Bias-Reduction in Variational Regularization

Eva-Maria Brinkmann, Martin Burger, Julian Rasch, Camille Sutour

Bias-Reduction in Variational Regularization

Eva-Maria Brinkmann, Martin Burger, Julian Rasch, Camille Sutour*

December 21, 2016

Abstract

The aim of this paper is to introduce and study a two-step debiasing method for variational regularization. After solving the standard variational problem, the key idea is to add a consecutive debiasing step minimizing the data fidelity on an appropriate space, the so-called model manifold. The latter is defined by Bregman distances or infimal convolutions thereof, using the subgradient appearing in the optimality condition of the variational method. For particular settings, such as anisotropic ℓ^1- and TV-type regularization, previously used debiasing techniques are shown to be special cases. The proposed approach is however easily applicable to a wider range of regularizations. The two-step debiasing is shown to be well-defined and to optimally reduce bias in a certain setting.

Besides visual and PSNR-based evaluations, different notions of bias and variance decompositions are investigated in numerical studies. The improvements offered by the proposed scheme are demonstrated and its performance is shown to be comparable to optimal results obtained with Bregman iterations.

1 Introduction

Variational regularization methods with non-quadratic functionals such as total variation or ℓ^1-norms have evolved to a standard tool in inverse problems [8, 32], image processing [11], compressed sensing [10], and recently related fields such as learning theory [13]. The popularity of such approaches stems from superior structural properties compared to other regularization approaches. ℓ^1-regularization for example leads to sparse solutions with very accurate or even exact reconstruction of the support of the true solution. On the other hand it is known that such methods suffer from a certain bias due to the necessary increased weighting of the regularization term with increasing noise. Two well-known examples are the loss of contrast in total variation regularization [8, 25] or shrinked small peak values in ℓ^1-regularization. Accordingly, quantitative values of the solutions have to be taken with care.

Several approaches to reduce or eliminate the bias of regularization methods have been considered in literature: For ℓ^1-regularization and similar sparsity-enforcing techniques an ad-hoc approach is to determine the support of the solution by the standard variational methods in a first step, then use a second debiasing step that minimizes the residual (or a general data fidelity) restricted to that support, also known as *refitting* [19, 21, 18]. A slightly more advanced approach consists in adding a sign-constraint derived from the solution of the variational regularization method in addition to the support condition. This means effectively that the solution of the debiasing step shares an ℓ^1-subgradient with the solution of the variational regularization method. A different and more general approach is to iteratively reduce the bias via Bregman iterations [25] or similar approaches [20, 33]. Recent results for the inverse scale space method in the case of ℓ^1-regularization (respectively certain polyhedral regularization functionals [6, 22, 4]) show that the inverse scale space performs some kind of debiasing. Even more, under certain conditions, the variational regularization method and the inverse scale space method provide the same subgradient at corresponding settings of the regularization parameters [5]. Together with a characterization of the solution of the inverse scale space method as a minimizer of the residual on the set of elements with the same subgradient, this implies a surprising equivalence to the approach of performing a debiasing step

*Institut für Numerische und Angewandte Mathematik, Westfälische Wilhelms-Universität (WWU) Münster. Einsteinstr. 62, D 48149 Münster, Germany.

with sign-constraints. Recently, bias and debiasing in image processing problems were discussed in a more systematic way by Deledalle et al. [14, 15]. They distinguish two different types of bias, namely method bias and model bias. In particular they suggest a debiasing scheme to reduce the former, which can be applied to some polyhedral one-homogeneous regularizations. The key idea of their approach is the definition of suitable spaces, called model subspaces, on which the method bias is minimized. The remaining model bias is considered as the unavoidable part of the bias which is linked to the choice of regularization and hence the solution space of the variational method. The most popular example is the staircasing effect that occurs for total variation regularization due to the assumption of a piecewise constant solution. In the setting of ℓ^1-regularization a natural model subspace is the set of signals with a given support, which yields consistency with the ad-hoc debiasing approach mentioned above.

Based on this observation, the main motivation of this paper is to further develop the approach in the setting of variational regularization and unify it with the above-mentioned ideas of debiasing for ℓ^1-regularization, Bregman iterations, and inverse scale space methods.

Let us fix the basic notations and give a more detailed discussion of the main idea. Given a bounded linear operator $A\colon \mathcal{X} \to \mathcal{Y}$ between Banach spaces, a convex regularization functional $J\colon \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ and a differentiable data fidelity $H : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, we consider the solution of the variational method

$$u_\alpha \in \arg\min_{u\in\mathcal{X}} \; H(Au, f) + \alpha J(u) \tag{1.1}$$

as a first step. Here $\alpha > 0$ is a suitably chosen regularization parameter. This problem has a systematic bias, as we further elaborate on below. The optimality condition is given by

$$A^*\partial H(Au_\alpha, f) + \alpha p_\alpha = 0, \; p_\alpha \in \partial J(u_\alpha). \tag{1.2}$$

Now we proceed to a second step, where we only keep the subgradient p_α and minimize

$$\hat{u}_\alpha \in \arg\min_{u\in\mathcal{X}} \; H(Au, f) \text{ s.t. } p_\alpha \in \partial J(u). \tag{1.3}$$

Obviously, this problem is only of interest if there is no one-to-one relation between subgradients and primal values u, otherwise we always obtain $\hat{u}_\alpha = u_\alpha$. The most interesting case with respect to applications is the one of J being absolutely one-homogeneous, i.e. $J(\lambda u) = |\lambda| J(u)$ for all $\lambda \in \mathbb{R}$, where the subdifferential can be multivalued at least at $u = 0$. The debiasing step can be reformulated in an equivalent way as

$$\min_{u\in\mathcal{X}} \; H(Au, f) \text{ s.t. } D_J^{p_\alpha}(u, u_\alpha) = 0, \tag{1.4}$$

with the (generalized) Bregman distance given by

$$D_J^p(u, v) = J(u) - J(v) - \langle p, u - v\rangle, \quad p \in \partial J(v).$$

We remark that for absolutely one-homogeneous J this simplifies to

$$D_J^p(u, v) = J(u) - \langle p, u\rangle, \quad p \in \partial J(v).$$

The reformulation in terms of a Bregman distance indicates a first connection to Bregman iterations, which we make more precise in the sequel of the paper.

Summing up, we examine the following two-step method:

1) Compute the (biased) solution u_α of (1.1) with optimality condition (1.2),

2) Compute the (debiased) solution $\hat{u}_\alpha$ as the minimizer of (1.3) or equivalently (1.4).

In order to relate further to the previous approaches of debiasing ℓ^1-minimizers given only the support and not the sign, as well as the approach with linear model subspaces, we consider another debiasing approach being blind against the sign. The natural generalization in the case of an absolutely one-homogeneous functional J is to replace the second step by

$$\min_{u\in\mathcal{X}} \; H(Au, f) \text{ s.t. } \mathrm{ICB}_J^{p_\alpha}(u, u_\alpha) = 0,$$

where

$$\mathrm{ICB}_J^{p_\alpha}(u, u_\alpha) := \left[D_J^{p_\alpha}(\cdot, u_\alpha) \square D_J^{-p_\alpha}(\cdot, -u_\alpha)\right](u)$$

denotes the infimal convolution between the Bregman distances $D_J^{p_\alpha}(\cdot, u_\alpha)$ and $D_J^{-p_\alpha}(\cdot, -u_\alpha)$, evaluated at $u \in \mathcal{X}$. The infimal convolution of two functionals F and G on a Banach space $\mathcal{X}$ is defined as

$$\begin{aligned}(F \square G)(u) &= \inf_{\substack{\phi,\psi\in\mathcal{X},\\ \phi+\psi=u}} F(\phi) + G(\psi) \\ &= \inf_{z\in\mathcal{X}} F(u - z) + G(z).\end{aligned}$$

For the sake of simplicity we carry out all analysis and numerical experiments in this paper for a least-squares data fidelity (related to i.i.d. additive Gaussian noise)

$$H(Au, f) = \frac{1}{2}\|Au - f\|_{\mathcal{Y}}^2$$

for some Hilbert space $\mathcal{Y}$, but the basic idea does not seem to change for other data fidelities and noise models.

We show that the spaces characterized by the constraints

$$D_J^{p_\alpha}(u, u_\alpha) = 0 \quad \text{and} \quad \mathrm{ICB}_J^{p_\alpha}(u, u_\alpha) = 0$$

constitute a suitable extension of the model subspaces introduced in [14] to general variational regularization. In particular, we use those subspaces to provide a theoretical basis to define the bias of variational methods and investigate the above approach as a method to reduce it. Moreover, we discuss its relation to the statistical intuition of bias. At this point it is important to notice that choosing a smaller regularization parameter will also decrease bias, but on the other hand strongly increase variance. The best we can thus achieve is to reduce the bias at fixed α by the two-step scheme while introducing only a small amount of variance.

The remainder of the paper is organized as follows: In Section 2 we motivate our approach by considering bias related to the well-known ROF-model [29] and we review a recent approach on debiasing [14]. In the next section we introduce our debiasing technique supplemented by some first results. Starting with a discussion of the classical definition of bias in statistics, we consider a deterministic characterization of bias in Section 4. We reintroduce the notion of model and method bias as well as model subspaces as proposed in [14] and extend it to the infinite-dimensional variational setting. We furthermore draw an experimental comparison between the bias we consider in this paper and the statistical notion of bias. This provides nice illustrations of different types of bias. Finally, we comment on the relation of the proposed debiasing to Bregman iterations [25] and inverse scale space methods [30, 20]. We complete the paper with a description of the numerical implementation via a first-order primal-dual method and show numerical results for signal deconvolution and image denoising.

2 Motivation

Let us start with an intuitive approach to bias and debiasing in order to further motivate our method. To do so, we recall a standard example for denoising, namely the well-known ROF-model [29] and rewrite a recent debiasing approach [14] in the setting of our method.

2.1 Bias of total variation regularization

As already mentioned in the introduction, variational regularization methods suffer from a certain bias. This systematic error becomes apparent when the regularization parameter is increased. Indeed this causes a shift of the overall energy towards the regularizer, and hence a deviation of the reconstruction from the data in terms of quantitative values. Intuitively, this can be observed from the classical ROF-model [29], i.e.

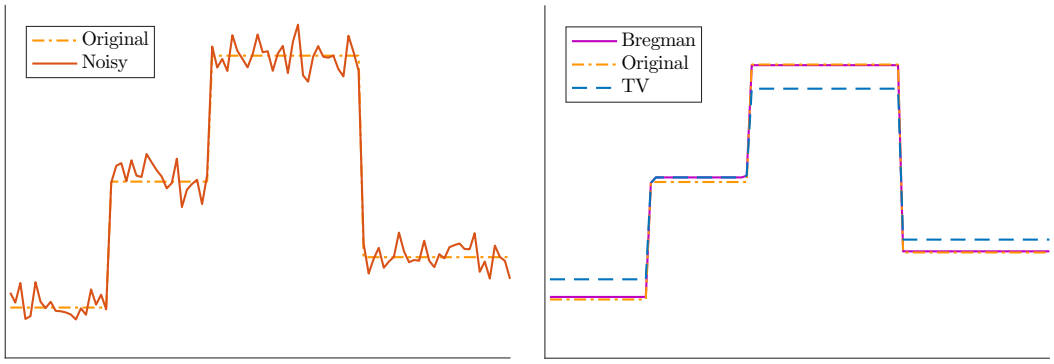
$$u_\alpha \in \arg\min_{u \in \mathbb{R}^n} \frac{1}{2}\|u - f\|_2^2 + \alpha\|\nabla u\|_1. \qquad (2.1)$$

It yields a piecewise constant signal u_α reconstructed from an observation $f \in \mathbb{R}^n$ corrupted by Gaussian noise (see Figure 1(a)). Figure 1(b) shows the solution of (2.1) together with the true, noiseless signal that we aimed to obtain. Even though the structure of the true signal is recovered, the quantitative values of the reconstruction do not match the true signal. Instead, jumps in the signal have a smaller height, which is often referred to as a loss of contrast. Without any further definition, one could intuitively consider this effect as the bias (or one part of the bias) of the ROF model. Hence, the goal of a bias reduction method would be to restore the proper signal height while keeping the (regularized) structure.

It has been shown in [25, 2] that this can be achieved by the use of Bregman iterations, i.e. by iteratively calculating

$$u_\alpha^{k+1} \in \arg\min_{u \in \mathbb{R}^n} \frac{1}{2}\|u - f\|_2^2 + \alpha D_J^{p_\alpha^k}(u, u_\alpha^k), \qquad (2.2)$$

where in our case $J(u) = \|\nabla u\|_1$, and $p_\alpha^k \in \partial J(u_\alpha^k)$ is a subgradient of the last iterate u_α^k. Since for total variation regularization the subgradient p_α^k essentially encodes the edge information of the last iterate, its iterative inclusion



(a) Original and noisy signal

(b) Denoising with TV regularization and Bregman iterations

Figure 1: Illustration of the bias of the ROF model on a 1D signal. (a) Original signal, and noisy signal corrupted by additive Gaussian noise. (b) Restoration of the noisy signal with TV regularization and Bregman iterations. The TV reconstruction recovers the structure of the signal but suffers from a loss of contrast, which is however well recovered with Bregman iterations.

allows to keep edges while restoring the right height of jumps between edges. We further elaborate on that in Section 4. And indeed, the reconstruction via (2.2) in Figure 1(b) shows an almost perfect recovery of the true signal even in terms of quantitative values. This implies that Bregman iterations are able to reduce or even eliminate our heuristically defined bias.

However, a theoretical basis and justification is still missing, i.e. a proper definition of the bias of variational methods, a proof that Bregman iterations indeed reduce the bias in that sense, and in particular a link to the statistical definition and understanding of bias. With this paper we aim to define a proper basis for this link, and in particular further establish the connection between bias reduction techniques and Bregman distances.

2.2 Recent debiasing and Bregman distances

In order to further motivate the use of Bregman distances for bias reduction let us recall and review a very recent approach on debiasing and work out its relation to Bregman distances. In [14], Deledalle et al. introduce a debiasing algorithm for anisotropic TV-type regularized problems

$$u_\alpha \in \arg\min_{u\in\mathbb{R}^n} \frac{1}{2}\|Au - f\|_2^2 + \alpha\|\Gamma u\|_1,$$

with a linear operator $A \in \mathbb{R}^{n\times d}$, a discrete gradient operator $\Gamma \in \mathbb{R}^{n\times m}$ and noisy data $f \in \mathbb{R}^d$. In [14] the authors argued that the loss of contrast characteristic for this kind of regularization is indeed bias in their sense. In order to correct for that error, the proposed debiasing method in [14] consists in looking for a debiased solution $\hat{u}_\alpha$ such that $\Gamma\hat{u}_\alpha$ and Γu_α share the same support, but $\hat{u}_\alpha$ features the right intensities. Mathematically, the solution $\hat{u}_\alpha$ of their debiasing problem is given by

$$\hat{u}_\alpha \in \arg\min_{u\in\mathbb{R}^n} \sup_{z\in F_\mathcal{I}} \frac{1}{2}\|Au - f\|_2^2 + \langle \Gamma u, z\rangle, \tag{2.3}$$

where $F_\mathcal{I} = \{z \in \mathbb{R}^m \mid z_\mathcal{I} = 0\}$, and $\mathcal{I}$ is the set of indices corresponding to nonzero entries of Γu_α. We can explicitly compute the supremum (the convex conjugate of the indicator function of the set $F_\mathcal{I}$), which is

$$\sup_{z\in F_\mathcal{I}} \langle \Gamma u, z\rangle = \begin{cases} \infty, & (\Gamma u)_i \neq 0 \text{ for some } i \notin \mathcal{I}, \\ 0, & \text{else.} \end{cases}$$

Hence, $\hat{u}_\alpha$ can only be a minimizer of (2.3) if $\operatorname{supp}(\Gamma\hat{u}_\alpha) \subset \operatorname{supp}(\Gamma u_\alpha)$, thus

$$\hat{u}_\alpha \in \arg\min_{u\in\mathbb{R}^n} \frac{1}{2}\|Au - f\|_2^2 \quad \text{s.t. } \operatorname{supp}(\Gamma\hat{u}_\alpha) \subset \operatorname{supp}(\Gamma u_\alpha). \tag{2.4}$$

We can also enforce this support condition using the infimal convolution of two ℓ^1-Bregman distances. Defining $J(u) = \|\Gamma u\|_1$, the subdifferential of J at u_α is given by

$$\partial J(u_\alpha) = \{\Gamma^T q_\alpha \in \mathbb{R}^n \mid \|q_\alpha\|_\infty \leq 1, \\ (q_\alpha)_i = \operatorname{sign}((\Gamma u_\alpha)_i) \text{ for } (\Gamma u_\alpha)_i \neq 0\}.$$

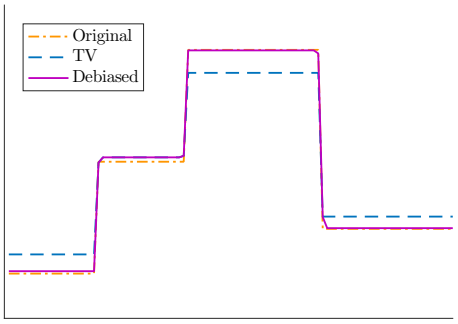


Figure 2: TV denoising of a one-dimensional noisy signal and debiasing using the proposed method.

In particular $|(q_\alpha)_i| = 1$ on the support of Γu_α. Let q_α be such a subgradient and consider the ℓ^1-Bregman distances $D^{q_\alpha}_{\|\cdot\|_1}(\cdot, \Gamma u_\alpha)$ and $D^{-q_\alpha}_{\|\cdot\|_1}(\cdot, -\Gamma u_\alpha)$. According to [23], their infimal convolution evaluated at Γu is given by:

$$\begin{aligned}
&\mathrm{ICB}^{q_\alpha}_{\|\cdot\|_1}(\Gamma u, \Gamma u_\alpha) \\
&= [D^{q_\alpha}_{\|\cdot\|_1}(\cdot, \Gamma u_\alpha) \square D^{-q_\alpha}_{\|\cdot\|_1}(\cdot, -\Gamma u_\alpha)](\Gamma u) \\
&= \sum_{i=1}^{m} (1 - |(q_\alpha)_i|)|(\Gamma u)_i|,
\end{aligned}$$

We observe that this sum can only be zero, if for all i either $|(q_\alpha)_i| = 1$ or $(\Gamma u)_i = 0$. Hence, given $p_\alpha = \Gamma^T q_\alpha \in \partial J(u_\alpha)$ with $|(q_\alpha)_i| < 1$ for $i \notin \mathcal{I}$, i.e. $|(q_\alpha)_i| = 1 \Leftrightarrow (\Gamma u_\alpha)_i \neq 0$, we can rewrite the above debiasing method (2.3) as

$$\min_{u \in \mathbb{R}^n} \frac{1}{2}\|Au - f\|_2^2 \text{ s.t. } \mathrm{ICB}^{q_\alpha}_{\|\cdot\|_1}(\Gamma u, \Gamma u_\alpha) = 0.$$

Note that the zero infimal convolution exactly enforces the support condition (2.4) only if the condition $|(q_\alpha)_i| = 1 \Leftrightarrow (\Gamma u_\alpha)_i \neq 0$ holds true. Intuitively, since the subdifferential is multivalued at $(\Gamma u_\alpha)_i = 0$, this leads to the question of how to choose q_α properly. However, our method does not depend on the choice of a particular q_α, but instead use a unique subgradient p_α coming from the optimality condition of the problem. We further comment on this in Section 4.

3 Debiasing

Inspired by the above observations, let us define the following two-step-method for variational regularization on Banach spaces. At first we compute a solution u_α of the standard variational method

$$1) \quad u_\alpha \in \arg\min_{u \in \mathcal{X}} \frac{1}{2}\|Au - f\|_{\mathcal{Y}}^2 + \alpha J(u), \tag{3.1}$$

where $A\colon \mathcal{X} \to \mathcal{Y}$ is a linear and bounded operator mapping from a Banach space $\mathcal{X}$ to a Hilbert space $\mathcal{Y}$, $J\colon \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ denotes a convex and one-homogeneous regularization functional and $f \in \mathcal{Y}$.

The first-order optimality condition of (3.1) reads:

$$p_\alpha = \frac{1}{\alpha} A^*(f - Au_\alpha),\ p_\alpha \in \partial J(u_\alpha), \tag{3.2}$$

and it is easy to show that this p_α is unique. We use this subgradient to carry over information about u_α to a second step. In the spirit of the previous paragraph the idea is to perform a constrained minimization of the data fidelity term only:

$$\begin{aligned}
2\text{ a)} \quad \hat{u}_\alpha \in \arg\min_{u \in \mathcal{X}} \ & \frac{1}{2}\|Au - f\|_{\mathcal{Y}}^2 \\
\text{s.t. } & \mathrm{ICB}^{p_\alpha}_J(u, u_\alpha) = 0.
\end{aligned} \tag{3.3}$$

If we reconsider the ad-hoc idea of ℓ^1- or TV-type debiasing from the introduction, it can be beneficial to add a sign or direction constraint to the minimization, rather than a support condition only. This can be achieved by the use of a single Bregman distance. Hence it is self-evident to define the following alternative second step:

$$\begin{aligned}
2\text{ b)} \quad \hat{u}_\alpha \in \arg\min_{u \in \mathcal{X}} \ & \frac{1}{2}\|Au - f\|_{\mathcal{Y}}^2 \\
\text{s.t. } & D^{p_\alpha}_J(u, u_\alpha) = 0.
\end{aligned} \tag{3.4}$$

We would like to point out that until now we only argued heuristically that the above method actually performs some kind of debiasing for specific problems. But since we are able to recover the debiasing method of [14] for $J(u) = \|\Gamma u\|_1$ as a special case, at least for this specific choice of regularization (and a finite-dimensional setting) our method is provably a debiasing in their sense.

However, our method is much more general. Since in contrast to [14] it does not depend on a specific representation of u_α, it can theoretically be carried out for any suitable regularizer J. In particular, the method does not even depend on the specific choice of the data term. In order to obtain a unique subgradient p_α from the optimality condition it is desirable e.g. to have a

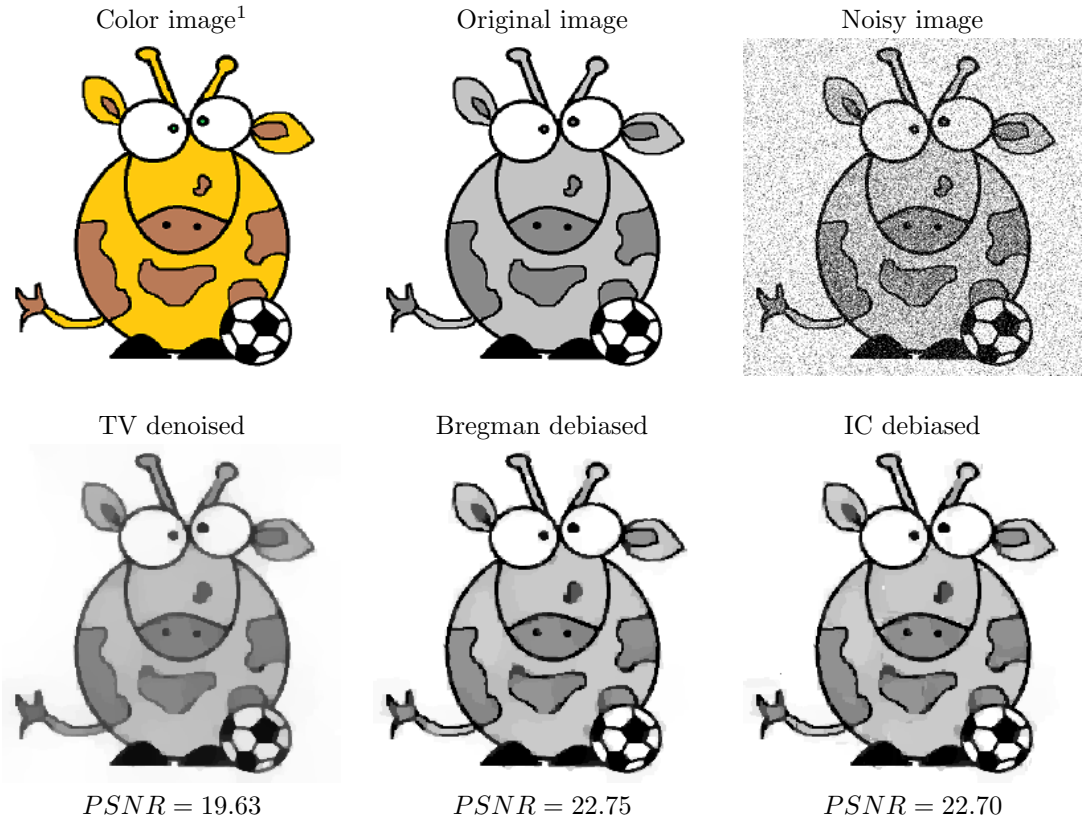


Figure 3: Denoising of a cartoon image. First row: original image, noisy image corrupted by Gaussian noise. Second row: TV reconstruction and debiasing using the Bregman distance and the infimal convolution, respectively. The TV reconstruction recovers well the structures of the images but suffers from a loss of contrast, while the debiased solutions allow for a more accurate dynamic.

[1] The color image is provided in order to point out that it is indeed a Giraffe and not a cow.

differentiable data fidelity, but if we drop that condition, the data term is theoretically arbitrary. Since this generalization requires a little more work, we focus on a squared Hilbert space norm in this paper.

Before we actually lay a theoretical foundation for our framework and prove that our method indeed is a debiasing method, we show some motivating numerical results and prove the well-definedness of the method.

3.1 A first illustration

To give a first glimpse of the proposed method, we revisit the ROF reconstruction model (2.1) from Section 2 and show some numerical results in one and two dimensions.

Taking the subgradient p_α of the TV reconstruction u_α of the one-dimensional signal and performing our debiasing method, we obtain the results in Figure 2. The second step restores the right height of the jumps and yields the same result as the Bregman iterations we performed in Section 2.

As a second example we perform denoising on a cartoon image corrupted by Gaussian noise. The first row of Figure 3 shows the original image in the middle and its noisy version on the right. The left image in the second row is the denoising result obtained with the ROF-model (2.1). We observe that noise has been reduced substantially, but some part of the contrast is lost. The second step of our method restores the contrast while keeping the structure of the first solution, yielding the two results on the right-hand side of the second row.

3.2 Well-definedness of the method

The aim of this section is to show that the method defined above is well-defined, i.e. that

there always exists at least one solution to the problem. By exploiting that p_α lies in the range of A^* we can prove coercivity and subsequently existence for problem (3.4). In fact, we can give a more general result.

Theorem 3.1. *Let $\mathcal{X}$ be the dual space of some Banach space $\mathcal{Z}$. Let $p \in \partial J(0) \subset \mathcal{Z} \subset \mathcal{X}^*$ be such that there exists w with*

$$J^*\left(\frac{p - A^*w}{\tau}\right) = 0$$

for some $\tau < 1$. Moreover, let the mapping $u \mapsto \|Au\| + J(u)$ be coercive on $\mathcal{X}$, and let J be the convex conjugate of a proper functional on the predual space $\mathcal{Z}$.
Then there exists a minimizer of

$$\min_{u \in \mathcal{X}} \|Au - f\|_{\mathcal{Y}} \ \text{ s.t. } \ J(u) - \langle p, u\rangle = 0.$$

Proof. Let $\mathcal{A} = \{u \in \mathcal{X} \mid J(u) - \langle p, u\rangle = 0\}$ be the admissible set. Since $0 \in \mathcal{A}$ we can look for a minimizer in the sublevel set

$$S = \big\{u \in \mathcal{A} \mid \|Au - f\|_{\mathcal{Y}} \leq \|f\|_{\mathcal{Y}}\big\}.$$

By the triangle inequality we have $\|Au\|_{\mathcal{Y}} \leq 2\|f\|_{\mathcal{Y}}$ on S, hence $u \mapsto Au$ is bounded. From the definition of the convex conjugate we know that for all $u \in \mathcal{X}, y \in \mathcal{Y}$ we have

$$\langle y, u\rangle \leq J^*(y) + J(u).$$

Hence for $u \in S$ we find

$$\begin{aligned} J(u) &= \langle p, u\rangle \\ &= \langle p - A^*w, u\rangle + \langle w, Au\rangle \\ &\leq \langle \frac{p - A^*w}{\tau}, \tau u\rangle + \|w\|_{\mathcal{Y}}\|Au\|_{\mathcal{Y}} \\ &\leq J^*\left(\frac{p - A^*w}{\tau}\right) + J(\tau u) + \|w\|_{\mathcal{Y}}\|Au\|_{\mathcal{Y}} \end{aligned}$$

which implies by the one-homogeneity of J that

$$J(u) \leq \frac{\|w\|_{\mathcal{Y}}\ \|Au\|_{\mathcal{Y}}}{1 - \tau}.$$

Thus we obtain the boundedness of $u \mapsto \|Au\|_{\mathcal{Y}} + J(u)$ on S. With the coercivity we obtain that S is bounded, hence the Banach-Alaoglu theorem implies that S is precompact in the weak-* topology.

It remains to show that the data fidelity and the constraint are weak-* lower semicontinuous (l.s.c.). Since norms are weak-* l.s.c. and A is linear and bounded we immediately deduce it for the objective functional. J is weak-* l.s.c. as a convex conjugate of a proper functional on $\mathcal{Z}$, and since $p \in \mathcal{Z}$ we obtain weak-* lower semicontinuity of the constraint.

Those arguments together directly yield the existence of a minimizer. □ □

Note that the assumptions always hold for $p = p_\alpha$ obtained from (3.2) with $w = \frac{1}{\alpha}(f - Au_\alpha)$ and τ arbitrarily small, hence we conclude the existence of a minimizer $\hat{u}_\alpha$ of (3.4). The assumption $p \in \mathcal{Z}$ is a typical regularity assumption on the operator A, namely the range $\mathcal{R}(A^*) \subset \mathcal{Z}$ [8].

The situation for (3.3) is less clear, since there is no similar way to obtain coercivity. As we shall see in Section 4, (3.3) consists in minimizing a quadratic functional over a linear subspace, which immediately implies the existence of $\hat{u}_\alpha$ if $\mathcal{X}$ has finite dimensions. In an infinite-dimensional setting we cannot provide an existence result in general, since there is neither a particular reason for the subspace to be closed nor for the quadratic functional to be coercive (in ill-posed problems we typically have an operator A with nonclosed range).

3.3 Optimal debiasing on singular vectors

In the following we work out the behavior of the debiasing method on singular vectors [2], which represent the extension of the concept of classical singular value decomposition to nonlinear regularization functionals. According to [2], $u^\lambda \in \mathcal{X}$ is a singular vector if for some $\lambda > 0$

$$\lambda A^* A u^\lambda \in \partial J(u^\lambda)$$

holds. Without going too much into detail, singular vectors can be considered as generalized "eigenfunctions" of the regularization functional J. As such, they describe a class of exact solutions to problem (3.1) in the following sense:

Let us consider a multiple cu^λ of such a singular vector for $c > \lambda\alpha$. According to [2], the solution u_α of the variational problem (3.1) for data $f = cAu^\lambda$ is given by

$$u_\alpha = (c - \alpha\lambda)u^\lambda,$$

and the subgradient from the optimality condition is

$$p_\alpha = \lambda A^* A u^\lambda \in \partial J(u_\alpha).$$

Hence u_α recovers cu^λ up to a (known) scalar factor $\alpha\lambda$ and shares a subgradient with u^λ. This means that the variational method leaves the singular vector basically untouched, which allows for its exact recovery. Intuitively, the quantity $-\lambda\alpha u^\lambda$ hence represents the bias of the variational method in this case, which should be removed by our debiasing method (3.4). And indeed we obtain $\hat{u}_\alpha = cu^\lambda$ as a minimizer of (3.4), since

$$\|A\hat{u}_\alpha - f\|_{\mathcal{Y}} = \|A(\hat{u}_\alpha - cu^\lambda)\|_{\mathcal{Y}} = 0$$

and since $\hat{u}_\alpha$ lies in the admissible set due to the shared subgradient. If A has trivial nullspace, $\hat{u}_\alpha$ is even unique. Hence, the debiasing strategy leads to the exact reconstruction of the solution and corrects the bias $-\lambda\alpha u^\lambda$. Note that this is indeed an important result, since if the debiasing method failed for singular vectors it would be doubtful whether the method is reliable in general.

Since the infimal convolution of Bregman distances is nonnegative and less or equal than one of the Bregman distances, it also vanishes at $\hat{u}_\alpha = cu^\lambda$. In particular

$$\begin{aligned}\mathrm{ICB}_J^{p_\alpha}(cu^\lambda, u_\alpha) &\leq D_J^{p_\alpha}(cu^\lambda, u_\alpha)\\ &= J(cu^\lambda) - \langle p_\alpha, cu^\lambda\rangle = 0.\end{aligned}$$

Consequently, $\hat{u}_\alpha$ is also a solution of (3.3).

4 Bias and Model Manifolds

In the following we provide a more fundamental discussion of bias and decompositions obtained by debiasing methods. An obvious point to start is the definition of bias, which is indeed not always coherent in the imaging literature with the one in statistics.

4.1 Definitions of bias

We first recall the classical definition of bias in statistics. Let f be a realization of a random variable modeling a random noise perturbation of clean data $f^* = Au^*$, such that $\mathbb{E}[f] = f^*$. If we consider a general canonical estimator $\hat{U}(f)$, the standard definition of bias in this setup is given by

$$\begin{aligned}\mathbb{B}^{\mathrm{stat}}(\hat{U}) &= \mathbb{E}[u^* - \hat{U}(f)]\\ &= u^* - \mathbb{E}[\hat{U}(f)].\end{aligned} \tag{4.1}$$

Unfortunately, this bias is hard to manipulate for nonlinear estimators. Hence, we consider a deterministic definition of bias, which relies on the clean data f^*:

$$\begin{aligned}\mathbb{B}^*(\hat{U}) &= \mathbb{E}[u^* - \hat{U}(f^*)] = u^* - \hat{U}(f^*)\\ &= u^* - \hat{U}(\mathbb{E}[f]).\end{aligned} \tag{4.2}$$

We immediately note the equivalence of the two definitions in the case of linear estimators, but our computational experiments do not show a significant difference between $\mathbb{B}^{\mathrm{stat}}$ and $\mathbb{B}^*$ even for highly nonlinear variational methods. In general, the purpose of debiasing is to reduce the quantitative bias B^d, i.e. here the error between u^* and $\hat{U}(f^*)$ in an appropriate distance measure d:

$$B^d(\hat{U}(f^*)) = d(\hat{U}(f^*), u^*).$$

Let us consider the specific estimator $u_\alpha(f^*)$, i.e. the solution of problem (3.1) with clean data f^*. As already argued in Section 2, it suffers from a certain bias due to the chosen regularization. Following [14], this bias can be decomposed into two parts. The first part is related to the regularization itself, and it occurs if the assumption made by the regularization does not match the true object that we seek to recover. For example, trying to recover a piecewise linear object using TV regularization leads to the staircasing effect due to the assumption of a piecewise constant solution. This part of the bias is not avoidable since it is inherent to the regularization, and it is referred to as *model bias*. In particular, we cannot hope to correct it.

However, even if the regularity assumption fits, the solution still suffers from a systematic error due to the weight on the regularization. For TV regularization for example, this is the loss of contrast observed in Section 2. This remaining part is referred to as *method bias*, and this is the part that we aim to correct. As we see in the remainder of the section, the estimator $u_\alpha(f^*)$ provides the necessary information to correct this bias. Deledalle et al. [14] define an appropriate linear model subspace related to that estimator, on which the debiasing takes place. It allows to define the model bias as the difference between u^* and its projection onto the model subspace. The remaining part of the difference between the reconstructed solution and u^* is then the method bias. In the following we reintroduce the notion of model subspaces provided by [14] and further generalize it

to the variational setting in infinite dimensions. The latter may imply the nonclosedness of the model subspace and hence nonexistence of the projection of u^* onto it. Moreover, it seems apparent that in some nonlinear situations it might be more suitable to consider a model manifold instead of a linear space and we hence generalize the definition in this direction.

Let us first assume that we are already given an appropriate model manifold.

Definition 4.1. Let $\mathcal{M}_{f^*}$ be a given model manifold and $d\colon \mathcal{X}\times\mathcal{X} \to [0,\infty)$ a distance measure. An estimator $\hat{U}(f^*)$ of u^* is a *debiasing* of $u_\alpha(f^*)$ if $\hat{U}(f^*) \in \mathcal{M}_{f^*}$ and

$$d(\hat{U}(f^*), u^*) < d(u_\alpha(f^*), u^*).$$

If there exists a minimizer

$$\hat{u}_\alpha(f^*) \in \arg\min_{v\in\mathcal{M}_{f^*}} d(v, u^*), \tag{4.3}$$

we call it an *optimal debiasing*. In any case, we define the *magnitude of the model bias* as

$$B^d_{\mathrm{mod}}(\mathcal{M}_{f^*}) = \inf_{v\in\mathcal{M}_{f^*}} d(v, u^*).$$

Obviously the model bias only depends on the model manifold and for a given $u_\alpha(f^*)$ it is hence, as already indicated, a fixed quantity that we cannot manipulate. Instead we want to perform the debiasing on the manifold only, so we consider another bias for elements of $\mathcal{M}_{f^*}$ only.

Definition 4.2. Given an optimal debiasing $\hat{u}_\alpha(f^*)$ on $\mathcal{M}_{f^*}$, we define the *magnitude of the method bias* of $v \in \mathcal{M}_{f^*}$ as

$$B^d_{\mathrm{meth}}(v) = d(v, \hat{u}_\alpha(f^*)).$$

The optimal debiasing $\hat{u}_\alpha(f^*)$ obviously does not suffer from method bias. Note that if the minimizer in (4.3) does not exist, which can happen in particular in ill-posed problems in infinite dimensions, then the magnitude of the method bias is not well-defined or has to be set to $+\infty$.

With these definitions at hand, we now aim to compute an optimal debiasing, i.e. the solution of (4.3). The remaining questions are how to choose an appropriate model manifold $\mathcal{M}_{f^*}$ and the distance measure d. We start with the latter.

An easy choice for the distance measure d is a squared Hilbert space norm. In finite dimensions, the optimal debiasing $\hat{u}_\alpha(f^*)$ is then the (unique) orthogonal projection of u^* onto $\mathcal{M}_{f^*}$. We even obtain an orthogonal decomposition of the bias B^d of any estimator $v \in \mathcal{M}_{f^*}$ into method and (constant) model bias:

$$\begin{aligned} B^d(v) &= \|v - u^*\|^2 \\ &= \| \underbrace{v - \hat{u}_\alpha(f^*)}_{\text{method bias}} \|^2 + \| \underbrace{\hat{u}_\alpha(f^*) - u^*}_{\text{model bias}} \|^2. \end{aligned}$$

In infinite dimensions, we obtain the same result if the minimizer of (4.3) exists, which is e.g. the case if $\mathcal{M}_{f^*}$ is convex and closed.

Unfortunately, for general inverse problems with an operator we do not know u^* and hence cannot compute its projection onto $\mathcal{M}_{f^*}$. Instead we have access to the data $f^* = Au^*$ (or rather to one noisy realization f of f^* in practice, which we discuss later). In order to make the bias (and the associated debiasing) accessible, we can consider bias through the operator A. Hence the optimal debiasing comes down to computing the minimizer of (4.3) over $A(\mathcal{M}_{f^*})$, i.e.

$$\begin{aligned} \hat{u}_\alpha(f^*) &\in \arg\min_{v\in A(\mathcal{M}_{f^*})} \|u^* - v\|^2 \\ &= \arg\min_{v\in\mathcal{M}_{f^*}} \|Au^* - Av\|^2 \\ &= \arg\min_{v\in\mathcal{M}_{f^*}} \|f^* - Av\|^2. \end{aligned} \tag{4.4}$$

Correspondingly, if such a minimizer $\hat{u}_\alpha(f^*)$ exists, we measure the magnitude of model and method bias in the output space, rather than in image space, i.e.

$$\begin{aligned} B^d_{\mathrm{mod}}(\mathcal{M}_{f^*}) &= \inf_{v\in\mathcal{M}_{f^*}} \|Av - f^*\|^2, \\ B^d_{\mathrm{meth}}(v) &= \|A\hat{u}_\alpha(f^*) - Av\|^2. \end{aligned}$$

We can hence at least guarantee that the optimal debiasing has no method bias in the output space. For denoising problems without any operator (A being the identity), or for A invertible on $\mathcal{M}_{f^*}$ we obtain the equivalence of both approaches. In ill-posed inverse problems it is usually rather problematic to measure errors in the output space, since noise can also be small in that norm. Notice however that we do not use the output space norm on the whole space, but on the rather small model manifold, on which - if chosen appropriately - the structural components dominate. On the latter the output space norm is reasonable.

The main advantage of this formulation is that we are able to compute a minimizer of

(4.4), since it is in fact a constrained least-squares problem with the data fidelity of (3.1). Its solution of course requires a proper choice of the underlying model subspace $\mathcal{M}_{f^*}$, which we discuss in the following.

4.2 Model manifolds

In general, a model subspace can be characterized as the space of possible solutions for the debiasing step following the first solution $u_\alpha(f)$ of the variational problem (3.1). As such it contains the properties of $u_\alpha(f)$ that we want to carry over to the debiased solution. In the context of sparsity-enforcing regularization this is basically a support condition on the debiased solution.

4.2.1 Differential model manifolds

Deledalle et al. [14] use the notion of Fréchet derivative to define their model subspace in a finite-dimensional setting. We naturally generalize this concept using the directional derivative instead, and further extend it to infinite dimensions. The following definitions can e.g. be found in [31].

Definition 4.3. Let $\mathcal{V}$ and $\mathcal{W}$ be Banach spaces. A mapping $F\colon \mathcal{V} \to \mathcal{W}$ is called Fréchet differentiable at $x \in \mathcal{V}$ if there exists a linear and bounded operator $\mathrm{D}F(x;\cdot)\colon \mathcal{V} \to \mathcal{W}$ such that

$$\lim_{\|g\|_\mathcal{V}\to 0} \frac{\|F(x+g) - F(x) + \mathrm{D}F(x;g)\|_\mathcal{W}}{\|g\|_\mathcal{V}} = 0.$$

Definition 4.4. A mapping $F\colon \mathcal{V} \to \mathcal{W}$ is called directionally differentiable in the sense of Gâteaux at $x \in \mathcal{V}$ if the limit

$$\mathrm{d}F(x;g) := \lim_{t\to 0^+} \frac{F(x+tg) - F(x)}{t}$$

exits for all $g \in \mathcal{V}$.

We can immediately deduce from the definition that, if the directional derivative $\mathrm{d}F(x;\cdot)$ exits, it is positively one-homogeneous in g, i.e.

$$\mathrm{d}F(x;\lambda g) = \lambda \mathrm{d}F(x;g)$$

for all $\lambda \geq 0$ and $g \in \mathcal{V}$. If it is linear in g, we call F Gâteaux differentiable at x.

Provided a unique and Fréchet differentiable map $f \mapsto u_\alpha(f)$, Deledalle et al. [14] introduce the tangent affine subspace

$$\mathcal{M}_f^\mathrm{F} = \big\{ u_\alpha(f) + \mathrm{D}u_\alpha(f;g) \mid g \in \mathcal{Y} \big\},$$

where $\mathrm{D}u_\alpha(f;\cdot)\colon \mathcal{Y} \to \mathcal{X}$ is the Fréchet derivative of $u_\alpha(f)$ at f. To be less restrictive, the easiest generalization of $\mathcal{M}_f^\mathrm{F}$ is to consider the directional derivative.

Definition 4.5. If the map $f \mapsto u_\alpha(f)$ is directionally differentiable with derivative $\mathrm{d}u_\alpha(f;\cdot)$, we define

$$\mathcal{M}_f^\mathrm{G} = \big\{ u_\alpha(f) + \mathrm{d}u_\alpha(f;g) \mid g \in \mathcal{Y} \big\}.$$

Note that if the map is Fréchet differentiable, $\mathcal{M}_f^\mathrm{G}$ is a linear space and coincides with the model subspace $\mathcal{M}_f^\mathrm{F}$.

We now derive a few illustrative examples that we use throughout the remainder of the paper. In order to keep it as simple as possible, the easiest transition from the finite-dimensional vector space setting to infinite dimensions are the ℓ^p-spaces of p-summable sequences:

Definition 4.6. For $1 \leq p < \infty$ we define the spaces ℓ^p of p-summable sequences with values in $\mathbb{R}^d$ by

$$\ell^p(\mathbb{R}^d) = \big\{ (x_i)_{i\in\mathbb{N}}, x_i \in \mathbb{R}^d : \sum_{i\in\mathbb{N}} |x_i|^p < \infty \big\},$$

where $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^d$. For $p = \infty$ we define

$$\ell^\infty(\mathbb{R}^d) = \big\{ (x_i)_{i\in\mathbb{N}}, x_i \in \mathbb{R}^d : \sup_{i\in\mathbb{N}} |x_i| < \infty \big\}.$$

It is easy to show that $\ell^p(\mathbb{R}^d) \subset \ell^q(\mathbb{R}^d)$ for $1 \leq p \leq q \leq \infty$. In particular for $d = 1$ we denote by ℓ^1, ℓ^2 and ℓ^∞ the spaces of summable, square-summable and bounded scalar-valued sequences.

Example 4.7. *Anisotropic shrinkage.* Let $f \in \ell^2$ be a square-summable sequence. The solution of

$$u_\alpha(f) \in \arg\min_{u\in\ell^1} \frac{1}{2}\|u - f\|_{\ell^2}^2 + \alpha\|u\|_{\ell^1} \quad (4.5)$$

for $\alpha > 0$ is given by

$$[u_\alpha(f)]_i = \begin{cases} f_i - \alpha \operatorname{sign}(f_i), & |f_i| \geq \alpha, \\ 0, & |f_i| < \alpha. \end{cases}$$

Its support is limited to where $|f_i|$ is above the threshold α. The directional derivative $\mathrm{d}u_\alpha(f;g)$ of $u_\alpha(f)$ into the direction $g \in \ell^2$ is

given by

$$[\mathrm{d}u_\alpha(f;g)]_i = \begin{cases} g_i, & |f_i| > \alpha \\ 0, & |f_i| < \alpha \\ g_i, & |f_i| = \alpha, \operatorname{sign}(f_i) = \operatorname{sign}(g_i) \\ 0, & |f_i| = \alpha, \operatorname{sign}(f_i) \neq \operatorname{sign}(g_i). \end{cases}$$

Proof. See Appendix 8.1. □

First, if we exclude the case $|f_i| = \alpha$, the directional derivative is linear, hence it is a Gâteaux derivative. In fact it is even an infinite-dimensional Fréchet derivative, and the resulting model subspace coincides with the model subspace defined in finite dimensions in [14]:

$$\mathcal{M}_f^{\mathrm{F}} = \{u \in \ell^2 \mid \operatorname{supp}(u) \subset \operatorname{supp}(u_\alpha(f))\}.$$

The model subspace carries over information about the support of the first solution $u_\alpha(f)$. Note that $\mathcal{M}_f^{\mathrm{F}}$ contains all elements of ℓ^2 which share the same support as $u_\alpha(f)$, but as well allows for zeros where $u_\alpha(f) \neq 0$. In that sense $u_\alpha(f)$ defines the maximal support of all $u \in \mathcal{M}_f^{\mathrm{F}}$. If we allow $|f_i|$ to be equal to α, we obtain a larger subspace which allows for support changes in the direction of f_i on the threshold:

$$u \in \mathcal{M}_f^{\mathrm{G}} \Leftrightarrow u_i = \begin{cases} \lambda \in \mathbb{R}, & |f_i| > \alpha, \\ 0, & |f_i| < \alpha, \\ \lambda \geq 0, & f_i = \alpha, \\ \lambda \leq 0, & f_i = -\alpha. \end{cases}$$

Note that the case $|f_i| > \alpha$ reveals a remaining shortcoming of the definition via the directional derivative, e.g. if $f_i > \alpha$ it is counter-intuitive to allow negative elements in $\mathcal{M}_f^{\mathrm{G}}$, while this is not the case for $f_i = \alpha$. The main reason appears to be the strong deviation of the linearization in such directions from the actual values of $[u_\alpha(f)]_i$, which is not controlled by the definition. However, minimizing the data term over $\mathcal{M}_f^{\mathrm{G}}$ for the debiasing in Eq. (4.4) forces the changes to have the right sign and the debiased solution $\hat{u}_\alpha(f)$ corresponds to hard-thresholding:

$$[\hat{u}_\alpha(f)]_i = \begin{cases} f_i, & |f_i| \geq \alpha, \\ 0, & |f_i| < \alpha. \end{cases}$$

Note that we as well maintain the signal directly on the threshold.

We obtain analogous results for isotropic shrinkage, i.e. if $f \in \ell^2(\mathbb{R}^d)$ for $d > 1$. Since the computation of the derivative requires a little more work, we provide the results in Appendix 8.2. A more interesting example is the model subspace related to anisotropic ℓ^1-regularized general linear inverse problems.

Example 4.8. *Anisotropic ℓ^1-regularization.* Let $A\colon \ell^1 \to \ell^2$ be a linear and bounded operator and $f \in \ell^2$. Consider the solution $u_\alpha(f)$ of the ℓ^1-regularized problem

$$u_\alpha(f) \in \arg\min_{u \in \ell^1} \frac{1}{2}\|Au - f\|_{\ell^2}^2 + \alpha\|u\|_{\ell^1}, \quad (4.6)$$

where we assume that the solution is unique for data in a neighborhood of f. Computing the directional derivative directly, is a more tedious task in this case, but computing the model subspace $\mathcal{M}_f^{\mathrm{G}}$ is actually easier via a slight detour.

Let $u_\alpha(f)$ be the solution for data f and $u_\alpha(\tilde{f})$ the solution for data $\tilde{f}$. The two subgradients from the optimality conditions

$$0 = A^*(Au_\alpha(f) - f) + \alpha p_\alpha, \;\; p_\alpha \in \partial\|u_\alpha(f)\|_{\ell^1}$$
$$0 = A^*(Au_\alpha(\tilde{f}) - \tilde{f}) + \alpha \tilde{p}_\alpha, \;\; \tilde{p}_\alpha \in \partial\|u_\alpha(\tilde{f})\|_{\ell^1}$$

satisfy a standard error estimate [9]

$$\|p_\alpha - \tilde{p}_\alpha\|_{\ell^\infty} \leq \frac{C}{\alpha}\|f - \tilde{f}\|_{\ell^2}.$$

In view of the subdifferential of the ℓ^1-norm,

$$\partial\|u\|_{\ell^1} = \{p \in \ell^\infty : \|p\|_{\ell^\infty} \leq 1, \\ p_i = \operatorname{sign}(u_i) \text{ for } u_i \neq 0\},$$

we have to consider several cases. If $[u_\alpha(f)]_i = 0$ and $|(p_\alpha)_i| < 1$ we derive from

$$\begin{aligned} |(\tilde{p}_\alpha)_i| &\leq |(\tilde{p}_\alpha)_i - (p_\alpha)_i| + |(p_\alpha)_i| \\ &\leq \frac{C}{\alpha}\|f - \tilde{f}\|_{\ell^2} + |(p_\alpha)_i|, \end{aligned}$$

that if $\|f - \tilde{f}\|_{\ell^2}$ is sufficiently small, then $|(\tilde{p}_\alpha)_i| < 1$. Hence $[u_\alpha(\tilde{f})]_i = 0$, and the derivative related to the perturbed data $\tilde{f}$ vanishes. In case $[u_\alpha(f)]_i = 0$ and $(p_\alpha)_i = 1$, by a similar argument $(\tilde{p}_\alpha)_i \neq -1$ and thus $[u_\alpha(\tilde{f})]_i \geq 0$ and $[\mathrm{d}u_\alpha(f;g)]_i \geq 0$. Analogously, $[\mathrm{d}u_\alpha(f;g)]_i \leq 0$ if $[u_\alpha(f)]_i = 0$ and $(p_\alpha)_i = -1$. If $[u_\alpha(f)]_i \neq 0$, the directional derivative is an arbitrary real number depending on the data perturbation.

Summing up we now know that every directional derivative is an element $v \in \ell^1$ fulfilling

$$v_i = \begin{cases} 0, & |(p_\alpha)_i| < 1, \\ \lambda \geq 0, & (u_\alpha)_i = 0, (p_\alpha)_i = 1, \\ \lambda \leq 0, & (u_\alpha)_i = 0, (p_\alpha)_i = -1. \end{cases} \tag{4.7}$$

On the other hand, for v satisfying (4.7), let $\tilde{u} = u_\alpha(f) + tv$ for $t > 0$ sufficiently small. Then p_α is a subgradient of $\tilde{u}$. Indeed, for example if $(u_\alpha)_i = 0$ and $(p_\alpha)_i = 1$, then $v_i \geq 0$, so $\tilde{u} \geq 0$, and hence $(p_\alpha)_i$ fulfills the requirement of a subgradient of $\tilde{u}$. The other cases follow analogously. Then from the optimality condition of $u_\alpha(f)$ we get:

$$\begin{aligned} &A^*(Au_\alpha(f) - f)) + \alpha p_\alpha = 0 \\ \Leftrightarrow\ &A^*(A(\underbrace{u_\alpha(f) + tv}_{\tilde{u}}) - (f + tAv)) + \alpha p_\alpha = 0. \end{aligned}$$

We then deduce that $\tilde{u}$ is a minimizer of problem (4.6) with data $\tilde{f} = f + tAv$. Hence, there exists a data perturbation such that v is the directional derivative of $u_\alpha(f)$. Putting these arguments together we now know that $u \in \mathcal{M}_f^{\mathrm{G}}$ if and only if

$$u_i = \begin{cases} \lambda \in \mathbb{R}, & [u_\alpha(f)]_i \neq 0, \\ 0, & [u_\alpha(f)]_i = 0, |(p_\alpha)_i| < 1, \\ \lambda \geq 0, & [u_\alpha(f)]_i = 0, (p_\alpha)_i = 1, \\ \lambda \leq 0, & [u_\alpha(f)]_i = 0, (p_\alpha)_i = -1. \end{cases}$$

It is not surprising that $\mathcal{M}_f^{\mathrm{G}}$ has a similar structure as the model manifold for the anisotropic shrinkage in Example 4.7. It allows for arbitrary changes on the support of $u_\alpha(f)$ and permits only zero values if $[u_\alpha(f)]_i = 0$ and $|(p_\alpha)_i| < 1$. If we exclude the case where $|(p_\alpha)_i| = 1$ even though $[u_\alpha(f)]_i$ vanishes, debiasing on $\mathcal{M}_f^{\mathrm{G}}$ effectively means solving a least-squares problem with a support constraint on the solution. But we again find an odd case where changes are allowed outside of the support of the initial solution $u_\alpha(f)$. It occurs when $|(p_\alpha)_i| = 1$ even though $[u_\alpha(f)]_i$ vanishes, which seems to be the indefinite case. However, it has been argued in [23] that a subgradient equal to ± 1 is a good indicator of support, hence it is reasonable to trust the subgradient in that case.

4.2.2 Variational model manifolds

As we have shown so far, the appropriate use of a derivative can yield suitable spaces for the debiasing. However, for already supposedly easy problems such as the latter example the explicit computation of such spaces or of the derivatives can be difficult or impossible. And even if it is possible, there remains the question of how to effectively solve the debiasing on those spaces, both theoretically and numerically.

On the other hand, the latter example implies that a subgradient of the first solution rather than the solution itself can provide the necessary information for the debiasing. This naturally leads us to the idea of Bregman distances in order to use the subgradient in a variational debiasing method. And indeed we show that the associated spaces are closely related, and that they link the concept of model manifolds to the already presented debiasing method from Section 3. Furthermore, this does not only provide a theoretical framework but also numerical solutions to perform debiasing in practice, even for more challenging problems.

In the following we introduce related manifolds motivated by the variational problem itself. The optimality condition of the variational problem (3.1) defines a unique map $f \mapsto p_\alpha \in \partial J(u_\alpha)$, which allows us to consider the following spaces. We drop the dependence of u_α on f for the sake of clarity.

Definition 4.9. For $p_\alpha \in \partial J(u_\alpha)$ defined by (3.2) we define

$$\begin{aligned} \mathcal{M}_f^{\mathrm{B}} &= \left\{ u \in \mathcal{X} \mid D_J^{p_\alpha}(u, u_\alpha) = 0 \right\}, \\ \mathcal{M}_f^{\mathrm{IC}} &= \left\{ u \in \mathcal{X} \mid \mathrm{ICB}_J^{p_\alpha}(u, u_\alpha) = 0 \right\}. \end{aligned}$$

In order to assess the idea of the above spaces, we first revisit the anisotropic shrinkage problem of Example 4.7.

Example 4.10. *Anisotropic shrinkage.* The optimality condition of problem (4.5) yields the subgradient

$$(p_\alpha)_i = \frac{f_i - (u_\alpha)_i}{\alpha} = \begin{cases} \operatorname{sign}(f_i), & |f_i| \geq \alpha, \\ \frac{f_i}{\alpha}, & |f_i| < \alpha, \end{cases} \tag{4.8}$$

and for $J = \|\cdot\|_{\ell^1}$ the Bregman distance takes the following form:

$$\begin{aligned} D_{\ell^1}^{p_\alpha}(u, u_\alpha) &= \|u\|_{\ell^1} - \langle p_\alpha, u \rangle \\ &= \sum_{i \in \mathbb{N}} |u_i| - (p_\alpha)_i u_i \\ &= \sum_{i \in \mathbb{N}} (\operatorname{sign}(u_i) - (p_\alpha)_i) u_i. \end{aligned}$$

A zero Bregman distance thus means that either $u_i = 0$ or $\operatorname{sign}(u_i) = (p_\alpha)_i$. Having a closer look at the subgradient (4.8), we observe that if $|f_i| < \alpha$, then $|(p_\alpha)_i| < 1$. Hence the latter condition cannot be fulfilled, so in this case u_i has to be zero. We can thus characterize the model manifold related to a zero Bregman distance as:

$$u \in \mathcal{M}_f^{\mathrm{B}} \Leftrightarrow u_i = \begin{cases} \lambda \operatorname{sign}(f_i), \lambda \geq 0, & |f_i| \geq \alpha, \\ 0, & |f_i| < \alpha. \end{cases}$$

As for $\mathcal{M}_f^{\mathrm{G}}$, the model manifold $\mathcal{M}_f^{\mathrm{B}}$ fixes the maximum support to where $|f_i| \geq \alpha$. However, $\mathcal{M}_f^{\mathrm{B}}$ only allows for values on the support sharing the same sign as f_i (respectively $(u_\alpha)_i$).

By adapting the proof of [23], we obtain a similar result for the infimal convolution of Bregman distances, without the restriction on the sign:

$$\begin{aligned} \mathrm{ICB}_{\ell^1}^{p_\alpha}(u, u_\alpha) &= [D_{\ell^1}^{p_\alpha}(\cdot, u_\alpha) \square D_{\ell^1}^{-p_\alpha}(\cdot, -u_\alpha)](u) \\ &= \sum_{i \in \mathbb{N}} (1 - |(p_\alpha)_i|)|u_i|. \end{aligned}$$

For this infimal convolution to be zero we need either $u_i = 0$ or $|(p_\alpha)_i| = 1$. By the structure of the subgradient p_α we thus find

$$u \in \mathcal{M}_f^{\mathrm{IC}} \Leftrightarrow u_i = \begin{cases} \lambda \in \mathbb{R}, & |f_i| \geq \alpha, \\ 0, & |f_i| < \alpha. \end{cases}$$

Hence we observe the following connection between the subspaces:

$$\mathcal{M}_f^{\mathrm{B}} \subset \mathcal{M}_f^{\mathrm{G}} \subset \mathcal{M}_f^{\mathrm{IC}}.$$

Note that the space $\mathcal{M}_f^{\mathrm{G}}$ related to the directional derivative seems to be the odd space of the three. While allowing for arbitrary sign for $|f| > \alpha$, it only allows for changes in the direction of f directly on the threshold. In that sense, $\mathcal{M}_f^{\mathrm{B}}$ and $\mathcal{M}_f^{\mathrm{IC}}$ seem to be more suitable in order to either include or exclude the sign-constraint. A second glance at the manifolds reveals that $\mathcal{M}_f^{\mathrm{IC}}$ is a linear space, as we further elaborate on in the next subsection. In this case it is actually even the span of $\mathcal{M}_f^{\mathrm{B}}$, which is however not true in general. This can e.g. be seen from the next example of isotropic TV-type regularization.

Example 4.11. *Isotropic TV-type regularization* Let $A \colon \ell^2(\mathbb{R}^n) \to \ell^2(\mathbb{R}^d)$ and $\Gamma \colon \ell^2(\mathbb{R}^n) \to \ell^1(\mathbb{R}^m)$ be linear and bounded operators and $J(u) = \|\Gamma u\|_{\ell^1(\mathbb{R}^m)}$ for $d, m, n \in \mathbb{N}$. We aim to find the variational model manifolds for the debiasing of the solution

$$u_\alpha \in \operatorname*{arg\,min}_{u \in \ell^2(\mathbb{R}^n)} \frac{1}{2}\|Au - f\|_{\ell^2(\mathbb{R}^d)} + \alpha \|\Gamma u\|_{\ell^1(\mathbb{R}^m)}.$$

Given the (unique) subgradient $p_\alpha \in \partial J(u_\alpha)$ from the optimality condition, the chain rule for subdifferentials [16, p. 27] implies the existence of a $q_\alpha \in \partial \|\cdot\|_{\ell^1(\mathbb{R}^m)}(\Gamma u_\alpha)$ such that $p_\alpha = \Gamma^* q_\alpha$ and

$$D_J^{p_\alpha}(u, u_\alpha) = D_{\ell^1(\mathbb{R}^m)}^{q_\alpha}(\Gamma u, \Gamma u_\alpha).$$

If we denote the angle between $(\Gamma u)_i$ and $(q_\alpha)_i$ by φ_i, the Bregman distance reads:

$$\begin{aligned} D_J^{p_\alpha}(u, u_\alpha) &= D_{\ell^1(\mathbb{R}^m)}^{q_\alpha}(\Gamma u, \Gamma u_\alpha) \\ &= \sum_{i \in \mathbb{N}} |(\Gamma u)_i| - (q_\alpha)_i \cdot (\Gamma u)_i \\ &= \sum_{i \in \mathbb{N}} |(\Gamma u)_i| \big(1 - \cos(\varphi_i) |(q_\alpha)_i|\big) \end{aligned}$$

For a zero Bregman distance we can distinguish two cases: If $|(q_\alpha)_i| < 1$, then $(\Gamma u)_i$ has to be zero. If $|(q_\alpha)_i| = 1$, then either $(\Gamma u)_i = 0$ or $\cos(\varphi_i) = 1$, hence $(\Gamma u)_i = \lambda (q_\alpha)_i$ for $\lambda \geq 0$. Hence the model manifold $\mathcal{M}_f^{\mathrm{B}}$ is given by

$$u \in \mathcal{M}_f^{\mathrm{B}} \Leftrightarrow (\Gamma u)_i = \begin{cases} \lambda (q_\alpha)_i, \lambda \geq 0, & |(q_\alpha)_i| = 1, \\ 0, & |(q_\alpha)_i| < 1. \end{cases}$$

In particular, if $(\Gamma u_\alpha)_i \neq 0$, then by the structure of the $\ell^1(\mathbb{R}^m)$-subdifferential we know that $(q_\alpha)_i = \frac{(\Gamma u_\alpha)_i}{|(\Gamma u_\alpha)_i|}$ and thus $(\Gamma u)_i = \mu (\Gamma u_\alpha)_i$ for some $\mu \geq 0$. So provided that $|(q_\alpha)_i| < 1$ whenever $(\Gamma u_\alpha)_i = 0$ we find

$$u \in \mathcal{M}_f^{\mathrm{B}} \Leftrightarrow (\Gamma u)_i = \begin{cases} \mu (\Gamma u_\alpha)_i, \mu \geq 0, & (\Gamma u_\alpha)_i \neq 0, \\ 0, & (\Gamma u_\alpha)_i = 0. \end{cases}$$

Performing the debiasing on the latter space hence means minimizing the data term with a support and direction constraint on the gradient of the solution. This in particular allows to restore the loss of contrast which we have observed for TV regularization in Section 2. Note that the condition $|(q_\alpha)_i| < 1 \Leftrightarrow (\Gamma u_\alpha)_i = 0$ excludes the odd case where the subgradient seems to contain more information than the first solution, as already seen in Example 4.8.

In the above illustration of the model manifold, the debiasing seems to rely on the choice of q_α, which is obviously *not* unique. However, in practice we still use the unique subgradient p_α from the optimality condition which avoids the issue of the choice of a "good" q_α.

The computation of $\mathcal{M}_f^{\mathrm{IC}}$ is a little more difficult in this case, since we cannot access an explicit representation of the functional $\mathrm{ICB}_J^{p_\alpha}(\cdot, u_\alpha)$. However, since

$$\mathrm{ICB}_{\ell^1(\mathbb{R}^m)}^{q_\alpha}(\Gamma u, \Gamma u_\alpha) \leq \mathrm{ICB}_J^{p_\alpha}(u, u_\alpha)$$

(see Appendix 8.3, Thm. 8.1), we can instead use the infimal convolution of two $\ell^1(\mathbb{R}^m)$-Bregman distances to illustrate the model manifold. We have (Appendix 8.3, Thm. 8.2)

$$\mathrm{ICB}_{\ell^1(\mathbb{R}^m)}^{q_\alpha}(\Gamma u, \Gamma u_\alpha) = \sum_{i\in\mathbb{N}} G((\Gamma u)_i, (q_\alpha)_i)$$

with $G\colon \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ defined as

$$G((\Gamma u)_i, (q_\alpha)_i) = \begin{cases} |(\Gamma u)_i|(1 - |\cos(\varphi_i)||(q_\alpha)_i|), \\ \qquad\qquad \text{if } |(q_\alpha)_i| < |\cos(\varphi_i)|, \\ |(\Gamma u)_i||\sin(\varphi_i)|\sqrt{1 - |(q_\alpha)_i|^2}, \\ \qquad\qquad \text{if } |(q_\alpha)_i| \geq |\cos(\varphi_i)|. \end{cases}$$

For G to be zero we once again distinguish two situations. If $|(q_\alpha)_i| < 1$, the first case of G can only vanish if $(\Gamma u)_i = 0$. In the second case, since $1 > |(q_\alpha)_i| \geq |\cos(\varphi_i)|$, we infer $\varphi_i \notin \{0, \pi\}$, and hence neither the sinus nor the square root can vanish. This means once again that $(\Gamma u)_i = 0$. If $|(q_\alpha)_i| = 1$ we can only be in the second case, which vanishes independently of $(\Gamma u)_i$. Thus $(\Gamma u)_i$ can be arbitrary. Putting the arguments together, we find

$$u \in \mathcal{M}_f^{\mathrm{IC}} \Rightarrow \mathrm{ICB}_{\ell^1(\mathbb{R}^m)}^{q_\alpha}(\Gamma u, \Gamma u_\alpha) = 0$$
$$\Leftrightarrow (\Gamma u)_i = \begin{cases} \lambda \in \mathbb{R}^m, & |(q_\alpha)_i| = 1, \\ 0, & |(q_\alpha)_i| < 1. \end{cases}$$

This is indeed not the span of $\mathcal{M}_f^{\mathrm{B}}$, but it instead allows for arbitrary elements if $|(q_\alpha)_i| = 1$. From this example, we cannot immediately state that $\mathcal{M}_f^{\mathrm{B}} \subset \mathcal{M}_f^{\mathrm{IC}}$, because so far we only know that $\mathcal{M}_f^{\mathrm{B}}$ as well as $\mathcal{M}_f^{\mathrm{IC}}$ are subsets of the set $\{u \in \mathcal{X} \mid \mathrm{ICB}_{\ell^1(\mathbb{R}^m)}^{q_\alpha}(\Gamma u, \Gamma u_\alpha) = 0\}$. However, in the next subsection we see that $\mathcal{M}_f^{\mathrm{B}} \subset \mathcal{M}_f^{\mathrm{IC}}$ is indeed true and it is actually a general property of the variational model manifolds.

Note that we gain the same support condition on the gradient as for $\mathcal{M}_f^{\mathrm{B}}$, but allow for arbitrary gradient directions on the support, which intuitively does not seem restrictive enough. However, in practice for the debiasing the direction is not arbitrary, but the data term decides, so we can expect a similar result for debiasing in $\mathcal{M}_f^{\mathrm{B}}$ and $\mathcal{M}_f^{\mathrm{IC}}$. Indeed the numerical studies in Section 6 confirm these expectations.

4.3 Properties of variational model manifolds

In the following we discuss some properties of the variational subspaces $\mathcal{M}_f^{\mathrm{B}}$ and $\mathcal{M}_f^{\mathrm{IC}}$. All results are general and do not depend on the particular choice of a subgradient, so we drop the dependence on f in the notation of the subspaces. Let $v \in \mathcal{X}$ and $p \in \partial J(v)$. We start with a result on the structure of $\mathcal{M}^{\mathrm{B}}$:

Theorem 4.12. *The set*

$$\mathcal{M}^{\mathrm{B}} = \{u \in \mathcal{X} \mid D_J^p(u, v) = 0\}$$

is a nonempty convex cone.

Proof. The map $u \mapsto D_J^p(u, v)$ is convex and nonnegative, hence

$$\{u \mid D_J^p(u, v) = 0\} = \{u \mid D_J^p(u, v) \leq 0\}$$

is convex as a sublevel set of a convex functional. Moreover, for each $c \geq 0$ we have

$$D_J^p(cu, v) = c\, D_J^p(u, v),$$

i.e. if u is an element of the set, then every positive multiple cu is an element, too. Hence it is a convex cone. Since $D_J^p(v, v) = 0$ it is not empty. □ □

The structure of $\mathcal{M}^{\mathrm{IC}}$ is even simpler; as announced in a special example above it is indeed a linear space:

Theorem 4.13. *The set*

$$\mathcal{M}^{\mathrm{IC}} = \{u \in \mathcal{X} | [D_J^p(\cdot, v) \square D_J^{-p}(\cdot, -v)](u) = 0\}$$

is a nonempty linear subspace of $\mathcal{X}$.

Proof. By analogous arguments as above we deduce the convexity and since

$$\mathrm{ICB}_J^p(0, v) = \inf_{\phi+\psi=0} D_J^p(\phi, v) + D_J^{-p}(\psi, -v)$$
$$\leq D_J^p(v, v) + D_J^{-p}(-v, -v) = 0$$

the set is not empty. For arbitrary $c \in \mathbb{R} \setminus \{0\}$ we have

$$\begin{aligned}&\mathrm{ICB}_J^p(cu, v)\\ &= \inf_z \{J(cu - z) + J(z) - \langle p, cu - 2z\rangle\}\\ &= |c| \inf_w \{J(u - w) + J(w) - \langle p, u - 2w\rangle\},\end{aligned}$$

where we use the one-to-one transform $z = cw$ for $c > 0$ and $z = c(u - w)$ for $c < 0$. This implies that $\mathrm{ICB}_J^p(cu, v) = 0$ if $\mathrm{ICB}_J^p(u, v) = 0$. Hence $\mathcal{M}^{\mathrm{IC}}$ is a linear subspace. □ □

As one may expect from the fact that the infimal convolution is a weaker distance than the original Bregman distance, we obtain an immediate inclusion between the corresponding manifolds:

Lemma 4.14. $\mathcal{M}^{\mathrm{B}} \subset \mathcal{M}^{\mathrm{IC}}$.

Proof. Let $u \in \mathcal{M}^{\mathrm{B}}$, i.e. $D_J^p(u, v) = 0$. For $c \geq 0$ we have

$$D_J^{-p}(-cu, -v) = c\, D_J^p(u, v).$$

Thus we deduce

$$\begin{aligned}\mathrm{ICB}_J^p(u, v) &= \inf_{\phi+\psi=u} D_J^p(\phi, v) + D_J^{-p}(\psi, -v)\\ &\leq D_J^p(2u, v) + D_J^{-p}(-u, -v)\\ &= 2D_J^p(u, v) + D_J^p(u, v) = 0.\end{aligned}$$

The assertion follows by the nonnegativity of the maps $u \mapsto D_J^p(u, v)$ and $u \mapsto D_J^{-p}(u, -v)$. Note that for $p \neq 0$ the subset is proper, since e.g. $-v \in \mathcal{M}^{\mathrm{IC}}$ but $-v \notin \mathcal{M}^{\mathrm{B}}$. □ □

We finally elaborate on the importance of absolute one-homogeneity of J for our approach (respectively also other debiasing approaches as in [14]), such that the subdifferential can be multivalued. Otherwise the model manifolds may just contain a single element and debiasing in this manifold cannot produce any other solution. This is e.g. the case for a strictly convex functional.

Lemma 4.15. *Let J be strictly convex. Then $\mathcal{M}^{\mathrm{B}}$ and $\mathcal{M}^{\mathrm{IC}}$ are singletons.*

Proof. For strictly convex J the mapping $u \mapsto D_J^p(u, v)$ is strictly convex as well, hence $D_J^p(u, v) = 0$ if and only if $u = v$ and $\mathcal{M}^{\mathrm{B}} = \{v\}$. The infimal convolution preserves strict convexity [1, p. 170] and since

$$\mathrm{ICB}_J^p(0, v) \leq D_J^p(v, v) + D_J^{-p}(-v, -v) = 0$$

we find $\mathcal{M}^{\mathrm{IC}} = \{0\}$. □ □

However, one can easily exclude this case since our assumption of J being one-homogeneous guarantees that it is not strictly convex.

4.4 Bias-variance estimates

Another justification for the deterministic definition of bias as well as our choice for the distance measure in Section 4.1 can be found in the variational model itself. In order to derive quantitative bounds for bias and variance in a variational model, we start with the Tikhonov regularization (Ridge regression) model related to the functional $J(u) = \frac{1}{2}\|u\|_{\mathcal{X}}^2$. The optimality condition for this problem is given by

$$A^*(Au_\alpha(f) - f) + \alpha u_\alpha(f) = 0.$$

We easily see that there exists $w_\alpha = \frac{1}{\alpha}(f - Au_\alpha(f))$ such that $u_\alpha(f) = A^* w_\alpha$ and

$$Au_\alpha(f) - Au^* + \alpha w_\alpha = f - Au^*.$$

Now let us assume that a source condition $u^* \in \mathrm{Im}[A^*]$ holds, i.e. $u^* = A^* w^*$ for some w^*. In this case we can subtract αw^* on both sides and take a squared norm to arrive at

$$\begin{aligned}&\|Au_\alpha(f) - Au^*\|_{\mathcal{Y}}^2 + \alpha^2\|w_\alpha - w^*\|_{\mathcal{Y}}^2\\ &\qquad + 2\alpha\langle Au_\alpha(f) - Au^*, w_\alpha - w^*\rangle\\ &= \|f - Au^*\|_{\mathcal{Y}}^2 + \alpha^2\|w^*\|_{\mathcal{Y}}^2 - 2\alpha\langle f - Au^*, w^*\rangle.\end{aligned}$$

Now taking the expectation on both sides and using $\mathbb{E}[f] = f^* = Au^*$ we find

$$\begin{aligned}&\mathbb{E}[\|Au_\alpha(f) - Au^*\|_{\mathcal{Y}}^2] + \alpha^2\mathbb{E}[\|w_\alpha - w^*\|_{\mathcal{Y}}^2]\\ &\qquad + 2\alpha\mathbb{E}[\|u_\alpha(f) - u^*\|_{\mathcal{X}}^2]\\ &= \mathbb{E}[\|f - Au^*\|_{\mathcal{Y}}^2] + \alpha^2\|w^*\|_{\mathcal{Y}}^2. \qquad (4.9)\end{aligned}$$

The left-hand side is the sum of three error terms for the solution measured in different norms: in the output space, the space of the source element, and the original space used for regularization. All of them can be decomposed in a bias and a variance term, e.g.

$$\begin{aligned}&\mathbb{E}[\|u_\alpha(f) - u^*\|_{\mathcal{X}}^2]\\ =&\|\mathbb{E}[u_\alpha(f)] - u^*\|_{\mathcal{X}}^2 + \mathbb{E}[\|u_\alpha(f) - \mathbb{E}[u_\alpha(f)]\|_{\mathcal{X}}^2].\end{aligned}$$

The term $\mathbb{E}[\|f - Au^*\|_{\mathcal{Y}}^2]$ in (4.9) is exactly the variance in the data. As a consequence $\alpha\|w^*\|_{\mathcal{Y}}$ measures the bias in this case. Note that in particular for zero variance we obtain a direct estimate of the bias via $\alpha\|w^*\|_{\mathcal{Y}}$.

In the case of the variational model (3.1) this can be generalized using recent approaches [4, 7, 9, 27] using the source condition $A^*w^* \in \partial J(u^*)$. Now completely analogous computations as above yield

$$\begin{aligned}&\mathbb{E}[\|Au_\alpha(f) - Au^*\|_{\mathcal{Y}}^2] + \alpha^2\mathbb{E}[\|w_\alpha - w^*\|_{\mathcal{Y}}^2] \\ &\qquad + 2\alpha\mathbb{E}[D_J^{\mathrm{sym}}(u_\alpha(f), u^*)] \\ &= \mathbb{E}[\|f - Au^*\|_{\mathcal{Y}}^2] + \alpha^2\|w^*\|_{\mathcal{Y}}^2,\end{aligned}$$

with the only difference that we now use the symmetric Bregman distance

$$D_J^{\mathrm{sym}}(u_\alpha(f), u^*) = \langle A^*w_\alpha - A^*w^*, u_\alpha(f) - u^*\rangle,$$

with $A^*w_\alpha \in \partial J(u_\alpha(f))$. The bias-variance decomposition on the right-hand side remains the same. In the noiseless case it is then natural to consider this (here, deterministic) estimate as a measure of bias:

$$\begin{aligned}&\|Au_\alpha(f^*) - Au^*\|_{\mathcal{Y}}^2 + \alpha^2\|w_\alpha - w^*\|_{\mathcal{Y}}^2 \\ &\qquad + 2\alpha D_J^{\mathrm{sym}}(u_\alpha(f^*), u^*) \\ &= \alpha^2\|w^*\|_{\mathcal{Y}}^2,\end{aligned}$$

Here, as already discussed in Section 4.1, we again consider a difference between the exact solution u^* and the estimator for $\mathbb{E}[f] = f^*$, i.e. the expectation of the noise, rather than the expectations of the estimators $u_\alpha(f)$ over all realizations of f (which coincide if J is quadratic). We observe that there are three natural distances to quantify the error and thus also the bias: a quadratic one in the output space and a predual space (related to w), and the symmetric Bregman distance related to the functional J. The first term $\|Au_\alpha(f^*) - Au^*\|_{\mathcal{Y}}^2$ is exactly the one we use as a measure of bias. The second term $\alpha^2\mathbb{E}[\|w_\alpha - w^*\|_{\mathcal{Y}}^2]$ is constant on the model subspace $\mathcal{M}_{f^*}^{\mathrm{B}}$, since by definition of the subspace $p_\alpha = A^*w_\alpha$ is a subgradient of all the elements in $\mathcal{M}_{f^*}^{\mathrm{B}}$. The third term $D_J^{\mathrm{sym}}(u_\alpha(f^*), u^*)$ is not easy to control; if the subspace is appropriate, meaning that $p_\alpha \in \partial J(u^*)$, then the symmetric Bregman distance vanishes for every element in $\mathcal{M}_{f^*}^{\mathrm{B}}$. In any other case, we do not have access to a subgradient $p^* \in \partial J(u^*)$, so we cannot control the Bregman distance for any element of the subspace. Hence, with our method we minimize the part of the bias that we can actually control. In fact, if the model subspace is right, we even minimize the whole bias.

4.5 Back to the proposed method

To sum up, the debiasing method we have introduced in Equations (3.3) and (3.4) comes down to debiasing over $\mathcal{M}_{f^*}^{\mathrm{IC}}$ and $\mathcal{M}_{f^*}^{\mathrm{B}}$, respectively, while the results of Section 3 guarantee the existence of the optimal debiasing $\hat{u}_\alpha(f^*)$ at least on $\mathcal{M}_{f^*}^{\mathrm{B}}$.

However in practice, we do not have access to the clean data f^*, but often only to one noisy realization f, which makes the regularization in (3.1) necessary in the first place. Instead of the true model subspace $\mathcal{M}_{f^*}$, we hence use an approximation $\mathcal{M}_f$ computed from the noisy data f to perform the debiasing of the reconstruction $u_\alpha(f)$ for noisy data. The following experiments show that $\mathcal{M}_f$ is a good approximation of $\mathcal{M}_{f^*}$ in terms of the resulting bias and bias reduction. They also relate the different definitions of bias we have considered. In particular, we distinguish between the statistical bias of Equation (4.1) which is the expectation over several noisy realizations f and the deterministic bias we define in Equation (4.2), which instead considers the outcome given the noiseless data f^*.

Figure 4 displays the TV denoising and debiasing (using the Bregman distance subspace) results obtained with noisy data f (first row) or clean data f^* (second row) with the same regularization parameter $\alpha = 0.3$. We have performed the experiments for either the cartoon *Giraffe* image, or the natural *Parrot* image[2]. First, for the Giraffe image we observe that the TV denoised solution $u_\alpha(f^*)$ for clean data suffers from a heavy loss of contrast, i.e. from method bias. The debiased solution $\hat{u}_\alpha(f^*)$ however is again close to the original data f^*. This shows that if the noiseless data is well represented by the choice of regularization (and hence $\mathcal{M}_{f^*}$), i.e. if there is no or little model bias, the debiasing procedure allows to recover the original signal almost perfectly. On the other hand, the same experiments on the natural *Parrot* image show the problem of model bias since the choice of regularization does not entirely match the data f^*. The debiasing allows to recover the lost contrast, but even the result for noiseless data still suffers from bias, i.e. the loss of small structures, which is model bias in that case.

Besides, if α is big enough to effectively remove noise during the denoising step, then the TV solutions $u_\alpha(f)$ and $u_\alpha(f^*)$ are close to each

[2] `http://r0k.us/graphics/kodak/`

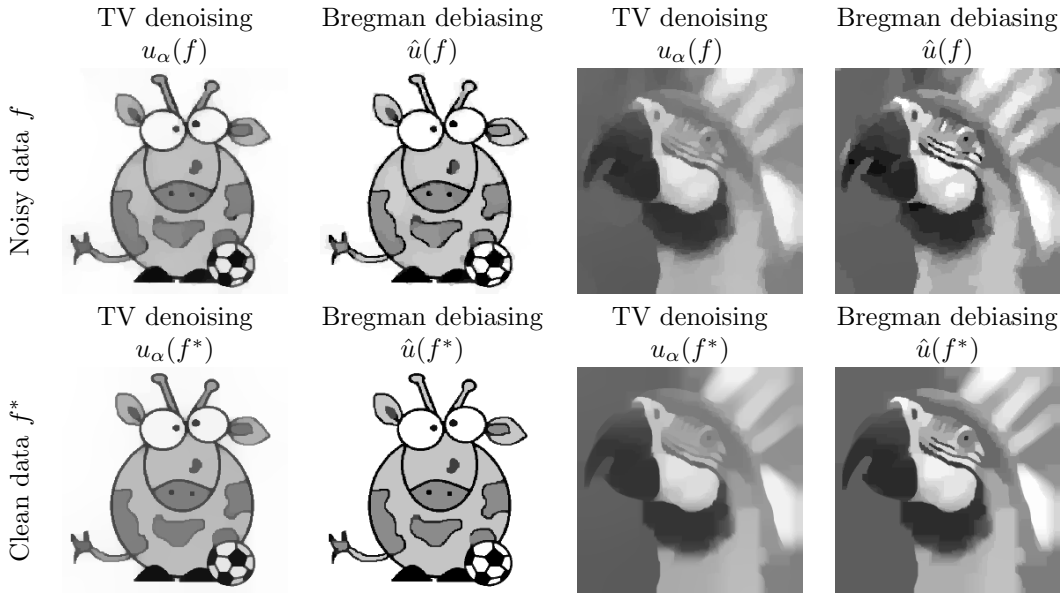


Figure 4: TV denoising and debiasing of the Giraffe and the Parrot images for either noisy data f or clean data f^*, with the same regularization parameter $\alpha = 0.3$.

other. This leads to comparable model subspaces and hence debiased solutions, which confirms that $\mathcal{M}_f$ is indeed a good approximation to $\mathcal{M}_{f^*}$.

Furthermore, we can assess the bias for both subspaces. On $\mathcal{M}_{f^*}$ we can only use the deterministic definition (4.2) of bias whereas on $\mathcal{M}_f$ we use the statistical definition (4.1). Figures 5 and 6 show the bias estimation on the *Giraffe* cartoon image and the natural *Parrot* image. The first row shows the estimations of the statistical bias $\mathbb{B}^{\text{stat}}$ for the two estimators $u_\alpha(f)$ and $\hat{u}_\alpha(f)$ for noisy data f. In the second row the bias $\mathbb{B}^*$ for the two estimators $u_\alpha(f^*)$ and $\hat{u}_\alpha(f^*)$ for clean data f^* is displayed. This deterministic bias can also be decomposed into the associated model and method bias, whereas such a decomposition has not been defined for the statistical bias. The overall deterministic bias $\mathbb{B}^*(u_\alpha(f^*)) = u^* - u_\alpha(f^*)$ for TV denoising appears to be really close to the statistical bias on noisy data in the first row. The same applies for the bias of the debiased solutions in the second column. This confirms that the estimation of the model subspace that we perform with noisy data is indeed a good approximation to the ideal model subspace for clean data, and that the resulting statistical and deterministic bias are closely related.

Besides, the difference $u^* - \hat{u}_\alpha(f^*)$ in the second row shows the remaining bias after the debiasing step, which is model bias. For the *Giraffe* image, this bias is small because the cartoon image is well approximated in the model subspace associated to TV regularization. The *Parrot* image however suffers from a heavier model bias, for example the loss of the small structures around the eye. Finally, in the third column, the difference $\hat{u}_\alpha(f^*) - u_\alpha(f^*)$ shows the error that has been removed by the debiasing step, which corresponds to the method bias. It is particularly interesting for the *Parrot* image. Here one can see the piecewise constant areas which correspond to the re-establishment of the lost contrast within the piecewise constant model provided by the model subspace.

4.6 Relation to inverse scale space methods

We finally comment on the relation of the debiasing approaches to Bregman iterations respectively inverse scale space methods, which are rather efficiently reducing bias as demonstrated in many examples [25, 30, 20]. The Bregman iteration is iteratively constructed by

$$u^{k+1} \in \arg\min_{u \in \mathcal{X}} \frac{1}{2}\|Au - f\|_{\mathcal{Y}}^2 + \alpha D_J^{p^k}(u, u^k),$$
$$p^{k+1} = p^k + \frac{1}{\alpha} A^*(f - Au^{k+1}) \in \partial J(u^{k+1}).$$

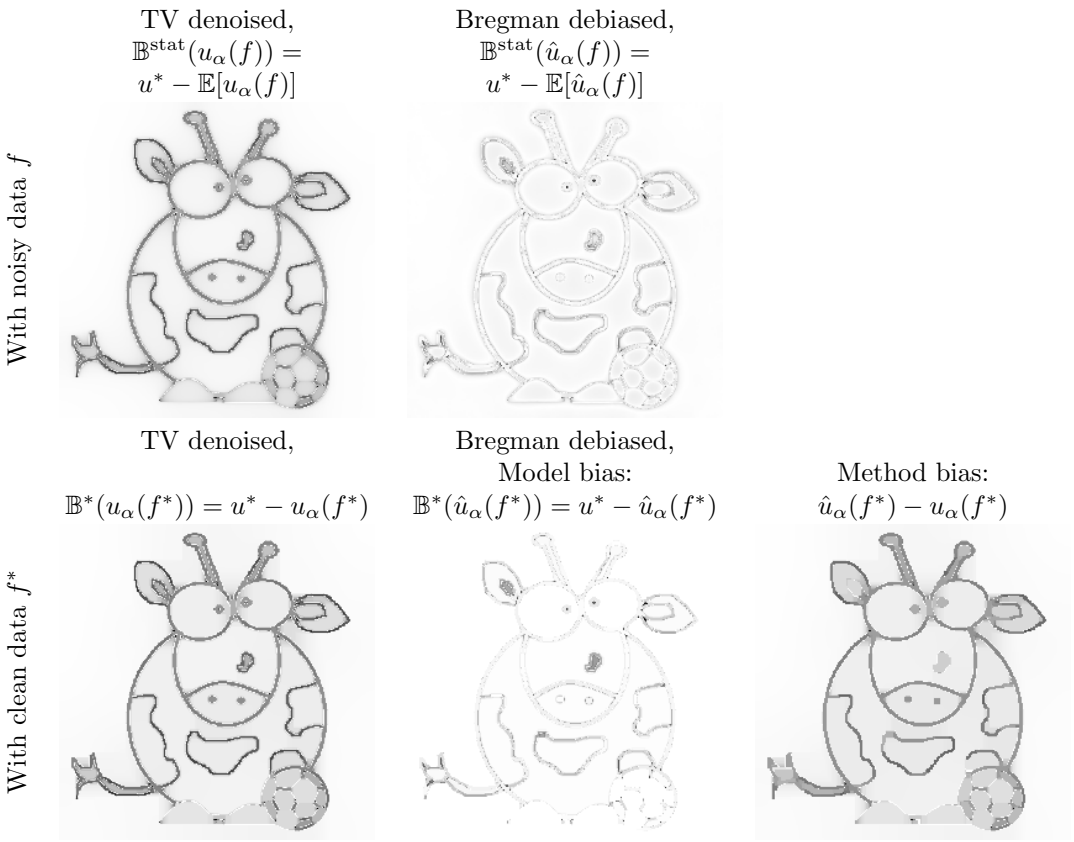


Figure 5: Bias estimation. First row: Statistical bias computed on five hundred noisy realizations of the *Giraffe* cartoon image. Second row: Deterministic bias computed between the clean data and the recovered solution from clean data f^*. In the first column, TV denoising leads to bias. In the second column, the debiasing that has been performed has reduced (or suppressed) the method bias. The remaining (small) model bias is due to the necessary regularization. In the third column, the difference between $\hat{u}_\alpha(f^*)$ and $u_\alpha(f^*)$ shows the bias that has been reduced by the debiasing step, hence the method bias.

In the limit $\alpha \to \infty$ we obtain the time continuous inverse scale-space method, which is the differential inclusion

$$\partial_t p(t) = A^*(f - Au(t)), p(t) \in \partial J(u(t)),$$

with initial values $u(0) = 0$, $p(0) = 0$. A strong relation to our debiasing approach comes from the characterization of the primal solution given $p(t)$ [6, 22, 24]

$$u(t) \in \arg\min_{u \in \mathcal{X}} \|Au - f\|_{\mathcal{Y}}^2 \quad \text{s.t. } p(t) \in \partial J(u(t)).$$

This reconstruction step is exactly the same as the variational debiasing step using the Bregman distance, however with a different preceding construction of the subgradient $p(t)$ (noticing that t corresponds to $\frac{1}{\alpha}$ for the variational method).

From the last observation it becomes apparent that the Bregman debiasing approach with (3.2) and (3.4) is exactly equivalent if the variational method yields the same subgradient as the inverse scale space method, i.e. $p_\alpha = p(\frac{1}{\alpha})$. This can indeed happen, as the results for singular vectors demonstrate [2]. Moreover, in some cases there is full equivalence for arbitrary data, e.g. in a finite-dimensional denoising setting investigated in [5]. It has been shown that for A being the identity and $J(u) = \|\Gamma u\|_1$ with $\Gamma\Gamma^*$ being diagonally dominant the identity $p_\alpha = p(\frac{1}{\alpha})$ holds, which implies that the Bregman debiasing approach and the inverse scale space method yield exactly the same solution. For other cases that do not yield a strict equivalence we include the Bregman iteration for comparison in numerical studies discussed below.

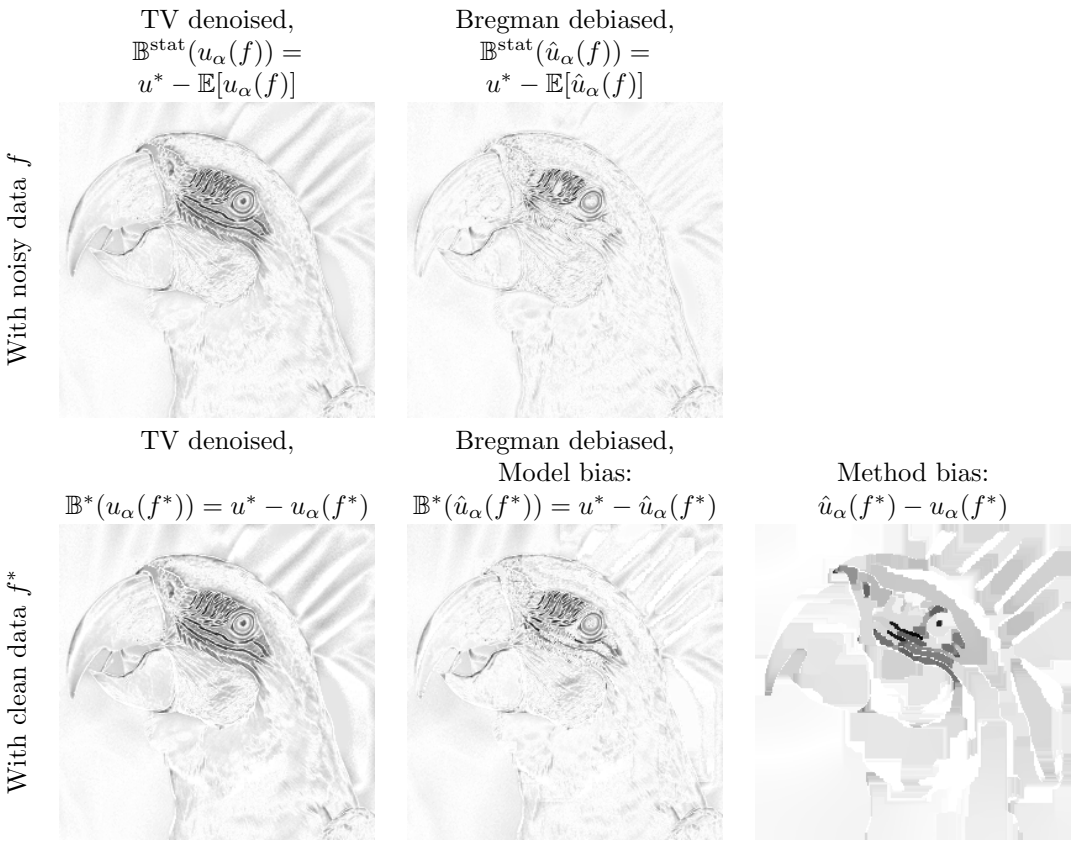


Figure 6: Bias estimation. First row: Statistical bias computed on five hundred noisy realizations of the *Parrot* natural image. Second row: Deterministic bias computed between the clean data and the recovered solution from clean data f^*. On the first column, TV denoising leads to both kinds of bias, model bias and method bias. On the second column, the debiasing that has been performed has reduced (or suppressed) the method bias, and the remaining bias is model bias. On the third column, the difference between $\hat{u}_\alpha(f^*)$ and $u_\alpha(f^*)$ shows the bias that has been reduced by the debiasing step, hence the method bias.

5 Numerical Implementation

In Section 3 we have introduced a two-step-method (cf. Eq. (3.1) – (3.4)) in order to compute a variationally regularized reconstruction with reduced method bias in the sense discussed in Section 4. Its solution requires the minimization of the data fidelity over a subspace defined by a zero Bregman distance or a zero infimal convolution thereof, respectively.

This constraint is difficult to realize numerically, but can be approximated by a rather standard variational problem. We can translate the hard constraint into a soft constraint such that for $\gamma > 0$ the reformulated problems read:

$$\text{a) } \hat{u}_\alpha \in \arg\min_{u\in\mathcal{X}} \frac{1}{2}\|Au - f\|_{\mathcal{Y}}^2 + \gamma D_J^{p_\alpha}(u, u_\alpha),$$

$$\text{b) } \hat{u}_\alpha \in \arg\min_{u\in\mathcal{X}} \frac{1}{2}\|Au - f\|_{\mathcal{Y}}^2 + \gamma \text{ICB}_J^{p_\alpha}(u, u_\alpha).$$

For $\gamma \to \infty$ we obtain the equivalence of the hard and soft constrained formulations. However, for the numerical realization already a moderately large γ is enough to enforce the constraint up to a satisfactory level. For our simulations we chose $\gamma = 1000$, but our tests showed that already for $\gamma \geq 500$ the value of the Bregman distance or its infimal convolution stays numerically zero. Of course the choice of the parameter γ depends on the specific problem we aim to solve and probably has to be adjusted slightly for different image sizes or operators.

Algorithm 1 Primal-Dual Algorithm for Variational Regularization (Step 1)

Input: f, $\alpha > 0$
Initialization: $\sigma, \tau > 0$, $u^0 = \bar{u}^0 = 0$, $y_1^0 = y_2^0 = 0$
  while not converged **do**
    $y_1^{k+1} = \frac{y_1^k + \sigma A u^k - \sigma f}{1+\sigma}$
    $y_2^{k+1} = \Pi_{B_\alpha^\infty}(y_2^k + \sigma \Gamma u^k)$
    $u^{k+1} = u^k - \tau(A^* y_1^{k+1} + \Gamma^* y_2^{k+1})$
    $\bar{u}^{k+1} = 2u^{k+1} - u^k$
  end while
  return $u_\alpha = u^{k+1}$, $p_\alpha = \frac{1}{\alpha} A^*(f - Au_\alpha)$ (c.f. (3.2))

Discretization

For our numerical experiments we choose the setting $\mathcal{X} = \mathbb{R}^n$, $\mathcal{Y} = \mathbb{R}^d$ and $J(u) = \|\Gamma u\|_1$. In general $\Gamma \in \mathbb{R}^{n \times m}$ denotes a discrete linear operator, for the experiments with total variation regularization we choose a discretization of the gradient with forward finite differences. For a general linear forward operator $A \in \mathbb{R}^{n \times d}$ we hence end up with the following discrete optimization problems:

$$
\begin{aligned}
&1. \qquad u_\alpha \in \arg\min_{u \in \mathbb{R}^n} \frac{1}{2}\|Au - f\|_2^2 + \alpha \|\Gamma u\|_1, \\
&2.\,a) \quad \hat{u}_\alpha \in \arg\min_{u \in \mathbb{R}^n} \frac{1}{2}\|Au - f\|_2^2 \\
&\qquad\qquad + \gamma \left(\|\Gamma u\|_1 - \langle p_\alpha, u \rangle \right), \\
&\quad b) \quad \hat{u}_\alpha \in \arg\min_{u \in \mathbb{R}^n} \frac{1}{2}\|Au - f\|_2^2 \\
&\qquad\qquad + \gamma \min_{z \in \mathbb{R}^n} \Big\{ \|\Gamma(u - z)\|_1 - \langle p_\alpha, u - z \rangle \\
&\qquad\qquad + \|\Gamma z\|_1 + \langle p_\alpha, z \rangle \Big\},
\end{aligned}
$$

where we leave out the particular spaces for the primal (and dual) variables for the sake of simplicity in the following. Taking a closer look at these minimization problems, we observe that we can exactly recover the optimization problem in the first step by means of problem 2. a) if we choose $\gamma = \alpha$ and $p_\alpha = 0$. We therefore concentrate on the minimization problems in the second step.

Primal-dual and dual formulation

Using the notion of convex conjugates [28], the corresponding primal-dual and dual formulations of our problems are given by

$$
\begin{aligned}
a) \quad & \min_u \max_{y_1, y_2} \langle y_1, Au \rangle + \langle y_2, \Gamma u \rangle - \gamma \langle p_\alpha, u \rangle \\
& \qquad - \frac{1}{2}\|y_1\|_2^2 - \langle y_1, f \rangle - \iota_{B_\gamma^\infty}(y_2) \\
= & \max_{y_1, y_2} -\frac{1}{2}\|y_1\|_2^2 - \langle y_1, f \rangle - \iota_{B_\gamma^\infty}(y_2) \\
& \qquad - \iota_{\gamma p_\alpha}(A^* y_1 + \Gamma^* y_2), \\
b) \quad & \min_{u,z} \max_{y_1, y_2, y_3} \langle y_1, Au \rangle + \langle y_2, \Gamma u - \Gamma z \rangle \\
& \qquad + \langle y_3, \Gamma z \rangle - \gamma \langle p_\alpha, u \rangle + 2\gamma \langle p_\alpha, z \rangle \\
& \qquad - \frac{1}{2}\|y_1\|_2^2 - \langle y_1, f \rangle \\
& \qquad - \iota_{B_\gamma^\infty}(y_2) - \iota_{B_\gamma^\infty}(y_3) \\
= & \max_{y_1, y_2, y_3} -\frac{1}{2}\|y_1\|_2^2 - \langle y_1, f \rangle \\
& \qquad - \iota_{B_\gamma^\infty}(y_2) - \iota_{B_\gamma^\infty}(y_3) \\
& \qquad - \iota_{\gamma p_\alpha}(A^* y_1 + \Gamma^* y_2) \\
& \qquad - \iota_{-2\gamma p_\alpha}(-\Gamma^* y_2 + \Gamma^* y_3),
\end{aligned}
$$

Solution with a primal-dual algorithm

In order to find a saddle point of the primal-dual formulations, we apply a version of the popular first-order primal-dual algorithms [26, 17, 12]. The basic idea is to perform gradient descent on the primal and gradient ascent on the dual variables. Whenever the involved functionals are not differentiable, here the ℓ^1-norm, this comes down to computing the corresponding proximal mappings. The specific updates needed for our method are summarized in Algorithm 1 for the first regularization problem, and Algorithm 2 and Algorithm 3 for the two different debiasing steps.

We comment on our choice of the stopping criterion. We consider the primal-dual gap of our saddle point problem, which is defined as the difference between the primal and the dual

Algorithm 2 Primal-Dual Algorithm for Bias-Reduction with $\mathcal{M}_f^{\mathrm{B}}$ (Step 2 a))

Input: f, $\gamma > 0$, p_α, which is obtained via Algorithm 1
Initialization: $\sigma, \tau > 0$, $u^0 = \bar{u}^0 = 0$, $y_1^0 = y_2^0 = 0$
  while not converged **do**
    $y_1^{k+1} = \frac{y_1^k + \sigma A u^k - \sigma f}{1+\sigma}$
    $y_2^{k+1} = \Pi_{B_\gamma^\infty}(y_2^k + \sigma \Gamma u^k)$
    $u^{k+1} = u^k - \tau(A^* y_1^{k+1} + \Gamma^* y_2^{k+1} - \gamma p_\alpha)$
    $\bar{u}^{k+1} = 2u^{k+1} - u^k$
  end while
  return $\hat{u}_\alpha = u^{k+1}$

Algorithm 3 Primal-Dual Algorithm for Bias-Reduction with $\mathcal{M}_f^{\mathrm{IC}}$ (Step 2 b))

Input: f, $\gamma > 0$ and p_α, which is obtained via Algorithm 1.
Initialization: $\sigma, \tau > 0$, $u^0 = z^0 = \bar{u}^0 = \bar{z}^0 = 0$, $y_1^0 = y_2^0 = y_3^0 = 0$
  while not converged **do**
    $y_1^{k+1} = \frac{y_1^k + \sigma A u^k - \sigma f}{1+\sigma}$
    $y_2^{k+1} = \Pi_{B_\gamma^\infty}(y_2^k + \sigma \Gamma(u^k - z^k))$
    $y_3^{k+1} = \Pi_{B_\gamma^\infty}(y_3^k + \sigma \Gamma z^k)$
    $u^{k+1} = u^k - \tau(A^* y_1^{k+1} + \Gamma^* y_2^{k+1} - \gamma p_\alpha)$
    $z^{k+1} = z^k - \tau(-\Gamma^* y_2^{k+1} + \Gamma^* y_3^{k+1} + 2\gamma p_\alpha)$
    $\bar{u}^{k+1} = 2u^{k+1} - u^k$
    $\bar{z}^{k+1} = 2z^{k+1} - z^k$
  end while
  return $\hat{u}_\alpha = u^{k+1}$

Parameters	
α	0.3
γ	1000
$\sigma = \tau$	$\frac{1}{\sqrt{8}}$
ϵ_1	10^{-5}
ϵ_2	10^{-6}
ϵ_3	10^{-6}

Table 1: Choice of parameters for a total variation denoising problem of an image of size 256x256 with values in $[0, 1]$, corrupted by Gaussian noise with variance 0.05.

problem for the current values of variables. As in the course of iterations the algorithm is approaching the saddle point, this gap converges to zero. Hence we consider our algorithm converged if this gap is below a certain threshold $\epsilon_1 > 0$. We point out that the indicator functions regarding the ℓ^∞-balls are always zero due to the projection of the dual variables in every update. Since the constraints with respect to the other indicator functions, for example

$$A^* y_1 + \Gamma^* y_2 - \gamma p_\alpha = 0$$

in case a), are hard to satisfy exactly numerically, we instead control that the norm of the left-hand side is smaller than a certain threshold ϵ_2 (respectively ϵ_3 for case b)). All in all we stop the algorithm if the current iterates satisfy:

$$\begin{aligned} \text{a)} \quad PD(u, y_1, y_2) = \big(&- \gamma \langle p_\alpha, u \rangle \\ &+ \frac{1}{2}\|Au - f\|_2^2 + \gamma \|\Gamma u\|_1 \\ &+ \frac{1}{2}\|y_1\|_2^2 + \langle y_1, f \rangle \big)/n < \epsilon_1 \end{aligned}$$

and

$$\|A^* y_1 + \Gamma^* y_2 - \gamma p_\alpha\|_1 / n < \epsilon_2$$

$$\begin{aligned} \text{b)} \quad PD(u, z, y_1, y_2) = \big(&- \gamma \langle p_\alpha, u \rangle + 2\gamma \langle p_\alpha, z \rangle \\ &+ \frac{1}{2}\|Au - f\|_2^2 \\ &+ \gamma \|\Gamma u - \Gamma z\|_1 + \gamma \|\Gamma z\|_1 \\ &+ \frac{1}{2}\|y_1\|_2^2 + \langle y_1, f \rangle \big)/n < \epsilon_1 \end{aligned}$$

and

$$\begin{aligned} &\|A^* y_1 + \Gamma^* y_2 - \gamma p_\alpha\|_1 / n < \epsilon_2, \\ &\| - \Gamma^* y_2 + \Gamma^* y_3 + 2\gamma p_\alpha\|_1 / n < \epsilon_3. \end{aligned}$$

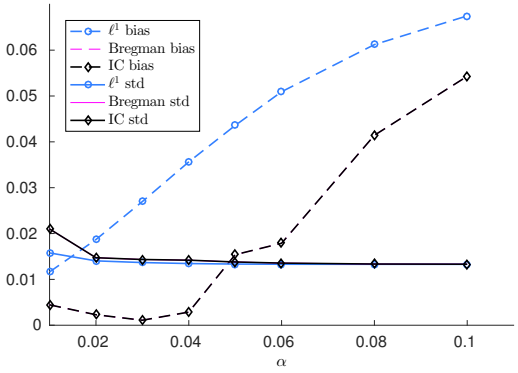


Figure 8: ℓ^1-deconvolution of a 1D signal. Average bias and variance computed over one thousand realizations of the noisy signal for the noisy, restored and debiased signals.

Note that we normalize the primal-dual gap and the constraints by the number of primal pixels n in order to keep the thresholds ϵ_1, ϵ_2 and ϵ_3 independent of varying image resolutions. We give an example for the specific choice of parameters for our total variation denoising problems in Table 5.

6 Numerical Results

This section provides further experiments and numerical results that illustrate the proposed debiasing method.

6.1 ℓ^1-deconvolution

The first application that we illustrate is the deconvolution of a one-dimensional signal using anisotropic shrinkage (4.6). Figure 7 displays the original signal, the blurry signal corrupted by additive Gaussian noise with standard deviation $\sigma = 0.05$, the ℓ^1-reconstructed signal and the debiased signals computed over the Bregman subspace $\mathcal{M}_f^{\mathrm{B}}$ and the infimal convolution subspace $\mathcal{M}_f^{\mathrm{IC}}$. The last two completely overlap on these two plots. One can see that provided that the ℓ^1-reconstruction finds the right peak locations, the debiasing method is able to restore the amplitude of the original signal.

Figure 8 displays the evolution of the average bias of the estimated signals as well as the standard deviation of the error. They were computed over one thousand noisy realizations for the noisy, ℓ^1-reconstructed and debiased signals, as a function of the regularization parameter α. These curves illustrate several behaviors: As expected, the residual variance decreases when the regularization parameter increases. For a very low value of α, the debiasing reintroduces some noise so the average variance is higher than for the ℓ^1-reconstructed signal, revealing the bias-variance trade-off that has to be settled. As α increases, the gap between the variance of the ℓ^1-reconstructed and debiased signal vanishes. On the other hand, the average bias is indeed smaller for the debiased signal than for the ℓ^1-reconstructed signal. Besides, for small values of the regularization parameter the average bias for the debiased signal is stable and close to zero, showing the effective reduction of the method bias. Then it increases by steps which correspond to the progressive vanishing of the peaks, related to model bias. All in all, these plots show the ability of the proposed approach to reduce the method bias (here, the loss of intensity on the peaks), hence allowing for more efficient noise reduction and reconstruction for a wider range of regularization parameters.

6.2 Anisotropic TV denoising

In this subsection we study debiasing by means of the discrete ROF model [29] given by:

$$u_\alpha(f) \in \arg\min_{u\in\mathbb{R}^n} \frac{1}{2}\|u-f\|_2^2 + \alpha\|\Gamma u\|_1, \quad (6.1)$$

where the 1-norm is anisotropic, i.e.

$$\|\Gamma u\|_1 = \sum_{i=1}^{m/2} |(\Gamma u)_{1,i}| + |(\Gamma u)_{2,i}|,$$

with $(\Gamma u)_1$ and $(\Gamma u)_2$ denoting the discrete gradient images in horizontal and vertical direction, respectively. We compare the original denoising result of Problem (6.1) to the proposed debiased solutions obtained with the Bregman subspace $\mathcal{M}_f^{\mathrm{B}}$ or the infimal convolution subspace $\mathcal{M}_f^{\mathrm{IC}}$.

6.2.1 Cartoon image

The *Giraffe* cartoon image has been designed not to have model bias; it is piecewise constant, which makes it suitable for TV denoising and allows us to study the reduction of the method bias only. It takes values in $[0,1]$ and has been artificially corrupted with additive Gaussian noise with zero mean and variance $\sigma^2 = 0.05$, reaching an initial PSNR of

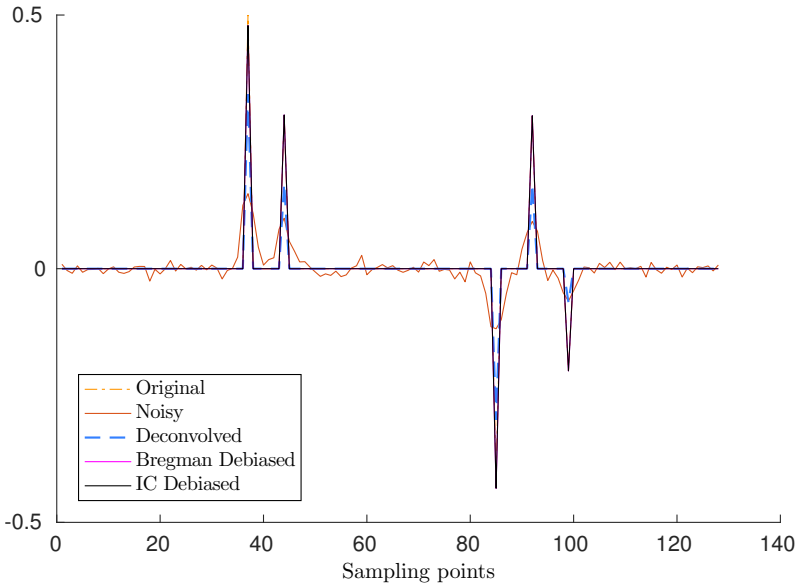


Figure 7: ℓ^1-deconvolution of a 1D signal. Original, noisy convolved, ℓ^1-reconstructed, Bregman debiased and Infimal convolution debiased signals.

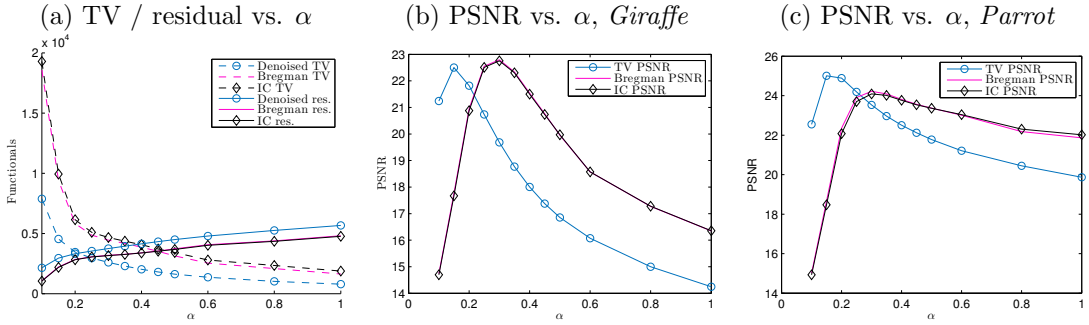


Figure 9: Evolution of (a) The total variation and the residual for the cartoon *Giraffe* image, (b) The average PSNR for TV denoising, Bregman debiasing and infimal convolution debiasing for the cartoon *Giraffe* image and (c) The average PSNR for TV denoising, Bregman debiasing and infimal convolution debiasing for the natural *Parrot* image as a function of the regularization parameter α.

about 13dB. The original image and a noisy realization are already displayed on the first line of Fig. 3 in Section 3.

Figure 10 displays the TV denoising result as well as the debiased solutions computed on the Bregman subspace $\mathcal{M}_f^{\mathrm{B}}$ or the infimal convolution subspace $\mathcal{M}_f^{\mathrm{IC}}$ for different values of the regularization parameter α. On the first line, $\alpha = 0.15$ is the optimal regularization parameter for TV denoising (in terms of PSNR, see Fig. 9-(b)). However, when performing the debiasing, noise is strongly amplified. On the second line, $\alpha = 0.3$ is the optimal regularization parameter for debiasing, and overall, (in terms of PSNR, see Fig. 9-(b)). On the third line $\alpha = 0.6$ leads to an oversmoothed solution, but the debiasing step still allows to recover a lot of the lost contrast.

Since we expect the variational method to systematically underestimate the value of the regularization functional and overestimates the residual (see [2] for a precise computation on singular values), we compare the corresponding quantities when varying α in Figure 9-(a). We observe that for a very large range of values of α there appears to be an almost constant offset between the values for the solution $u_\alpha(f)$ and the debiased solution $\hat{u}_\alpha(f)$ (except for very

small values of α, when noise dominates). This seems to be due to the fact that the debiasing step can correct the bias in the regularization functional (here total variation) and residual to a certain extent. This corresponds well to the plot of PSNR vs. α in Fig. 9-(b), which confirms that the PSNR after the debiasing step is significantly larger than the one in $u_\alpha(f)$ for a large range of values of α, which contains the ones relevant in practice. The fact that the PSNR is decreased by the debiasing step for very small α corresponds to the fact that indeed the noise is amplified in such a case, visible also in the plots for the smallest value of α in Figure 10.

Altogether, these results show that the proposed debiasing approach improves the denoising of the cartoon image both visually and quantitatively.

6.2.2 Natural image

The debiasing can also be evaluated on natural images such as the *Parrot* picture. TV denoising on such images leads to both method bias and model bias. We expect to reduce the former with the proposed method, while the latter is due to the piecewise constant approximation associated with the ROF model. The *Parrot* image takes values in $[0, 1]$ and has been artificially corrupted with additive Gaussian noise with zero mean and variance $\sigma = 0.05$, reaching an initial PSNR of about 13dB. The original image and a noisy realization are displayed on the first line of Figure 13.

Analogously to Figure 10, Figure 13 also displays the TV denoising result as well as the debiased solutions computed on the Bregman subspace or the infimal convolution subspace for different values of the regularization parameter α. On the second line, $\alpha = 0.15$ is the optimal regularization parameter for TV denoising (in terms of PSNR, see Fig. 9-(c)). However, when performing the debiasing, the remaining noise is strongly amplified. On the third line, $\alpha = 0.3$ is the optimal regularization parameter for debiasing (in terms of PSNR, see Fig. 9-(c)). On the fourth line $\alpha = 0.6$ leads to an oversmoothed solution but the debiasing step still allows to recover the lost contrast.

Note that in the *Parrot* case, the optimal result in terms of PSNR is obtained for the TV denoising, for $\alpha = 0.15$. However, the debiasing obtained with $\alpha = 0.3$ visually provides a smoother result on the background, while preserving the fine structures such as the stripes around the eye.

Note also that in each case the artifacts of TV denoising such as staircasing remain and even become more apparent. This however seems natural as the contrast is increased. Since these issues are in fact model bias they are not dealt with by the debiasing method we perform here, but could be reduced by an appropriate choice of regularization such as total generalized variation [3].

6.2.3 Statistical behavior

For both images, the statistical behavior of the proposed debiasing methods can be evaluated by computing the statistical bias $\mathbb{E}[u^* - \hat{U}]$ as well as the variance $\mathbb{V}\mathrm{ar}[u^* - \hat{U}]$ between the true image u^* and an estimator $\hat{U}$. In our case this is either the solution of the ROF-model (6.1) or the corresponding debiased result. Figure 11 displays the evolution of the estimated statistical bias and standard deviation of the TV, Bregman debiased and infimal convolution debiased estimators for the cartoon *Giraffe* and natural *Parrot* images, as a function of the regularization parameter α. These curves reflect some interesting behaviors: As expected, the residual variance decreases as the regularization parameter increases. Besides, the variance is always slightly higher for the debiased solutions, which reflects the bias-variance compromise that has to be settled. However, as the regularization parameter increases, the gap between the denoised and debiased variance decreases. On the other hand, as the regularization parameter grows, the bias increases for each method, and it always remains higher for the denoised solutions than for the debiased solutions. One interesting fact is the behavior of the bias curve for the cartoon *Giraffe* image: for low values of the regularization parameter (up to $\alpha \approx 0.3$), the evolution of the bias for the debiased solutions is relatively stable. This means that for those values, one can increase the regularization parameter in order to reduce the variance without introducing too much (at this point, method) bias. Then, for higher regularization parameters the bias increases in a steeper way, parallel to the evolution of the original bias for the TV denoised image. This reflects the evolution of the model bias from this point on, when the high regularization parameter provides a model subspace whose elements are too smooth compared to the true

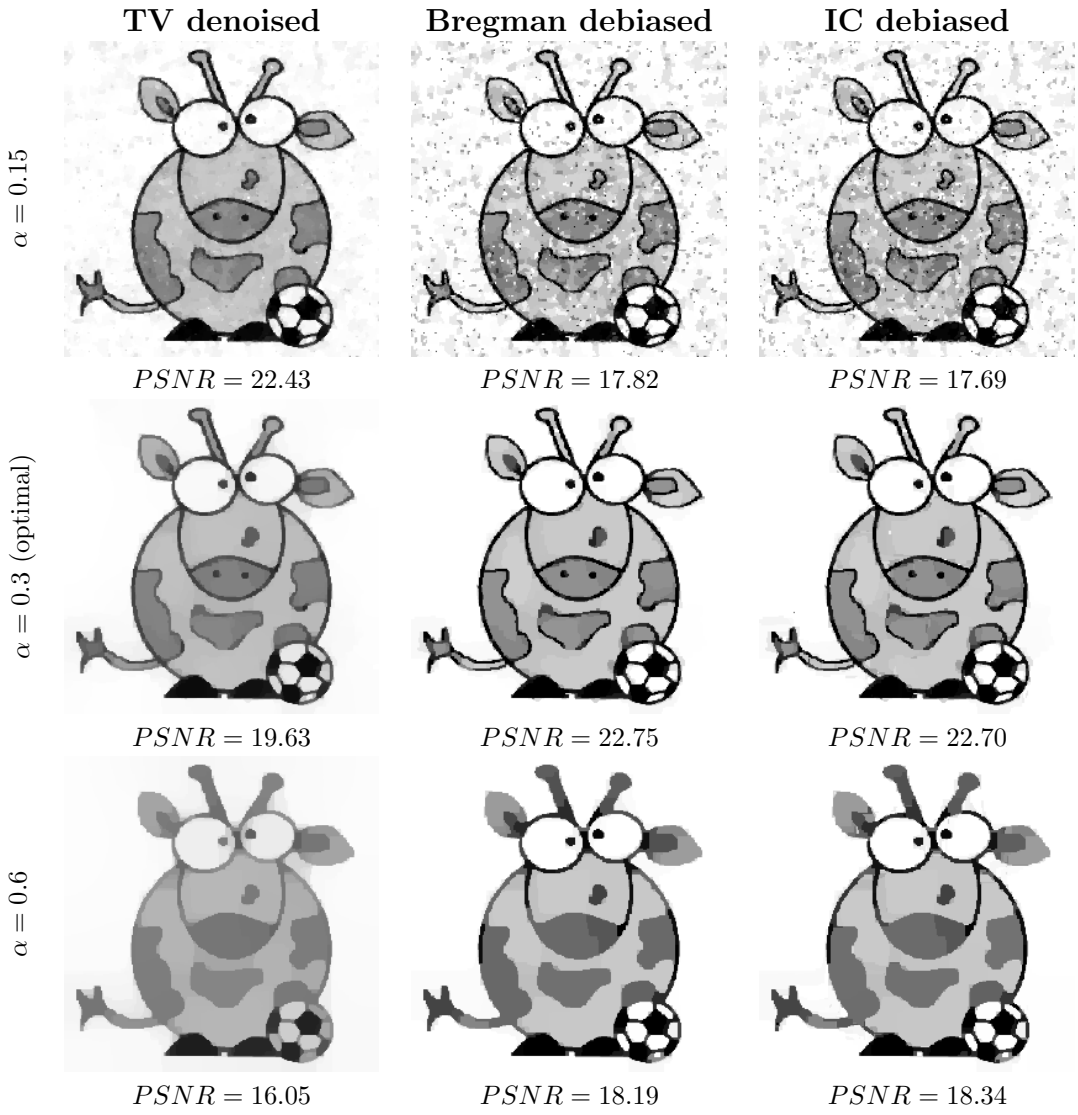


Figure 10: Denoising of the *Giraffe* cartoon image for different values of the regularization parameter α. First column: TV denoising. Second column: Debiasing on the Bregman subspace. Third column: Debiasing on the infimal convolution subspace.

image. For the natural *Parrot* image, the model bias occurs even for small values of the regularization parameter, because the model subspace provided by the TV regularization does not properly fit the image prior.

These curves also illustrate the optimal bias-variance balance that can be achieved with or without the debiasing procedure. Intuitively, one would expect the optimal bias-variance trade-off to be reached when the bias and the standard deviation curves intersect each other. This is indeed confirmed by the PSNR curves from Fig. 9-(b) and 9-(c). Looking at those intersection points on both curves for the TV denoised solution on the one hand and for the debiased solutions on the other hand, one can see that the optimal compromise for the debiasing is reached for a higher regularization parameter than for the denoising. This offers more denoising performance, and it leads to a smaller (for the *Giraffe* image) or equal (for the *Parrot* image) average bias and standard deviation.

6.3 Isotropic TV denoising

Finally, we extend the examples presented in [14] with a few numerical results for isotropic

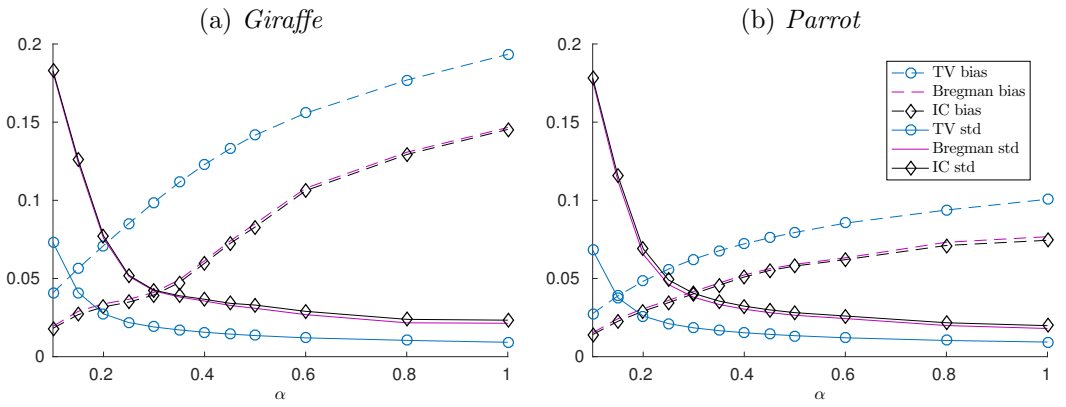


Figure 11: Evolution of the average residual bias and standard deviation computed over 500 noisy realizations of (a) *Giraffe* and (b) *Parrot* for TV denoising, Bregman debiasing and infimal convolution debiasing.

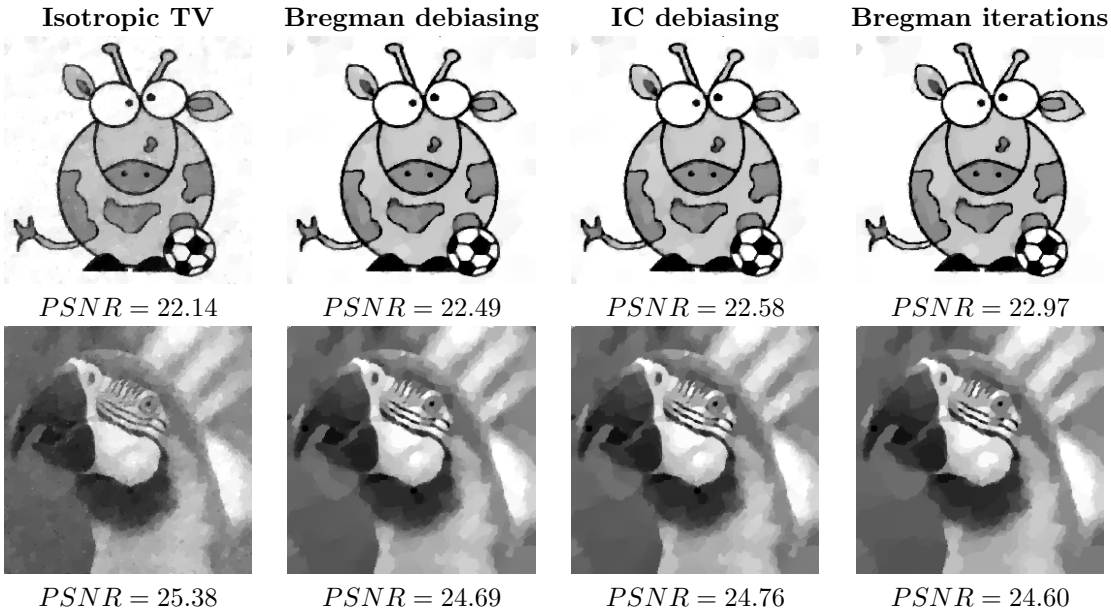


Figure 12: Isotropic TV denoising and debiasing of the cartoon *Giraffe* and natural *Parrot* images, and comparison to Bregman iterations.

TV denoising:

$$\|\Gamma u\|_1 = \sum_{i=1}^{m/2} \sqrt{|(\Gamma u)_{1,i}|^2 + |(\Gamma u)_{2,i}|^2}.$$

We then compare the denoising result to the solutions provided by the two alternative second steps of our debiasing method. Moreover, we also compare them to the result obtained from Bregman iterations. Figure 12 displays the optimal (in terms of PSNR) denoising and debiasing for the *Giraffe* and *Parrot* images. The regularization parameter has been set to $\alpha = 0.2$ for the denoising result and to $\alpha = 0.3$ for the debiasing. Similarly to the anisotropic case, the debiasing both visually and quantitatively improves the quality of the cartoon *Giraffe* image. For the natural *Parrot* image, even though the PSNR is not improved by the debiasing process, one can still observe that the higher regularization parameter offers a better denoising of the background, while the debiasing guarantees that the fine structures around the eye are preserved with a good contrast. Besides, the proposed debiasing approach offers similar results to Bregman iterations, displayed in the fourth column.

However, the interesting aspect of our debiasing approach is that we only apply a two-step procedure, while Bregman iterations have to be performed iteratively with a sufficiently high number of steps.

7 Conclusion

We have introduced two variational debiasing schemes based on Bregman distances and their infimal convolution, which are applicable for nonsmooth convex regularizations and generalize known debiasing approaches for ℓ^1- and TV-type regularization. Based on a recent axiomatic approach to debiasing by Deledalle and coworkers [14], which we further generalized towards infinite-dimensional problems, we were able to provide a theoretical basis of our debiasing approach and work out meaningful model subspaces for variational methods. Moreover, we were able to relate the approach to Bregman iterations and inverse scale space methods.

From the numerical experiments we observe that the debiasing scheme improves the results for a wide range of regularization parameters, which includes the ones providing optimal results. Surprisingly, we often find visually optimal choices of the regularization parameters in the range where bias and standard deviation of the debiased solution are approximately of the same size.

Various questions remain open for future studies: An obvious one is the generalization to other regularization schemes such as total generalized variation (cf. [3]) or nonlocal methods for improved results on natural images. As already indicated in the introduction, the method is theoretically not restricted to squared Hilbert-space norms. Instead, it can be carried out for any suitable data fidelity H and we expect it to improve the results. From a theoretical, and in particular from a statistical viewpoint, the question is then how to relate the method to actual bias reduction, and how to properly motivate and define bias in this setting.

Another further improvement might be achieved by only approximating the model subspace by tuning the parameter γ without letting it tend to infinity.

We acknowledge a very recent and related work on the topic from another perspective, which has been developed in parallel to this work [15]. It will be interesting to investigate the connections in future work.

8 Appendix

We have included some examples and proofs in the Appendix in order not to interrupt the flow of the paper. These are in particular the proof for anisotropic shrinkage and the calculation of the corresponding derivatives in Example 4.7, the same result for isotropic shrinkage and the calculation of the infimal convolution of two ℓ^1-Bregman distances in Example 4.11.

8.1 Proofs of Example 4.7

Anisotropic shrinkage. Let $f \in \ell^2$ be a square-summable sequence. Then the solution of

$$u_\alpha \in \arg\min_{u\in\ell^1} \frac{1}{2}\|u-f\|_{\ell^2}^2 + \alpha\|u\|_{\ell^1} \tag{8.1}$$

for $\alpha > 0$ is given by

$$[u_\alpha(f)]_i = \begin{cases} f_i - \alpha\ \mathrm{sign}(f_i), & |f_i| \geq \alpha \\ 0, & |f_i| < \alpha. \end{cases}$$

Proof. By the dual definition of the ℓ^1-norm we find

$$\begin{aligned} &\min_{u\in\ell^1} \frac{1}{2}\|u-f\|_{\ell^2}^2 + \alpha\|u\|_{\ell^1} \\ &= \min_{u\in\ell^1} \sup_{\substack{r\in\ell^\infty \\ \|r\|_{\ell^\infty}\leq\alpha}} \frac{1}{2}\|u-f\|_{\ell^2}^2 + \langle r, u\rangle \\ &= \sup_{\|r\|_{\ell^\infty}\leq\alpha} \min_{u\in\ell^1} \frac{1}{2}\|u-f\|_{\ell^2}^2 + \langle r, u\rangle. \end{aligned}$$

We can explicitely compute the minimizer for u as $u = f - r$ and hence

$$\begin{aligned} &\sup_{\|r\|_{\ell^\infty}\leq\alpha} \min_{u\in\ell^1} \frac{1}{2}\|u-f\|_{\ell^2}^2 + \langle r, u\rangle \\ &= \sup_{\|r\|_{\ell^\infty}\leq\alpha} -\frac{1}{2}\|r\|_{\ell^2}^2 + \langle r, f\rangle. \end{aligned}$$

This supremum can be computed explicitely pointwise with the corresponding Lagrangian

$$\mathcal{L}(r_i, \lambda) = -\frac{1}{2}r_i^2 + r_i f_i + \lambda(|r_i|^2 - \alpha^2)$$

with $\lambda \leq 0$. The optimality condition yields

$$f_i - r_i + 2\lambda r_i = 0$$

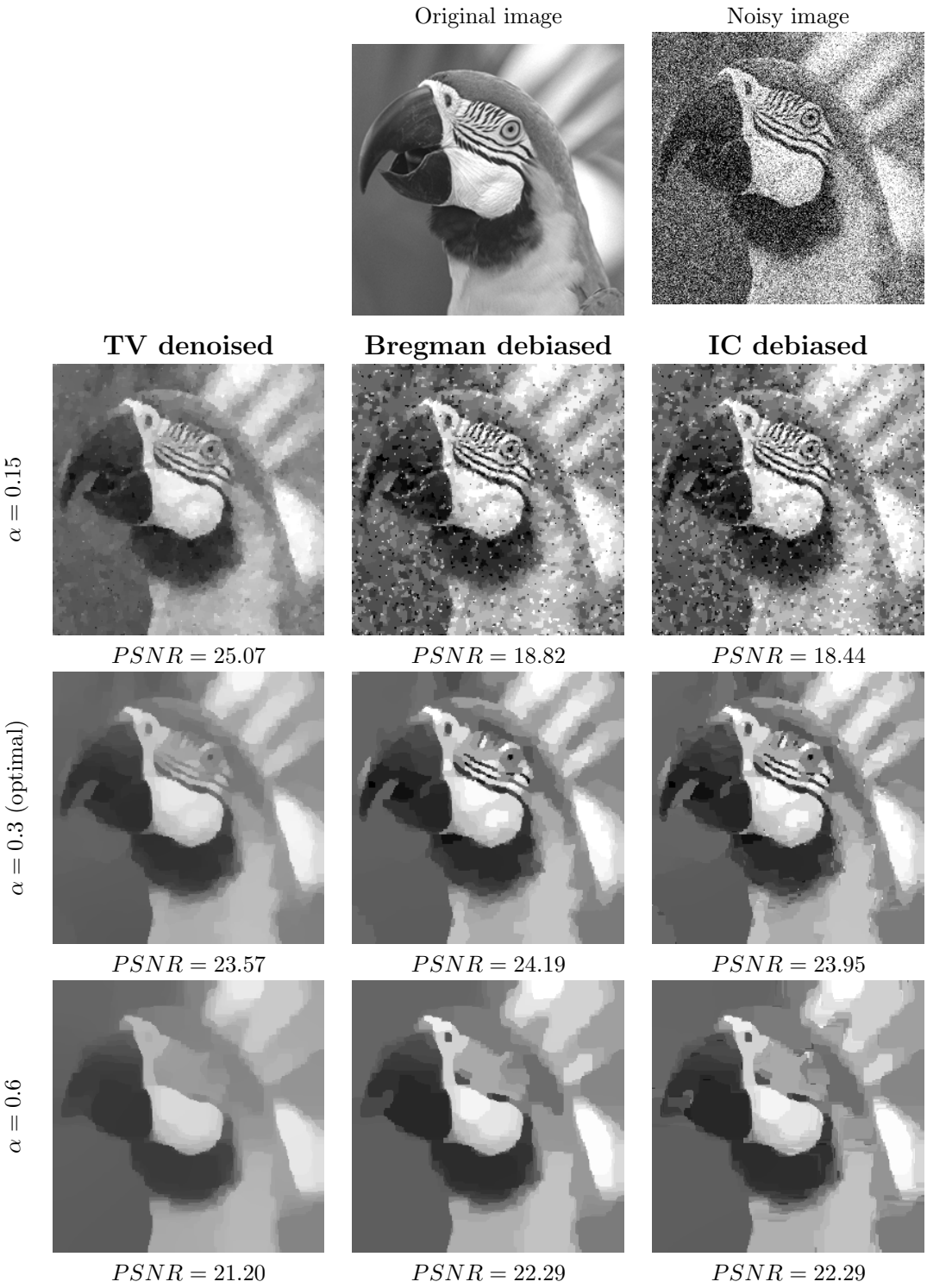


Figure 13: Denoising of the *Parrot* image for different values of the regularization parameter α. First column: TV denoising. Second column: Debiasing on the Bregman subspace. Third column: Debiasing on the infimal convolution subspace.

and hence

$$r_i = \frac{f_i}{1-2\lambda}.$$

We distinguish three cases: If $r_i = \alpha$, then $u_i = f_i - \alpha$ and $f_i = (1-2\lambda)\alpha \Leftrightarrow \lambda = \frac{\alpha - f_i}{2\alpha}$, which by the nonpositivity of λ implies that $f_i \geq \alpha$. Analogously we find for $r_i = -\alpha$ that $u_i = f_i + \alpha$ and $f_i \leq -\alpha$. In case $|r_i| < \alpha$, we obtain that $\lambda = 0$ and hence $r_i = f_i$ and $u_i = 0$ when $|f_i| < \alpha$. Summarizing the first two cases yields the assertion. □ □

The directional derivative $\mathrm{d}u_\alpha(f;g)$ of $u_\alpha(f)$ into the direction $g \in \ell^2$ is given by

$$[\mathrm{d}u_\alpha(f;g)]_i = \begin{cases} g_i, & |f_i| > \alpha \\ 0, & |f_i| < \alpha \\ g_i, & |f_i| = \alpha, \operatorname{sign}(f_i) = \operatorname{sign}(g_i) \\ 0, & |f_i| = \alpha, \operatorname{sign}(f_i) \neq \operatorname{sign}(g_i). \end{cases}$$

Proof. For all $i \in \mathbb{N}$ we have to compute

$$[\mathrm{d}u_\alpha(f;g)]_i = \lim_{t\to 0^+} \frac{[u_\alpha(f+tg)]_i - [u_\alpha(f)]_i}{t}$$

and distinguish four cases: If $|f_i| > \alpha$ and t is small enough, then $|f_i + tg_i| > \alpha$ and $\operatorname{sign}(f_i + tg_i) = \operatorname{sign}(f_i)$ and we obtain

$$\begin{aligned} &[\mathrm{d}u_\alpha(f;g)]_i \\ &= \lim_{t\to 0^+} \frac{1}{t}\Big(f_i + tg_i - \alpha \operatorname{sign}(f_i + tg_i) \\ &\qquad\qquad - (f_i - \alpha \operatorname{sign}(f_i))\Big) \\ &= \lim_{t\to 0^+} \frac{tg_i - \alpha \operatorname{sign}(f_i) + \alpha \operatorname{sign}(f_i)}{t} = g_i. \end{aligned}$$

If $|f_i| < \alpha$ and t is small enough, then $|f_i + tg_i| < \alpha$ and we immediately find

$$[\mathrm{d}u_\alpha(f;g)]_i = 0.$$

If $|f_i| = \alpha$, we have to distinguish

$$\operatorname{sign}(f_i) = \operatorname{sign}(g_i),$$
$$\operatorname{sign}(f_i) \neq \operatorname{sign}(g_i).$$

For the former case, let $f_i = \alpha$. Then $|f_i + \alpha g_i| > \alpha$ and

$$\begin{aligned} &[\mathrm{d}u_\alpha(f;g)]_i \\ &= \lim_{t\to 0^+} \frac{f_i + tg_i - \alpha \operatorname{sign}(f_i + tg_i)}{t} \\ &= \lim_{t\to 0^+} \frac{\alpha + tg_i - \alpha}{t} \\ &= g_i. \end{aligned}$$

We find the same result for $f_i = -\alpha$. In the latter case, let again $f_i = \alpha$. Then $g_i < 0$ and for t small enough we have $|f_i + tg_i| < \alpha$, hence

$$[\mathrm{d}u_\alpha(f;g)]_i = 0.$$

We analogously derive the same result for $f_i = -\alpha$ which completes the proof. □ □

8.2 Isotropic shrinkage

Let $f \in \ell^2(\mathbb{R}^d)$ be a vector-valued signal for $d \in \mathbb{N}$. Analogously to Example 4.7 or 8.1 the solution of the isotropic shrinkage problem

$$u_\alpha(f) \in \operatorname*{arg\,min}_{u \in \ell^1(\mathbb{R}^d)} \frac{1}{2}\|u - f\|^2_{\ell^2(\mathbb{R}^d)} + \alpha \|u\|_{\ell^1(\mathbb{R}^d)}$$

is given by the isotropic soft-thresholding

$$[u_\alpha(f)]_i = \begin{cases} (1 - \frac{\alpha}{|f_i|}) f_i, & |f_i| > \alpha, \\ 0, & |f_i| \leq \alpha. \end{cases}$$

Here, the computation of the directional derivative requires a little more work. At first, let us compute the directional derivative of the function $F\colon \mathbb{R}^d \to \mathbb{R}$, $x \mapsto \frac{1}{|x|}$ into the direction $g \in \mathbb{R}^d$. We define $G\colon \mathbb{R}^d \to \mathbb{R}$, $x \mapsto \frac{1}{|x|^2}$ and calculate

$$\begin{aligned} \mathrm{d}G(x;g) &= \lim_{t\to 0^+} \frac{G(x+tg) - G(x)}{t} \\ &= \lim_{t\to 0^+} \frac{1}{t}\left(\frac{1}{|x+tg|^2} - \frac{1}{|x|^2}\right) \\ &= \lim_{t\to 0^+} \frac{1}{t}\left(\frac{|x|^2 - |x+tg|^2}{|x|^2|x+tg|^2}\right) \\ &= \lim_{t\to 0^+} \frac{1}{t}\left(\frac{-2tx\cdot g - t^2|g|^2}{|x|^2|x+tg|^2}\right) \\ &= -2\frac{x\cdot g}{|x|^4}. \end{aligned}$$

Then by the chain rule we obtain

$$\begin{aligned} \mathrm{d}F(x;g) = \mathrm{d}\sqrt{G}(x;g) &= \frac{\mathrm{d}G(x;g)}{2\sqrt{G(x)}} \\ &= -2\frac{x\cdot g}{|x|^4}\frac{|x|}{2} = -\frac{x\cdot g}{|x|^3}. \end{aligned}$$

Let us further define the projection of a vector $x \in \mathbb{R}^d$ onto another vector $y \in \mathbb{R}^d$ as

$$\Pi_y(x) = \frac{y\cdot x}{|y|^2} y.$$

As for the anisotropic case we now have to compute

$$[\mathrm{d}u_\alpha(f;g)]_i = \lim_{t\to 0^+} \frac{1}{t}\big([u_\alpha(f+tg)]_i - [u_\alpha(f)]_i\big)$$

and we can again distinguish four cases. Let at first $|f_i| > \alpha$. Then for t small enough we have $|f_i + tg_i| > \alpha$ and hence

$$\begin{aligned}
&\lim_{t\to 0^+} \frac{1}{t}\big([u_\alpha(f+tg)]_i - [u_\alpha(f)]_i\big)\\
&= \lim_{t\to 0^+} \frac{1}{t}\left(\left(1 - \frac{\alpha}{|f_i+tg_i|}\right)(f_i+tg_i) \right.\\
&\qquad\qquad \left. - \left(1-\frac{\alpha}{|f_i|}\right) f_i\right)\\
&= \lim_{t\to 0^+} \frac{1}{t}\left(f_i + tg_i - \alpha\frac{f_i+tg_i}{|f_i+tg_i|} - f_i + \alpha\frac{f_i}{|f_i|}\right)\\
&= \lim_{t\to 0^+} \frac{1}{t}\left(tg_i - \frac{\alpha t g_i}{|f_i+tg_i|} \right.\\
&\qquad\qquad \left. - \alpha f_i\left(\frac{1}{|f_i+tg_i|} - \frac{1}{|f_i|}\right)\right)\\
&= g_i - \alpha\frac{g_i}{|f_i|} + \alpha f_i \frac{f_i\cdot g_i}{|f_i|^3}\\
&= g_i + \frac{\alpha}{|f_i|}\left(\Pi_{f_i}(g_i) - g_i\right).
\end{aligned}$$

For $|f_i| < \alpha$ and t small enough we easily find $|f_i + tg_i| < \alpha$ and hence

$$[\mathrm{d}u_\alpha(f;g)]_i = 0.$$

In case $|f_i| = \alpha$ we need to distinguish whether $|f_i + tg_i| > \alpha$ or $|f_i + tg_i| \le \alpha$ for arbitrarily small t. We hence compute

$$\begin{aligned}
&|f_i + tg_i| && > \alpha\\
\Leftrightarrow\ &|f_i + tg_i|^2 && > \alpha^2\\
\Leftrightarrow\ &|f_i|^2 + 2tf_i\cdot g_i + t^2|g_i|^2 && > \alpha^2\\
\Leftrightarrow\ &2f_i\cdot g_i + t|g_i|^2 && > 0,
\end{aligned}$$

which for arbitrarily small t is true only if $f_i\cdot g_i \ge 0$. Analogously we find that $|f_i + tg_i| < \alpha$ for small t is only true if $f_i\cdot g_i < 0$. Hence let now $|f_i| = \alpha$ and $f_i\cdot g_i \ge 0$. Then we obtain

$$\begin{aligned}
[\mathrm{d}u_\alpha(f;g)]_i &= \lim_{t\to 0^+}\frac{1}{t}\big([u_\alpha(f+tg)]_i\big)\\
&= \lim_{t\to 0^+}\frac{1}{t}\left(\left(1-\frac{\alpha}{|f_i+tg_i|}\right)(f_i+tg_i)\right).
\end{aligned}$$

Using $\alpha = |f_i|$, we find

$$\begin{aligned}
&\lim_{t\to 0^+} \frac{|f_i|f_i}{t}\left(\frac{1}{|f_i|} - \frac{1}{|f_i+tg_i|}\right) + g_i - \frac{|f_i|g_i}{|f_i+tg_i|}\\
&= |f_i|f_i\frac{f_i\cdot g_i}{|f_i|^3}\\
&= \Pi_{f_i}(g_i).
\end{aligned}$$

In the last case $|f_i| = \alpha$ and $f_i\cdot g_i < 0$, we find

$$[\mathrm{d}u_\alpha(f;g)]_i = \lim_{t\to 0^+}\frac{1}{t}\big([u_\alpha(f+tg)]_i\big) = 0.$$

Summing up we have

$$[\mathrm{d}u_\alpha(f;g)]_i = \begin{cases} g_i + \frac{\alpha}{|f_i|}\left(\Pi_{f_i}(g_i) - g_i\right), & |f_i| > \alpha,\\ 0, & |f_i| < \alpha,\\ \Pi_{f_i}(g_i), & |f_i| = \alpha,\ f_i\cdot g_i > 0,\\ 0, & |f_i| = \alpha,\ f_i\cdot g_i \le 0. \end{cases}$$

Hence, the corresponding model subspace is given by

$$u \in \mathcal{M}_f^{\mathrm{G}} \Leftrightarrow u_i = \begin{cases} v\in\mathbb{R}^d, & |f_i| > \alpha,\\ 0, & |f_i| < \alpha,\\ \lambda f_i,\ \lambda\ge 0, & |f_i| = \alpha. \end{cases}$$

Analogously to the anisotropic case, the model subspace allows for arbitrary elements, here even including the direction, if the magnitude $|f_i|$ of the signal is strictly above the threshold parameter α. As already discussed in Example 4.7, $|f_i| = \alpha$ is the odd case of the three, since in contrast to $|f_i| > \alpha$ it only allows for changes into the direction of the signal f_i. If we exclude that case, we again find a linear derivative, hence a Gâteaux derivative and even a Fréchet derivative. Accordingly the isotropic shrinkage is the immediate generalization of the anisotropic shrinkage, which we can find as a special case for $d = 1$.

Summing up, the debiasing procedure on this subspace again yields the solution of hard thresholding:

$$[\hat{u}(f)]_i = \begin{cases} f_i, & |f_i| \ge \alpha,\\ 0, & |f_i| < \alpha. \end{cases}$$

Note that we again maintain the signal directly on the threshold.

8.3 Infimal convolution of ℓ^1 Bregman distances

Theorem 8.1. *Let $\Gamma\colon \ell^2(\mathbb{R}^n) \to \ell^1(\mathbb{R}^m)$ be linear and bounded and $J(u) = \|\Gamma u\|_{\ell^1(\mathbb{R}^m)}$ for*

$d, n \in \mathbb{N}$. *Let further* $q_\alpha \in \partial\|\cdot\|_{\ell^1(\mathbb{R}^m)}(\Gamma u_\alpha)$ *such that* $p_\alpha = \Gamma^* q_\alpha$. *Then*

$$\mathrm{ICB}^{q_\alpha}_{\ell^1(\mathbb{R}^m)}(\Gamma u, \Gamma u_\alpha) \leq \mathrm{ICB}^{p_\alpha}_J(u, u_\alpha).$$

Proof.

$$\begin{aligned}
&\mathrm{ICB}^{p_\alpha}_J(u, u_\alpha)\\
&= \inf_{z\in\ell^2(\mathbb{R}^n)} D^{p_\alpha}_J(u-z, u_\alpha) + D^{-p_\alpha}_J(z, -u_\alpha)\\
&= \inf_{z\in\ell^2(\mathbb{R}^n)} \|\Gamma(u-z)\|_{\ell^1(\mathbb{R}^m)} - \langle p_\alpha, u-z\rangle\\
&\quad + \|\Gamma z\|_{\ell^1(\mathbb{R}^m)} + \langle p_\alpha, z\rangle\\
&= \inf_{z\in\ell^2(\mathbb{R}^n)} \|\Gamma(u-z)\|_{\ell^1(\mathbb{R}^m)} - \langle q_\alpha, \Gamma(u-z)\rangle\\
&\quad + \|\Gamma z\|_{\ell^1(\mathbb{R}^m)} + \langle q_\alpha, \Gamma z\rangle\\
&= \inf_{\Gamma z\in\ell^1(\mathbb{R}^m)} \|\Gamma(u-z)\|_{\ell^1(\mathbb{R}^m)} - \langle q_\alpha, \Gamma(u-z)\rangle\\
&\quad + \|\Gamma z\|_{\ell^1(\mathbb{R}^m)} + \langle q_\alpha, \Gamma z\rangle\\
&= \inf_{w\in\ell^1(\mathbb{R}^m)} \|\Gamma u - w\|_{\ell^1(\mathbb{R}^m)} - \langle q_\alpha, \Gamma u - w\rangle\\
&\quad + \|w\|_{\ell^1(\mathbb{R}^m)} + \langle q_\alpha, w\rangle\\
&= \inf_{w\in\ell^1(\mathbb{R}^m)} D^{q_\alpha}_{\ell^1(\mathbb{R}^m)}(\Gamma u - w, \Gamma u_\alpha)\\
&\quad + D^{-q_\alpha}_{\ell^1(\mathbb{R}^m)}(w, -\Gamma u_\alpha)\\
&= \mathrm{ICB}^{q_\alpha}_{\ell^1(\mathbb{R}^m)}(\Gamma u, \Gamma u_\alpha).
\end{aligned}$$

□ □

Note that we get equality for surjective Γ.

Theorem 8.2. *Let* $v, u \in \ell^1(\mathbb{R}^m)$ *and* $q \in \partial\|v\|_{\ell^1(\mathbb{R}^m)}$. *Then*

$$\mathrm{ICB}^{q_\alpha}_{\ell^1(\mathbb{R}^m)}(u, v) = \sum_{i\in\mathbb{N}} G(u_i, q_i)$$

with $G\colon \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ *defined as*

$$G(u_i, q_i) = \begin{cases} |u_i|(1 - |\cos(\varphi_i)||q_i|), & |q_i| < |\cos(\varphi_i)|,\\ |u_i||\sin(\varphi_i)|\sqrt{1-|q_i|^2}, & |q_i| \geq |\cos(\varphi_i)|. \end{cases}$$

where φ_i *denotes the angle between* u_i *and* q_i, *i.e.* $\cos(\varphi_i)|u_i||q_i| = u_i \cdot q_i$ *with* $\varphi_i := 0$ *for* $q_i = 0$ *or* $u_i = 0$.

Proof. Let

$$\begin{aligned}
f_1(u) = D^q_{\ell^1(\mathbb{R}^m)}(u, v) &= \|u\|_{\ell^1(\mathbb{R}^m)} - \langle q, u\rangle,\\
f_2(u) = D^{-q}_{\ell^1(\mathbb{R}^m)}(u, -v) &= \|u\|_{\ell^1(\mathbb{R}^m)} + \langle q, u\rangle.
\end{aligned}$$

f_1, f_2 are proper, l.s.c. and convex so the Fenchel-Moreau theorem [16, p. 18] allows us to compute the infimal convolution as

$$f_1 \square f_2 = (f_1^* + f_2^*)^*.$$

We have

$$f_1^*(z) = \iota_{B^\infty(1)}(z+q),\ f_2^*(z) = \iota_{B^\infty(1)}(z-q),$$

where $\iota_{B^\infty(1)}$ denotes the characteristic function of the $\ell^\infty(\mathbb{R}^m)$-ball

$$B^\infty(1) = \big\{z \in \ell^\infty(\mathbb{R}^m) \mid \|z\|_{\ell^\infty(\mathbb{R}^m)} \leq 1\big\}.$$

Thus

$$\begin{aligned}
(f_1 \square f_2)(u) = &\sup_{z\in\ell^\infty(\mathbb{R}^m)} \langle u, z\rangle\\
&\text{s.t. } \|z+q\|_{\ell^\infty(\mathbb{R}^m)} \leq 1, \|z-q\|_{\ell^\infty(\mathbb{R}^m)} \leq 1.
\end{aligned}$$

Taking into account the specific form of these constraints, we can carry out the computation pointwise and set up the corresponding Lagrangian for every $i \in \mathbb{N}$

$$\begin{aligned}
\mathcal{L}(z_i, \lambda_i, \mu_i) = &-z_i \cdot u_i + \lambda_i(|z_i - q_i| - 1)\\
&+ \mu_i(|z_i + q_i| - 1),
\end{aligned} \tag{8.2}$$

where we from now on leave out the dependence on i for simplicity. Every optimal point of (8.2) has to fulfill the four Karush-Kuhn-Tucker conditions, namely

$$\begin{aligned}
&\frac{\partial}{\partial z}\mathcal{L}(z, \lambda, \mu) = 0, && \lambda(|z-q| - 1) = 0,\\
&\lambda, \mu \geq 0, && \mu(|z+q| - 1) = 0,
\end{aligned}$$

and Slater's condition implies the existence of Lagrange multipliers for a KKT-point of (8.2). The first KKT-condition yields

$$-u + \lambda\frac{z-q}{|z-q|} + \mu\frac{z+q}{|z+q|} = 0. \tag{8.3}$$

Let us first remark that the case $u = 0$ causes the objective function to vanish, hence in the following $u \neq 0$. Then let us address the case $q = 0$ in which (8.3) yields

$$u = (\lambda + \mu)\frac{z}{|z|}.$$

In case $|z| < 1$, we find that $(\lambda + \mu) = 0$ and hence $u = 0$. In case $|z| = 1$, we obtain that $u = (\lambda + \mu)z$ and $(\lambda + \mu) = |u|$, hence $z = \frac{u}{|u|}$. We infer

$$z \cdot u = \frac{u \cdot u}{|u|} = |u|.$$

If $q \neq 0$, we can distinguish three cases:
1st case: $|z - q| < 1, |z + q| = 1$.
Thus $\lambda = 0$ and (8.3) yields

$$u = \mu(z + q).$$

Since $|z+q|=1$ we deduce $\mu=|u|$, so

$$z=\frac{u}{|u|}-q$$

and finally for the value of the objective function

$$z\cdot u=\left(\frac{u}{|u|}-q\right)\cdot u=|u|-q\cdot u.$$

2nd case: $|z+q|<1, |z-q|=1$
We analogously find

$$z\cdot u=|u|+q\cdot u.$$

The first two cases thus occur whenever (insert z in the conditions)

$$\left|\frac{u}{|u|}\pm 2q\right|<1,$$

meaning that either of the conditions has to be fulfilled. We calculate

$$\begin{aligned}&\left|\frac{u}{|u|}-2q\right|^2<1\\ \Leftrightarrow\quad &|q|^2<q\cdot\frac{u}{|u|}\\ \Leftrightarrow\quad &|q|\ <\cos(\varphi).\end{aligned}$$

Hence $q\cdot u>0$ and

$$|u|-q\cdot u=|u|-|q\cdot u|.$$

In the second case we analogously find

$$|q|<-\cos(\varphi),$$

hence $q\cdot u<0$ and

$$|u|+q\cdot u=|u|-|q\cdot u|,$$

so we may summarize the first two cases as

$$z\cdot u=|u|-|q\cdot u|=|u|(1-|\cos(\varphi)||q|),$$

if $|q|<|\cos(\varphi)|$.

3rd case: $|z-q|=1, |z+q|=1$
At first we observe that from

$$|z+q|^2=|z-q|^2$$

we may deduce that $z\cdot q=0$. Therefore we have

$$|z+q|=1\Rightarrow|z|=\sqrt{1-|q|^2}.$$

In the third case the optimality condition (8.3) reads

$$u=\lambda(z-q)+\mu(z+q). \tag{8.4}$$

We multiply the optimality condition by q and obtain

$$\begin{aligned}&u\cdot q &&=\lambda(z-q)\cdot q+\mu(z+q)\cdot q\\ \Leftrightarrow\ &u\cdot q &&=(\mu-\lambda)\ |q|^2\\ \Leftrightarrow\ &(\mu-\lambda) &&=u\cdot\frac{q}{|q|^2}.\end{aligned}$$

Multiplying (8.4) by z yields

$$u\cdot z=(\lambda+\mu)|z|^2 \tag{8.5}$$

and another multiplication of (8.4) by u yields

$$\begin{aligned}|u|^2&=(\lambda+\mu)z\cdot u+(\mu-\lambda)q\cdot u\\ &=(\lambda+\mu)^2|z|^2+(\mu-\lambda)^2|q|^2\\ &=(\lambda+\mu)^2|z|^2+\left(u\cdot\frac{q}{|q|}\right)^2,\end{aligned}$$

where we inserted the previous results in the last two steps. We rearrange and find

$$(\lambda+\mu)=\sqrt{|u|^2-\left(u\cdot\frac{q}{|q|}\right)^2}\,|z|^{-1},$$

which finally leads us to

$$\begin{aligned}u\cdot z&=(\lambda+\mu)|z|^2\\ &=\sqrt{|u|^2-\left(u\cdot\frac{q}{|q|}\right)^2}\,|z|\\ &=|u|\sqrt{\left(1-\left(\frac{u}{|u|}\cdot\frac{q}{|q|}\right)^2\right)(1-|q|^2)}\\ &=|u|\sqrt{(1-|\cos(\varphi)|^2)\,(1-|q|^2)}\\ &=|u||\sin(\varphi)|\sqrt{(1-|q|^2)},\end{aligned}$$

which yields the assertion. Note that in case $|z-q|<1, |z+q|<1$, the first KKT-condition yields $u=0$, which may only occur if the objective function $z\cdot u$ vanishes anyway. □ □

Acknowledgements

This work was supported by ERC via Grant EU FP 7 - ERC Consolidator Grant 615216 LifeInverse. MB acknowledges support by the German Science Foundation DFG via EXC 1003 Cells in Motion Cluster of Excellence, Münster, Germany.

References

[1] H- Bauschke, and P. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer Publishing Company, Incorporated, (2011)

[2] M. Benning, and M. Burger, *Ground states and singular vectors of convex variational regularization methods*, Methods and Applications of Analysis, 20, 295–334 (2013)

[3] K. Bredies, K. Kunisch, and T. Pock, *Total generalized variation*, SIAM J. Imaging Sciences, 3, 492–526 (2010)

[4] M. Burger, *Bregman distances in inverse problems and partial differential equations* in: J.Hiriard-Urrurty, A.Korytowski, H.Maurer, M.Szymkat, eds., Advances in Mathematical Modeling, Optimization, and Optimal Control, Springer, 2016.

[5] M. Burger, G. Gilboa, M. Moeller, L. Eckardt, and D. Cremers, *Spectral decompositions using one-homogeneous functionals*, arxiv 1601.02912 (2016), and submitted.

[6] M. Burger, M. Moeller, M. Benning, and S. Osher, *An Adaptive inverse scale space method for compressed sensing*, Mathematics of Computation, 82, 269–299 (2013)

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

[8] M. Burger, and S. Osher, *A guide to the TV zoo*, In: Level Set and PDE-based Reconstruction Methods in Imaging, 1–70, Springer, Berlin (2013)

[9] M. Burger, E. Resmerita, and L. He, *Error estimation for Bregman iterations and inverse scale space methods in image restoration*, Computing, 81, 109–135 (2007)

[10] E. J. Candes, and Y. Plan, *A probabilistic and RIPless theory of compressed sensing*, , IEEE Transactions on Information Theory, 57(11), 7235–7254 (2011)

[11] V. Caselles, A. Chambolle, and M. Novaga, *Total variation in imaging*, in: Handbook of Mathematical Methods in Imaging, Springer Science & Business Media, 1016–1057 (2011)

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

[13] P.L. Combettes, S. Salzo, and S. Villa, *Regularized learning schemes in Banach Space*, Preprint, arXiv:1410.6847 (2014)

[14] C.A. Deledalle, N. Papadakis, and J. Salmon, *On debiasing restoration algorithms: applications to total-variation and nonlocal-means*, in: Scale Space and Variational Methods in Computer Vision 2015, Lecture Notes in Computer Science, 9087, 129–141 (2015)

[15] C.A. Deledalle, N. Papadakis, J. Salmon, and S. Vaiter, *CLEAR: Covariant LEAst-square Re-fitting with applications to image restoration*, Preprint, arXiv:1606.05158 (2016)

[16] I.Ekeland, and R.Temam, *Convex Analysis and Variational Problems*, SIAM, Philadelphia (1999)

[17] E. Esser, X. Zhang, and T. Chan *A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science*, SIAM J. Imaging Sci., 3(4), 1015–1046 (2010)

[18] J. Min, C. Vonesch, H. Kirshner, L. Carlini, N. Olivier, S. Holden, S. Manley, J. C. Ye, and M. Unser, *FALCON: fast and unbiased reconstruction of high-density super-resolution microscopy data*, Scientific Reports, 4 (2014)

[19] M. Figueiredo, R. Nowak, and S. Wright, *Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems*, IEEE J. Sel. Top. Signal Process., 1(4), 586–598 (2007)

[20] M. Burger, G. Gilboa, S. Osher, and J. Xu, *Nonlinear inverse scale space methods*, Comm. Math. Sci., 4, 178–212 (2006)

[21] J. Lederer, *Trust, but verify: benefits and pitfalls of least-squares refitting in high dimensions*, Preprint, arXiv:1306.0113v1 (2013)

[22] M. Moeller, and M. Burger, *Multiscale methods for polyhedral regularizations*, SIAM J. Optim., 23, 1424–1456 (2013)

[23] M. Moeller, E.-M. Brinkmann, M. Burger and T. Seybold, *Color Bregman TV*, SIAM J. Imaging Sciences, 7(4), 2771–2806 (2014)

[24] S. Osher, F. Ruan, J. Xiong, Y. Yao, and W. Yin, *Sparse recovery via differential inclusions*, Applied and Computational Harmonic Analysis (2016)

[25] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, *An iterative regularization method for total variation-based image restoration*, SIAM Multiscale Model. Simul., 4, 460–489 (2005)

[26] T. Pock, D. Cremers, H. Bischof, and A. Chambolle, *An algorithm for minimizing the Mumford-Shah functional*, in: IEEE 12th International Conference on Computer Vision 2009, 1133–1140 (2009)

[27] E. Resmerita, *Regularization of ill-posed problems in Banach spaces: convergence rates*, Inverse Problems, 21(4), 1303 (2005)

[28] R. T. Rockafellar,*Convex Analysis*, Princeton Landmarks in Mathematics, Princeton University Press, Princeton, NJ (1997).

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

[30] O. Scherzer and C. Groetsch, *Inverse scale space theory for inverse problems*, Scalespace 2106 (2001)

[31] A. Shapiro, *On Concepts of Directional Differentiability*, Journal of Optimization Theory and Applications, 66(3), 477–487 (1990)

[32] T. Schuster, B. Kaltenbacher, B. Hofmann, and K.S. Kazimierski, *Regularization Methods in Banach Spaces*, DeGruyter, Berlin (2012)

[33] E. Tadmor, S. Nezzar, and L. Vese, *A multiscale image representation using hierarchical (BV,L2) decompositions*, Multiscale Modeling and Simulations 2, 554–579 (2004)