane Schönlieb
Department of Applied Mathematics and Theoretical Physics (DAMTP)
Centre for Mathematical Sciences
Wilberforce Road
Cambridge CB3 0WA
UNITED KINGDOM

Prof. Dr. Jin Keun Seo
School of Mathematics and Computing
(Computational Science and
Engineering)
Yonsei University
Seoul 120-749
KOREA, REPUBLIC OF

Prof. Dr. Samuli Siltanen
Department of Mathematics and
Statistics
University of Helsinki
P.O. Box 68
00014 University of Helsinki
FINLAND

Dr. Antonio Stanziola
Department of Medical Physics and
Biomedical Engineering
University College London
Gower Street
London WC1E 6BT
UNITED KINGDOM

Prof. Dr. Andrew M. Stuart
Mathematics Institute
University of Warwick
Coventry CV4 7AL
UNITED KINGDOM