




Implicit Regularization Effects of the Sobolev Norms in Image Processing

Bowen Zhu¹ · Jingwei Hu² · Yifei Lou³ · Yunan Yang⁴ 

Received: 25 July 2022 / Revised: 27 July 2023 / Accepted: 17 October 2023

© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2023

Abstract

In this paper, we propose to use the general L^2 -based Sobolev norms, i.e., H^s norms where $s \in \mathbb{R}$, to measure the data discrepancy due to noise in image processing tasks that are formulated as optimization problems. As opposed to a popular trend of developing regularization methods, we emphasize that an *implicit* regularization effect can be achieved through the class of Sobolev norms as the data-fitting term. Specifically, we analyze that the implicit regularization comes from the weights that the H^s norm imposes on different frequency contents of an underlying image. We further analyze the underlying noise assumption of using the Sobolev norm as the data-fitting term from a Bayesian perspective, build the connections with the Sobolev

J. Hu was partially supported by National Science Foundation through CAREER grant DMS-2153208. Y. Lou was partially supported by National Science Foundation through CAREER grant DMS-1846690. Y. Yang was partially supported by National Science Foundation through grant DMS-1913129. Y. Yang acknowledges support from Dr. Max Rössler, the Walter Haefner Foundation and the ETH Zürich Foundation. This paper was initiated in the Summer Research Program for Women in Mathematics in Summer 2021. All authors acknowledge the generous support from the Mathematical Sciences Research Institute (MSRI).

✉ Yunan Yang
yy837@cornell.edu

Bowen Zhu
bz1010@nyu.edu

Jingwei Hu
hujw@uw.edu

Yifei Lou
yflou@unc.edu

¹ New York University, New York, NY 10012, USA

² Department of Applied Mathematics, University of Washington, Seattle, WA 98195, USA

³ Department of Mathematics & School of Data Science and Society, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

⁴ Department of Mathematics, Cornell University, Ithaca 14850, USA

gradient-based methods, and discuss the preconditioning effects on the convergence rate of the gradient descent algorithm, leading to a better understanding of functional spaces/metrics and the optimization process involved in image processing. Numerical results in two geophysical applications of image denoising and full waveform inversion demonstrate the implicit regularization effects.

Keywords H^s norm · Frequency bias · Image processing · Inverse problem · Implicit regularization

Mathematics Subject Classification 65K10 · 46E36 · 68U10 · 49N45 · 92C55 · 49Q22

1 Introduction

Digital images provide a powerful and intuitive way to represent the physical world. Unfortunately, noise is inevitable in the data that is taken or transmitted. When recovering an underlying image from its corrupted measurements, one requires a *fidelity* term to properly model the discrepancy of an imaging formation model as well as a *regularization* term to refine the solution space of this inverse problem. The choice of such data fidelity term often depends on specific applications, specifically on the assumption of the noise distribution [12]. For example, a standard approach for additive Gaussian noise is the least-squares fitting. Using the maximum a posteriori (MAP) estimation, Aubert and Aujol [4] formulated a non-convex data fidelity term for multiplicative noise, which can be solved via a difference-of-convex algorithm [1, 38]. In photon-counting devices such as x-ray computed tomography (CT) [24, 34] and positron emission tomography (PET) [60], the number of photons collected by a device follows a Poisson distribution, thus referred to as Poisson noise. Following the MAP of Poisson statistics, the data discrepancy for Poisson noise can be modeled by a log-likelihood form [17, 18, 37]. Since the nonlinearity of such data fidelity causes computational difficulties, a popular approach in CT reconstruction adopts a weighted least-squares model [58] as the data-fitting term.

To date, major research interests in image processing community have focused on developing regularization methods by exploiting the prior knowledge and/or the special structures of an imaging problem. For instance, the classic Tikhonov regularization [59] returns a smooth output in an attempt to remove the noise, however, at the cost of smearing out important structures and edges. Total variation (TV) [52] is an edge-preserving regularization in that it tends to diffuse along the edges, rather than across, but TV causes a staircasing (blocky) artifact. As remedies, total generalized variation (TGV) [10] and fractional-order TV (FOTV) [68] were proposed to preserve higher-order smoothness. In addition, non-local regularizations [44, 69] based on patch similarities [11] work well for textures and repetitive patterns in an image.

Instead of proposing explicit regularization models, we reveal in this paper that implicit regularization effects can be achieved by using only the L^2 -based Sobolev norms as a data fidelity term. Specifically, we propose to minimize the following data-fitting term,

$$\Phi_{H^s}(u) := \frac{1}{2} \| \mathcal{A}u - f_\sigma \|_{H^s}^2, \quad (1)$$

where f_σ denotes the noisy measurements with an additive Gaussian noise of standard deviation σ and \mathcal{A} denotes a degradation operator. Recall that a Sobolev space is a vector space of functions equipped with a norm that combines the L^p norms of the function and its derivatives up to a given order. We are particularly interested in the L^2 -based Sobolev spaces, often referred to as the H^s spaces for $s \in \mathbb{R}$. They are Hilbert spaces, and the inner products involve the Laplacian operator, thus easy to implement. After discretization, the squared H^s norm as the objective function becomes a weighted least-squares error, and the quadratic nature makes it efficient for gradient computation. Its associated norm is naturally equipped with a particular form of weighting in the Fourier domain. Both the order of biasing (e.g., toward either low or high frequencies) and the strength of biasing can be controlled by the choice of $s \in \mathbb{R}$. When $s = 0$, it reduces to the standard L^2 norm with equal weights on all the frequencies due to Parseval's identity. Since H^s is a generalization of the L^2 norm, using the H^s norm undoubtedly leads to improved results when the parameter s is appropriately chosen according to the prior information, e.g., noise spectra. On the other hand, the H^s norms offer additional flexibility by choosing s to achieve either smoothing ($s < 0$) or sharpening ($s > 0$) effects depending on the noise type in an input image. It was analyzed in [26] that using different H^s norms as the objective function is equivalent to modifying the spectral property of the forward problem, thus exhibiting different stability with respect to data noises. In [67], a particular frequency bias of the H^s norm was utilized to accelerate fixed-point iterations when seeking numerical solutions to elliptic partial differential equation (PDEs).

The introduction of Sobolev spaces was significant for the development of functional analysis [55] and various applications related to PDEs [28] such as the finite element method [57]. There have been relevant works to the Sobolev norms in image processing and inverse problems. For example, the H^{-1} semi-norm is closely related to the quadratic Wasserstein (W_2) metric from optimal transportation [62] under both the asymptotic regime [49] and the non-asymptotic regime [51]. The asymptotic regime refers to the fact that the two datasets under comparison are close enough such that one of them can be considered as a small perturbation of the other, while the non-asymptotic regime does not have a such assumption. The connections between the W_2 metric and the H^{-1} semi-norm have been utilized in many applications [26, 50] such as Bayesian inverse problems [23]. Another close connection comes from works on the Sobolev gradient [45], in which the gradient of a given functional is taken with respect to the inner product induced by the underlying Sobolev norm [14, 56] with demonstrated effects in sharpening and edge-preserving.

In this paper, we illustrate the implicit regularization effects of the H^s norm as a data-fitting term on a toy example of deblurring a square image, together with two geophysical applications of image denoising and full waveform inversion. In those examples, we use only the H^s norm as a data fidelity term in the objective function without any regularization term. The final reconstructions mitigate the impact of the noise, reflecting the implicit regularization effects. This approach is particularly effective when the spectral contents of the noise are well separated from the spectral

contents of the actual image. Since some natural images have a broad bandwidth with spectral contents spreading out in the frequency domain, the implicit regularization by H^s alone may not effectively preserve the important features. In those scenarios, it is beneficial to incorporate, for example, the total variation as a regularization term together with the H^s norm as the data fidelity. We acknowledge that using the H^s norm as the data fidelity term together with the total variation regularization has been studied in [39, 48] for image decomposition. In this work, we generalize their approaches by considering s as a tunable hyperparameter in practical implementations and proposing a more efficient algorithm by the alternating direction method of multipliers (ADMM) [9, 31].

The main contributions of this work are threefold. First, we propose to use the H^s norms as a novel data-fitting term to effectively utilize their implicit regularization effects for noise removal. Second, we analyze the underlying noise assumption of using the H^s norms as the objective function from a Bayesian perspective, its connections to the Sobolev gradient flow, and the resulting preconditioning effects on the convergence rate. Such analysis contributes to a better understanding of the advantages by using the L^2 -based Sobolev norms in image processing. Lastly, we present a series of computational approaches to calculating the H^s norms under different setups.

The rest of the paper is organized as follows. Section 2 devotes to the analysis of the Sobolev norms, including the implicit regularization effects, the noise assumption from a Bayesian perspective, the connections to the W_2 distance, the Sobolev gradient, and the preconditioning effects. We describe three approaches for computing the H^s norm in Sect. 3 under different boundary conditions and choices of s . In Sect. 4, we conduct experiments on geographical examples to demonstrate different scenarios where the weak norm ($s < 0$) and the strong norm ($s > 0$) are preferred, respectively. Section 5 revisits the H^s +TV model [39, 48] with a tunable parameter s and an efficient algorithm for image deblurring. Conclusions follow in Sect. 6.

2 Analysis on Sobolev Norms

In this section, we briefly review the definitions and properties of the L^2 -based Sobolev norms, followed by discussing the implicit regularization effects in Sect. 2.2. We draw connections of the Sobolev norms to a Bayesian interpretation of data fidelity in Sect. 2.3, the quadratic Wasserstein distance [62] in Sect. 2.4, and the Sobolev gradient [14] in Sect. 2.5. Lastly in Sect. 2.6, we discuss how the choice of the Sobolev norm can affect the convergence rate of the gradient descent algorithm.

2.1 H^s Sobolev Space

There are two common ways to define the L^2 -based Sobolev norm. One is based on the Sobolev space $W^{k,p}(\mathbb{R}^d)$ for a nonnegative integer k ; see Definition 1.

Definition 1 (Sobolev Space $W^{k,p}(\mathbb{R}^d)$) Let $1 \leq p < \infty$ and k be a nonnegative integer. If a function f and its weak derivatives $D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$, $|\alpha| \leq k$ all lie

in $L^p(\mathbb{R}^d)$, where α is a multi-index and $|\alpha| = \sum_{i=1}^d \alpha_i$, we say $f \in W^{k,p}(\mathbb{R}^d)$ and define the $W^{k,p}(\mathbb{R}^d)$ norm of f as

$$\|f\|_{W^{k,p}(\mathbb{R}^d)} := \left(\sum_{|\alpha| \leq k} \|D^\alpha f\|_{L^p(\mathbb{R}^d)}^p \right)^{1/p}. \tag{2}$$

In this work, we focus on the L^2 -based Sobolev space $W^{k,2}$, which is a Hilbert space.

While Definition 1 is concerned with integer derivatives (in contrast to fractional derivatives), there exists a natural extension to a more general L^2 -based Sobolev space $W^{s,2}(\mathbb{R}^d)$ for an arbitrary scalar $s \in \mathbb{R}$ through the Fourier transform. This leads to the second definition of the Sobolev space. Specifically, we define by $\mathcal{F} : \mathcal{S}'(\mathbb{R}^d) \mapsto \mathcal{S}'(\mathbb{R}^d)$ the Fourier transform where

$$\mathcal{F}f(\xi) = \hat{f}(\xi) = (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} f(x)e^{-ix \cdot \xi} dx, \quad \forall f \in \mathcal{S}'(\mathbb{R}^d). \tag{3}$$

We further denote \mathcal{F}^{-1} as the inverse Fourier transform, I as the identity operator, $\langle \xi \rangle := \sqrt{1 + |\xi|^2}$, and $\mathcal{S}'(\mathbb{R}^d)$ as the space of tempered distributions [2, Sec. 6.4].

Definition 2 (Sobolev Space $H^s(\mathbb{R}^d)$) Let $s \in \mathbb{R}$, the Sobolev space H^s over \mathbb{R}^d is given by

$$H^s(\mathbb{R}^d) := \left\{ f \in \mathcal{S}'(\mathbb{R}^d) : \mathcal{F}^{-1} [\langle \xi \rangle^s \mathcal{F}f] \in L^2(\mathbb{R}^d) \right\}. \tag{4}$$

The space $H^s(\mathbb{R}^d)$ is equipped with the norm

$$\|f\|_{H^s(\mathbb{R}^d)} := \left\| \mathcal{F}^{-1} [\langle \xi \rangle^s \mathcal{F}f] \right\|_{L^2(\mathbb{R}^d)} = \|\mathcal{P}_s f\|_{L^2(\mathbb{R}^d)}, \tag{5}$$

where the operator $\mathcal{P}_s := (I - \Delta)^{s/2}$.

When $s = 0$, the $H^s(\mathbb{R}^d)$ space (norm) reduces to the standard L^2 space (norm). One can show that $W^{k,2}(\mathbb{R}^d) = H^k(\mathbb{R}^d)$ for any integer k [2]. We remark that $\|f\|_{H^k(\mathbb{R}^d)} \neq \|f\|_{W^{k,2}(\mathbb{R}^d)}$ for the same k in general, but the two norms are equivalent, which can be shown through Fourier transforms. Hereafter, we focus on $H^s(\mathbb{R}^d)$ for $s \in \mathbb{R}$ since it allows for fraction-valued s which can be a tunable parameter in image processing applications.

2.2 Implicit Regularization Effects of the H^s Norms

Without loss of generality, we consider the following data formation model based on a linear inverse problem,

$$f_\sigma = Au + n_\sigma, \tag{6}$$

where f_σ denotes the noisy measurements with an additive Gaussian noise n_σ of standard deviation σ , and \mathcal{A} denotes a degradation operator. A general inverse problem is posted as recovering an underlying image u from the data f_σ with the knowledge of \mathcal{A} . If \mathcal{A} is the identity operator, i.e., $\mathcal{A} = I$, this problem is referred to as *denoising*. If \mathcal{A} can be formulated as a convolution operator with a blurring kernel, it is called image *deblurring* or *deconvolution*.

We assume that the linear operator \mathcal{A} is injective and asymptotically diagonal in the Fourier domain such that there exist two constants $C_1, C_2 > 0$, and

$$C_1 \langle \xi \rangle^{-\alpha} \hat{u}(\xi) \leq \widehat{\mathcal{A}u}(\xi) \leq C_2 \langle \xi \rangle^{-\alpha} \hat{u}(\xi), \tag{7}$$

where $\alpha \in \mathbb{R}$, and the hat symbol denotes the Fourier transform with frequency coordinate ξ .

When $\alpha > 0$, we say the operator \mathcal{A} is ‘‘smoothing.’’ The value of α can describe to some extent the degree of ill-conditionedness (or difficulty) of solving an inverse problem [5, 25] in the sense that the larger the α is, the more ill-posed the associated inverse problem becomes.

We examine the regularization effects of using the H^s norm defined in (5) to quantify the data misfit. In other words, we seek a solution of the inverse problem (6) by minimizing

$$\frac{1}{2} \| \mathcal{A}u - f_\sigma \|_{H^s}^2 = \frac{1}{2} \| \mathcal{P}_s(\mathcal{A}u - f_\sigma) \|_{L^2}^2 = \frac{1}{2} \int_{\mathbb{R}^d} \langle \xi \rangle^{2s} | \widehat{\mathcal{A}u}(\xi) - \widehat{f_\sigma}(\xi) |^2 d\xi, \tag{8}$$

without any additional regularization term. The minimizer of $\Phi_{H^s}(u)$ has a closed-form solution, i.e.,

$$u = (\mathcal{P}_s \mathcal{A})^\dagger \mathcal{P}_s f_\sigma, \quad (\mathcal{P}_s \mathcal{A})^\dagger = \left(\mathcal{A}^* \mathcal{P}_s^* \mathcal{P}_s \mathcal{A} \right)^{-1} \mathcal{A}^* \mathcal{P}_s^*, \tag{9}$$

where the superscript \dagger denotes the Moore–Penrose inverse operator [64, Chapter 11] and \mathcal{A}^* is the adjoint operator of \mathcal{A} under the L^2 inner product. Note that $\mathcal{P}_s^* = \mathcal{P}_s$ as \mathcal{P}_s is self-adjoint. By comparing (9) with the standard least-squares solution, we conclude that the H^s -based inversion can be seen as a weighted least-squares method if $s \neq 0$.

Remark 1 A variant of (8) is to use the \dot{H}^s semi-norm instead of the standard H^s norm. That is, we replace $\langle \xi \rangle^{2s} = (1 + |\xi|^2)^s$ by $|\xi|^{2s}$, and the objective function becomes

$$\Phi_{\dot{H}^s}(u) = \frac{1}{2} \| \mathcal{A}u - f_\sigma \|_{\dot{H}^s}^2 := \frac{1}{2} \int_{\mathbb{R}^d} |\xi|^{2s} | \widehat{\mathcal{A}u}(\xi) - \widehat{f_\sigma}(\xi) |^2 d\xi. \tag{10}$$

The frequency bias from $\Phi_{\dot{H}^s}$ is more straightforward to analyze than the one from $\Phi_{H^s}(u)$, as the weight in front of each frequency is precisely an algebraic factor $|\xi|^s$. If $f \in H^s$ for $s > 0$, we have $\|f\|_{\dot{H}^s} < \infty$. However, this is not the case for $s < 0$. For example, a function f may have a finite H^{-1} norm, but if $\int f dx \neq 0$, it does not have a well-defined \dot{H}^{-1} norm.

Remark 2 If two scalars $s_1 < s_2$, then $H^{s_2} \subset H^{s_1}$ is continuously embedded. In other words, we specify the order among all H^s spaces, e.g., $H^2 \subset H^1 \subset L^2 \subset H^{-1} \subset H^{-2}$.

We consider the following three scenarios to illustrate the implicit regularization effects of Φ_{H^s} as an objective function. A similar analysis applies to $\Phi_{\dot{H}^s}$.

- When $s = 0$, the solution (9) reduces to the standard least-squares solution, i.e., $u = \mathcal{A}^\dagger f_\sigma$. Without any regularization term, this solution inevitably overfits the noise in the observation f_σ .
- When $s > 0$, \mathcal{P}_s can be regarded as a differential operator, which amplifies high-frequency contents of f_σ . If the noise in f_σ is also high frequency, the overfitting phenomenon caused by \mathcal{P}_s is even worse than the standard least-squares solution. On the other hand, if f_σ is corrupted by lower-frequency noise, the weighted least-squares would avoid overfitting.
- When $s < 0$, \mathcal{P}_s is an integral operator, meaning that applying \mathcal{P}_s to f_σ suppresses high-frequency components. The noisy content in f_σ does not fully “propagate” into the reconstructed solution u . The inverse problem is less sensitive to the high-frequency noise in f_σ , indicating the improved well-posedness. Again, this property becomes disadvantageous if f_σ is subject to lower-frequency noise.

Based on the above three different types of scenarios, it is clear that the H^s norm causes a particular weight on the frequency contents of the input function according to the choice of s . We will later refer to this property as the *spectral bias* of the H^s norm. For the higher frequency of the noise, the smaller negative s should be used to suppress it. On the other hand, for the lower frequency of the noise, the bigger positive s is more effective in neutralizing its effect. Later in Sect. 4, we will use this as an intuitive way to choose the proper s for different imaging tasks.

Remark 3 To summarize, if the data are polluted with high-frequency noise, using a weak norm as the objective function alone improves the posedness of the inverse data-fitting problem without the help of any regularization term. On the other hand, a potential disadvantage of the weaker norm is that the objective function not only implicitly suppresses higher-frequency noisy contents but also higher-frequency components of the noise-free data. Consequently, the reconstruction loses the high-frequency resolution, as illustrated in [26, Figure 4].

Remark 4 One can also generalize (6) to a nonlinear inverse problem. The main properties of the H^s norm will remain, but the analysis would be less straightforward. In Sect. 4.2, we present such a nonlinear example and numerically demonstrate the benefits of using the H^s norm.

Next, we demonstrate the aforementioned properties regarding $s = 0$, $s > 0$, and $s < 0$ through numerical examples of reconstructing a (discrete) image u from (6) by minimizing the discretized objective function

$$\Phi_{H^s}(u) = \frac{1}{2} \|P_s(Au - f_\sigma)\|_{L^2}^2,$$

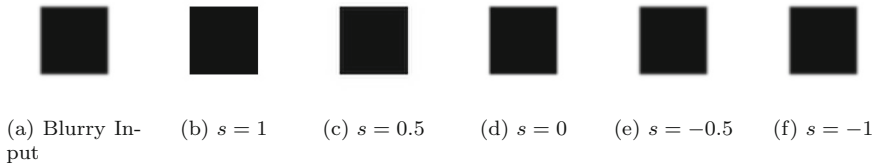


Fig. 1 Effects of minimizing Φ_{H^s} with different choices of s . The reconstructed solutions gradually transition from sharp to blurry after the same number of gradient descent iterations, showing that strong norms ($s > 0$) are better at sharpening

where P_s is a proper discretization of the continuous operator \mathcal{P}_s , and A denotes the linear operator \mathcal{A} in the matrix form; please refer to Sect. 3 for discretization details. Applying the gradient descent algorithm with a fixed step size η to minimize the objective function $\Phi_{H^s}(u)$ yields the following iterative step:

$$u^{(n+1)} = u^{(n)} - \eta \nabla \Phi(u^{(n)}) = u^{(n)} - \eta A^T P_s^T P_s (A u^{(n)} - f_\sigma). \quad (11)$$

For the sake of convergence, the step size η should be chosen smaller than $1/L$ where $L = \lambda_{\max}$, the largest eigenvalue of $A^T P_s^T P_s A$ (also its matrix 2-norm). As L depends on the choice of s , we choose η small enough such that all H^s methods converge for a fair comparison.

We apply the gradient descent iteration (11) to a simple example of image deblurring. Consider a binary image of size 100×100 with a black square in the middle to be the ground truth, referred to as the Square image. We can consider the image as a discretization of a 2D function $u : \mathbb{R}^2 \mapsto \mathbb{R}$ on a regular mesh and it is typical to assume that $u \in L^2(\mathbb{R}^2)$. The continuous linear mapping $A : L^2(\mathbb{R}^2) \mapsto L^2(\mathbb{R}^2)$ is a convolution operator. Its discrete version can be formulated as a convolution with 15×15 Gaussian kernel of standard deviation 1, which can be implemented through `fspecial('gaussian', 15, 1)` in Matlab. The blurry image is further corrupted by an additive Gaussian noise with standard deviation σ .

When $\sigma = 0$, the input image is blurry but not noisy, as seen in Fig. 1a. We show reconstructed images by minimizing Φ_{H^s} with different choices of s via (11). We fix the step size $\eta = 10^{-4}$ for all methods such that the gradient descent algorithms converge for all. The five values of s in Fig. 1 cover all scenarios: $s = 0$, $s > 0$, and $s < 0$. After running 100 iterations of the gradient descent algorithm (11), we observe in Fig. 1 a gradual transition from sharp to blurry reconstruction results as s decreases from $s = 1$ to $s = -1$. This is aligned with our earlier discussion that the operator \mathcal{P}_s for positive s is a differential operator, which boosts the higher-frequency content of $A^T(Au^{(n)} - f_\sigma)$, corresponding to the gradient when the L^2 norm becomes the objective function. Another way to interpret the results is the change in convergence rate. The H^s objective function with $s > 0$ increases the eigenvalue for the oscillatory modes of the image, which dominates the initial residual. Thus, these error modes are eliminated faster than the case of $s \leq 0$; see more details in Sect. 2.6. Consequently, it accelerates the gradient descent algorithm to converge to the sharp ground truth,

as the only missing information in the blurry input is precisely in the high-frequency domain. In summary, strong norms ($s > 0$) are good at sharpening.

We then examine the influence of noise on the reconstructions by minimizing the Φ_{H^s} functional. For this purpose, we add different amounts of noises, i.e., $\sigma = 0.1$ and $\sigma = 0.5$, to the same blurry image (shown in Fig. 1a), leading to noisy and blurry data shown in Fig. 2a and Fig. 2g, respectively. Again, we reconstruct the images by running 100 iterations of gradient descent with the same step size $\eta = 1$, which is much larger than the earlier example. Here, we use weak H^s norms to reduce the noise impact. The Lipschitz constant of the gradient induced by the weak H^s norm is much smaller than the ones based on the strong norms, which thus affects the choice of the step size (to be larger). The top row of Fig. 2 corresponds to a smaller noise level ($\sigma = 0.1$). The L^2 -based method, i.e., $s = 0$, clearly suffers from overfitting the noise, as the reconstruction is even noisier than the input. The best result is achieved at $s = -0.5$, while the reconstructed images are over smooth as s decreases. This set of tests shows both advantages and potential limitations of weak norms ($s < 0$) as addressed in Remark 3. The bottom row of Fig. 2 corresponds to a larger noise level ($\sigma = 0.5$), when the overfitting phenomenon is more severe not only for the L^2 norm, but also for the cases of $s = -0.5$ and $s = -0.25$. The best reconstruction occurs at $s = -1$, where the spectral bias of the objective function toward lower-frequency contents of the residual (the difference between the current iterate and the input image) is the strongest. That is, the weighting coefficients on the low-frequency components are much bigger in contrast to the ones on the high-frequency ones due to the rapid decay of function $\langle \xi \rangle^{-1}$ compared to $\langle \xi \rangle^{-0.5}$. The comparison between two noise levels also implies that the best choice of s is data-dependent. One heuristic principle is that the noisier the input is, the weaker objective function (smaller s) one should choose to avoid overfitting the noise.

In Fig. 3, we show the cross-sections of 2D images; the location of the cross-section is indicated by the red lines in Fig. 2a and Fig. 2g. In Fig. 3a, the 1D plots clearly show the over-smoothing artifact for $s = -1$, and the construction of $s = -0.5$ is closest to the ground truth. In contrast, the case $s = -0.5$ is no longer good enough to “smooth” out the stronger noise in Fig. 3b, and the result from $s = -1$ turns out to be the best fit.

We remark that two critical factors for observing different results in Fig. 1 and Fig. 2 are the use of iterative scheme (11) and the maximum number of iterations. For the noise-free setup in Fig. 1, the analytic least-squares solutions (9) for different s coincide, while they have different rates of convergence in the iterative scheme. See Fig. 4a where we plot the convergence history of the relative error defined by $\|u^{(n)} - u^*\|_{L^2} / \|u^*\|_{L^2}$, where u^* is the closed-form solution and the function L^2 norm becomes the matrix Frobenius norm after discretization.

In the noisy setup, the iterative gradient descent formula (11) can be regarded as the Landweber iteration [25, Sec. 6.1],

$$u^{(n+1)} = u^{(n)} + T^\top (y^\delta - Tu^{(n)}),$$

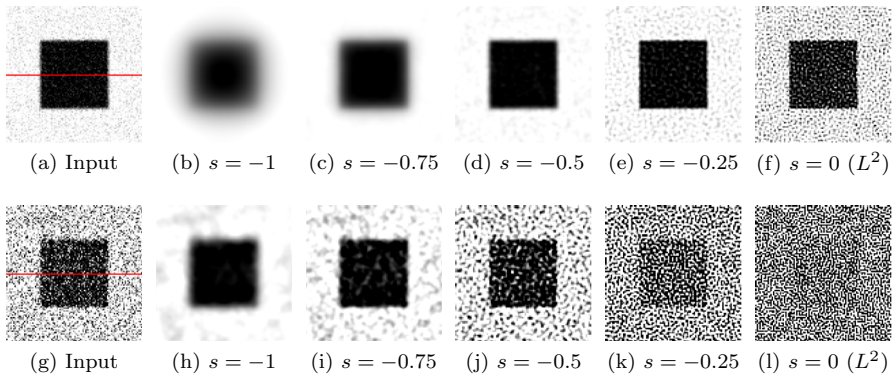


Fig. 2 Deblurring the Square image by minimizing $\Phi_{H^s}(u)$. The top row presents the blurry noisy input with $\sigma = 0.1$ and reconstruction results of different s values. A noisier case ($\sigma = 0.5$) is illustrated in the bottom row

where $T = \sqrt{\eta}P_s A$ and $y^\delta = \sqrt{\eta}P_s f_\sigma$. As stated in [25], “many iterative methods exhibit a ‘self-regularizing property’ in the sense that early termination of the iterative process has a regularizing effect” (P. 154). In Fig. 4b and Fig. 4c, we show the change in the relative error for a very large number of iterations so that we can observe the crucial impact of early stopping. Based on [25, Sec. 6.1], a good stopping criterion is to choose the smallest n such that $\|y^\delta - Tu^{(n)}\|_{L^2} < 2\sqrt{\eta}\delta_s$ where $\delta_s := \|P_s(f_\sigma - f)\|_{L^2} \approx \|f_\sigma - f\|_{H^s}$ and f denotes the noise-free blurred data.

We remark that $\|y^\delta - Tu^{(n)}\|_{L^2}$ monotonically decreases with respect to n since its square is the scaled objective function Φ_{H^s} . The value δ_s is the noise power measured in the H^s norm. The optimal stopping iteration is $\mathcal{O}(\delta_s^{-2})$ [25, Proposition 6.4]. Note that for the noises in Fig. 2, $\delta_s = \mathcal{O}(|\xi|^s)$ with the noise frequency ξ . Thus, the optimal stopping iteration is $\mathcal{O}(|\xi|^{-2s})$, monotonically decreasing as s increases, precisely as illustrated in Fig. 4. Moreover, even if we can use an a priori estimate to choose the optimal stopping iteration, Fig. 4c shows that a weaker H^s norm can achieve a smaller optimal relative error. This is because, when n is the optimal iteration number,

$$\|u^* - u^{(n)}\|^2 \leq \|u^* - u^{(0)}\|^2(1 - 2/\delta_s).$$

Since $u^{(0)}$ is fixed for any choice of s , the smaller the δ_s , the smaller the optimal error in the reconstructed solution.

2.3 A Bayesian Interpretation

The choice of the data fidelity term in image processing can be derived from a Bayesian approach under a proper assumption on the noise distribution of the data [12]. In this subsection, we present the noise assumption associated with the proposed H^s data fidelity term (8) under the Bayesian framework.

One major advantage of the Bayesian approach is to account for the uncertainty in the data which will be propagated to the solution to the inverse problem. It combines

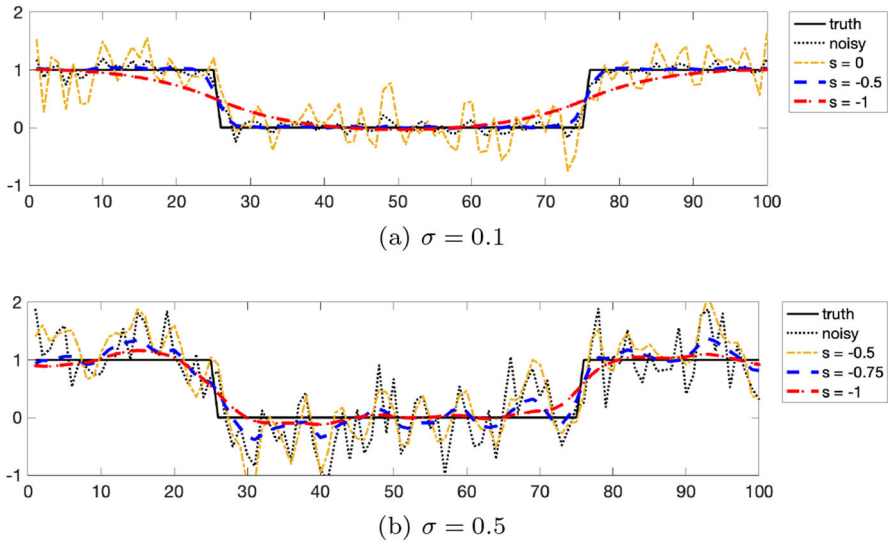


Fig. 3 The zoom-in view for different choice of s at the cross-section (the red line) illustrated in Fig. 2a and Fig. 2g, respectively

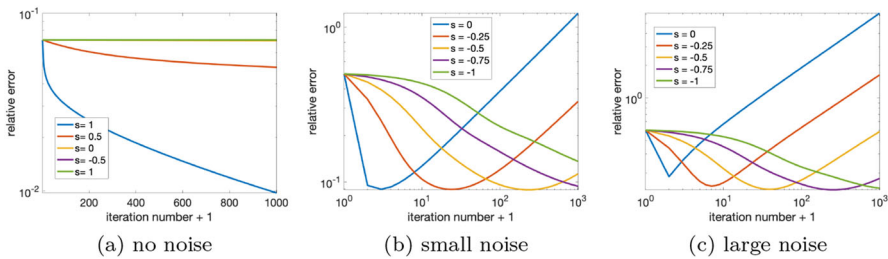


Fig. 4 The relative error in Frobenius norm $\|u^{(n)} - u^*\|_F / \|u^*\|_F$ with respect to the number of iterations for setups in Fig. 1 and Fig. 2

a probabilistic model for the observed data f_σ with a density function $\mathbb{P}(f_\sigma|u)$ and a probability distribution $\mathbb{P}(u)$ representing the prior knowledge regarding the unknown u . Bayes’ theorem provides a way to construct the posterior distribution, denoted as $\mathbb{P}(u|f_\sigma)$, where

$$\mathbb{P}(u|f_\sigma) = \frac{\mathbb{P}(f_\sigma|u)\mathbb{P}(u)}{\mathbb{P}(f_\sigma)}. \tag{12}$$

The posterior distribution $\mathbb{P}(u|f_\sigma)$ can be regarded as the solution to the Bayesian inverse problem, which differs from the deterministic framework of solving inverse problems that returns a single value of u , e.g., the minimizer of (8).

Although the Bayesian and the deterministic approaches are quite different, there are connections when we try to find the maximum a posteriori (MAP) estimation. Without loss of generality, we consider that the prior distribution $\mathbb{P}(u)$ follows the normal

distribution $\mathcal{N}(0, C)$ and C is invertible. Then maximizing the posterior distribution $\mathbb{P}(u|f_\sigma)$ is equivalent to the following minimization problem [22, Sec. 4.3],

$$u^* = \operatorname{argmin}_u \mathcal{E}(u; f_\sigma) + \frac{1}{2} \langle u, C^{-1}u \rangle_{L^2}, \tag{13}$$

where $\mathcal{E}(u; f_\sigma) = -\log \mathbb{P}(f_\sigma|u)$ is commonly known as the negative log-likelihood function. Consider the inverse problem model (6) where we assume the additive noise $n_\sigma \sim \mathcal{N}(0, \Gamma)$. We then have

$$\mathbb{P}(f_\sigma|u) = \mathcal{N}(\mathcal{A}u, \Gamma) \propto \exp\left(-\frac{1}{2} \|\mathcal{A}u - f_\sigma\|_\Gamma^2\right), \quad \langle \cdot, \cdot \rangle_\Gamma := \langle \cdot, \Gamma^{-1} \cdot \rangle_{L^2},$$

after ignoring the normalizing constant. If the inverse covariance operator $\Gamma^{-1} = \mathcal{P}_s^* \mathcal{P}_s$ for \mathcal{P}_s defined in Sect. 2.2, we then have

$$\mathcal{E}(u; f_\sigma) = -\log \mathbb{P}(f_\sigma|u) \propto \frac{1}{2} \|\mathcal{A}u - f_\sigma\|_\Gamma^2 = \frac{1}{2} \|\mathcal{P}_s(\mathcal{A}u - f_\sigma)\|_{L^2}^2,$$

which reduces to our H^s objective function (8).

Based on the above Bayesian interpretation, using (8) as the data fidelity term is equivalent to a data noise assumption $n_\sigma \sim \mathcal{N}(0, (\mathcal{P}_s^* \mathcal{P}_s)^{-1})$ in the Bayesian framework. Note that the L^2 norm corresponds to $n_\sigma \sim \mathcal{N}(0, I)$, the standard Gaussian. This perspective again demonstrates that we can enforce prior information to achieve implicit regularization effects through the data fidelity (likelihood function) term. For example, $n_\sigma \sim \mathcal{N}(0, (I - \Delta)^{-1})$ ($s = 1$) supposes a smooth additive noise, while $n_\sigma \sim \mathcal{N}(0, I - \Delta)$ ($s = -1$) assumes that the noise lacks of smoothness. This interpretation also extends to the semi-norm $\Phi_{\dot{H}^s}$ (10), which corresponds to $n_\sigma \sim \mathcal{N}(0, (-\Delta)^{-s})$. These noise models are often studied in the context of Elliptic Gaussian Process [6, 7] or Fractional Gaussian Field [41] for non-integer choices of s . There have been many works on sampling such noise models through solving the (fractional) elliptic PDEs; for example, see [8]. We also remark that in [23], the authors have shown that the quadratic Wasserstein metric (W_2) as a likelihood function in Bayesian inference is asymptotically equivalent to assuming a multiplicative Gaussian noise model where the covariance operator is a weighted Laplacian operator.

2.4 Relationship with the W_2 Distance

Here, we review a connection between the Sobolev norms and the quadratic Wasserstein (W_2) distance [62] to provide a better understanding of both metrics. The Wasserstein distance defined below is associated to the cost function $c(x, y) = |x - y|^p$ in the optimal transportation problem.

Definition 3 (Wasserstein Distance) We denote by $\mathcal{P}_p(\Omega)$ the set of probability measures with finite moments of order p . For $1 \leq p < \infty$,

$$W_p(\mu, \nu) = \left(\inf_{T_{\mu, \nu} \in \mathcal{M}} \int_{\Omega} |x - T_{\mu, \nu}(x)|^p d\mu(x) \right)^{\frac{1}{p}}, \quad \mu, \nu \in \mathcal{P}_p(\Omega), \quad (14)$$

where $T_{\mu, \nu}$ is a push-forward map such that $T_{\mu, \nu} \# \mu = \nu$ [62], and \mathcal{M} is the set of all maps that push forward μ into ν . Note that W_2 corresponds to the case $p = 2$.

An asymptotic connection between the W_2 metric and the H^s norm was first provided in [49] given the two probability distributions under comparison are close enough such that the linearization error is small. Consider μ as the probability measure and $d\pi$ as an infinitesimal perturbation that has zero total mass. Then

$$W_2(\mu, \mu + d\pi) = \|d\pi\|_{\dot{H}_{(d\mu)}^{-1}} + \mathcal{O}(d\pi). \quad (15)$$

We remark that $\dot{H}_{(d\mu)}^{-1}$ is the weighted \dot{H}^{-1} semi-norm. We refer readers to [62, Sec. 7.6] for its detailed definition.

A connection between W_2 and \dot{H}^{-1} under a non-asymptotic regime was later presented in [51]. Let f and g be the probability densities for the measure μ and ν , respectively. Provided that $c_1 \leq f, g \leq c_2$, we have the following non-asymptotic equivalence between W_2 and \dot{H}^{-1} [51],

$$\frac{1}{c_2} \|f - g\|_{\dot{H}^{-1}} \leq W_2(\mu, \nu) \leq \frac{1}{c_1} \|f - g\|_{\dot{H}^{-1}}. \quad (16)$$

Note that in both the asymptotic and the non-asymptotic regimes, the W_2 metric shares a similar spectral bias as the \dot{H}^{-1} semi-norm, up to a weighting function. Thus, the implicit regularization properties for the case $s = -1$ discussed in Sect. 2.2 can extend to the quadratic Wasserstein metric. This finding explains the improved stability of the Wasserstein metric in inverse problems from various applied fields, including machine learning [3], parameter identification [66], and full waveform inversion [65].

2.5 Relationship with the Sobolev Gradient Flow

The well-known heat equation $u_t = \Delta u$ where $u : \Omega \mapsto \mathbb{R}$ (Ω is an open subset of \mathbb{R}^2 with smooth boundary $\partial\Omega$) can be seen as the gradient flow of the energy functional

$$E(u) = \frac{1}{2} \int_{\Omega} |\nabla u|^2 dx = \frac{1}{2} \|\nabla u\|_{L^2}^2,$$

with respect to the L^2 inner product $\langle v, w \rangle_{L^2} = \int_{\Omega} v w dx$. A different gradient flow can be derived from a more general inner product, for example, based on the Hilbert space H^s in Definition 2 for any $s \in \mathbb{R}$. An inner product on the Sobolev space $H^1(\Omega)$ [28, 56] can be defined as

$$g_{\lambda}(v, w) = (1 - \lambda)\langle v, w \rangle_{L^2} + \lambda\langle v, w \rangle_{H^1} = \langle v, w \rangle_{L^2} + \lambda\langle v, w \rangle_{\dot{H}^1},$$

for any $\lambda > 0$ and $\langle v, w \rangle_{\dot{H}^1} = \langle \nabla v, \nabla w \rangle_{L^2}$. If we are only interested in periodic functions on the domain Ω , the gradient operators considered here are equipped with the periodic boundary condition. When $\lambda = 0$, $g_\lambda(v, w)$ reduces the conventional L^2 inner product, and when $\lambda = 1$, it becomes the standard H^1 inner product: $\langle v, w \rangle_{H^1} = \langle v, w \rangle_{L^2} + \langle \nabla v, \nabla w \rangle_{L^2}$. Calder et al. [14] exploited a general Sobolev gradient flow for image processing and established the well-posedness of the Sobolev gradient flow $u_t = (I - \lambda \Delta)^{-1} \Delta u$ in both the forward and the backward directions of minimizing $E(u)$. Specifically worth noticing is that the backward direction can be regarded as a sharpening operator [40, 42].

Without loss of generality, we set $\lambda = 1$ when studying a connection between the Sobolev gradient and the gradient of the H^s norm as the energy functional. Given any energy (objective) functional $E(u)$, an inner product based on the Sobolev metric $H^1(\Omega)$ gives a specific gradient formula

$$\nabla_{H^1} E(u) = (I - \Delta)^{-1} \nabla_{L^2} E(u), \tag{17}$$

such that $\forall v \in H^1(\Omega) \subset L^2(\Omega)$, we have

$$\langle \nabla_{H^1} E(u), v \rangle_{H^1} = \langle \nabla_{L^2} E(u), v \rangle_{L^2} = \lim_{\varepsilon \rightarrow 0} \frac{E(u + \varepsilon v) - E(u)}{\varepsilon}, \tag{18}$$

If we consider the energy functionals $\Phi_{L^2}(u)$ (i.e., $\Phi_{H^0}(u)$) and $\Phi_{H^{-1}}(u)$ defined in (8), we have

$$\begin{aligned} \nabla_{L^2}(\Phi_{L^2}(u)) &= \mathcal{A}^*(\mathcal{A}u - f_\sigma), \\ \nabla_{H^1}(\Phi_{L^2}(u)) &= (I - \Delta)^{-1} \mathcal{A}^*(\mathcal{A}u - f_\sigma), \\ \nabla_{L^2}(\Phi_{H^{-1}}(u)) &= \mathcal{A}^*(I - \Delta)^{-1}(\mathcal{A}u - f_\sigma). \end{aligned}$$

Correspondingly, we have the following three gradient flow equations:

$$u_t = -\mathcal{A}^*(\mathcal{A}u - f_\sigma) \quad (L^2 \text{ gradient flow of } \Phi_{L^2}(u)), \tag{19}$$

$$u_t = -(I - \Delta)^{-1} \mathcal{A}^*(\mathcal{A}u - f_\sigma) \quad (H^1 \text{ gradient flow of } \Phi_{L^2}(u)), \tag{20}$$

$$u_t = -\mathcal{A}^*(I - \Delta)^{-1}(\mathcal{A}u - f_\sigma) \quad (L^2 \text{ gradient flow of } \Phi_{H^{-1}}(u)). \tag{21}$$

If \mathcal{A}^* shares the same set of eigenfunctions as the Laplace operator Δ , then $\mathcal{A}^*(I - \Delta)^{-1} = (I - \Delta)^{-1} \mathcal{A}^*$, and hence (20) is exactly equivalent to (21). Even if \mathcal{A}^* does not commute with $(I - \Delta)^{-1}$, one can still view $(I - \Delta)^{-1}$ as a smoothing (integral) preconditioning operator upon the residual $\mathcal{A}u - f_\sigma$, which we wish to reduce to zero no matter the objective function is $\Phi_{L^2}(u)$ or $\Phi_{H^{-1}}(u)$. To sum up, (20) and (21) are similar in nature in terms of the spectral bias of the resulting gradient descent dynamics, which demonstrates the equivalence between the change of the gradient flow and the change of the objective function under certain circumstances. In contrast to (19), both (20) and (21) are equipped with the smoothing property due to the additional $(I - \Delta)^{-1}$ operator.

2.6 Changing the Rate of Convergence

So far, our analysis has been focusing on how the H^s norm is related to the data noise n_σ and its regularization effects during the optimization process. In this section, we address another interesting property of the H^s norm as the objective function: it may improve the rate of convergence in gradient descent.

Extending the L^2 gradient flow (21) to a general $\Phi_{H^s}(u)$ energy functional, we obtain a gradient flow equation with respect to u :

$$u_t = -\mathcal{A}^* \mathcal{P}_s^* \mathcal{P}_s (\mathcal{A}u - f_\sigma), \tag{22}$$

where $\mathcal{P}_s = (I - \Delta)^{s/2}$. Minimizing the $\Phi_{H^s}(u)$ energy functional (8) is equivalent to reducing the H^s norm of the residual $\mathcal{R} := \mathcal{A}u - f_\sigma$. Based on (22), we have that

$$\mathcal{R}_t = \mathcal{A}u_t = -\mathcal{A} \mathcal{A}^* \mathcal{P}_s^* \mathcal{P}_s \mathcal{R}. \tag{23}$$

The decay rate of the residual \mathcal{R} is directly determined by the spectral property of the linear operator $\mathcal{A} \mathcal{A}^* \mathcal{P}_s^* \mathcal{P}_s$. After discretization, (23) becomes

$$R^{(k)} = (I - \eta E_s) R^{(k-1)} = (I - \eta E_s)^k R^{(0)}, \quad E_s = \mathcal{A} \mathcal{A}^\top P_s^\top P_s,$$

where I is the identity matrix and η is a properly chosen step size in gradient descent. As a result,

$$\|R^{(k)}\|_2 = \|(I - \eta E_s)^k R^{(0)}\|_2 \leq (1 - \eta \lambda_{\min})^k \|R^{(0)}\|_2,$$

where λ_{\min} is the minimum eigenvalue of E_s , which consequently depends on the choice of s . Given a fixed forward operator \mathcal{A} , by properly choosing s , we may improve the convergence rate by increasing λ_{\min} . For example, if $\mathcal{A}u = \Delta u$, choosing the H^{-2} norm as the objective function yields the fastest convergence among the class of H^s norms [67].

3 Numerical Computation of the H^s Norms

In this section, we present three numerical methods for computing the general H^s norms of any $s \in \mathbb{R}$. The first one (in Sect. 3.1) applies to periodic functions defined on a domain, which is either the entire \mathbb{R}^d or a compact subset of \mathbb{R}^d , denoted by Ω . We are mainly interested in periodic functions to align with a fast implementation of convolution that assumes the periodic boundary condition. In addition, we discuss the functions with zero Neumann boundary condition in Sect. 3.2 and integer-valued s in Sect. 3.3.

3.1 Through the Discrete Fourier Transform

Recall that the Hilbert space $H^s(\mathbb{R}^n)$, $s \in \mathbb{R}$, is equipped with the norm (5). If we compute the H^s norm of a periodic function $f \in H^s$ defined on the entire \mathbb{R}^d , or equivalently, defined on $\Omega \subset \mathbb{R}^d$, we have

$$\|f\|_{H^s(\mathbb{R}^n)} = \|\mathcal{P}_s f\|_{L^2(\mathbb{R}^n)} \approx \|P_s f\|_{L^2(\mathbb{R}^n)}, \tag{24}$$

where $\mathcal{P}_s f = \mathcal{F}^{-1} [(1 + |\xi|^2)^{s/2} \mathcal{F} f]$ and “ \approx ” indicates the approximation by discretization. The discretization of the linear operator \mathcal{P}_s , denoted as P_s , can be computed explicitly through diagonalization, or implicitly, through the fast Fourier transform. For the former, the discretization of \mathcal{F} is the discrete Fourier transform (DFT) matrix, while the discretization of \mathcal{F}^{-1} is its conjugate transpose. The discretization of $(1 + |\xi|^2)^{s/2}$ is correspondingly a diagonal matrix.

3.2 Through the Discrete Cosine Transform

If we are interested in computing the H^s norm of non-periodic functions on the domain Ω that is a compact subset of \mathbb{R}^d , we adopt the zero Neumann boundary condition [53] as the boundary condition for the Laplacian operator. As a result, rather than DFT, a consistent definition is through the discrete cosine transform (DCT) due to its relationship with the discrete Laplacian on a regular grid associated with the zero Neumann boundary condition, i.e.,

$$\|f\|_{H^s(\Omega)} \approx \|\widehat{P}_s f\|_{L^2(\Omega)}, \quad \widehat{P}_s = C^{-1}(I - \Lambda)^{s/2}C, \tag{25}$$

where C and C^{-1} are matrices representing the DCT and its inverse, respectively, and Λ is a diagonal matrix whose diagonal entries are eigenvalues of the discrete Laplacian with the zero Neumann boundary condition. One may observe that (25) shares great similarity with (5) except for the facts that DFT is replaced with DCT and the diagonal matrix also varies according to eigenvectors and eigenvalues of the discrete Laplacian with different boundary conditions.

3.3 Through Solving a Partial Differential Equation

Let $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz-smooth domain. The Hilbert space $H^s(\Omega)$ is the same as the Sobolev space $W^{s,2}(\Omega)$ for all integers $s \in \mathbb{Z}$; see [2, Sec. 7], i.e.,

$$W^{s,2}(\Omega) = \{f|_\Omega : f \in W^{s,2}(\mathbb{R}^d)\} = \{f|_\Omega : f \in H^s(\mathbb{R}^d)\} = H^s(\Omega).$$

Consequently, we can define an equivalent norm for functions in $H^s(\Omega)$ through $\|\cdot\|_{W^{s,2}(\Omega)}$, which involves differential operators with the zero Neumann boundary conditions [53]. When $s \in \mathbb{N}$, the computation of the $W^{s,2}(\Omega)$ norm should follow its definition in Definition 1, while the differential operator involved should be handled

with the zero Neumann boundary condition. In this case, one explicit definition of $\|f\|_{H^{-s}(\Omega)}$ via the Laplace operator [53, 67] is given by

$$\|f\|_{H^{-s}(\Omega)} = \|u\|_{H^s(\Omega)}, \tag{26}$$

where $u(x)$ is the solution to the following partial differential equation with the zero Neumann boundary condition [53, Section 3],

$$\begin{cases} \mathfrak{L}^s u(x) = f(x), & x \in \Omega, \\ \nabla u \cdot \mathbf{n} = 0, & x \in \partial\Omega, \end{cases} \tag{27}$$

for $\mathfrak{L}^s = \sum_{|\alpha| \leq s} (-1)^{|\alpha|} D^{2\alpha}$.

We may define the operator \mathfrak{L}^{-s} by setting $u = \mathfrak{L}^{-s} f$. Combining (26) and (27), we have

$$\|f\|_{H^{-s}(\Omega)}^2 = \langle u, f \rangle_{L^2(\Omega)} = \langle \mathfrak{L}^{-s} f, f \rangle_{L^2(\Omega)} = \|\tilde{\mathcal{P}}_s f\|_2^2, \tag{28}$$

where $\tilde{\mathcal{P}}_s^* \tilde{\mathcal{P}}_s = \mathfrak{L}^{-s}$. We may also denote $\tilde{\mathcal{P}}_s = \mathfrak{L}^{-s/2}$. The numerical discretization of $\tilde{\mathcal{P}}_s$ is denoted as \tilde{P}_s .

Note that (5) and (26) do not yield precisely the same norm given $f \in H^s(\mathbb{R}^d)$ with $s \in \mathbb{Z}$. For example, when $s = -2$ and $d = 2$, the definition (5) depends on the integral operator $(I - \Delta)^{-1}$ based on the definition of the $H^{-s}(\Omega)$ norm, while the definition (26) depends on the integral operator $(I - \Delta + \Delta^2)^{-1/2}$ based on the definition of the $W^{-s,2}(\Omega)$ norm in (1). However, the leading terms in both definitions match. Thus, they are equivalent norms for functions that belong to the same functional space $H^s(\Omega) = W^{s,2}(\Omega)$ given a fixed s . We remark that the H^s norms with non-integer s cannot be calculated through PDEs; instead, one should refer to Sect. 3.2.

4 Experiments

In this section, we first present the denoising results of low-frequency noise arisen in geographical images in Sect. 4.1, followed by a nonlinear geophysical inverse problem in Sect. 4.2. In both examples, there is no regularization term in the objective function, so the implicit regularization effects purely come from the H^s norm as the data fidelity term. In a concrete application, an intuitive way of choosing s is based on comparing the data noise oscillation frequency as suggested in Sect. 2.2, followed by minor tuning of s .

4.1 Geophysical Image Denoising

We present a denoising example from a seismic application, in which the noise is mostly of low frequencies. The noisy image is the output of the so-called reverse-

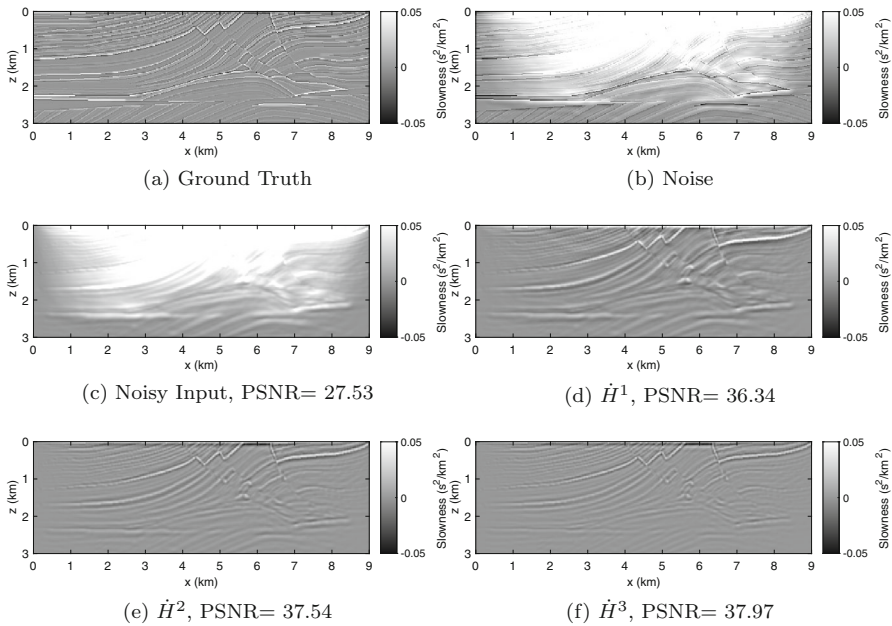


Fig. 5 Marmousi RTM image denoising using different \dot{H}^s semi-norms as the data fidelity term

time migration (RTM) [20], which is a wave equation migration method to illustrate complex structures of the earth, especially strong contrast geological interfaces. The output image is the zero time-lag cross-correlation between the source and the receiver wavefields.

However, artifacts are produced by the cross-correlation of source-receiver wavefields propagating in the same direction. Specifically, migration artifacts appear at shallow depths, above strong reflectors, and severely mask the migrated structures; see Fig. 5c. They are generated by the cross-correlation of reflections, backscattered waves, head waves, and diving waves [70]. The ground truth image is shown in Fig. 5a. We are interested in reducing the noise (see Fig. 5b), strongly dominated by the low-frequency components, in the input image Fig. 5c by minimizing the objective function (10), where the linear operator \mathcal{A} is the identity. We regard the noisy image as a piece-wise constant discretization of an $H^3(\Omega)$ function where Ω is the rectangle domain.

Based on the discussion in Sect. 2.2, it is beneficial to use strong norms (i.e., $s > 0$) to suppress the low-frequency noise. Here, we consider \dot{H}^1 , \dot{H}^2 , and \dot{H}^3 with the corresponding results shown in Figs. 5d-5f, respectively. We quantitatively measure the reconstruction performance in terms of the peak signal-to-noise ratio (PSNR), which is defined by

$$\text{PSNR}(u^*, \tilde{u}) := 20 \log_{10} \frac{NM}{\|u^* - \tilde{u}\|_F^2},$$

where u^* is the restored image, \tilde{u} is the ground truth, and N , M are the number of pixels and the maximum peak value of \tilde{u} , respectively. According to PSNR, using the

\dot{H}^3 norm as the objective function produces the best recovery. We also demonstrate that all the three strong semi-norms can effectively suppress the low-frequency noise in Fig. 5c without changing the reflecting features of the underlying image.

4.2 Full Waveform Inversion

Here we present a full waveform inversion (FWI) example. It is a nonlinear inverse problem where one aims to invert parameter u (usually the wave velocity) given the observed data g (usually the wave pressure field) through a nonlinear relationship $F(u) = g$. The forward operator is implicitly given through the wave equation constraint. More specifically, we denote by $D \subset \mathbb{R}^{d-1} \times \mathbb{R}^+ \subset \mathbb{R}^d$ the upper half space, and $T \subset [0, \infty)$ the temporal domain with $|T| < \infty$. Let $s(x, t) = w(t)\delta(x - x_s)$ be a point source, where the source location $x_s \in D$, and the time-dependent function $w(t) \in H^1(T) \subset L^\infty(T)$. We consider the unknown parameter (squared slowness) $u \in L^\infty(D)$. Let $v \in L^2(D; L^2(T))$ solve the following wave equation

$$\begin{cases} u(x) \frac{\partial^2 v}{\partial t^2}(x, t) - \Delta v(x, t) = s(x, t) & (x, t) \in D \times T, \\ v(x, 0) = 0, \frac{\partial v}{\partial t}(x, 0) = 0 & x \in D, \\ \nabla_x v(x, t) \cdot \mathbf{n} = 0 & x \in \partial D. \end{cases} \tag{29}$$

Let $R : L^2(D; L^2(T)) \mapsto L^2(D_0; L^2(T))$ be a bounded linear observation operator where $D_0 \subset D$ denotes the set of receivers. The observed data $g = Rv$. Thus, the forward operator $F : L^\infty(D) \mapsto L^2(D_0; L^2(T))$ maps u to g and is implicitly given through the wave equation.

The nonlinear inverse problem is often reformulated as a PDE-constrained optimization problem where one aims to find the optimal u by minimizing the difference between the observed data g and the simulated data $F(u)$ evaluated at the current prediction of u . While the least-squares method, i.e., using the squared L^2 norm as the data fidelity term, has been the conventional choice, and an additional regularization term is often added, here we consider only the general H^s data-fitting term as the objective function. That is,

$$\min_{u \in L^\infty(D)} \frac{1}{2} \|F(u) - g\|_{H^s}^2. \tag{30}$$

Since the data $F(u), g \in L^2(D_0 \times T)$, (30) is well defined for $s \leq 0$. We perform optimization using different s values and demonstrate its impacts on the inversion.

The true velocity parameter is presented in Fig. 6a and all the tests start with the same initial guess shown in Fig. 6b. We use the L-BFGS method [46] to solve for (30) and manually stop the iterative process after 200 iterations. The inversion result using the L^2 norm (corresponding to $s = 0$) is shown in Fig. 6c. It converges to a local minimum with many wrong features compared to the ground truth. Similarly, when using the $H^{-0.5}$ norm, the layers in the recovered subsurface image in Fig. 6d do not match their true locations, despite a slight improvement from the L^2 -based result. When using the H^{-1} norm, the reconstruction is qualitatively much better as

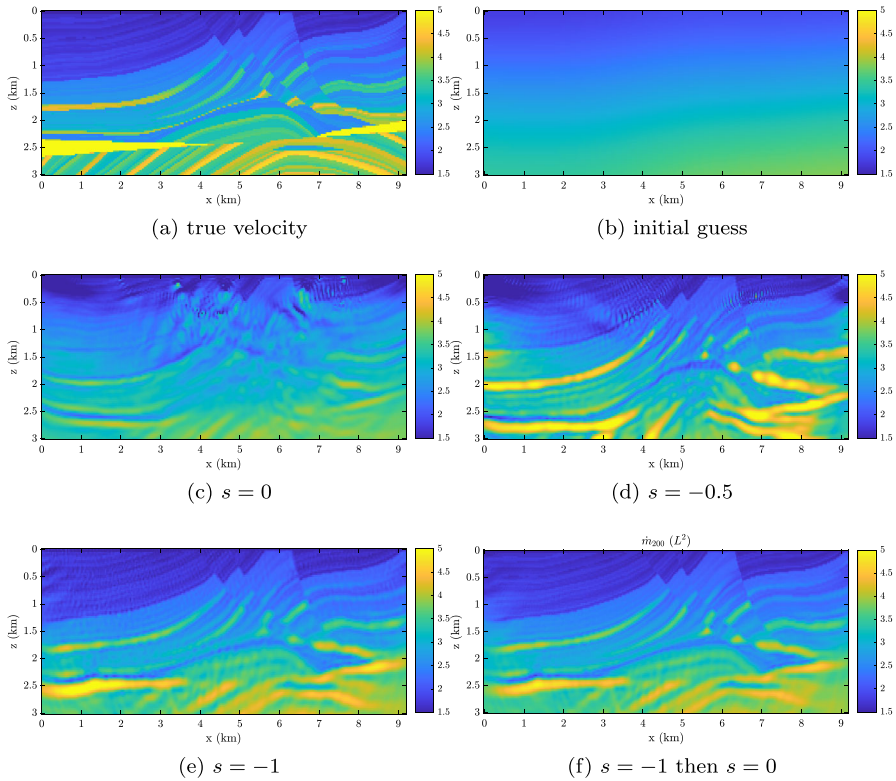


Fig. 6 FWI example for Sect.4.2: (a) true velocity; (b) initial guess; (c)-(e) reconstructions after 200 iterations using the L^2 , $H^{-0.5}$, and the H^{-1} norms, respectively; (f) reconstruction using the H^{-1} in the first 100 iterations followed by another 100 iterations using the L^2 norm

the structural properties of the inverted velocity image become very close to the ground truth, as one can see in Fig. 6e.

Since this is a nonlinear inverse problem, the resulting optimization problem (30) is highly non-convex. The problem that the iterates are trapped at the local minima is often referred to as cycle skipping in FWI [63]. We expect that the change of the objective function modifies the optimization landscape. It is well known that low-frequency components of the wave data are less likely to suffer from cycle skipping [13]. As we have discussed in Sect.2.2, when $s < 0$, the H^s norm has a natural bias toward the low-frequency content of the input, and the smaller the s , the stronger the bias. Hence, it is not surprising to see that with the same initial guess, H^{-1} norm as the objective function can converge to the global minima, while the L^2 norm and the $H^{-0.5}$ norm get stuck at local minima.

On the other hand, the H^{-1} inversion in Fig. 6e lacks high resolution despite having most of the correct features. Again, it is a property of the weak norm ($s < 0$). It is usually the high-frequency components of the data g that resolve the sharp features in the reconstructed parameter u . However, the high-frequency components of the data, including both the useful physical information and the high-frequency noise, are given

much smaller weight in a weak norm, resulting in a low-resolution reconstruction. We have commented on this phenomenon earlier in Remark 3. This dilemma can be mitigated by performing a transition of the objective function. For example, one can first use the H^{-1} norm as the objective function to take advantage of the bigger basin of attraction. Once the iterate is close to the ground truth, one can switch to stronger norms such as the L^2 . In Fig. 6f, we perform a transition of the objective function from H^{-1} after 100 iterations of L-BFGS to the L^2 norm for another 100 iterations. The resolution of the reconstruction is visibly improved compared to 200 iterations of the H^{-1} norm alone as shown in Fig. 6e. A more rigorous analysis on how to adaptively update s will be left to future work.

5 A Case Study of Using Total Variation

The natural implicit regularization effects of the H^s norm could be further enhanced by combining with a regularization term, such as the TV regularization. There have been two main directions related to the combination in the literature.

First, minimizing the total variation energy under the general H^s Sobolev space has been studied both numerically and theoretically [29, 30, 35, 54]. In such frameworks, the objective function is solely the TV energy, while the model parameter u is assumed to belong to the H^s functional space. Our work here is different from the literature since we fix the parameter space to be L^2 and consider the objective function to be H^s or possibly H^s together with a regularization term. As a result, the objective function explicitly includes the H^s norm, equivalent to the assumption that the data space (as opposed to the model parameter space) is H^s .

The second main direction in the literature is more relevant to our work. Combining the H^{-1} data-fitting term with the TV regularization was first studied in [48] and later generalized to any negative Sobolev norm in [39]. The literature mainly focuses on image decomposition by using TV to single out a cartoon (piece-wise constant) image and the H^s norms for oscillatory components like textures and noises. Two recent works [19, 33] further propose to decompose a signal or an image into three components: a piece-wise constant component, a smooth (low-oscillating) component, and a high oscillatory component, the last of which is modeled by H^{-1} .

We advocate using the data fidelity term of the H^s norm by itself as an implicit regularization effect on images. However, the frequency biases induced by the H^s norm do not work so well on natural images due to complicated structures that spread out the entire frequency domain. As a result, image restoration requires an explicit regularization term to ensure satisfactory results. To this end, we present a proof-of-concept idea by incorporating the TV regularization together with the H^s -based data fidelity term. As H^s reduces to the L^2 metric for $s = 0$, we expect any regularization term combined with the H^s would outperform the one with the standard least-squares misfit by treating s as a tunable hyperparameter.

We also present a new algorithm to minimize the H^s norm with the TV regularization based on ADMM, as detailed in Sect. 5.1. Under this efficient algorithmic framework, we then numerically investigate the power of combining the H^s data-fitting term together with the TV regularization by presenting the deblurring examples

in Sect. 5.2. The numerical results demonstrate that H^s +TV, as a more general framework, outperforms the traditional L^2 +TV, making it a promising choice in image processing.

5.1 An Numerical Algorithm for Minimizing TV regularization and H^s Data-Fitting Term

We revisit the celebrated TV regularization [52] for image restoration that minimizes the following energy functional over the bounded-variation (BV) space [61]

$$\min_{u \in \text{BV}(\Omega)} J(u) = \frac{\lambda}{2} \|\mathcal{A}u - f_\sigma\|_{H^s}^2 + \mu \|\nabla u\|_{L^1}, \tag{31}$$

where $\lambda, \mu \in \mathbb{R}^+$ are scalars balancing the data-fitting term, $\Omega = [0, 1]^2 \subset \mathbb{R}^2$ is a unit square, and the regularization term. We include two parameters λ, μ for the ease of disabling either one of them in experiments. We consider the linear operator $\mathcal{A} : \text{BV}(\Omega) \mapsto L^2(\Omega)$ a convolution operator, and f_σ is the noisy blurry data. Osher, Solé, and Vese first proposed the framework (31) for the case $s = -1$ [48], which was later generalized by Lieu and Vess in [39] to any $s < 0$. Here, we extend the framework and apply it to any $s \in \mathbb{R}$. Moreover, we regard s as a tunable hyperparameter, together with λ and μ in (31).

We discuss the discretization of the model (31). Suppose a two-dimensional (2D) image function is defined on an $m \times n$ Cartesian grid. By using a standard linear index, we can represent a 2D image as a vector, i.e., the $((i - 1)m + j)$ -th component denotes the intensity value at pixel (i, j) . We define a discrete gradient operator,

$$\mathbf{D}u := \begin{bmatrix} D_x \\ D_y \end{bmatrix} u, \tag{32}$$

where D_x, D_y are the finite forward difference operator with the periodic boundary condition in the horizontal and vertical directions, respectively. We adopt the periodic boundary condition for the finite difference scheme to align with the periodic boundary condition when implementing the discrete convolution operator A by the fast Fourier transform (FFT). We denote $N := mn$ and the Euclidean spaces by $\mathcal{X} := \mathbb{R}^N, \mathcal{Y} := \mathbb{R}^{2N}$, then $u \in \mathcal{X}, Au \in \mathcal{X}$, and $\mathbf{D}u \in \mathcal{Y}$.

The H^s norm can be expressed in terms of the weighted norm, which is equivalent to the multiplication of \mathbf{P}_s , the discrete representation of the operator \mathcal{P}_s . Given the choice of s and the particular boundary condition, we can select a preferable way of implementing \mathbf{P}_s as any of the three types of matrices P_s, \widehat{P}_s , and \widetilde{P}_s discussed in Sect. 3. To align with the periodic boundary condition used for \mathbf{D} and A , we choose $\mathbf{P}_s = P_s$. In summary, we obtain the following objective function in a discrete form,

$$J(u) = \frac{\lambda}{2} \|\mathbf{P}_s(Au - f_\sigma)\|_2^2 + \mu \|\mathbf{D}u\|_1. \tag{33}$$

There are a number of optimization algorithms available to minimize $J(u)$ in order to find the optimal solution u , such as primal-dual algorithms [16, 27], dual projection method [15], and Bregman iterations [32, 47]. Here, we present the alternating direction method of multipliers (ADMM) [9, 31], by introducing an auxiliary variable d and studying an equivalent form of (33)

$$\min_{u \in \mathcal{X}, d \in \mathcal{Y}} \mu \|d\|_1 + \frac{\lambda}{2} \|\mathbf{P}_s(Au - f_\sigma)\|_2^2 \quad \text{s.t.} \quad d = \mathbf{D}u. \tag{34}$$

Here, we keep both μ and λ to include the scenario without the TV term, i.e., $\mu = 0$. The corresponding augmented Lagrangian function is expressed as

$$\mathcal{L}(u, d; v) = \mu \|d\|_1 + \frac{\lambda}{2} \|\mathbf{P}_s(Au - f_\sigma)\|_2^2 + \langle \rho v, \mathbf{D}u - d \rangle + \frac{\rho}{2} \|d - \mathbf{D}u\|_2^2, \tag{35}$$

with a dual variable v and a positive parameter ρ . The ADMM framework involves the following iterations,

$$\begin{cases} u^{(k+1)} = \arg \min_u \mathcal{L}(u, d^{(k)}; v^{(k)}), \\ d^{(k+1)} = \arg \min_d \mathcal{L}(u^{(k+1)}, d; v^{(k)}), \\ v^{(k+1)} = v^{(k)} + \mathbf{D}u^{(k+1)} - d^{(k+1)}. \end{cases} \tag{36}$$

By taking the derivative of \mathcal{L} with respect to u , we obtain a closed-form solution of the u -subproblem in (36), i.e.,

$$u^{(k+1)} = \left(\lambda A^T \mathbf{P}_s^T \mathbf{P}_s A + \rho \mathbf{D}^T \mathbf{D} \right)^{-1} \left(\lambda A^T \mathbf{P}_s^T \mathbf{P}_s f_\sigma + \mathbf{D}^T (d^{(k)} - \rho v^{(k)}) \right). \tag{37}$$

We remark that $-\mathbf{D}^T \mathbf{D}$ is the discrete Laplacian operator with the periodic boundary condition. In this case, the discrete operators (matrices), A , A^T , $\mathbf{P}_s^T \mathbf{P}_s$, and $\mathbf{D}^T \mathbf{D}$ all have the discrete Fourier modes as eigenvectors. As a result, the matrix $\lambda A^T \mathbf{P}_s^T \mathbf{P}_s A + \rho \mathbf{D}^T \mathbf{D}$ in (37) shares the Fourier modes as eigenvectors, and its inverse can be computed efficiently by FFT.

The d -subproblem in (36) has also a closed-form solution given by

$$\mathbf{d}^{(k+1)} = \mathbf{shrink} \left(\nabla u^{(k+1)} + \mathbf{v}^{(k)}, \frac{\mu}{\rho} \right), \tag{38}$$

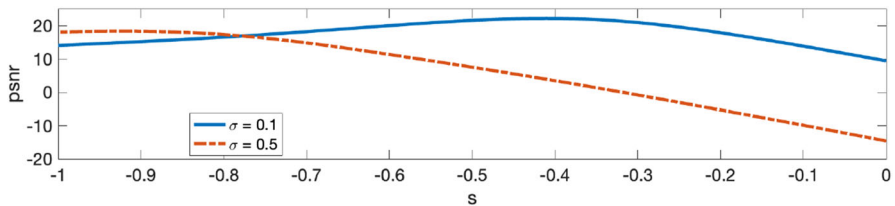
where $\mathbf{shrink}(\mathbf{v}, \beta) = \text{sign}(\mathbf{v}) \circ \max\{|\mathbf{v}| - \beta, 0\}$ with the Hadamard (elementwise) product \circ . Finally, $v^{(k+1)}$ is updated based on $u^{(k+1)}$ and $d^{(k+1)}$. The iterative process continues until reaching the stopping criteria or the maximum number of iterations.

5.2 Image Deblurring

We start this subsection by expanding the deblurring example in Sect. 2.2. In particular, we conduct a comprehensive study of the H^s norms with different choices of s

Table 1 Deblurring the Square image comparison among different H^s norms in terms of PSNR. Visual results corresponding to the second and the third rows are shown in Fig. 2

Add TV	Noise σ	input	$s = 0$	$s = -0.25$	$s = -0.5$	$s = -0.75$	$s = -1$
No	0	24.25	194.57	194.57	194.57	194.57	194.57
No	0.1	18.61	9.46	19.62	21.63	17.38	14.09
No	0.5	5.95	-14.57	-3.05	7.54	16.29	18.12
Yes	0.1	18.61	39.03	39.49	39.85	40.16	40.39
Yes	0.5	5.95	27.67	27.99	28.23	28.39	28.44

**Fig. 7** Illustrating how the PSNR value depends on different H^s norms for deblurring the Square image without regularization. The optimal s varies with the noise intensity. For a larger noise variance, it is preferable to select a weaker norm (corresponding to a smaller s)

under a variety of noise levels and whether the TV regularization term is included in the objective function or not. We remark that the noise here is high-frequency Gaussian noise. The PSNR values in different settings of deblurring the Square image are recorded in Table 1.

The first row of Table 1 is about the reconstruction without using TV from noise-free data, i.e., $\sigma = 0$. All the PSNR values are all over 190, which implies the perfect recovery (subject to numerical round-off errors). In this noise-free case, the reconstruction is a standard (weighted) least-squares solution. Furthermore, the choice of the data-fitting term does not affect the minimizer of the optimization problem, though the convergence rate may differ. As seen in Fig. 1, the same number of gradient descent iterations yields different sharpness when s varies.

Still without the regularization term, we examine the results using the noisy blurry data and record the PSNR values in the second and the third rows of Table 1. These quantitative values reflect that the reconstruction results after a fixed number of gradient descent iterations (11) differ drastically with respect to different s values, as also illustrated in Fig. 2. We plot the PSNR values with more s values in Fig. 7 than those documented in Table 1, which further illustrates that the optimal choice of s depends on the noise level.

The effect of the TV regularization is presented in the last two rows of Table 1. On one hand, TV significantly improves the results over the model without TV. On the other hand, using the optimal H^s norm as the data-fitting term together with TV outperforms the classic TV with the L^2 norm, as the former has an extra degree of freedom.

We further test on two images: Circles and Cameraman, for image deblurring. The blurring kernel is fixed as a 7×7 Gaussian function with the standard deviation of

Table 2 We fix $\mu = 1$ and list other optimal parameters for TV and the proposed method

Test image	Noise level σ	TV		proposed		
		λ	ρ	s	λ	ρ
Circles	0.1	10	200	-1.6	18	380
	0.2	5	500	-2.5	8	500
Cameraman	0.1	40	2,500	-0.2	40	50
	0.2	12.5	50	0.5	10	50

Table 3 Image deblurring comparison in terms of PSNR

Test image	σ	Input	TV	Hyper	BM3D	WAI	Proposed
Circles	0.1	19.78	32.56	30.61	32.52	31.96	32.93
	0.2	13.91	29.84	28.10	29.97	29.78	30.03
Cameraman	0.1	18.96	24.52	24.54	25.49	24.40	24.53
	0.2	13.65	22.89	22.75	23.53	22.92	22.96

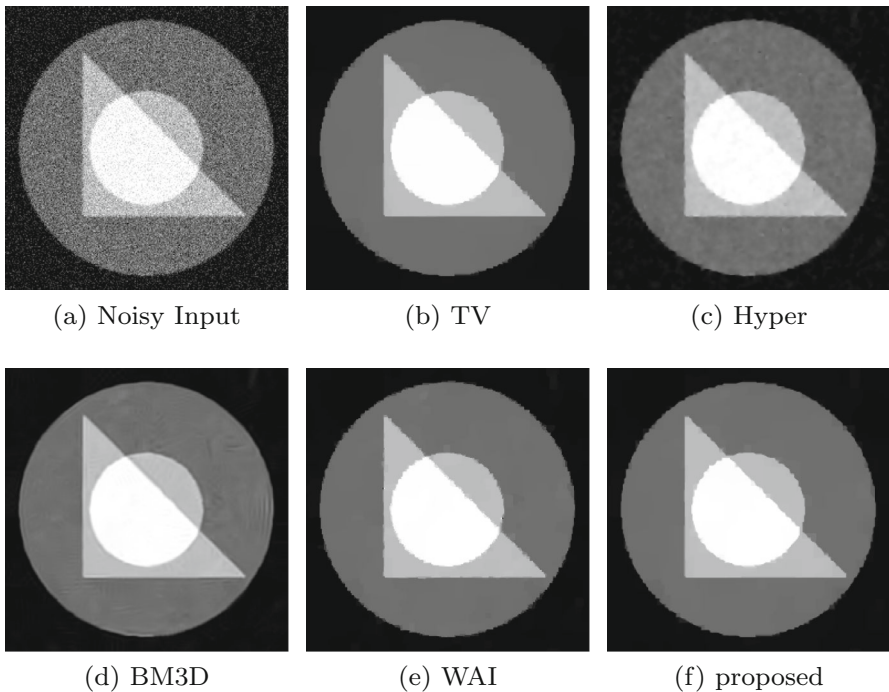


Fig. 8 Comparison of deblurring the Circles image with a 7×7 Gaussian blur and additive Gaussian noise of $\sigma = 0.1$

1. By assuming the periodic boundary condition and using the Convolution Theorem, the linear operator A can be implemented by FFT. We also consider two noise levels: $\sigma = 0.1$ and 0.2 as the standard deviation of the additive Gaussian random noise. We compare the proposed approach H^s+TV with TV, a hyper-Laplacian model (Hyper)

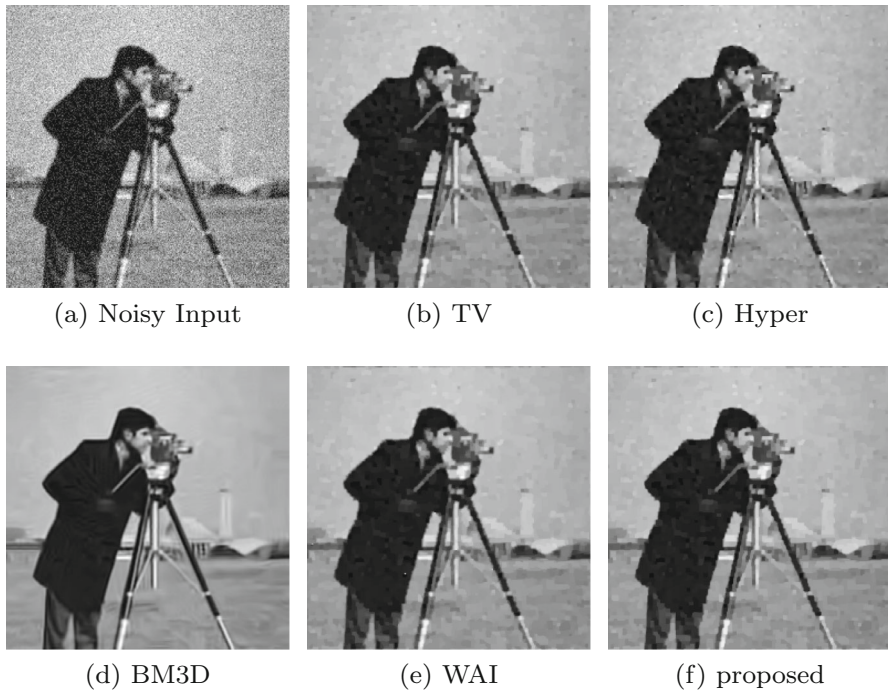


Fig. 9 Comparison of deblurring the Cameraman image with 7×7 Gaussian blur and additive Gaussian noise of $\sigma = 0.1$

[36], a modification of BM3D from denoising to deblurring [21], and a weighted anisotropic and isotropic (WAI) regularization proposed in [43]. We use the online codes of the competing methods: Hyper, BM3D, and WAI with their default parameter choices. For TV and the proposed approach, we fix $\mu = 1$ and tune the parameters of λ , ρ , and s so that they can achieve the highest PSNR for each combination of testing image and noise level. Typically, the stronger the noise, the smaller the optimal s value. However, for the Cameraman image polluted by the large noise, the optimal s value is positive to counterbalance the smoothing incurred by the TV regularization term. We provide the optimal parameter values in Table 2 and record the PSNR values in Table 3. Visually we present the deblurring results under a lower noise level ($\sigma = 0.1$) in Fig. 8-9. The proposed approach works particularly well for images with simple geometries such as Circles, and is comparable to the state-of-the-art deblurring methods for the Cameraman image. Motivated by the BM3D algorithm, we think using a state-dependent s instead of a fixed global s can improve the performance.

6 Conclusions

In this paper, we proposed a novel idea of using the Sobolev (H^s) norms as a data fidelity term for imaging applications. We revealed implicit regularization effects offered by the proposed data-fitting term rather than the commonly used regularization

term. Specifically, we shall choose a weak norm ($s < 0$) for high-frequency noises and a strong norm ($s > 0$) for low-frequency noises. The more oscillatory the high-frequency noise, the smaller the s . The smoother the low-frequency noise, the larger the s . This rule of thumb helps us choose an initial guess for the parameter s , which is further locally tuned in the experiment. We discussed the connections between the Sobolev norm and the Sobolev gradient flow. From a Bayesian inference perspective, we analyzed the underlying noise assumption for a Sobolev norm as the data fidelity term. We further revealed that one could choose a proper Sobolev norm as an objective function to improve the convergence rate in gradient descent, achieving preconditioning effects. We presented three numerical schemes to compute the H^s norms under different domains and boundary conditions. Experimental results showed that the H^s data-fitting term alone as the objective function has implicit regularization effects on the performance of various inverse problems. Furthermore, the H^s data-fitting term combined with the TV regularization, i.e., H^s+TV , works particularly well for images with simple geometries and always outperforms the standard L^2+TV . In the framework of ADMM, one can efficiently minimize the H^s+TV model with a tunable parameter s .

References

1. An, L.T.H., Tao, P.D.: The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Ann. Oper. Res.* **133**(1), 23–46 (2005)
2. Arbogast, T., Bona, J.L.: *Methods of applied mathematics*, The University of Texas at Austin. Lecture Notes in Applied Mathematics (2008)
3. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: *International Conference on Machine Learning*, pp. 214–223. PMLR (2017)
4. Aubert, G., Aujol, J.F.: A variational approach to removing multiplicative noise. *SIAM J. Appl. Math.* **68**(4), 925–946 (2008)
5. Bal, G.: *Introduction to Inverse Problems*. Lecture Notes-Department of Applied Physics and Applied Mathematics, Columbia University, New York (2012)
6. Benassi, A., Roux, D., Jaffard, S.: Elliptic gaussian random processes. *Rev. Matemática Iberoamericana* **13**(1), 19–90 (1997)
7. Benfatto, G., Gallavotti, G., Nicolò, F.: Elliptic equations and gaussian processes. *J. Funct. Anal.* **36**(3), 343–400 (1980)
8. Bolin, D., Kirchner, K.: The rational SPDE approach for Gaussian random fields with general smoothness. *J. Comput. Graph. Stat.* **29**(2), 274–285 (2020)
9. Boyd, S., Parikh, N., Chu, E.: *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. Now Publishers Inc, Hanover (2011)
10. Bredies, K., Kunisch, K., Pock, T.: Total generalized variation. *SIAM J. Imaging Sci.* **3**(3), 492–526 (2010)
11. Buades, A., Coll, B., Morel, J.M.: A review of image denoising algorithms, with a new one. *Multiscale Model. Simul.* **4**(2), 490–530 (2005)
12. Bungert, L., Burger, M., Korolev, Y., Schönlieb, C.B.: Variational regularisation for inverse problems with imperfect forward operators and general noise models. *Inverse Prob.* **36**(12), 125014 (2020)
13. Bunks, C., Saleck, F.M., Zaleski, S., Chavent, G.: Multiscale seismic waveform inversion. *Geophysics* **60**(5), 1457–1473 (1995)
14. Calder, J., Mansouri, A., Yezzi, A.: Image sharpening via Sobolev gradient flows. *SIAM J. Imaging Sci.* **3**(4), 981–1014 (2010)
15. Chambolle, A.: An algorithm for total variation minimization and applications. *J. Math. Imaging Vis.* **20**, 89–97 (2004)
16. Chan, T.F., Golub, G.H., Mulet, P.: A nonlinear primal-dual method for total variation-based image restoration. *SIAM J. Sci. Comput.* **20**(6), 1964–1977 (1999)

17. Chowdhury, M.R., Qin, J., Lou, Y.: Non-blind and blind deconvolution under Poisson noise using fractional-order total variation. *J. Math. Imaging Vis.* **62**(9), 1238–1255 (2020)
18. Chowdhury, M.R., Zhang, J., Qin, J., Lou, Y.: Poisson image denoising based on fractional-order total variation. *Inverse Probl. Imaging* **14**(1), 77–96 (2020)
19. Cicone, A., Huska, M., Kang, S.H., Morigi, S.: Jot: a variational signal decomposition into jump, oscillation and trend. In: *IEEE Transactions on Signal Processing* (2022 to appear)
20. Claerbout, J.F.: Toward a unified theory of reflector mapping. *Geophysics* **36**(3), 467–481 (1971)
21. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image restoration by sparse 3d transform-domain collaborative filtering. In: *Image Processing: Algorithms and Systems VI*, Vol. 6812, p. 681207. International Society for Optics and Photonics (2008)
22. Dashti, M., Stuart, A.M.: *The Bayesian Approach to Inverse Problems*, pp. 311–428. Springer, Cham (2017)
23. Dunlop, M.M., Yang, Y.: Stability of Gibbs posteriors from the wasserstein loss for Bayesian full waveform inversion. arXiv preprint [arXiv:2004.03730](https://arxiv.org/abs/2004.03730) (2020)
24. Elbakri, I.A., Fessler, J.A.: Statistical image reconstruction for polyenergetic x-ray computed tomography. *IEEE Trans. Med. Imaging* **21**(2), 89–99 (2002)
25. Engl, H.W., Hanke, M., Neubauer, A.: *Regularization of Inverse Problems*, vol. 375. Springer, New York (1996)
26. Engquist, B., Ren, K., Yang, Y.: The quadratic Wasserstein metric for inverse data matching. *Inverse Prob.* **36**(5), 055001 (2020)
27. Esser, E., Zhang, X., Chan, T.F.: 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)
28. Evans, L.C.: *Partial Differential Equations*. American Mathematical Society, Providence, RI (1998)
29. Giga, M.H., Giga, Y.: Very singular diffusion equations: second and fourth order problems. *Jpn. J. Ind. Appl. Math.* **27**(3), 323–345 (2010)
30. Giga, Y., Muszkieta, M., Rybka, P.: A duality based approach to the minimizing total variation flow in the space H^{-s} . *Jpn. J. Ind. Appl. Math.* **36**(1), 261–286 (2019)
31. Glowinski, R., Marroco, A.: Sur l’approximation, par éléments finis d’ordre un, et la résolution, par pénalisation-dualité d’une classe de problèmes de dirichlet non linéaires. *ESAIM* **9**(R2), 41–76 (1975)
32. Goldstein, T., Osher, S.: The split Bregman method for L^1 -regularized problems. *SIAM J. Imaging Sci.* **2**(2), 323–343 (2009)
33. Huska, M., Kang, S.H., Lanza, A., Morigi, S.: A variational approach to additive image decomposition into structure, harmonic, and oscillatory components. *SIAM J. Imaging Sci.* **14**(4), 1749–1789 (2021)
34. Kak, A.C., Slaney, M.: *Principles of Computerized Tomographic Imaging*. SIAM, New York (2001)
35. Kim, Y., Vese, L.A.: Image recovery using functions of bounded variation and Sobolev spaces of negative differentiability. *Inverse Probl. Imaging* **3**(1), 43 (2009)
36. Krishnan, D., Fergus, R.: Fast image deconvolution using hyper-Laplacian priors. *Adv. Neural. Inf. Process. Syst.* **22**, 1033–1041 (2009)
37. Le, T., Chartrand, R., Asaki, T.J.: A variational approach to reconstructing images corrupted by Poisson noise. *J. Math. Imaging Vis.* **27**(3), 257–263 (2007)
38. Li, Z., Lou, Y., Zeng, T.: Variational multiplicative noise removal by DC programming. *J. Sci. Comput.* **68**(3), 1200–1216 (2016)
39. Lieu, L.H., Vese, L.A.: Image restoration and decomposition via bounded total variation and negative Hilbert-Sobolev spaces. *Appl. Math. Optim.* **58**(2), 167–193 (2008)
40. Liu, J., Lou, Y., Ni, G., Zeng, T.: An image sharpening operator combined with framelet for image deblurring. *Inverse Probl.* **36**(4), 045015 (2020)
41. Lodhia, A., Sheffield, S., Sun, X., Watson, S.S.: Fractional gaussian fields: a survey. *Probab. Surv.* **13**, 1–56 (2016)
42. Lou, Y., Kang, S.H., Soatto, S., Bertozzi, A.L.: Video stabilization of atmospheric turbulence distortion. *Inverse Probl. Imaging* **7**(3), 839 (2013)
43. Lou, Y., Zeng, T., Osher, S., Xin, J.: A weighted difference of anisotropic and isotropic total variation model for image processing. *SIAM J. Imaging Sci.* **8**(3), 1798–1823 (2015)
44. Lou, Y., Zhang, X., Osher, S., Bertozzi, A.: Image recovery via nonlocal operators. *J. Sci. Comput.* **42**(2), 185–197 (2010)
45. Neuberger, J.: *Sobolev Gradients and Differential Equations*. Springer, New York (2009)
46. Nocedal, J., Wright, S.: *Numerical Optimization*. Springer, New York (2006)

47. Osher, S., Burger, M., Goldfarb, D., Xu, J., Yin, W.: An iterated regularization method for total variation-based image restoration. *Multiscale Model. Simul.* **4**, 460–489 (2005)
48. Osher, S., Solé, A., Vese, L.: Image decomposition and restoration using total variation minimization and the H^{-1} norm. *Multiscale Model. Simul.* **1**(3), 349–370 (2003)
49. Otto, F., Villani, C.: Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal.* **173**(2), 361–400 (2000)
50. Papadakis, N., Peyré, G., Oudet, E.: Optimal transport with proximal splitting. *SIAM J. Imaging Sci.* **7**(1), 212–238 (2014)
51. Peyre, R.: Comparison between W_2 distance and \dot{H}^{-1} norm, and localization of Wasserstein distance. *ESAIM Control Optim. Calc. Var.* **24**(4), 1489–1501 (2018)
52. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. *Physica D* **60**(1–4), 259–268 (1992)
53. Schechter, M.: Negative norms and boundary problems. *Ann. Math.* **72**, 581–593 (1960)
54. Schönlieb, C.B.: *Partial Differential Equation Methods for Image Inpainting*, vol. 29. Cambridge University Press, Cambridge (2015)
55. Sobolev, S.L.: *Applications of Functional Analysis in Mathematical Physics*, vol. 7. American Mathematical Society, Washington, DC (1963)
56. Sundaramoorthi, G., Yezzi, A., Mennucci, A.C.: Sobolev active contours. *Int. J. Comput. Vis.* **73**(3), 345–366 (2007)
57. Szabó, B., Babuska, I.: *Finite Element Analysis*. Wiley, New York (1991)
58. Thibault, J.B., Sauer, K.D., Bouman, C.A., Hsieh, J.: A three-dimensional statistical approach to improved image quality for multislice helical ct. *Med. Phys.* **34**(11), 4526–4544 (2007)
59. Tikhonov, A.N.: On the stability of inverse problems. *Dokl. Akad. Nauk SSSR* **39**, 195–198 (1943)
60. Vardi, Y., Shepp, L., Kaufman, L.: A statistical model for positron emission tomography. *J. Am. Stat. Assoc.* **80**(389), 8–20 (1985)
61. Vese, L.A., Le Guyader, C.: *Variational Methods in Image Processing*. CRC Press, Boca Raton (2016)
62. Villani, C.: *Topics in Optimal Transportation*, vol. 58. American Mathematical Society, Providence, RI (2003)
63. Virieux, J., Operto, S.: An overview of full-waveform inversion in exploration geophysics. *Geophysics* **74**(6), WCC16–WCC2 (2009)
64. Wang, G., Wei, Y., Qiao, S., Lin, P., Chen, Y.: *Generalized Inverses: Theory and Computations*, vol. 53. Springer, New York (2018)
65. Yang, Y., Engquist, B., Sun, J., Hamfeldt, B.F.: Application of optimal transport and the quadratic Wasserstein metric to full-waveform inversion. *Geophysics* **83**(1), R43–R62 (2018)
66. Yang, Y., Nurbekyan, L., Negrini, E., Martin, R., Pasha, M.: Optimal transport for parameter identification of chaotic dynamics via invariant measures. *SIAM J. Appl. Dyn. Syst.* **22**(1), 269–310 (2023)
67. Yang, Y., Townsend, A., Appelö, D.: Anderson acceleration based on the H^{-s} Sobolev norm for contractive and noncontractive fixed-point operators. *J. Comput. Appl. Math.* **403**, 113844 (2022)
68. Zhang, J., Chen, K.: A total fractional-order variation model for image restoration with nonhomogeneous boundary conditions and its numerical solution. *SIAM J. Imaging Sci.* **8**(4), 2487–2518 (2015)
69. Zhang, X., Burger, M., Bresson, X., Osher, S.: Bregmanized nonlocal regularization for deconvolution and sparse reconstruction. *SIAM J. Imaging Sci.* **3**(3), 253–276 (2010)
70. Zhang, Y., Sun, J.: Practical issues in reverse time migration: True amplitude gathers, noise removal and harmonic source encoding. *First break* **27**(1) (2009)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.