1.6 Data-driven Regularization and Quantitative Imaging

Moritz Flaschel, Michael Hintermüller, Clemens Sirotenko, and Karsten Tabelow

Inverse problems are ubiquitous in all areas of science where measurements and data play a role. Numerous examples can be found in physics, economics, engineering, or medical imaging, which is the particular subject of this article. In general, inverse problems focus on the reconstruction of quantities $u_{true} \in X$ from measured noisy and degraded data $f \in Y$. Here, X, Y are typically normed vector spaces. It is often assumed that the process that generates the data is at least partially known a priory and can be described by a possibly nonlinear function $A : X \to Y$ called *forward mapping*. In the case of image reconstruction, we often consider an image domain $\Omega \subset \mathbf{R}^d$, e.g., a rectangle in $\mathbf{R}^2$. A continuous image on this domain is a function $u : \Omega \to \mathbf{R}^m$ where $u(x)$ describes, for instance, the image intensity or the color at the location $x \in \Omega$. To account for unknown errors in the data, we introduce the variable $\eta \in Y$. While in the field of nonparametric statistics, this error is modeled as a random variable following some known distribution $\eta \sim \mathbf{P}^Y$ on Y, the deterministic viewpoint assumes η just to be a norm-bounded highly oscillating function in Y. The overall data-generating process can eventually be described by the equation

$$f = A(u_{true}) + \eta \tag{1}$$

with f being either a Y-valued random variable or simply an element of Y, when the noise distribution is not modeled explicitly.

Model-based reconstruction. A natural approach to recovering u_{true} from measured data f could be to solve the least-squares problem

$$A^\dagger(f) = \operatorname*{argmin}_{u \in X} \frac{1}{2}\|A(u) - f\|_Y^2, \tag{2}$$

which defines an operator called *pseudo-inverse* $A^\dagger : R(A) \to X$ on the range $R(A)$ of the operator A. However, it is well known that this operator can be set valued, and even for bounded and linear A, the pseudo-inverse $A^\dagger$ is known to be discontinuous in many applications. This results in amplified errors and unstable reconstruction processes. The classical way to overcome these problems is to regularize the problem. Regularization generally describes the process of using prior knowledge about the appearance of the true solution. These are often properties like sparsity or additional smoothness. This approach leads to optimization problems of the type

$$\operatorname*{minimize}_{u \in X} D(A(u), f) + \mathcal{R}(\alpha, u). \tag{3}$$

Here, $D : Y \times Y \to \mathbf{R}$ denotes a distance-like function on Y, also called *data fidelity*, which is chosen depending on the noise distribution. The term $\mathcal{R}(\alpha, u)$ is used to penalize undesired behavior of the solution, e.g., fast oscillations. The parameter $\alpha \in U$ balances the influence of the regularizer $\mathcal{R}(\alpha, u)$ and the data-fidelity term $D(A(u), f)$. Often, $\alpha > 0$ is chosen to be scalar, but in order to account also for different regularization in different parts of the image domain, we allow that α lies in some general vector space U. The approach of representing estimators of clean images as minimization problems of type (3) is often called *variational* or *model-based approach*.

For a comparison of the least-squares and a regularized solution using a total variation (TV) penalty with scalar parameter; see Figure 1.

Weighted magnetic resonance imaging. The main examples that we want to focus on are weighted magnetic resonance imaging (MRI) and quantitative MRI (qMRI). The reconstruction of sensor data for complex weighted MRI can be mathematically described by the continuous Fourier transform. The forward operator reads in this case $A : L^2(\Omega, \mathbf{C}) \to L^2(\Omega_k, \mathbf{C})$ defined as $Au = P \circ \mathcal{F}u$ where $\mathcal{F}$ denotes the continuous Fourier transform, and P is a projection-type linear operator that selects a set of frequencies $\Omega_k \subset \mathbf{R}^2$ in the space of possible all frequencies, also called k*-space*. In Figure 1, a prototypical MRI setup and two reconstructions are depicted. A sampling mask is shown on the right-hand side, in which the white lines represent the frequencies in Ω_k that are actually sampled. If there are only a few white lines, this is referred to as *highly under-sampled* images. A practical disadvantage of classical weighted MRI is that the images depend on acquisition parameters that are not comparable across time and site. This problem will be addressed later by qMRI.

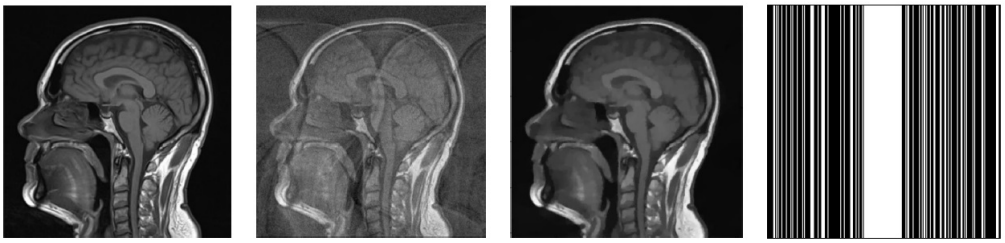

Fig. 1: *Different solutions of the MRI inverse problem. From left to right: Clean image u_{true}, least-squares solution $A^\dagger f$, total variation regularization, sampling mask.*

Data-driven regularization

While the model-based approach features nice theoretical properties, there might be some underlying structure in the data that is not captured by these handcrafted regularizers. One possible approach to incorporate additional knowledge, which is hidden in the data, is proposed in [2] where the classical TV minimization and its second-order extension, the total generalized variation (TGV), are combined with spatial adaptive regularization parameter choice; see also Figure 1 for an example of TV reconstruction with scalar parameters. In fact, we consider for the linear forward operator $A : X \to Y$ the problem

$$\underset{u \in X}{\text{minimize}} \quad \frac{1}{2}\|Au - f\|_Y^2 + \int_\Omega \alpha(x)\, \mathrm{d}|Du|\,. \tag{4}$$

Recall that $|Du|$ denotes the total variation of the measure-valued distributional derivative Du of u. For the sake of simplicity, we will concentrate here on TV only. The idea is to choose the spatially-dependent regularization parameter $\alpha \in C(\overline{\Omega})_{>0}$ adapted to the data f. In [2], the following unsupervised loss function is proposed:

$$L_{k,\sigma}(u, f) = \int_\Omega \max(k * (Au - f)^2 - \overline{\sigma}^2, 0) + \min(k * (Au - f)^2 - \underline{\sigma}^2, 0)\, \mathrm{d}x. \tag{5}$$

Here, the true variance $\sigma^2 > 0$ of $\eta \in Y$ must be known in advance, and $k : \mathbf{R}^n \to \mathbf{R}$ denotes a suitable L^1-normalized kernel with small support including zero. The function $L_{k,\sigma}(u, f)$ locally penalizes deviations of the averaged squared residual $(Au - f)^2$ from a small neighborhood $(\underline{\sigma}^2, \overline{\sigma}^2)$ around the known true variance $\sigma^2 > 0$. The approach results in the bilevel optimization problem

$$\begin{cases} \underset{u \in X, \alpha \in U_{ad}}{\text{minimize}} & L_{k,\sigma}(u, f) + r(\alpha) \\ \text{s.t.} & u = u(\alpha, f) \in \underset{u \in X}{\text{argmin}} \ \frac{1}{2}\|Au - f\|_Y^2 + \int_\Omega \alpha(x)\, \mathrm{d}|Du|, \end{cases} \tag{6}$$

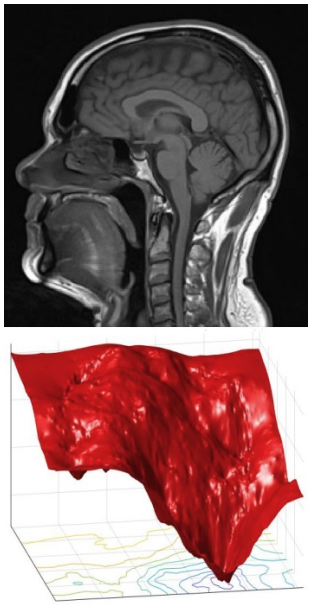

Fig. 2: *Spatially-dependent parameter for TGV regularization, which is found using the approach in [2]*

where $r : U \to \mathbf{R}$ enforces additional smoothness on α, and $U_{ad} \subset U$ is the set of admissible parameters. While this approach does not require training data, we still face a nonconvex, non-differentiable bilevel problem with solution in a nonreflexive Banach space. This problem class is notoriously difficult to solve due to the lack of standard constraint qualifications. To overcome these issues, additional smoothing techniques are presented in [2], which are known from optimal control theory to eventually find stationary points of an optimization problem where the constraints are given by a quasi-linear elliptic partial differential equation (PDE). These stationary points are found by utilizing a projected gradient descent approach, whose analysis is presented in an infinite-dimensional setup. The results of the proposed method are presented in Figure 2. Note that the regularization parameter, found by (6), is low in areas around the neck, where fine structures must be preserved, and high in smoother areas of the image. A natural extension of this work is presented in [3]. Here, the lower-level problem is approximated by k steps of an iterative solver that is known to converge to a solution of (4). Let the k-th step of this solver be denoted by $S_k(\alpha, f, u_0)$ when initialized at u_0. The spatially-adaptive regularization parameter $\alpha \in C(\overline{\Omega})$ is then replaced by a learnable structure, e.g., a neural network $\alpha_\theta : X \to C(\overline{\Omega})$ with parameters $\theta \in \Theta$. The resulting learning problem then reads

$$\underset{\theta \in \Theta}{\text{minimize}} \quad \frac{1}{2M} \sum_{i=1}^{M} \|u_i^\dagger - u_i\|_X^2 + \frac{1}{2}\|\theta\|^2 \quad \text{s.t. } u_i = S_k(\alpha_\theta(A^\dagger f_i), f_i, A^\dagger f_i)\,, \tag{7}$$

where training pairs $(u_i^\dagger, f_i)_{i=1}^M$ are used that are generated according to $f_i = Au_i^\dagger + \eta_i$. The overall unrolled optimization algorithm has the structure of a multi-layer neural network, and (7) is a classical supervised learning problem whose solution can be approximated by using of the shelf stochastic optimizers like Adam (Kingma & Ba, 2014) alongside techniques from automatic differentiation. The setup is analyzed in [3] in finite dimensions for an unrolled primal-dual splitting algorithm (PDHG), and the convergence of the solutions of (7) to solutions of a suitable bilevel problem is addressed for $k \to \infty$.

Quantitative imaging

In contrast to weighted MRI, the goal of qMRI is the voxel-wise reconstruction of tissue parameters such as proton-spin density ρ and relaxation times T_1, T_2. These are sensitive to the biological tissue properties, allow for a quantitative comparison across different scanners and can be used as non-invasive disease markers. The general idea is to collect not only one, but multiple highly

under-sampled Fourier images $f_i = P_i \mathcal{F} u_i, i = 1, \dots L$ at subsequent time points $t_1 < \dots < t_L$. The images $u = (u_1, \dots, u_L)$ are implicitly connected to the physical quantity q by an equation $e(u, q) = 0$, where $e : X \times Q \to Z$ is an operator describing some physical law, and $q \in Q$ denotes the physical quantity of interest. In many cases, $e(u, q) = 0$ consists of partial or ordinary differential equations. This type of problem setting particularly fits for qMRI, which has been intensively investigated recently in [1]. Here, $u_i(x) = \rho(x)[m_1(x, t_i) + i m_2(x, t_i)] \in \mathbf{C}$, where $\rho(x)$ is the proton-spin density, and $m(x, t) = (m_1, m_2, m_3)(x, t)$ denotes the average magnetization of the hydrogen protons located at x at time t under the influence of some externally controlled magnetic field $B(x, t)$. The differential equation that describes this process is called Bloch equation:

$$\partial_t m(x,t) = m(x,t) \times \gamma B(x,t) - \left(\frac{m_1(x,t)}{T_2(x)}, \frac{m_2(x,t)}{T_2(x)}, \frac{m_3(x,t) - m_{eq}}{T_1(x)} \right)^\top . \qquad \text{(Bloch)}$$

Here $q = (\rho, T_1, T_2)(x)^\top$ denotes the physical quantity of interest. In [1] and forthcoming articles, a series of problems is analyzed that aim at extracting $q \in Q$ from the measurements f. Starting from the smooth constrained optimization problem

$$\underset{q \in \mathcal{C}_{ad}}{\text{minimize}} \quad \frac{1}{2} \| A\Pi(q) - f \|_Y^2 + \frac{\alpha}{2} \| q \|_Q^2 , \qquad (8)$$

fast solvers are developed and analyzed in function space using a smooth solution operator $\Pi : Q \to X$ of the equation $e(u, q) = 0$. However, this assumption is only rarely met in practice, where not every Bloch equation has an explicit solution operator. Moreover, the Bloch equation is only a mathematical model that uses simplified experimental assumptions that are far more complicated in reality. In order to account for this model uncertainty, the equation $e(u, q) = 0$ is replaced by some learned equation $e_\theta(u, y) = 0$ in which some parts or even the whole operator $e_\theta(\cdot, \cdot)$ is learned a priori from training data using a neural network with weights θ. The resulting optimization problem eventually reads in this case

$$\underset{u \in X, q \in Q}{\text{minimize}} \quad \frac{1}{2} \| Au - f \|_Y^2 + \frac{\alpha}{2} \| q \|_Q^2 \quad \text{subject to} \quad e_\theta(u, q) = 0 \text{ and } q \in \mathcal{C}_{ad}. \qquad (9)$$

As above, $\mathcal{C}_{ad} \subset Q$ denotes a convex, closed set of admissible parameters. This problem is much more involved than (8), due to the nondifferentiability that is introduced by using typical non-smooth activation functions of the neural network. In [1], stationarity concepts for these types of problems in infinite dimensions are developed and a sequential quadratic programming-type descent algorithm is proposed. One set of recovered physical parameters from noisy data is depicted in Figure 3.

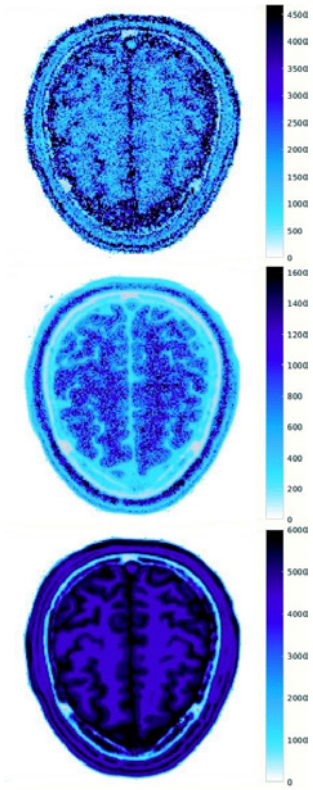

Fig. 3: *The results from the work [1]. The parameters $T1, T2$, and ρ are presented from top to bottom.*

Estimation bias

Even if we know the operator A from physics exactly or are convinced that it describes an appropriate model for the observed data f, the noise distribution η and the operator A do influence whether the estimates of u_{true} using, e.g., Eq. (2), are biased due to some nonlinearity of A or the noise η being non-Gaussian. As a real-world example, we consider measurements S from

diffusion-weigthed MRI (dMRI) which can be modeled by a diffusion kurtosis model (DKI):

$$S_{b,\vec{g}} = S_0 \cdot \exp\left(-bD + \frac{b^2}{6}\left(\frac{\mathrm{Tr}(D)}{3}\right)^2 W\right) + \eta\,, \tag{10}$$

where $D = \sum_{i,j=1}^{3} g_i g_j D^{ij}$, $W = \sum_{i,j,k,l=1}^{3} g_i g_j g_k g_l W^{ijkl}$, and acquisition parameters $b, \vec{g}$. Here, $u = (D^{ij}, W^{ijkl})$ are the components of the diffusion and kurtosis tensor and form the basic parameters of interest. This general DKI model can be simplified assuming a symmetry that can be justified by biophysical considerations. Moreover, the noise η is not Gaussian but Rician or, more generally, noncentral χ distributed as it refers to magnitude data from complex Gaussian noise. The resulting estimation bias has been examined in [4, 5], which shows different severity for different parameters from u due to the non-Gaussian noise, the nonlinearity of the operator, and the violation of assumptions for real data. Consequently, for parameter estimation not only noise reduction, e.g., through regularization, is important, but also appropriate bias reduction; see, e.g., Polzehl and Tabelow, *Low SNR in diffusion MRI models*, J. Amer. Statist. Assoc., **111** (2016), pp. 1480–1490.

Conclusions and outlook

The development and mathematical analysis of quantitative imaging methodologies is still at the beginning stage. Relevant topics for the application of qMRI range from the development of more sophisticated physical models encoded in the equation $e(u, q) = 0$ to the usage of modern data-driven regularization techniques, such as (coupled) dictionary learning or plug-and-play methods, that are able to incorporate prior knowledge from data while retaining interpretability and which can also be combined with handcrafted regularization techniques. Usually, these regularizers lead to high-dimensional nonconvex and nondifferentiable optimization problems for which the analysis is delicate. Current mathematical research focuses on the development and analysis of fast solvers for such problems and on the analysis of their regularization properties in dependence of the sampling methodology. Moreover, algorithms involving learned structures often are used in a finite-dimensional context. However, an infinite-dimensional setup is particularly important in physical imaging. Not only because the algorithms should work resolution-independently, but also because the physical models are provided in the continuous regime. For this reason, there is a natural demand to extend these data-driven regularization techniques to an infinite-dimensional setup. This topic is also of particular relevance in the current work.

References

[1] G. Dong, M. Hintermüller, K. Papafitsoros, *Optimization with learning-informed differential equation constraints and its applications*, ESAIM Control Optim. Calc. Var., **28** (2022), pp. 3/1–3/44.

[2] M. Hintermüller, K. Papafitsoros, C. Rautenberg, H. Sun, *Dualization and automatic distributed parameter selection of total generalized variation via bilevel optimization*, Numer. Funct. Anal. Optim., **43**:8 (2022), pp. 887–932.

[3] A. Kofler, F. Altekrüger, F.A. Ba, Ch. Kolbitsch, E. Papoutsellis, D. Schote, C. Sirotenko, F.F. Zimmermann, K. Papafitsoros, *Learning regularization parameter-maps for variational image reconstruction using deep neural networks and algorithm unrolling*, arXiv Preprint no. 2301.05888, 2023.

[4] J.M. Oeschger, K. Tabelow, S. Mohammadi, *Axisymmetric diffusion kurtosis imaging with Rician bias correction: A simulation study*, Magn. Reson. Med., **89**:2 (2023), pp. 787–799.

[5] ———, *Investigating apparent differences between standard DKI and axisymmetric DKI and its consequences for biophysical parameter estimates*, bioRxiv Preprint no. 2023.06.21.545891, 2023.