IEEE Open Journal of Signal Processing

Received 27 December 2021; revised 3 March 2022; accepted 3 March 2022. Date of publication 7 March 2022; date of current version 28 March 2022. The review of this article was arranged by Associate Editor Angshul Majumdar.

Digital Object Identifier 10.1109/OJSP.2022.3157082

Sparsest Univariate Learning Models Under Lipschitz Constraint

SHAYAN AZIZNEJAD, THOMAS DEBARRE, AND MICHAEL UNSER (Fellow, IEEE)

Biomedical Imaging Group, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

CORRESPONDING AUTHOR: SHAYAN AZIZNEJAD (e-mail: shayan.aziznejad@epfl.ch). *(Shayan Aziznejad and Thomas Debarre contributed equally to this work.)*

This work was supported in part by the European Research Council (ERC Project FunLearn) under Grant 101020573 and in part by the Swiss National Science Foundation under Grant 200020_184646/1.

ABSTRACT Beside the minimization of the prediction error, two of the most desirable properties of a regression scheme are *stability* and *interpretability*. Driven by these principles, we propose continuous-domain formulations for one-dimensional regression problems. In our first approach, we use the Lipschitz constant as a regularizer, which results in an implicit tuning of the overall robustness of the learned mapping. In our second approach, we control the Lipschitz constant explicitly using a user-defined upper-bound and make use of a sparsity-promoting regularizer to favor simpler (and, hence, more interpretable) solutions. The theoretical study of the latter formulation is motivated in part by its equivalence, which we prove, with the training of a Lipschitz-constrained two-layer univariate neural network with rectified linear unit (ReLU) activations and weight decay. By proving representer theorems, we show that both problems admit global minimizers that are continuous and piecewise-linear (CPWL) functions. Moreover, we propose efficient algorithms that find the sparsest solution of each problem: the CPWL mapping with the least number of linear regions. Finally, we illustrate numerically the outcome of our formulations.

INDEX TERMS Robust learning, sparsity, Lipschitz regularity, continuous and piecewise linear functions, representer theorems.

I. INTRODUCTION

The goal of a regression model is to learn a mapping $f : \mathcal{X} \to \mathcal{Y}$ from a collection of data points $(x_m, y_m) \in \mathcal{X} \times \mathcal{Y}$, $m = 1, \ldots, M$, such that $y_m \approx f(x_m)$, while avoiding the problem of overfitting [1]–[3]. Here, $\mathcal{X}$ denotes the input domain and $\mathcal{Y}$ is the set of possible outcomes. A common way of carrying out this task is to solve a minimization problem of the form

$$\min_{f \in \mathcal{F}} \left(\sum_{m=1}^{M} E\left(f(x_m), y_m\right) + \mathcal{R}(f) \right), \tag{1}$$

where $\mathcal{F}$ is the underlying search space, the convex loss function $E : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ enforces the consistency of the learned mapping with the given data points, and the regularization functional $\mathcal{R} : \mathcal{F} \to \mathbb{R}_{\geq 0}$ injects prior knowledge on the form of the mapping f, which is designed to alleviate the problem of overfitting.

A. NONPARAMETRIC REGRESSION

In some cases, the optimization can be performed over an infinite-dimensional function space—see, for example, [4]–[6] for applications in signal and image processing. A prominent example is the family of reproducing-kernel Hilbert spaces (RKHS) $\mathcal{F} = \mathcal{H}(\mathbb{R}^d)$, $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$ [7], [8], in which the regression problem is formulated as

$$\min_{f \in \mathcal{H}(\mathbb{R}^d)} \left(\sum_{m=1}^{M} E\left(f(\boldsymbol{x}_m), y_m\right) + \lambda \|f\|_{\mathcal{H}}^2 \right). \tag{2}$$

The fundamental result in RKHS theory is the kernel representer theorem [9], [10], which states that the unique solution of (2) admits the kernel expansion

$$f(\cdot) = \sum_{m=1}^{M} a_m \mathrm{k}(\cdot, \boldsymbol{x}_m), \tag{3}$$

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

where $\mathrm{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the unique reproducing kernel of $\mathcal{H}(\mathbb{R}^d)$ and $a_m \in \mathbb{R}, m = 1, \ldots, M$, are learnable parameters. The expansion (3) allows one to recast the infinite-dimensional problem (2) into a finite-dimensional one and to use standard computational tools of convex analysis to solve it. Many classical kernel-based schemes are based on this approach, including support-vector machines and radial-basis functions [11]–[13].

B. PARAMETRIC REGRESSION

In cases when (1) cannot be recast as a finite-dimensional optimization problem, another common approach is to restrict the search space $\mathcal{F}$ to a subspace that admits a parametric representation. This approach is used in deep neural networks (DNNs), which have become a prominent tool in machine learning and data science in recent years [14], [15]. They outperform classical kernel-based methods for various image-processing tasks. In particular, they have become state-of-the-art for image classification [16], inverse problems [17], and image segmentation [18]. However, most published works are empirical, and the outstanding performance of DNNs is yet to be fully understood. To this end, many recent works are directed towards studying DNNs from a theoretical perspective. Unsurprisingly, *stability* and *interpretability*, which are key principles in machine learning, play a central role in these works. For example, the stability of state-of-the-art deep-learning-based methods has been dramatically challenged in image classification [19], [20] and image reconstruction [21]. Attempts have also been made to understand and interpret DNNs from different perspectives, such as rate-distortion theory [22], [23]). However, the community is still far from reaching a global understanding and these questions are still active areas of research.

C. OUR CONTRIBUTIONS

In this paper, we introduce two variational formulations for regressing one-dimensional data that favor "stable" and "simple" regression models. Similar to RKHS theory, the latter are nonparametric continuous-domain problems in the sense that $\mathcal{F}$ in (1) is an infinite-dimensional function space. Inspired by the stability principle, we focus on the development of regression schemes with controlled Lipschitz regularity. This is motivated by the observation that many analyses in deep learning require assumptions on the Lipschitz constant of the learned mapping [24]–[26]. Likewise, in the context of so-called "plug-and-play" methods—*i.e.,* when a trainable module is inserted into an iterative-reconstruction framework—, the rate of convergence of the overall scheme often depends on the Lipschitz constant of this module [27]–[32].

In our first formulation, we use the Lipschitz constant of the learned mapping as a regularization term. Specifically, we consider the minimization problem

$$\min_{f \in \mathrm{Lip}(\mathbb{R})} \left(\sum_{m=1}^{M} E\left(f(x_m), y_m\right) + \lambda L(f) \right), \tag{4}$$

where $\mathrm{Lip}(\mathbb{R})$ is the space of Lipschitz-continuous real functions and $L(f)$ denotes the Lipschitz constant of $f \in \mathrm{Lip}(\mathbb{R})$. In this formulation, one can implicitly control the Lipschitz regularity of the learned function by varying the regularization parameter λ. We prove a representer theorem that characterizes the solution set of (4). In particular, we prove that the global minimum is achieved by a continuous and piecewise-linear (CPWL) mapping. Next, motivated by the simplicity principle, we find the mapping with the minimal number of linear regions. Note that many previous works study problems similar to (4) in more general settings, typically using the Lipschitz constant of the nth derivative $\mathcal{R}(f) = L(f^{(n)})$ with $n \geq 0$ as the regularization term [33]–[38]. More recently, [39] has studied the classification problem over metric spaces and derived a parametric form for a solution of this problem. Our work complements this interesting line of research by providing an in-depth analysis of the $n = 0$ case which is related to second-order total-variation minimization, and by focusing more on computational aspects of (4). More precisely, we propose a two-step algorithm to reach the sparsest CPWL solution of (4). The first step consists in solving a discrete problem with ℓ_∞ regularization, and the second is a sparsification step proposed in [40] that reaches the sparsest solution.

In the second scenario, we explictly control the Lipschitz constant of the learned mapping by imposing a hard constraint. Inspired by the theoretical insights of the first problem, we add a second-order total-variation (TV) regularization term that is known to promote sparse CPWL functions [40], [41]. This leads to the minimization problem

$$\min_{f \in \mathrm{BV}^{(2)}(\mathbb{R})} \left(\sum_{m=1}^{M} E\left(f(x_m), y_m\right) + \lambda \mathrm{TV}^{(2)}(f) \right), \quad \text{s.t.} \quad L(f) \leq \overline{L}, \tag{5}$$

where $\mathrm{BV}^{(2)}(\mathbb{R})$ is the space of functions with bounded second-order TV and $\overline{L}$ is the user-defined upper-bound for the desired Lipschitz regularity of the learned mapping. The interesting aspect of (5) is that the simplicity and stability of the learned mapping can be adjusted by tuning the parameters $\lambda > 0$ and $\overline{L} > 0$, respectively. In this case as well, we prove a representer theorem which guarantees the existence of CPWL solutions. We propose a two-step algorithm to find the sparsest CPWL solution which is similar to that of the first scenario. The main difference is the first step, where the discrete problem has a ℓ_1 regularization term and a ℓ_∞ constraint.

D. CONNECTION TO NEURAL NETWORKS

Another major motivation for this work is to further elucidate the tight connection between CPWL functions and neural networks. It is well known that the input-output mapping of any feed-forward DNN with linear spline (*e.g.,* the rectified linear unit, also known as ReLU) activations is a CPWL function [42], [43]. This is due to the fact that these mappings are compositions of affine transformations and pointwise activations. Hence, since the ReLU activation is itself a CPWL

function and the CPWL property of functions is preserved by composition, the full input-output mapping is CPWL. Conversely, any CPWL function can be represented *exactly* by a DNN with linear-spline activations [44]. This establishes a direct link with spline theory, as first highlighted by Poggio *et al.* [45] and then further explored in various works [41], [46]–[49].

When it comes to shallow networks, the connection with our framework becomes even more explicit. It is well known in the literature that the standard training (*i.e.,* with weight decay) of a two-layer univariate ReLU network is equivalent to solving a TV-based variational problem such as (5) without the Lipschitz constraint [47], [50]. Specifically, the weight-decay penalty can be shown to be equal to the second-order TV of the input-output mapping of the full network at the optimum [47, Proposition 18]. As we demonstrate, these results can be readily extended to prove the equivalence between the training of a Lipschitz-constrained two-layer univariate ReLU network and our formulation (5). Our description of the solution set of Problem (5) thus provides insights into the training of Lipschitz-aware neural networks.

E. OUTLINE

The paper is organized as follows: we review the required mathematical background in Section II. In Section III, we introduce our supervised-learning formulations and we state their corresponding representer theorems. We then propose our algorithms for finding the corresponding sparsest CPWL solution in Section IV. Finally, we provide numerical illustrations and discussions in Section V.

II. MATHEMATICAL PRELIMINARIES

A. WEAK DERIVATIVES

Schwartz' space of smooth and compactly supported test functions is denoted by $\mathcal{D}(\mathbb{R})$. It is known that the nth-order derivative is a continuous mapping over $\mathcal{D}(\mathbb{R})$, which we denote as $\mathrm{D}^n : \mathcal{D}(\mathbb{R}) \to \mathcal{D}(\mathbb{R})$ [51]. By duality, this allows one to extend the derivative operator to the whole class $\mathcal{D}'(\mathbb{R})$ of distributions. The extended operator is called the nth-order weak derivative and will be denoted by $\mathrm{D}^n : \mathcal{D}'(\mathbb{R}) \to \mathcal{D}'(\mathbb{R})$. For any $w \in \mathcal{D}'(\mathbb{R})$, the distribution $\mathrm{D}^n\{w\} \in \mathcal{D}'(\mathbb{R})$ is defined via its action on a generic test function $\varphi \in \mathcal{D}(\mathbb{R})$ as $\langle \mathrm{D}^n w, \varphi \rangle = (-1)^n \langle w, \mathrm{D}^n \varphi \rangle$. The fundamental property is that the weak derivative of any Schwartz test function $\varphi \in \mathcal{D}(\mathbb{R}) \subseteq \mathcal{D}'(\mathbb{R})$ is well-defined and coincides with the classical notion of derivative (see [52, Section 3.3.2.] for more details on the extension by duality).

B. BANACH SPACES

A Banach space is a normed topological vector space that is complete in its norm topology. The prototypical examples of Banach spaces are $L_p(\mathbb{R})$ for $p \in [1, +\infty]$ which are the spaces of Lebesgue measurable functions with finite L_p norm. For $p \neq +\infty$, this reads as

$$L_p(\mathbb{R}) = \left\{ f : \mathbb{R} \to \mathbb{R} \text{ measurable: } \|f\|_{L_p} < +\infty \right\}, \quad (6)$$

where $\|f\|_{L_p} = \left(\int_{\mathbb{R}} |f(x)|^p \mathrm{d}x\right)^{\frac{1}{p}}$. Alternatively, one can define $L_p(\mathbb{R}) = \overline{(\mathcal{D}(\mathbb{R}), \|\cdot\|_{L_p})}$ as the completion of $\mathcal{D}(\mathbb{R})$ with respect to the L_p norm for $p \in [1, +\infty)$. The case $p = +\infty$ is particular. Indeed, the L_∞ norm is defined as $\|f\|_{L_\infty} = \operatorname{ess\,sup}_{x \in \mathbb{R}} |f(x)|$, where the essential supremum extracts an upper-bound that is valid almost everywhere. Contrarily to the other L_p spaces, the space $\mathcal{D}(\mathbb{R})$ is not dense in $L_\infty(\mathbb{R})$; in fact, the completion of $\mathcal{D}(\mathbb{R})$ with respect to the L_∞ norm is the space $\mathcal{C}_0(\mathbb{R})$ of continuous functions that vanish at infinity [53].

Finally, we denote the space of bounded Radon measures by $\mathcal{M}(\mathbb{R})$. Following the Riesz-Markov theorem, we view $\mathcal{M}(\mathbb{R})$ as the continuous dual of $\mathcal{C}_0(\mathbb{R})$. This allows us to define the total-variation norm over this space as [54]

$$\|w\|_{\mathcal{M}} = \sup_{\substack{\varphi \in \mathcal{C}_0(\mathbb{R}) \\ \|\varphi\|_{L_\infty} = 1}} |\langle w, \varphi \rangle| = \sup_{\substack{\varphi \in \mathcal{D}(\mathbb{R}) \\ \|\varphi\|_{L_\infty} = 1}} |\langle w, \varphi \rangle|, \quad (7)$$

where the last equality follows from the denseness of $\mathcal{D}(\mathbb{R})$ in $\mathcal{C}_0(\mathbb{R})$. Interestingly, the total-variation norm is a generalization of the L_1 norm. In fact, the space $L_1(\mathbb{R})$ is included in $\mathcal{M}(\mathbb{R})$ and, for any function $f \in L_1(\mathbb{R})$, we have that $\|f\|_{L_1} = \|f\|_{\mathcal{M}}$. Moreover, the space $\mathcal{M}(\mathbb{R})$ contains shifted Dirac impulses with $\|\delta(\cdot - x_0)\|_{\mathcal{M}} = 1$. Finally, for any absolutely summable sequence $\boldsymbol{a} = (a_n) \in \ell_1(\mathbb{Z})$ and distinct locations x_n, $n \in \mathbb{Z}$, we have that

$$w_{\boldsymbol{a}} = \sum_{n \in \mathbb{Z}} a_n \delta(\cdot - x_n) \in \mathcal{M}(\mathbb{R}) \quad \text{and} \quad \|w_{\boldsymbol{a}}\|_{\mathcal{M}} = \|\boldsymbol{a}\|_{\ell_1}. \quad (8)$$

This property establishes a tight link between the total-variation norm and the discrete ℓ_1 norm which is known to promote sparsity and is the key element in the field of compressed sensing [55]–[57]. This enabled researchers to interpret the total-variation norm as a sparsity-promoting norm in the continuous domain. Since then, additional connections have been drawn between optimization problems that involve the total-variation norm and many areas of research such as super resolution [58]–[60], kernel methods, [61], [62], and splines [48], [63]–[66]. The computational aspects of this framework have also been investigated, leading to the development of practical algorithms in various settings [67]–[69].

C. LIPSCHITZ CONSTANT

We denote by $\mathrm{Lip}(\mathbb{R})$, the space of Lipschitz-continuous functions $f : \mathbb{R} \to \mathbb{R}$ with a finite Lipschitz constant, satisfying

$$L(f) = \sup_{x_1 \neq x_2} \frac{|f(x_1) - f(x_2)|}{|x_1 - x_2|} < +\infty. \quad (9)$$

Following Rademacher's theorem, any Lipschitz-continuous function $f \in \mathrm{Lip}(\mathbb{R})$ is differentiable almost everywhere with a measurable and essentially bounded derivative. The Lipschitz constant of the function then corresponds to the essential supremum of its derivative, so that

$$L(f) = \|\mathrm{D}\{f\}\|_{L_\infty} = \operatorname*{ess\,sup}_{x \in \mathbb{R}} |f'(x)|. \quad (10)$$

Conversely, any distribution $f \in \mathcal{D}'(\mathbb{R})$ whose weak derivative lies in $L_\infty(\mathbb{R})$ is indeed a Lipschitz-continuous function [70, Theorem 1.36]. In other words, we have that

$$\mathrm{Lip}(\mathbb{R}) = \{f \in \mathcal{D}'(\mathbb{R}) : \mathrm{D}\{f\} \in L_\infty(\mathbb{R})\}, \tag{11}$$

which allows us to view $\mathrm{Lip}(\mathbb{R})$ as the native Banach space associated to the pair $(L_\infty(\mathbb{R}), \mathrm{D})$ in the sense of [71].

D. SECOND-ORDER TOTAL-VARIATION

To conclude this section, we introduce the space $\mathrm{BV}^{(2)}(\mathbb{R})$ of functions with finite second-order total-variation, defined as

$$\mathrm{TV}^{(2)}(f) = \|\mathrm{D}^2\{f\}\|_{\mathcal{M}} = \sup_{\substack{\varphi \in \mathcal{D}(\mathbb{R}) \\ \|\varphi\|_{L_\infty}=1}} \langle \mathrm{D}^2 f, \varphi \rangle \tag{12}$$

$$= \sup_{\substack{\varphi \in \mathcal{D}(\mathbb{R}) \\ \|\varphi\|_{L_\infty}=1}} \int_{\mathbb{R}} f(x)\varphi''(x)\mathrm{d}x. \tag{13}$$

Analogous to the famous total-variation regularization of Rudin-Osher-Fatemi [72], which promotes piecewise-constant functions and causes the notorious staircase effect, the second-order total variation favors CPWL functions. In dimension $d = 1$, this coincides with the known class of nonuniform linear splines which has been extensively studied from an approximation-theoretical point of view [73], [74]. Motivated by this, the $\mathrm{TV}^{(2)}$ regularization has been exploited to learn activation functions of deep neural networks [41], [75]. In a similar vein, the identification of the sparsest CPWL solutions of $\mathrm{TV}^{(2)}$-regularized problems has been thoroughly studied in [40].

III. LIPSCHITZ-AWARE FORMULATIONS FOR SUPERVISED LEARNING

We now introduce our formulations for supervised learning that are based on controlling the Lipschitz constant of the learned mapping. Let us first mention that the Lipschitz constant can be indirectly controlled using a $\mathrm{TV}^{(2)}$-type regularizer. Indeed, the two seminorms are connected, as demonstrated in Theorem 1.

Theorem 1: Any function with second-order bounded-variation is Lipschitz continuous. Moreover, for any $f \in \mathrm{BV}^{(2)}(\mathbb{R})$, we have the upper-bound

$$L(f) \leq \mathrm{TV}^2(f) + \ell(f) \tag{14}$$

for the Lipschitz constant of f, where

$$\ell(f) = \inf_{x_1 \neq x_2} \frac{|f(x_1) - f(x_2)|}{|x_1 - x_2|} \geq 0. \tag{15}$$

Finally, (14) is saturated if and only if f is monotone and convex/concave.

The proof of Theorem 1 is given in Appendix A. A weaker version of this theorem is proven in [76], where $\ell(f)$ is replaced with $|f(1) - f(0)|$, which is clearly an upper-bound. The importance of the updated bound is that it is sharp in the sense that it is an equality for monotone and convex/concave functions.

A weaker version of (14) motivated the authors of [76] to provide a global bound for the Lipschitz constant of deep neural networks and to regularize it during training. Although this is an interesting approach to control the Lipschitz constant of the learned mapping, the obtained guarantee is too conservative. This is due to the fact that, as soon as f has some oscillations, the difference between the two sides of (14) dramatically increases and the bound becomes loose. Here, by contrast, we shall ensure the global stability of the learned mapping by directly controlling the Lipschitz constant itself.

A. LIPSCHITZ REGULARIZATION

We first consider the Lipschitz constant as a regularizer and study the minimization problem

$$\mathcal{V}_{\mathrm{Lip}} = \underset{f \in \mathrm{Lip}(\mathbb{R})}{\arg\min} \left(\sum_{m=1}^{M} E(f(x_m), y_m) + \lambda L(f) \right), \tag{16}$$

where $E : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a strictly convex and coercive function and where $\lambda > 0$ is the regularization parameter. We also assume, without loss of generality, that the data points x_m are sorted in the increasing order $x_1 < x_2 < \cdots < x_M$. In Theorem 2, we state our main theoretical contributions regarding the minimization problem (16).

Theorem 2: Regarding the minimization problem (16), the following statements hold.

1) The solution set $\mathcal{V}_{\mathrm{Lip}}$ is a nonempty, convex and weak*-compact subset of $\mathrm{Lip}(\mathbb{R})$.
2) There exists a unique vector $\mathbf{z} = (z_m) \in \mathbb{R}^M$ such that

$$\mathcal{V}_{\mathrm{Lip}} = \underset{f \in \mathrm{Lip}(\mathbb{R})}{\arg\min}\, L(f), \quad \text{s.t.} \quad \mathrm{f}(\mathrm{x_m}) = \mathrm{z_m}, \quad \forall \mathrm{m}. \tag{17}$$

3) The optimal Lipschitz constant has the closed-form expression

$$L_{\min} = \max_{2 \leq m \leq M} \left| \frac{z_m - z_{m-1}}{x_m - x_{m-1}} \right|. \tag{18}$$

Consequently, any $L_{\min}$-Lipschitz function f that satisfies $f(x_m) = z_m, m = 1, \ldots, M$ is a solution of (16).

4) Let $\mathcal{E} \subseteq \mathbb{R}^2$ be the union of the graphs of all solutions of (16), defined as

$$\mathcal{E} = \left\{(x, y) \in \mathbb{R}^2 : \exists f \in \mathcal{V}_{\mathrm{Lip}}, y = f(x)\right\}. \tag{19}$$

Let us also define the right and left planar cones $\mathcal{R}, \mathcal{L} \subseteq \mathbb{R}^2$ as

$$\mathcal{R} = \{\alpha_1(1, L_{\min}) + \alpha_2(1, -L_{\min}) : \alpha_1, \alpha_2 \geq 0\}, \tag{20}$$

and $\mathcal{L} = -\mathcal{R}$. With the convention that $\mathcal{R}_0 = \mathcal{L}_{M+1} = \mathbb{R}^2$, we have that

$$\mathcal{E} = \bigcup_{m=1}^{M+1} (\mathcal{R}_{m-1} \cap \mathcal{L}_m), \tag{21}$$

where the $\mathcal{R}_m$ and $\mathcal{L}_m$ are shifted versions of $\mathcal{R}$ and $\mathcal{L}$, with

$$\mathcal{R}_m = (x_m, z_m) + \mathcal{R}, \quad \mathcal{L}_m = (x_m, z_m) + \mathcal{L}, \forall m. \tag{22}$$

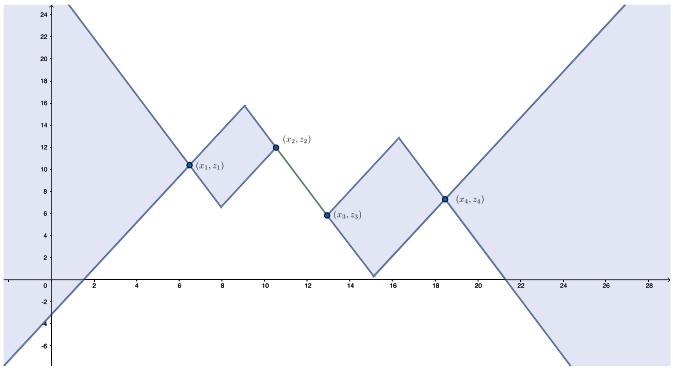


FIGURE 1. The union of the graphs of all solutions in a simple example with four data points. Note that all solutions must directly connect (x_2, z_2) to (x_3, z_3), since the slope of this segment is $L_{\min}$ whose formula is given in (18).

5) Any solution of the constrained minimization problem

$$\min_{f\in \mathrm{BV}^{(2)}(\mathbb{R})} \mathrm{TV}^{(2)}(f), \quad \text{s.t.} \quad f(x_m) = z_m,\ 1 \le m \le M \tag{23}$$

is included in $\mathcal{V}_{\mathrm{Lip}}$. In particular, the solution set of (16) always includes a continuous and piecewise-linear function.

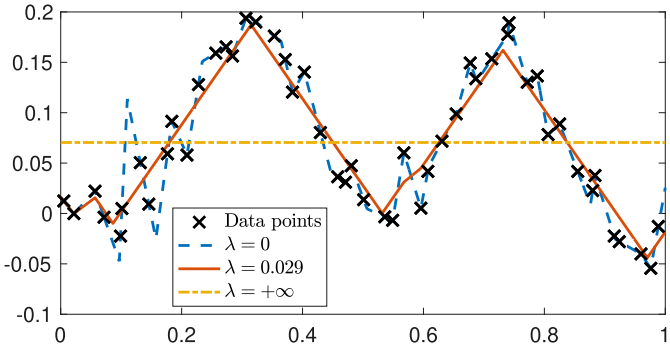
The proof of Theorem 2 is given in Appendix B.. Items 1 and 2 are classical results that hold for a general class of variational problems (see [77] for a generic result). Their practical implication is Item 3, which provides a way to identify solutions of (16). The solution set $\mathcal{V}_{\mathrm{Lip}}$ is further explored in Item 4, where a geometrical insight is given (see Fig. 1). Finally, the result that has the greatest practical relevance is stated in Item 5 which creates an interesting link with $\mathrm{TV}^{(2)}$ minimization problems and hence guarantees the existence of CPWL solutions.

B. LIPSCHITZ CONSTRAINT

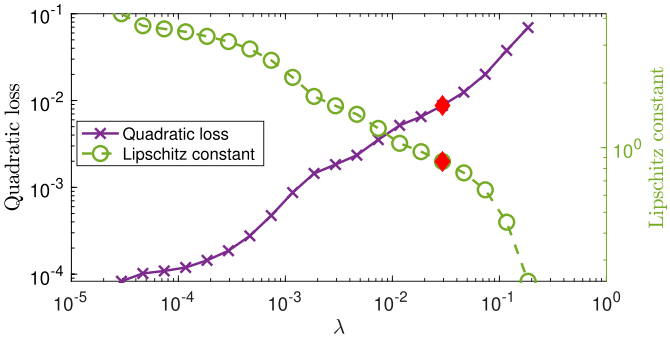
While the first formulation is interesting on its own right and results in learning CPWL mappings with tunable Lipschitz constants, it does not necessarily yield a sparse (and, hence, interpretable) solution. In fact, the learned mapping can have undesirable oscillations as illustrated in Fig. 3. This observation motivates us to propose a second formulation that combines $\mathrm{TV}^{(2)}$ regularization with a constraint over the Lipschitz constant, as expressed by

$$\mathcal{V}_{\mathrm{hyb}} = \arg\min_{f\in \mathrm{BV}^{(2)}(\mathbb{R})} \left(\sum_{m=1}^{M} \mathrm{E}(f(x_m), y_m) + \lambda \mathrm{TV}^{(2)}(f) \right), \quad \text{s.t.} \quad L(f) \le \bar{L}. \tag{24}$$

The quantity $\bar{L}$ is the maximal value allowed for the Lipschitz constant of the learned mapping. In this way, the stability is directly controlled by the user, while the regularization term removes undesired oscillations (tunable with $\lambda > 0$). The solution set $\mathcal{V}_{\mathrm{hyb}}$ is characterized in Theorem 3, from which we also deduce the existence of CPWL solutions.



(a) Reconstructions for different values of λ. Number of linear regions: 10 for $\lambda = 0.029$ versus 37 for $\lambda = +\infty$.



(b) Evolution of the training error and the Lipschitz regularity with respect to λ. The diamond corresponds to $\lambda = 0.029$ (shown in Figure 2a).

FIGURE 2. Example of our first formulation (16) for $M = 50$ data points.

Theorem 3: The solution set $\mathcal{V}_{\mathrm{hyb}}$ of Problem (24) is a nonempty, convex, and weak*-compact subset of $\mathrm{BV}^{(2)}(\mathbb{R})$ whose extreme points are linear splines with at most $(M-1)$ linear regions. Moreover, there exists a unique vector $\mathbf{z} = (z_m)$ such that

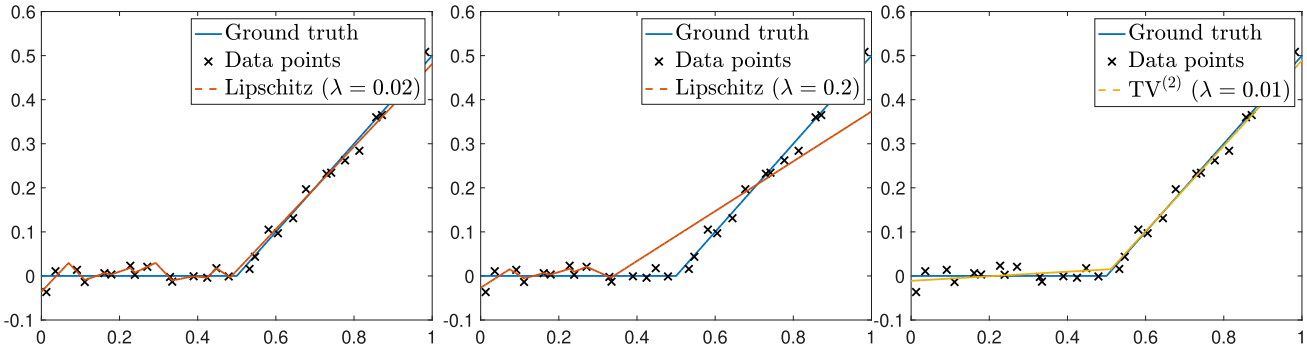
$$\mathcal{V}_{\mathrm{hyb}} = \arg\min_{f\in \mathrm{BV}^{(2)}(\mathbb{R})} \mathrm{TV}^{(2)}(f), \quad \text{s.t.} \quad f(x_m) = z_m,\ 1 \le m \le M. \tag{25}$$

Finally, the optimal $\mathrm{TV}^{(2)}$ cost has the closed-form expression

$$\mathrm{TV}_{\min} = \sum_{m=2}^{M-1} \left| \frac{z_m - z_{m-1}}{x_m - x_{m-1}} - \frac{z_m - z_{m+1}}{x_m - x_{m+1}} \right|. \tag{26}$$

The proof is given in Appendix D.. The proof involves the weak*-closedness of the constraint box $L(f) \le \bar{L}$ which is essential to prove existence. Once the existence of a minimizer is guaranteed, we can invoke the results of Debarre *et al.* in [40] for $\mathrm{TV}^{(2)}$ minimization to deduce the remaining parts. We also remark that the Lipschitz constraint only affects the vector $\mathbf{z}$ in (25), which forces its entries to satisfy the inequalities

$$\left| \frac{z_m - z_{m-1}}{x_m - x_{m-1}} \right| \le \bar{L}, \quad m = 2, \ldots, M. \tag{27}$$



(a) Lipschitz regularization. Number of linear regions: 13.

(b) Lipschitz regularization. Number of linear regions: 9.

(c) TV$^{(2)}$ regularization. Number of linear regions: 2.

FIGURE 3. Reconstructions with a ReLU ground truth and $M = 30$ data points.

C. CONNECTION TO NEURAL NETWORKS

In this part, we show that our second formulation (24) is equivalent to training a two-layer neural network with weight decay and a Lipschitz constraint. Let us recall that a univariate ReLU network with two layers and skip connections is a mapping $f_{\boldsymbol{\theta}} : \mathbb{R} \to \mathbb{R}$ of the form

$$f_{\boldsymbol{\theta}}(x) = c_0 + c_1 x + \sum_{k=1}^{K} v_k \mathrm{ReLU}(w_k x - b_k), \qquad (28)$$

where $c_1 \in \mathbb{R}$ is the weight of the skip connection, $K \in \mathbb{N}$ is the width of the network, $v_k, w_k \in \mathbb{R}, k = 1, \ldots, K$ are the linear weights and $b_k \in \mathbb{R}$, $k = 1, \ldots, K$ and $c_0 \in \mathbb{R}$ are the bias terms of the first and second layers, respectively. These parameters are concatenated in a single vector $\boldsymbol{\theta} = (K, \mathbf{v}, \mathbf{w}, \mathbf{b}, \mathbf{c})$, and we denote by $\boldsymbol{\Theta}$ the set of all possible parameter vectors $\boldsymbol{\theta}$. Thus, the training problem with Lipschitz constraint and weight decay is formulated as

$$\mathcal{V}_{NN} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\min} \left(\sum_{m=1}^{M} \mathrm{E}(f_{\boldsymbol{\theta}}(x_m), y_m) + \lambda \mathrm{R}(\boldsymbol{\theta}) \right), \quad \text{s.t.} \quad L(f_{\boldsymbol{\theta}}) \leq \bar{L}, \qquad (29)$$

where $\mathrm{R}(\boldsymbol{\theta}) = \sum_{k=1}^{K} (\frac{|v_k|^2 + |w_k|^2}{2})$ is the regularization term corresponding to weight decay. In Proposition 1, we show the equivalence between this training problem and our Lipschitz-constrained formulation (24).

Proposition 1: For any solution $\boldsymbol{\theta}^*$ of (29), $f_{\boldsymbol{\theta}^*}$ is a CPWL solution of (24). Moreover, any CPWL solution of (24) can be expressed as a two-layer ReLU network $f_{\boldsymbol{\theta}^*}$ with skip connections whose parameter vector is optimal in the sense of (29), *i.e.*, $\boldsymbol{\theta}^* \in \mathcal{V}_{NN}$.

Proposition 1, whose proof is given in Appendix E., is an extension of the results of [47], [50], where this equivalence is proved in the absence of a Lipschitz constraint. These works rely on a result (*e.g.,* [50, Corollary C.2]) that describes the energy propagation in the training of feed-forward neural networks with weight decay, which can easily be extended to the Lipschitz-constrained case (Lemma 2). Proposition 1 provides a functional framework to study the training of Lipschitz-aware neural networks, which is a nontrivial task. To this end, Proposition 1 allows us to deploy our proposed algorithm (introduced in Section IV).

IV. FINDING THE SPARSEST CPWL SOLUTION

Using the theoretical results of Section III, we propose an algorithm to find the sparsest CPWL solution of Problems (16) and (24). To that end, we first compute the vector $\mathbf{z}$ of the value of the optimal function at the data points $x_1, \ldots, x_m$. Using this vector, we then deploy the sparsification algorithm of [40], whose use in the present method is motivated by the following theorem.

Theorem 4: Let $(x_m, z_m) \in \mathbb{R}^2$, $m = 1, \ldots, M$ be a collection of ordered data points with $x_1 < \cdots < x_M$. Then, the output f_{sparse} of the sparsification algorithm of Debarre *et al.* in [40] is the sparsest linear-spline interpolator of the data points. In other words, f_{sparse} is the CPWL interpolator with the fewest number of linear regions.

The proof is given in Appendix C.. Theorem 4 is a strong enhancement of [40, Theorem 4] where it is merely established that f_{sparse} is the sparsest CPWL solution of (23). In Theorem 4, we prove that f_{sparse} is in fact the sparsest of *all* CPWL interpolants of the data points (x_m, z_m), without restricting the search to the solutions of (23). This is a remarkable result in its own right, as it gives a nontrivial answer to the seemingly simple question: how to interpolate data points with the minimum number of lines? Here, we invoke Theorem 4 to deduce that, with the vector z defined in Item 2 of Theorem 2, f_{sparse} is the sparsest CPWL solution of (17). Similarly, with the vector z defined in Theorem 3, f_{sparse} is the sparsest CPWL solution of (24).

In the remaining part of this section, we detail our computation of the vectors $\mathbf{z}$ defined in Theorems 2 and 3. Let us define the empirical loss function $F : \mathbb{R}^M \to \mathbb{R}_{\geq 0}$ as

$$F(\mathbf{z}) = \sum_{m=1}^{M} E(z_m, y_m). \qquad (30)$$

For simplicity, we assume that F is differentiable; the prototypical example is the quadratic loss $F(\mathbf{z}) = \frac{1}{2}\sum_{m=1}^{M}(z_m - y_m)^2$. Following this notation and using (18), the vector $\mathbf{z}$ in Problem (17) is solution to the minimization problem

$$\min_{\mathbf{z}} \in \mathbb{R}^M \left(F(\mathbf{z}) + \lambda\|\mathbf{L}_{\text{inf}}\mathbf{z}\|_\infty\right), \tag{31}$$

where the matrix $\mathbf{L}_{\text{inf}} \in \mathbb{R}^{(M-1)\times M}$ is given by

$$[\mathbf{L}_{\text{inf}}]_{m,n} = \begin{cases} -v_{m+1}, & n = m \\ v_{m+1}, & n = m+1 \\ 0, & \text{otherwise} \end{cases} \tag{32}$$

where $v_m = (x_m - x_{m-1})^{-1}$, $m = 2, \ldots, M$. To solve (31), we use the well-known alternating-direction method of multipliers (ADMM) [78] by defining the augmented Lagrangian as

$$J(\mathbf{z}, \mathbf{u}, \mathbf{w}) = F(\mathbf{z}) + \lambda\|\mathbf{u}\|_\infty + \frac{\rho}{2}\|\mathbf{L}_{\text{inf}}\mathbf{z} - \mathbf{u}\|_2^2 + \mathbf{w}^T(\mathbf{L}_{\text{inf}}\mathbf{z} - \mathbf{u}), \tag{33}$$

where $\rho > 0$ is a tunable parameter. The principle of ADMM is to sequentially update the unknown variables $\mathbf{z} \in \mathbb{R}^M$ and $\mathbf{u}, \mathbf{w} \in \mathbb{R}^{M-1}$. Precisely, its kth iteration is given explicitly by

$$\mathbf{z}^{(k+1)} = \arg\min_{\mathbf{z}\in\mathbb{R}^M} J\left(\mathbf{z}, \mathbf{u}^{(k)}, \mathbf{w}^{(k)}\right), \tag{34}$$

$$\mathbf{u}^{(k+1)} = \arg\min_{\mathbf{u}\in\mathbb{R}^{M-1}} J\left(\mathbf{z}^{(k+1)}, \mathbf{u}, \mathbf{w}^{(k)}\right), \tag{35}$$

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \rho\left(\mathbf{L}_{\text{inf}}\mathbf{z}^{(k+1)} - \mathbf{u}^{(k+1)}\right). \tag{36}$$

The benefit of these sequential updates is that Problem (34) has a differentiable cost and hence, can be efficiently solved using gradient-based methods. (In the case of the quadratic loss $E(z, y) = \frac{1}{2}(z - y)^2$, one can even obtain a closed-form solution.) Unfortunately, the cost in (35) is not differentiable. However, one can rewrite the augmented Lagrangian as

$$J(\mathbf{z}_k, \mathbf{u}, \mathbf{w}_k) = \frac{\rho}{2}\left\|\mathbf{u} - \mathbf{L}_{\text{inf}}\mathbf{z}_k - \frac{1}{\rho}\mathbf{w}_k\right\|_2^2 + \lambda\|\mathbf{u}\|_\infty + \text{Cnst.}, \tag{37}$$

where the constant term accounts for all terms that do not depend on $\mathbf{u}$. Then, by defining the vector $\mathbf{v}_k = (\mathbf{L}_{\text{inf}}\mathbf{z}_k + \frac{1}{\rho}\mathbf{w}_k)$, we rewrite (35) as

$$\begin{aligned}\mathbf{u}^{(k+1)} &= \arg\min_{\mathbf{u}\in\mathbb{R}^{M-1}}\left(\frac{1}{2}\|\mathbf{u} - \mathbf{v}_k\|_2^2 + \frac{\lambda}{\rho}\|\mathbf{u}\|_\infty\right) \\ &= \text{prox}_{\frac{\lambda}{\rho}\|\cdot\|_\infty}(\mathbf{v}_k),\end{aligned} \tag{38}$$

by definition of the proximal operator. The proximal operator of the ℓ_∞-norm has computationally cheap implementations (see, for example, [79, Section 6.5.2]), which can be used to update $\mathbf{u}$ via (38).

Similarly and using (26), we formulate the search for the vector $\mathbf{z}$ associated to the Problem (24) as

$$\min_{\mathbf{z}\in\mathbb{R}^M}\left(F(\mathbf{z}) + \lambda\|\mathbf{L}_1\mathbf{z}\|_1 + i_{\|\mathbf{L}_{\text{inf}}\mathbf{z}\|_\infty \le \overline{L}}\right), \tag{39}$$

where i_E denotes the indicator function of the set E and $\mathbf{L}_1 \in \mathbb{R}^{(M-2)\times M}$ with

$$[\mathbf{L}_1]_{m,n} = \begin{cases} -v_{m+1}, & n = m \\ (v_{m+1} + v_{m+2}), & n = m+1 \\ -v_{m+2}, & n = m+2, \\ 0, & \text{otherwise} \end{cases} \tag{40}$$

for all $m = 1, \ldots, M-2$ and $n = 1, \ldots, M$. In this case, the augmented Lagrangian takes the form

$$\begin{aligned}J(\mathbf{z}, \mathbf{u}_1, \mathbf{u}_{\text{inf}}, \mathbf{w}_1, \mathbf{w}_{\text{inf}}) &= F(\mathbf{z}) \\ &+ \frac{\rho_1}{2}\|\mathbf{L}_1\mathbf{z} - \mathbf{u}_1\|_2^2 + \mathbf{w}_1^T(\mathbf{L}_1\mathbf{z} - \mathbf{u}_1) + \|\mathbf{u}_1\|_1 \\ &+ \frac{\rho_{\text{inf}}}{2}\|\mathbf{L}_{\text{inf}}\mathbf{z} - \mathbf{u}_{\text{inf}}\|_2^2 + \mathbf{w}_{\text{inf}}^T(\mathbf{L}_{\text{inf}}\mathbf{z} - \mathbf{u}_{\text{inf}}) + i_{\|\mathbf{u}_{\text{inf}}\|_\infty \le \overline{L}}.\end{aligned} \tag{41}$$

At the kth iteration, we then solve sequentially the following optimization problems

$$\mathbf{z}^{(k+1)} = \arg\min_{\mathbf{z}\in\mathbb{R}^M} J(\mathbf{z}, \mathbf{u}_1^{(k)}, \mathbf{u}_{\text{inf}}^{(k)}, \mathbf{w}_1^{(k)}, \mathbf{w}_{\text{inf}}^{(k)}), \tag{42}$$

$$\mathbf{u}_1^{(k+1)} = \arg\min_{\mathbf{u}_1\in\mathbb{R}^{M-2}} J(\mathbf{z}^{(k+1)}, \mathbf{u}_1, \mathbf{u}_{\text{inf}}^{(k)}, \mathbf{w}_1^{(k)}, \mathbf{w}_{\text{inf}}^{(k)}), \tag{43}$$

$$\mathbf{u}_{\text{inf}}^{(k+1)} = \arg\min_{\mathbf{u}_{\text{inf}}\in\mathbb{R}^{M-1}} J\left(\mathbf{z}^{(k+1)}, \mathbf{u}_1^{(k+1)}, \mathbf{u}_{\text{inf}}, \mathbf{w}_1^{(k)}, \mathbf{w}_{\text{inf}}^{(k)}\right), \tag{44}$$

$$\mathbf{w}_1^{(k+1)} = \mathbf{w}_1^{(k)} + \rho_1\left(\mathbf{L}_1\mathbf{z}^{(k+1)} - \mathbf{u}_1^{(k+1)}\right), \tag{45}$$

$$\mathbf{w}_{\text{inf}}^{(k+1)} = \mathbf{w}_{\text{inf}}^{(k)} + \rho_{\text{inf}}\left(\mathbf{L}_{\text{inf}}\mathbf{z}^{(k+1)} - \mathbf{u}_{\text{inf}}^{(k+1)}\right). \tag{46}$$

The cost function of Problem (42) is differentiable and so, we can solve it using gradient-based methods. For Problem (43), we invoke the proximal operator of the ℓ_1-norm that is known to be soft-thresholding [79, Section 6.5.2.]. Finally and for (44), the proximal operator of the indicator function $i_{\|\cdot\|_\infty \le \overline{L}}$ is the projection over the ℓ_∞ ball which has the simple separable expression

$$\left[\text{prox}_{i_{\|\cdot\|_\infty \le \overline{L}}}(\mathbf{v})\right]_n = \begin{cases} \overline{L}, & v_n > \overline{L} \\ v_n, & |v_n| \le \overline{L} \\ -\overline{L}, & v_n < -\overline{L}. \end{cases} \tag{47}$$

V. NUMERICAL EXAMPLES AND DISCUSSIONS

A. EXPERIMENTAL SETUP

In all our experiments, we consider the standard quadratic loss $E(y, z) = \frac{1}{2}(y - z)^2$. We draw the data-point locations x_m randomly in the interval $[0, 1]$. The values y_m are then generated as $y_m = f_0(x_m) + n_m$, where f_0 is some known CPWL function (gold standard) and n_m is drawn i.i.d. from a zero-mean normal distribution with variance σ^2.

B. EXAMPLE OF LIPSCHITZ REGULARIZATION

In this first experiment, we illustrate our first formulation (16). We take $M = 50$ data points, a CPWL ground-truth f_0 with 6 linear regions, and a noise level $\sigma = 0.02$.

The results are shown in Fig. 2. In Fig. 2(a), we show the reconstructions for extreme values of λ. On one hand, $\lambda \to 0$ corresponds to the exact interpolation Problem (17). On the other hand, $\lambda = +\infty$ corresponds to constant regression. Obviously, neither is very satisfactory: interpolation leads to overfitting (the reconstruction has 37 linear regions), and the constant regression to underfitting. We show an example of a more satisfactory reconstruction for $\lambda = 0.029$ (10 linear regions), which is visually acceptable. In Fig. 2(b), we show the evolution of the quadratic loss $\frac{1}{2}\sum_{m=1}^{M}(f^*(x_m) - y_m)^2$ and the Lipschitz constant $L(f^*)$, for various values of λ. With the aid of such curves, the user can choose what is considered acceptable for either of these costs and select a suitable value of λ.

C. LIMITATIONS OF LIPSCHITZ-ONLY REGULARIZATION

Despite its interesting theoretical properties, Problem (16) does not always yield satisfactory reconstructions. This is because it does not enforce a sparse reconstruction in the problem formulation, despite the fact that our algorithm reconstructs (one of) the sparsest elements of $\mathcal{V}_{\text{lip}}$. This leads to learned mappings with too many linear regions and, consequently, poor interpretability.

One such example is shown in Fig. 3, where we consider the shifted ReLU function $f_0(\cdot) = (\cdot - \frac{1}{2})_+$ as the ground-truth mapping. We also fix the standard deviation of the noise to $\sigma = 0.02$. Fig. 3(a) shows a reconstruction that solves Problem (16) with the regularization parameter $\lambda = 0.02$. Although the reconstruction is satisfactory in the active section ($x > 1/2$), it has many linear regions in the flat section ($x < 1/2$) that are not present in f_0. This is due to the fact that the active section forces the Lipschitz constant of the reconstruction to be around 1, while oscillations with a slope smaller than 1 in the flat section are not penalized by the regularization. This problem clearly cannot be fixed by a simple increase in the regularization parameter: with $\lambda = 0.2$ (Fig. 3(b)), not only there are still too many linear regions in the flat section (the reconstruction has 9 linear regions in total), but also the active section is poorly reconstructed because the Lipschitz constant is penalized too heavily by the regularization.

Hence, to reconstruct such a ground truth accurately, it is necessary to enforce the sparsity of the reconstruction, which is exactly the purpose of the $\text{TV}^{(2)}$ regularization. The reconstruction result of the $\text{TV}^{(2)}$-regularized problem (*i.e.,* Problem (24) with a relatively large Lipschitz bound) with $\lambda = 0.01$ is also shown in Fig. 3(c); it is clearly much more satisfactory than any of the Lipschitz-penalized reconstructions since it is very close to the ground truth and has the same sparsity (two linear regions).

D. ROBUSTNESS TO OUTLIERS OF THE LIPSCHITZ-CONSTRAINED FORMULATION

In this final experiment, we demonstrate the pertinence of our second formulation (Problem (24)). More precisely, we examine the increased robustness to outliers of our second formulation (24) with respect to $\text{TV}^{(2)}$ regularization. To that end, we generate the CPWL ground truth f_0 with 6 linear regions and $M = 50$ data points. We then consider an additive Gaussian-noise model with low standard deviation $\sigma = 10^{-3}$ for 90% of the data, and a much stronger $\sigma' = 3.5 * 10^{-2}$ for the remaining 10%, which can be considered outliers.

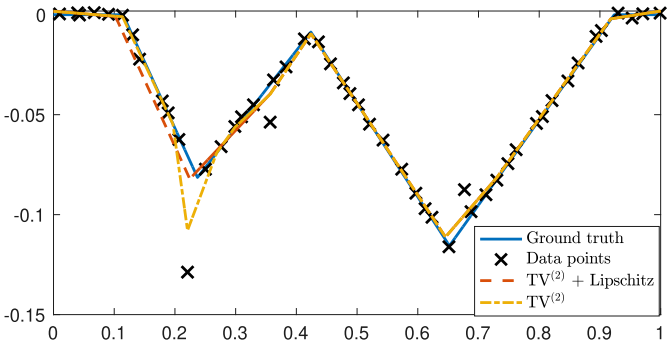


FIGURE 4. Reconstruction of $M = 50$ data points for $\lambda = 10^{-4}$. Our second formulation with $\bar{L} = 0.66$ produces 9 linear regions. We compare it to that of $\text{TV}^{(2)}$ which produces 12 linear regions.

We show in Fig. 4 the reconstruction results using our second formulation with $\lambda = 10^{-4}$ and $\bar{L} = 0.66$. The latter is quite satisfactory despite the presence of a strong outlier around $x_m = 0.22$. This is due to the fact that the Lipschitz constant is constrained. When using $\text{TV}^{(2)}$-regularization alone, at same regularization parameter, the reconstruction is very similar in most regions but is much more sensitive to this outlier which leads to an unwanted sharp peak and to the high Lipschitz constant $L(f^*) = 2.21$. Moreover, our reconstruction is more satisfactory in terms of sparsity (9 linear regions compared to 12, which is closer to the 6 linear regions of the target function f_0).

VI. CONCLUSION

We have proposed two schemes for the learning of one-dimensional continuous and piecewise-linear (CPWL) mappings with tunable Lipschitz constant. In the first scheme, we directly use the Lipschitz constant as a regularization term. We establish a representer theorem that allows us to deduce the existence of a CPWL solution for this continuous-domain optimization problem. In the second scheme, we use the second-order total-variation seminorm as the regularization term to which we add a Lipschitz constraint. Again, we proved the existence of a CPWL solution for this problem. Finally, we proposed an efficient algorithm to find the sparsest CPWL solution of each problem. We illustrated the outcome of each scheme via numerical examples. A potential application of the proposed algorithm is to design stable CPWL activation functions with a minimum number of linear regions in deep neural networks. This can, for example, be useful to train a denoising module in the context of plug-and-play methods for image reconstruction, whose convergence rates typically depend on the Lipschitz constant of the trainable denoising module [32].

APPENDIX A

A. PROOF OF THEOREM 1

Proof: For any $h > 0$ and $\mathbf{p} = (p_1, p_2) \in \mathbb{R}^2$ with $p_1 < p_2$, let us first define the test function $\varphi_h(\cdot; \mathbf{p}) \in \mathcal{C}_0(\mathbb{R})$ as

$$\varphi_h(x; \mathbf{p}) = h^{-1} \left(\mathrm{ReLU}\left(x - (p_1 - h)\right) - \mathrm{ReLU}(x - p_1) + \mathrm{ReLU}\left(x - (p_2 + h)\right) - \mathrm{ReLU}(x - p_2)\right).$$

This function will be used on several occasions throughout the proof. In particular, we use the explicit form of its second-order derivative given by

$$\mathrm{D}^2\varphi_h(\cdot; \mathbf{p}) = h^{-1} \left(\delta\left(\cdot - (p_1 - h)\right) - \delta(\cdot - p_1) + \delta\left(\cdot - (p_2 + h)\right) - \delta(\cdot - p_2)\right). \quad (48)$$

Upper-Bound: Similar to (10), we have that $\ell(f) = \operatorname{ess\,inf}_{x\in\mathbb{R}} |f'(x)|$. For a fixed $\epsilon > 0$, by definition of the essential supremum and infimum, there exist $\bar{x}, \underline{x} \in \mathbb{R}$ at which f is differentiable with $|f'(\bar{x})| \geq (L(f) - \epsilon)$ and $|f'(\underline{x})| \leq (\ell(f) + \epsilon)$. Without loss of generality, we assume that $\bar{x} < \underline{x}$. Following the limit definition of the derivative, we then consider a small radius $h > 0$ such that

$$\left|\frac{f(\bar{x} + h) - f(\bar{x})}{h}\right| \geq |f'(\bar{x})| - \epsilon \geq L(f) - 2\epsilon,$$

$$\left|\frac{f(\underline{x} + h) - f(\underline{x})}{h}\right| \leq |f'(\underline{x})| + \epsilon \leq \ell(f) + 2\epsilon.$$

Now, let us consider the test function $\varphi = \varphi_h(\cdot; (\bar{x} + h, \underline{x}))$. Following the definition of the total-variation norm (7) together with $\|\varphi\|_\infty = 1$, we deduce that $\mathrm{TV}^{(2)}(f) \geq |\langle \mathrm{D}^2 f, \varphi\rangle| = |\langle f, \mathrm{D}^2\varphi\rangle|$, where the last equality follows from the self-adjointness of the second-order derivative. Using (48), we thus have that

$$\begin{aligned}\mathrm{TV}^{(2)}(f) &\geq h^{-1}|f(\bar{x}) - f(\bar{x} + h) + f(\underline{x} + h) - f(\underline{x})| \\ &\geq \frac{|f(\bar{x} + h) - f(\bar{x})|}{h} - \frac{|f(\underline{x} + h) - f(\underline{x})|}{h} \\ &\geq L(f) - 2\epsilon - \ell(f) - 2\epsilon = L(f) - \ell(f) - 4\epsilon.\end{aligned}$$

Finally, by letting $\epsilon \to 0$, we deduce the desired upper-bound.

Saturation—Sufficient Conditions: Assume that $f \in \mathrm{BV}^{(2)}(\mathbb{R})$ is convex and increasing; we denote its second-order weak derivative by $w = \mathrm{D}^2 f$. Note that, in this case, the functions $(-f(\cdot))$, $f(-\cdot)$, and $(-f(-\cdot))$ are concave/decreasing, convex/decreasing, and concave/increasing, respectively. Hence, we only need to prove the saturation for f and the other cases immediately follow.

For a fixed $\epsilon > 0$, from (13) there exists a test function $\psi \in \mathcal{D}(\mathbb{R})$ with compact support $K = \mathrm{supp}(\psi)$ such that $\|\psi\|_{L_\infty} = 1$ and $\langle w, \psi\rangle \geq (\mathrm{TV}^{(2)}(f) - \epsilon)$. For any $T > 0$, we consider the test function $\psi_T = \varphi_1(\cdot; (-T, T))$. From (48), we obtain that

$$\begin{aligned}\langle w, \psi_T\rangle &= \langle f, \mathrm{D}^2\psi_T\rangle \\ &= (f(T + 1) - f(T)) - (f(-T) - f(-T - 1)) \\ &\leq L(f) - \ell(f),\end{aligned}$$

where we have used the increasing assumption to deduce that $f(T + 1) \geq f(T)$ and $f(-T) \geq f(-T - 1)$. By choosing T large enough so that $K \subseteq [-T, T]$, we ensure that $(\psi_T - \psi)$ is a nonnegative function, since for all $x \in K$, we will have that $\psi_T(x) = 1 = \|\psi\|_{L_\infty} \geq \psi(x)$. Next, the convexity of f implies that $w = \mathrm{D}^2 f$ is a positive measure. Hence,

$$0 \leq \langle w, \psi_T - \psi\rangle \leq L(f) - \ell(f) - \mathrm{TV}^{(2)}(f) + \epsilon. \quad (49)$$

By letting $\epsilon \to 0$, we deduce that $\mathrm{TV}^{(2)}(f) \leq (L(f) - \ell(f))$, which implies the saturation of (14).

Saturation—Necessary Conditions: Let $f \in \mathrm{BV}^{(2)}(\mathbb{R})$ be a function for which (14) is saturated.

Monotonicity: Assume by contradiction that f is not monotone. Hence, there exists $x_n \in \mathbb{R}$ such that $f'(x_n) < 0$. Indeed, if f' were a positive distribution, then for any $a, b \in \mathbb{R}$ with $a < b$, we would have that $(f(b) - f(a)) = \langle f, {}' \mathbb{1}_{[a,b]}\rangle \geq 0$, which contradicts the assumption of non-monotonicity. Similarly, there exists $x_p \in \mathbb{R}$ such that $f'(x_p) > 0$.

Next, consider a point $x_L \in \mathbb{R}$, distinct from x_n and x_p, such that $|f'(x_L)| > (L(f) - \epsilon) > 0$, where $0 < \epsilon < \frac{\min(-f'(x_n), f'(x_p))}{3}$ is a small constant. Without loss of generality, let us assume that $f'(x_L) > 0$. There exists a small radius $h \in (0, \frac{|x_L - x_n|}{2})$ such that

$$\frac{f(x_n + h) - f(x_n)}{h} \leq f'(x_n) + \epsilon < 0,$$

$$\frac{f(x_L + h) - f(x_L)}{h} \geq f'(x_L) - \epsilon > 0.$$

By considering the test function

$$\varphi = \begin{cases} \varphi_h(\cdot; (x_n + h, x_L)) & \text{if } x_n < x_L \\ \varphi_h(\cdot; (x_L, x_n + h)) & \text{if } x_n > x_L \end{cases}$$

and using (13) once again, we deduce that

$$\begin{aligned}\mathrm{TV}^{(2)}(f) &\geq h^{-1} |f(x_n) - f(x_n + h) + f(x_L + h) - f(x_L)| \\ &= \frac{f(x_L + h) - f(x_L)}{h} - \frac{f(x_n + h) - f(x_n)}{h} \\ &\geq f'(x_L) - \epsilon - f'(x_n) - \epsilon \geq L(f) - f'(x_n) - 3\epsilon \\ &> L(f),\end{aligned}$$

which contradicts the original assumption that (14) is saturated. For the case $f'(x_L) < 0$, the same arguments can be applied to x_p instead of x_n. This proves that f is monotone. In the following, we consider the case where f is an increasing function; the decreasing case can be deduced by symmetry.

Convexity/Concavity: We first consider the canonical decomposition $f = \mathrm{D}_\phi^{-2} w + p$, where $w = \mathrm{D}^2 f$, D_ϕ^{-2} is a right inverse of the second-order derivative, and $p(x) = ax + b$ is an affine term [41, Proposition 9]. We then use the Jordan decomposition of $w = \mathrm{D}^2 f$ as $w = (w_+ - w_-)$, where $w_+, w_- \in \mathcal{M}(\mathbb{R})$ are positive measures such that $\|w\|_{\mathcal{M}} = \|w_+\|_{\mathcal{M}} + \|w_-\|_{\mathcal{M}}$. This allows us to form the decomposition $f = (f_+ - f_-)$, where $f_s = \mathrm{D}_\phi^{-2} w_s + p_s, \quad s \in \{+, -\}$, $p_+(x) = (A + a)x + b$, and $p_-(x) = Ax$ with $A > 0$ being a sufficiently large constant such that the functions f_+ and f_-

are both convex and strictly increasing. Hence, they both satisfy the sufficient conditions for saturation, which implies that $\mathrm{TV}^{(2)}(f_s) = (L(f_s) - \ell(f_s))$ for $s \in \{+, -\}$.

Assume by contradiction that $w_s \neq 0$ for $s \in \{+, -\}$ and let $\epsilon < \frac{\min(\mathrm{TV}^{(2)}(f_+), \mathrm{TV}^{(2)}(f_-))}{2}$ be a small positive constant. Let $\bar{x}, \underline{x} \in \mathbb{R}$ such that $f'(\bar{x}) \geq (L(f) - \epsilon)$ and $f'(\underline{x}) \leq (\ell(f) + \epsilon)$. Using these inequalities, we deduce that

$$\begin{aligned}\mathrm{TV}^{(2)}(f) &= L(f) - \ell(f) \\ &\leq f'(\bar{x}) - f'(\underline{x}) + 2\epsilon \\ &= \big(f'_+(\bar{x}) - f'_-(\bar{x})\big) - \big(f'_+(\underline{x}) - f'_-(\underline{x})\big) + 2\epsilon \\ &= A_+ - A_- + 2\epsilon, \end{aligned} \tag{50}$$

where $A_s = (f'_s(\bar{x}) - f'_s(\underline{x}))$ for $s \in \{+, -\}$. We now consider two cases:

Case I: $\bar{x} > \underline{x}$: The convexity of f_- implies that $A_- \geq 0$. Moreover, we have that $A_+ = (f'_+(\bar{x}) - f'_+(\underline{x})) \leq (L(f_+) - \ell(f_+)) = \mathrm{TV}^{(2)}(f_+)$. Using (50), this yields that $\mathrm{TV}^{(2)}(f) \leq \mathrm{TV}^{(2)}(f_+) + 2\epsilon$, which can be rewritten as $2\epsilon \geq \mathrm{TV}^{(2)}(f_-)$. However, our original choice of ϵ implies that $\epsilon < \mathrm{TV}^{(2)}(f_-)/2$, which leads to a contradiction.

Case II: $\bar{x} \leq \underline{x}$: Similarly to the first case, we deduce that $A_+ \leq 0$ and $-A_- \leq \mathrm{TV}^{(2)}(f_-)$. Hence, we get that $2\epsilon \geq \mathrm{TV}^{(2)}(f_+)$ which contradicts the assumption $\epsilon < \mathrm{TV}^{(2)}(f_+)/2$. Since both cases lead to a contradiction, we have $w_- = 0$ or $w_+ = 0$, which implies that f is either convex or concave. ■

B. PROOF OF THEOREM 2

Proof: Items 1 and 2: The first step is to show that the sampling functional $\delta(\cdot - x_0) : f \mapsto f(x_0)$ is weak*-continuous in $\mathrm{Lip}(\mathbb{R})$. To that end, we identify the predual Banach space $\mathcal{X}$ such that $\mathrm{Lip}(\mathbb{R}) = \mathcal{X}'$ and then show that shifted Dirac impulses are included in $\mathcal{X}$, which is equivalent to weak*-continuity. We recall that following (11), we can view $\mathrm{Lip}(\mathbb{R})$ as the native Banach space associated to the pair $(L_\infty(\mathbb{R}), \mathrm{D})$. This allows us to deploy the machinery of [71] to identify its predual space. In short, it follows from [71] that the predual space has the direct-sum structure $\mathcal{X} = \mathrm{D}(L_1(\mathbb{R})) \oplus \mathrm{span}(\mathrm{e}^{-(\cdot)^2})$. In other words, any function $f \in \mathcal{X}$ can be decomposed as $f = \mathrm{D}\{g\} + c\mathrm{e}^{-(\cdot)^2}$, where $g \in L_1(\mathbb{R})$ and $c \in \mathbb{R}$. One can formally verify that $\delta = \mathrm{D}\{\mathrm{sgn} - \mathrm{erf}\} + \frac{2}{\sqrt{\pi}}\mathrm{e}^{-(\cdot)^2}$, where sgn is the sign function and erf is the Gauss error function. Due to the rapid decay of the erf function at $t = -\infty$ and the symmetry of $(\mathrm{sgn} - \mathrm{erf})$, we deduce that $\mathrm{sgn} - \mathrm{erf} \in L_1(\mathbb{R})$ and, hence, that $\delta \in \mathcal{X}$. Finally, due to the shift-invariant structure of $\mathcal{X}$, we deduce the weak*-continuity of the sampling functional $\delta(\cdot - x_0)$ for any $x_0 \in \mathbb{R}$.

Next, we invoke the general representer theorem for Banach semi-norms [80, Theorem 3] to deduce that the solution set $\mathcal{V}_{\mathrm{Lip}}$ of (16) is a nonempty, convex, weak*-compact set whose elements all pass through a fixed set of points. Put differently, the vector $\mathbf{z} = (z_m)$ with $z_m = f(x_m)$ is invariant to the choice of $f \in \mathcal{V}_{\mathrm{Lip}}$. This means that adding the constraints $z_m = f(x_m), m = 1, \ldots, M$ does not change the solution set, *i.e.,*

$$\begin{aligned}\mathcal{V}_{\mathrm{Lip}} &= \underset{f \in \mathrm{Lip}(\mathbb{R})}{\arg\min} \left(\sum_{m=1}^{M} E(f(x_m), y_m) + \lambda L(f) \right), \\ &\qquad \text{s.t.} \quad f(x_m) = z_m, \ m = 1, \ldots, M \\ &= \underset{f \in \mathrm{Lip}(\mathbb{R})}{\arg\min} L(f), \quad \text{s.t.} \quad f(x_m) = z_m, \end{aligned}$$

where the last equality is obtained by observing that $\sum_{m=1}^{M} E(f(x_m), y_m)$ is constant within the solution set $\mathcal{V}_{\mathrm{Lip}}$. Consequently, we can represent $\mathcal{V}_{\mathrm{Lip}}$ as a solution set of a constrained problem of the form (17).

Item 3: Let us first define the canonical CPWL interpolant of a collection of 1D data points.

Definition 1: For a series of data points $(x_m, z_m), m = 1, \ldots, M$, the canonical interpolant $f_{\mathrm{cano}} : \mathbb{R} \to \mathbb{R}$ is the unique CPWL function that passes through these points and is differentiable over $\mathbb{R}\backslash\{x_2, \ldots, x_{M-1}\}$.

We first prove that f_{cano} is a solution of (17). Clearly, the Lipschitz constant of f_{cano} is equal to $L(f_{\mathrm{cano}}) = L_{\min}$, where $L_{\min}$ is given in (18). Moreover, any function f that passes through the data points (x_m, z_m) necessarily has a Lipschitz constant greater than or equal to $L_{\min}$. This implies that f_{cano} is a solution of (17) and $L_{\min}$ is the minimal value of the Lipschitz constant. Consequently, any function that satisfies the interpolation constraints and is $L_{\min}$-Lipschitz is a solution of (17).

Item 4: Consider a generic point $(x, y) \in \mathcal{E}$, and let m be such that $x \in (x_{m-1}, x_m)$. By definition of $\mathcal{E}$, there exists a function $f \in \mathcal{V}_{\mathrm{Lip}}$ such that $y = f(x)$. From Item 3, we deduce that $L(f) = L_{\min}$. Hence, we have the inequalities

$$\left| \frac{y - z_{m-1}}{x - x_{m-1}} \right|, \left| \frac{y - z_m}{x - x_m} \right| \leq L_{\min}. \tag{51}$$

These inequalities can readily be translated into the inclusion $(x, y) \in \mathcal{R}_{m-1} \cap \mathcal{L}_m$, which implies that $\mathcal{E} \subseteq \bigcup_{m=1}^{M} (\mathcal{R}_{m-1} \cap \mathcal{L}_m)$. To show the reverse inclusion, consider a point in $(x, y) \in \mathcal{R}_{m-1} \cap \mathcal{L}_m$ for some $m \in \{1, \ldots, M+1\}$ and denote by $\tilde{f}_{\mathrm{cano}}$ the canonical interpolant of $\{(x_m, z_m)\}_{m=1}^{M} \cup \{(x, y)\}$. Following Item 3, the Lipschitz constant of $\tilde{f}_{\mathrm{cano}}$ is given by

$$L(\tilde{f}_{\mathrm{cano}}) = \max \left(L_{\min}, \left| \frac{y - z_{m-1}}{x - x_{m-1}} \right|, \left| \frac{y - z_m}{x - x_m} \right| \right) = L_{\min}, \tag{52}$$

where we establish the last equality by translating the inclusion $(x, y) \in \mathcal{R}_{m-1} \cap \mathcal{L}_m$ into the inequalities in (51). This implies that $\tilde{f}_{\mathrm{cano}}$ is a solution of (17) and so, by definition, we have that $(x, y) \in \mathcal{E}$.

Item 5: By [40, Proposition 5], f_{cano} is also a solution of (23). We therefore need to prove that any solution f_{opt} of (23) has the same Lipschitz constant $L(f_{\mathrm{opt}}) = L(f_{\mathrm{cano}}) = L_{\min}$. Due to the interpolation constraints, we necessarily have that $L(f_{\mathrm{opt}}) \geq L(f_{\mathrm{cano}})$; we must now prove the reverse inequality $L(f_{\mathrm{opt}}) \leq L(f_{\mathrm{cano}})$. By [40, Theorem 2], f_{opt} must follow f_{cano} in $\mathbb{R}\backslash[x_2, x_{M-1}]$. Moreover, in each interval $[x_m, x_{m+1}]$

for $m \in \{2, \dots, M-2\}$, f_{opt} either follows f_{cano} or is concave or convex over the interval $[x_{m-1}, x_{m+2}]$. Hence, it suffices to prove that, for any $m \in \{2, \dots, M-2\}$, we have that $L_m(f_{\text{opt}}) \leq L(f_{\text{cano}})$, where $L_m(f)$ denotes the Lipschitz constant of f restricted to the interval $[x_m, x_{m+1}]$.

Let m be an index for which f_{opt} need not follow f_{opt} in $[x_m, x_{m+1}]$. (If no such index exists, then the result is trivially true.) Assume that f_{opt} is convex in the interval $[x_{m-1}, x_{m+2}]$; the concave scenario is derived in a similar fashion. This implies that, in this interval, the function $(\tilde{x}_1, \tilde{x}_2) \mapsto \frac{f_{\text{opt}}(\tilde{x}_2) - f_{\text{opt}}(\tilde{x}_1)}{\tilde{x}_2 - \tilde{x}_1}$ is increasing in both its variables.

Hence, for any $\tilde{x}_1, \tilde{x}_2 \in [x_m, x_{m+1}]$ with $\tilde{x}_1 \neq \tilde{x}_2$, we have that $\frac{z_m - z_{m-1}}{x_m - x_{m-1}} \leq \frac{f_{\text{opt}}(\tilde{x}_2) - f_{\text{opt}}(\tilde{x}_1)}{\tilde{x}_2 - \tilde{x}_1} \leq \frac{z_{m+2} - z_{m+1}}{x_{m+2} - x_{m+1}}$. This directly implies the desired result $L_m(f_{\text{opt}}) \leq L(f_{\text{cano}})$. ■

C. PROOF OF THEOREM 4

Proof:

Let f^* be the output of [40, Algorithm 1]. It is thus a CPWL solution of Problem (17) with the minimum number of linear regions. We prove that *any* CPWL interpolant f of the data points $\text{P}_m = (x_m, z_m)$, $m = 1, \dots, M$—not necessarily a minimizer of $\text{TV}^{(2)}(f)$—has at least as many linear regions as f^*. Our proof is based on induction over the number M of data points. The initialization $M = 2$ trivially holds, since f^* then has a single linear region—it is simply the line connecting the two data points. Next, let $M > 2$ and assume that Theorem 4 holds for $(M-1)$ or less data points (the induction hypothesis). The canonical interpolatant f_{cano} introduced in Definition 1 can be expressed as

$$f_{\text{cano}}(x) = \alpha_1 x + \alpha_2 + \sum_{m=2}^{M-1} a_m (x - x_m)_+ \tag{53}$$

for some coefficients $\alpha_1, \alpha_2, a_m \in \mathbb{R}$. There are three possible scenarios:

1) all a_m's are positive (or negative);
2) at least one of them is zero;
3) there are two consecutive coefficients with opposite signs, so that $a_m a_{m+1} < 0$ for some m.

We analyze each case separately and use the induction hypothesis to deduce the desired result. In this proof, we refer to singularities of CPWL functions (*i.e.,* the boundary points between linear regions) as *knots*.

Case 1: In this case, it is known that f^* has $K = (\lceil \frac{M}{2} \rceil - 1)$ knots [40, Theorem 4]. Assume by contradiction that there exists a CPWL interpolant f with fewer knots and consider the K disjoint intervals (x_{2k-1}, x_{2k+1}) for $1 \leq k \leq (\lceil \frac{M}{2} \rceil - 1) = K$. We deduce that there exists an interval (x_{2k-1}, x_{2k+1}) in which f has no knots. This in turn implies that the data points P_{2k-1}, P_{2k}, and P_{2k+1} are aligned, and so that $a_{2k} = 0$, which yields a contradiction.

Case 2: Let $m \in \{2, M-1\}$ be such that $a_m = 0$. Consider the collection of $m < M$ data points $(\text{P}_{m'})_{1 \leq m' \leq m}$; by the induction hypothesis, f^* interpolates them with the minimal number K_1 of knots. The same applies to the collection of $(M - m + 1) < M$ points $(\text{P}_{m'})_{m \leq m' \leq M}$ with K_2 knots. Let f be a CPWL interpolant of all the M data points with the minimal number of knots. By definition of the K_i, f must have at least K_1 knots in the interval (x_1, x_m) and K_2 knots in the interval (x_m, x_M). Since these intervals are disjoint, f must have at least $K_1 + K_2$ knots in total. Yet, f^* has exactly $(K_1 + K_2)$ knots: indeed, f^* follows f_{cano} in the interval $[x_{m-1}, x_{m+1}]$, which has no knot at x_m since $a_m = 0$ (the points P_{m-1}, P_m, and P_{m+1} are aligned). This concludes that f^* has the minimum number of knots.

Case 3: Let $m \in \{2, M-2\}$ be such that $a_m a_{m+1} < 0$. Consider the collection of $(m+1) < M$ data points $(\text{P}_{m'})_{1 \leq m' \leq m+1}$; by the induction hypothesis, f^* interpolates them with the minimal number K_1 of knots. Similarly, f^* interpolates the $(M - m + 1) < M$ points $(\text{P}_{m'})_{m \leq m' \leq M}$ with the minimal number K_2 of knots. Let f be a CPWL interpolant of all the M data points with the minimal number of knots. We now state a useful lemma whose proof is given below.

Lemma 1: Let $m \in \{2, \dots, M-2\}$ be such that $a_m a_{m+1} < 0$. Then, any CPWL interpolant f of the data points $(\text{P}_{m'})_{1 \leq m' \leq M}$ can be modified to become another CPWL interpolant $\tilde{f}$ with as many (or fewer) knots such that $\tilde{f}$ has no knot in the interval (x_m, x_{m+1}).

By Lemma 1, it can be modified to become another interpolant $\tilde{f}$ with the same total number of knots and none in the interval (x_m, x_{m+1}). By definition of the K_i, $\tilde{f}$ must have at least K_1 knots in the interval (x_1, x_{m+1}) and K_2 knots in the interval (x_m, x_M). Yet, $\tilde{f}$ has no knots in the interval (x_m, x_{m+1}), so it must have at least K_1 knots in $(x_1, x_m]$ and K_2 knots in $[x_{m+1}, x_M)$. Since these intervals are disjoint, $\tilde{f}$ must have at least $(K_1 + K_2)$ knots in total. Yet, f^* follows f_{cano} in the interval $[x_{m-1}, x_{m+2}]$ and thus also has no knot in the interval (x_1, x_{m+1}). Therefore, by the induction hypothesis, f^* has K_1 knots in $(x_1, x_m]$ and K_2 knots in $[x_{m+1}, x_M)$, for a total of $(K_1 + K_2)$ knots. Since this is no more than $\tilde{f}$, f^* has the minimal number of knots, which proves the induction. ■

Proof of Lemma 1: Let f be a CPWL interpolant of the data points $(\text{P}_{m'})_{1 \leq m' \leq M}$ with P knots. In what follows, we consider a CPWL function $\tilde{f}$ that follows f outside this interval and (x_{m-1}, x_{m+2}), and we modify it inside this interval in order to remove all knots in (x_m, x_{m+1}) without increasing the total number of knots.

We consider the case $a_m > 0$ and $a_{m+1} < 0$ without loss of generality. Let $s^- = f'(x_{m-1}^-)$ and $s^+ = f'(x_{m+2}^+)$ be the slopes of f before and after the interval of interest (x_{m-1}, x_{m+2}), respectively, and we let $s_{\text{cano}}^- = f'_{\text{cano}}(x_{m-1}^-)$ and $s_{\text{cano}}^+ = f'_{\text{cano}}(x_{m+2}^+)$ be those of f_{cano}. We also introduce the linear functions $f^-(x) = z_{m-1} + s^-(x - x_{m-1})$ and $f^+(x) = z_{m+2} + s^+(x - x_{m+2})$. They prolong f in a straight line after P_{m-1} and before P_{m+2}, respectively. We now distinguish cases based on s^- and s^+.

Case I: $s^- \leq s_{cano}^-$ and $s^+ \leq s_{cano}^+$ Graphically, this corresponds to f lying in none of the gray regions in Fig. 5. In this case, the line $(\text{P}_m \text{P}_{m+1})$ intersects the linear function f^- at

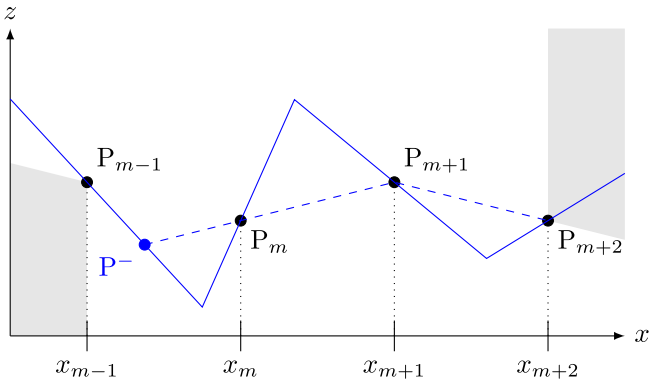


FIGURE 5. Illustration of Lemma 1 in the case $a_m > 0$ and $a_{m+1} < 0$. The interpolant f (solid line) satisfies $s^+ > s^+_{\text{cano}}$ and $s^- \le s^-_{\text{cano}}$. The modified interpolant $\tilde{f}$ (dashed line) also has three knots P^-, P_{m+1}, and P_{m+2}, but none in (x_m, x_{m+1}).

some point $P^- = (x^-, z^-)$ where $x^- \in (x_{m-1}, x_m)$, and with f^+ at some point $P^+ = (x^+, z^+)$ with $x^+ \in (x_{m+1}, x_{m+2})$. This is obvious graphically (see Fig. 5 as an illustration for P^-), and is due to the fact that $a_m > 0$ and $a_{m+1} < 0$. Hence, by taking an $\tilde{f}$ that connects the points P_{m-1}, P^-, P^+, and P_{m+2}, then $\tilde{f}$ has two knots in $[x_{m-1}, x_{m+2}]$ and its knots satisfy $x^-, x^+ \notin (x_m, x_{m+1})$. Since f clearly cannot have fewer than two knots in this interval, this proves the desired result.

Case II: $s^+ > s^+_{cano}$ and $s^- > s^-_{cano}$: In this case, f lies in both gray regions in Fig. 5. To pass through P_m, f must have at least one knot in $[x_{m-1}, x_m)$; let $P^- = (x^-, z^-)$ be the first of those knots (with $x^- < x_m$). Similarly, to pass through P_{m+1}, f must have a knot in $(x_{m+1}, x_{m+2}]$; let $P^+ = (x^+, z^+)$ be the last of those knots (with $x^+ > x_{m+1}$). Then, f must pass through the points P^-, P_m, P_{m+1}, P^+. Yet, the lines (P^-P_m) and $(P_{m+1}P^+)$ clearly cannot intersect in the interval $[x_m, x_{m+1}]$, which implies that at least two knots are needed in the interval (x^-, x^+). We conclude that f must have at least four knots in the interval $[x_{m-1}, x_{m+2}]$. Hence, we take an $\tilde{f}$ that simply connects the points P_{m-1}, P_m, P_{m+1}, and P_{m+2} and follows f elsewhere; the latter has four knots in $[x_{m-1}, x_{m+2}]$, which is no more than f and thus fulfills the requirements of the proof.

Case III: $s^+ > s^+_{cano}$ and $s^- \le s^-_{cano}$: This case is illustrated in Fig. 5: f is outside the gray region on the left, and inside the one on the right. With a similar argument as in Case II, f must have a least three knots in the interval $[x_{m-1}, x_{m+2}]$. The fact that $a_m > 0$ implies that the line (P_mP_{m+1}) intersects the linear function f^- at some point $P^- = (x^-, z^-)$ where $x^- \in (x_{m-1}, x_m)$. We then take an $\tilde{f}$ that connects the points P_{m-1}, P^-, P_{m+1}, and P_{m+2} and follows f elsewhere. The interpolant $\tilde{f}$ has three knots at x^-, x_{m+1}, and x_{m+2} in $[x_{m-1}, x_{m+2}]$ and thus satisfies the requirements of the proof.

Case IV: $s^+ \le s^+_{cano}$ and $s^- > s^-_{cano}$: This is similar to Case III, and can be readily deduced by symmetry, thus completing the proof of Lemma 1. ■

D. PROOF OF THEOREM 3

Proof: Existence: We rewrite the problem in (24) as an unconstrained minimization problem

$$\mathcal{V}_{\text{hyb}} = \operatorname*{arg\,min}_{f \in \mathcal{M}_{\text{D}^2}(\mathbb{R})} \sum_{m=1}^{M} \text{E}(f(x_m), y_m) + \lambda \text{TV}^{(2)}(f) + i_{L(f) \le \bar{L}}, \tag{54}$$

where i_E denotes the characteristic function of the set E and is defined as

$$i_E(f) = \begin{cases} 0, & f \in E \\ +\infty, & \text{otherwise.} \end{cases} \tag{55}$$

To prove the existence of a minimizer, we use a standard technique in convex analysis which involves the generalized Weierstrass theorem [81] to show that the cost functional of (54) is coercive and lower semicontinuous (in the weak*-topology), which is a sufficient condition for the existence of a solution.

The cost functional in (24) consists of three terms: (i) an empirical loss term $H(f) = \sum_{m=1}^{M} \text{E}(f(x_m), y_m)$; (ii) a second-order total-variation regularization term $R(f) = \lambda \text{TV}^{(2)}(f)$; and (iii) a Lipschitz constraint i_E, where $E = \{L(f) \le \bar{L}\}$. It is known (see [82] for a more general statement) that the functional $H(f) + R(f)$ is coercive and weak*-lower semincontinuous. This, together with the non-negativity of i_E, yields the coercivity of the total cost. The only missing item is the weak*-lower semicontinuity of i_E, for which it is sufficient to prove that E is a closed set for the weak*-topology.

Let $f_n \in \text{BV}^{(2)}(\mathbb{R})$ be a sequence of functions with $L(f_n) \le \bar{L}$ converging in the weak*-topology to $f_{\text{lim}} \in \text{BV}^{(2)}(\mathbb{R})$. To prove the weak*-closedness of E, we need to show that $L(f_{\text{lim}}) \le \bar{L}$, which is equivalent to $|f_{\text{lim}}(a) - f_{\text{lim}}(b)| \le \bar{L}|a - b|$ for any $a, b \in \mathbb{R}$.

For any $n \in \mathbb{N}$, we have that

$$\begin{aligned} |f_{\text{lim}}(a) - f_{\text{lim}}(b)| \le |f_{\text{lim}}(a) - f_n(a)| + |f_n(a) - f_n(b)| \\ + |f_n(b) - f_{\text{lim}}(b)|. \end{aligned} \tag{56}$$

Using the weak*-continuity of the sampling functionals $\delta(\cdot - a)$ and $\delta(\cdot - b)$ in $\text{BV}^{(2)}(\mathbb{R})$ (see, for example, [41]), we deduce that $f_n(a) \to f_{\text{lim}}(a)$ and $f_n(b) \to f_{\text{lim}}(b)$. Moreover, we have the estimate $|f_n(a) - f_n(b)| \le \bar{L}|a - b|$ for any $n \in \mathbb{N}$. Using these and letting the limit $n \to +\infty$ in (56), we get the desired bound.

Form of the Solution Set: Now that we have proved the existence of a solution $f_0^* \in \mathcal{V}_{\text{hyb}}$, we can apply a standard argument based on the strict convexity of $E(\cdot, \cdot)$ (see, for example, [83, Lemma 1]) to deduce that for any $f^* \in \mathcal{V}_{\text{hyb}}$, we have that $f^*(x_m) = f_0^*(x_m)$ for $m = 1, \ldots, M$. Hence, the original Problem (24) is equivalent to

$$\mathcal{V}_{\text{hyb}} = \operatorname*{arg\,min}_{f \in \text{BV}^{(2)}(\mathbb{R})} \text{TV}^{(2)}(f),$$

$$\text{s.t.} \quad \begin{cases} L(f) \leq \bar{L}, \\ f(x_m) = f_0^*(x_m), m = 1, \ldots, M. \end{cases} \tag{57}$$

Since $f_0^* \in \mathcal{V}_{\text{hyb}}$, we deduce that

$$L_0 \stackrel{\triangle}{=} \max_{2 \leq m \leq M} \left| \frac{f_0^*(x_m) - f_0^*(x_{m-1})}{x_m - x_{m-1}} \right| \leq L(f_0^*) \leq \bar{L}.$$

Yet, Item 5 in Theorem 2 implies that any solution f^* of the problem

$$\begin{aligned} &\underset{f \in \text{BV}^{(2)}(\mathbb{R})}{\arg\min} \ \text{TV}^{(2)}(f), \\ &\text{s.t.} \quad f(x_m) = f_0^*(x_m), \ m = 1, \ldots, M, \end{aligned} \tag{58}$$

is a solution of (17) with $z_m = f_0^*(x_m)$. Hence, by Item 3 of Theorem 2, we have that $L(f^*) = L_0 \leq \bar{L}$. This means that adding the Lipschitz constraint $L(f) \leq \bar{L}$ does not change the solution set of Problem (58). Hence, we have that

$$\begin{aligned} \mathcal{V}_{\text{hyb}} = \ &\underset{f \in \text{BV}^{(2)}(\mathbb{R})}{\arg\min} \ \text{TV}^{(2)}(f), \\ &\text{s.t.} \quad f(x_m) = f_0^*(x_m), \ m = 1, \ldots, M. \end{aligned} \tag{59}$$

The solution set of (59) has been fully described in [40], which yields the announced characterization. ■

E PROOF OF PROPOSITION 1

We start by proving a useful lemma.

Lemma 2: For any $\boldsymbol{\theta}^* = (K^*, \mathbf{v}^*, \mathbf{w}^*, \mathbf{b}^*, \mathbf{c}^*) \in \mathcal{V}_{\text{NN}}$, we have that $|v_k^*| = |w_k^*|$ for any $k = 1, \ldots, K$.

Proof: Let $\boldsymbol{\theta}^* = (K^*, \mathbf{v}^*, \mathbf{w}^*, \mathbf{b}^*, \mathbf{c}^*) \in \mathcal{V}_{NN}$ and $1 \leq k \leq K$. For any $\epsilon \in (-1, 1)$, we define a perturbed parameter vector $\boldsymbol{\theta}_\epsilon = (K^*, \mathbf{v}_\epsilon, \mathbf{w}_\epsilon, \mathbf{b}_\epsilon, \mathbf{c}^*)$, where for any $k' = 1, \ldots, K$ we have that

$$v_{\epsilon,k'} = \begin{cases} v_{k'}^*, & k' \neq k \\ (1+\epsilon)^{\frac{1}{2}} v_k^*, & k' = k \end{cases}, \tag{60}$$

$$w_{\epsilon,k'} = \begin{cases} w_{k'}^*, & k' \neq k \\ (1+\epsilon)^{-\frac{1}{2}} w_k^*, & k' = k \end{cases}, \tag{61}$$

$$b_{\epsilon,k'} = \begin{cases} b_{k'}^*, & k' \neq k \\ (1+\epsilon)^{-\frac{1}{2}} b_k^*, & k' = k. \end{cases} \tag{62}$$

Due to the positive homogeneity of the ReLU, one readily deduces from (28) that $f_{\boldsymbol{\theta}^*} = f_{\boldsymbol{\theta}_\epsilon}$ for any $\epsilon \in (-1, 1)$. This together with the optimality of $\boldsymbol{\theta}^*$ in Problem (29) implies that

$$v_k^{*2} + w_k^{*2} \leq (1+\epsilon) v_k^{*2} + (1+\epsilon)^{-1} w_k^{*2}, \quad \forall \epsilon \in (-1, 1).$$

Multiplying both sides of the above inequality by $(1+\epsilon) > 0$ yields

$$\epsilon w_k^{*2} \leq \epsilon(1+\epsilon) v_k^{*2}, \quad \forall \epsilon \in (-1, 1).$$

Letting $\epsilon \to 0^+$ yields $w_k^{*2} \leq v_k^{*2}$ and $\epsilon \to 0^-$ yields $w_k^{*2} \geq v_k^{*2}$, which proves that $|w_k^*| = |v_k^*|$. ■

Proof of Proposition 1: Using Lemma 2, we observe that for any $\boldsymbol{\theta}^* \in \mathcal{V}_{\text{NN}}$, we have that

$$\text{R}(\boldsymbol{\theta}^*) = \frac{1}{2} \sum_{k=1}^{K} (v_k^{*2} + w_k^{*2}) = \sum_{k=1}^{K} |v_k^*||w_k^*| = \text{TV}^{(2)}(f_{\boldsymbol{\theta}^*}),$$

where the last inequality comes from the simple observation that $\text{TV}^{(2)}(v\text{ReLU}(w \cdot -b)) = |v||w|$ for any $v, w, b \in \mathbb{R}$. Hence, one can rewrite the solution set $\mathcal{V}_{\text{NN}}$ as

$$\begin{aligned} \mathcal{V}_{\text{NN}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\text{red}}}{\arg\min} &\left(\sum_{m=1}^{M} \text{E}(f_{\boldsymbol{\theta}}(x_m), y_m) + \lambda \text{TV}^{(2)}(f_{\boldsymbol{\theta}}) \right), \\ &\text{s.t.} \quad L(f_{\boldsymbol{\theta}}) \leq \bar{L}, \end{aligned}$$

where $\boldsymbol{\Theta}_{\text{red}} = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \text{R}(\boldsymbol{\theta}) = \text{TV}^{(2)}(f_{\boldsymbol{\theta}})\}$ is the reduced parameter space. To prove the announced equivalence, it remains to show that the mapping $\boldsymbol{\Theta}_{\text{red}} \to \text{BV}^{(2)}(\mathbb{R}) : \boldsymbol{\theta} \mapsto f_{\boldsymbol{\theta}}$ is a bijection onto the CPWL members of $\text{BV}^{(2)}(\mathbb{R})$ with finitely many linear regions.

For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\text{red}}$, the function $f_{\boldsymbol{\theta}}$ is a CPWL member of $\text{BV}^{(2)}(\mathbb{R})$ with finitely many linear regions. To prove the converse, let $f \in \text{BV}^{(2)}(\mathbb{R})$ be a CPWL function with finitely many linear regions. Using the canonical representation of f, there exist $c_0, c_1 \in \mathbb{R}$, $K \in \mathbb{N}$ and $a_k, \tau_k \in \mathbb{R}$ with $a_k \neq 0$ for $k = 1, \ldots, K$ such that

$$f(x) = c_0 + c_1 x + \sum_{k=1}^{K} a_k \text{ReLU}(x - \tau_k).$$

Now by defining $v_k = \frac{a_k}{\sqrt{|a_k|}}$, $w_k = \sqrt{|a_k|}$ and, $b_k = \sqrt{|a_k|}\tau_k$ for $k = 1, \ldots, K$, the homogeneity of the ReLU yields $f = f_{\boldsymbol{\theta}}$ with $\theta = (K, \mathbf{c}, \mathbf{v}, \mathbf{w}, \mathbf{b}) \in \boldsymbol{\Theta}_{\text{red}}$, where the latter inclusion is due to the equalities $|v_k| = |w_k|$ for $k = 1, \ldots, K$. ■

REFERENCES

[1] G. Wahba, *Spline Models for Observational Data*, Philadelphia, PA, USA: SIAM, 1990.

[2] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. New York, NY, USA: Springer, 2006.

[3] T. Hastie, R. Tibshirani, and J. Friedman, "Overview of supervised learning," in *The Elements of Statistical Learning*. New York, NY, USA: Springer, 2009, pp. 9–41.

[4] V. Katkovnik, K. Egiazarian, and J. Astola, "A spatially adaptive nonparametric regression image deblurring," *IEEE Trans. Image Process.*, vol. 14, no. 10, pp. 1469–1478, Oct. 2005.

[5] V. Katkovnik, "A multiresolution nonparametric regression for spatially adaptive image de-noising," *IEEE Signal Process. Lett.*, vol. 11, no. 10, pp. 798–801, Oct. 2004.

[6] G. Mateos and G. B. Giannakis, "Robust nonparametric regression via sparsity control with application to load curve data cleansing," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1571–1584, Apr. 2012.

[7] T. Poggio and F. Girosi, "Networks for approximation and learning," *Proc. IEEE*, vol. 78, no. 9, pp. 1481–1497, Sep. 1990.

[8] T. Poggio and F. Girosi, "Regularization algorithms for learning that are equivalent to multilayer networks," *Science*, vol. 247, no. 4945, pp. 978–982, 1990.

[9] G. Kimeldorf and G. Wahba, "Some results on Tchebycheffian spline functions," *J. Math. Anal. Appl.*, vol. 33, no. 1, pp. 82–95, 1971.

[10] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Proc. Int. Conf. Comput. Learn. Theory*, Berlin, Heidelberg: Springer, 2001, pp. 416–426.

[11] B. Schölkopf and A. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[12] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Adv. Comput. Math.*, vol. 13, no. 1, pp. 1–50, 2000.

[13] I. Steinwart and A. Christmann, *Support Vector Machines*. New York, NY, USA: Springer, 2008.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge: MIT Press, 2016.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[17] K. Jin, M. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4509–4522, Sep. 2017.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, Cham, Switzerland: Springer, 2015, pp. 234–241.

[19] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2574–2582.

[20] A. Fawzi, S.-M. Moosavi-Dezfooli, and P. Frossard, "The robustness of deep networks: A geometrical perspective," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 50–62, Nov. 2017.

[21] V. Antun, F. Renna, C. Poon, B. Adcock, and A. C. Hansen, "On instabilities of deep learning in image reconstruction and the potential costs of AI," *Proc. Nat. Acad. Sci.*, vol. 117, no. 48, pp. 30088–30095, 2020.

[22] J. Macdonald, S. Wäldchen, S. Hauch, and G. Kutyniok, "A rate-distortion framework for explaining neural network decisions," 2019, *arXiv:1905.11092*.

[23] C. Heiß, R. Levie, C. Resnick, G. Kutyniok, and J. Bruna, "In-distribution interpretability for challenging modalities," 2020, *arXiv:2007.00758*.

[24] M. Arjovsky, S. Chintala, and L. Bottou, "GAN Wasserstein," 2017, *arXiv:1701.07875*.

[25] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 537–546.

[26] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, "Exploring generalization in deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5947–5956.

[27] H. Gupta, K. H. Jin, H. Q. Nguyen, M. T. McCann, and M. Unser, "CNN-based projected gradient descent for consistent CT image reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1440–1453, Jun. 2018.

[28] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[29] Y. Sun, J. Liu, and U. S. Kamilov, "Block coordinate regularization by denoising," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 908–921, May 2020.

[30] J. Liu, Y. Sun, C. Eldeniz, W. Gan, H. An, and U. S. Kamilov, "RARE: Image reconstruction using deep priors learned without ground truth," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 6, pp. 1088–1099, Oct. 2020.

[31] Z. Wu, Y. Sun, A. Matlock, J. Liu, L. Tian, and U. S. Kamilov, "SIMBA: Scalable inversion in optical tomography using deep denoising priors," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 6, pp. 1163–1175, Oct. 2020.

[32] P. Bohra, A. Goujon, D. Perdios, S. Emery, and M. Unser, "Learning Lipschitz-controlled activation functions in neural networks for plug-and-play image reconstruction methods," in *Proc. NeurIPS Workshop Deep Learn. Inverse Problems*, 2021.

[33] D. E. McClure, "Perfect spline solutions of l_∞ extremal problems by control methods," *J. Approximation Theory*, vol. 15, no. 3, pp. 226–242, 1975.

[34] S. Karlin, "Interpolation properties of generalized perfect splines and the solutions of certain extremal problems. I," *Trans. Amer. Math. Soc.*, vol. 206, pp. 25–66, 1975.

[35] C. de Boor, "How small can one make the derivatives of an interpolating function ?," *J. Approximation Theory*, vol. 13, no. 2, pp. 105–116, 1975.

[36] C. A. Micchelli, T. J. Rivlin, and S. Winograd, "The optimal recovery of smooth functions," *Numerische Mathematik*, vol. 26, no. 2, pp. 191–200, 1976.

[37] C. de Boor, "On "best" interpolation," *J. Approximation Theory*, vol. 16, no. 1, pp. 28–42, 1976.

[38] A. Pinkus, "On smoothest interpolants," *SIAM J. Math. Anal.*, vol. 19, no. 6, pp. 1431–1441, 1988.

[39] U. von Luxburg and O. Bousquet, "Distance-based classification with Lipschitz functions," *J. Mach. Learn. Res.*, vol. 5, no., pp. 669–695, 2004.

[40] T. Debarre, Q. Denoyelle, M. Unser, and J. Fageot, "Sparsest continuous piecewise-linear representation of data," *J. Comput. Appl. Math.*, vol. 406, 2022.

[41] M. Unser, "A representer theorem for deep neural networks," *J. Mach. Learn. Res.*, vol. 20, no. 110, pp. 1–30, 2019.

[42] R. Pascanu, G. Montufar, and Y. Bengio, "On the number of response regions of deep feed forward networks with piece-wise linear activations," 2013, *arXiv:1312.6098*.

[43] G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio, "On the number of linear regions of deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2924–2932.

[44] R. Arora, A. Basu, P. Mianjy, and A. Mukherjee, "Understanding deep neural networks with rectified linear units," 2016, *arXiv:1611.01491*.

[45] T. Poggio, L. Rosasco, A. Shashua, N. Cohen, and F. Anselmi, "Notes on hierarchical splines, DCLNs and i-theory," Center Brains, Minds Mach. (CBMM), Tech. Rep., 2015.

[46] R. Balestriero and R. G. Baraniuk, "Mad max: Affine spline insights into deep learning," *Proc. IEEE*, vol. 109, no. 5, pp. 704–727, May 2021.

[47] R. Parhi and R. D. Nowak, "The role of neural network activation functions," *IEEE Signal Process. Lett.*, vol. 27, pp. 1779–1783, Sep. 2020.

[48] R. Parhi and R. D. Nowak, "Banach space representer theorems for neural networks and ridge splines," *J. Mach. Learn. Res.*, vol. 22, no. 43, pp. 1–40, 2021.

[49] R. Parhi and R. D. Nowak, "What kinds of functions do deep neural networks learn? insights from variational spline theory," 2021, *arXiv:2105.03361*.

[50] I. Savarese, P. Evron, D. Soudry, and N. Srebro, "How do infinite width bounded norm networks look in function space ?," in *Proc. 32nd Conf. Learn. Theory*, 2019, pp. 2667–2690.

[51] L. Schwartz, *Théorie des Distributions*. Paris, France: Hermann Paris, 1966.

[52] M. Unser and P. Tafti, *An Introduction to Sparse Stochastic Processes*. Cambridge, U.K.: Cambridge Univ. Press, 2014, p.367.

[53] W. Rudin, *Functional Analysis*. New York, NY, USA: McGraw-Hill Sci., Eng. Math., 1991.

[54] W. Rudin, *Real and Complex Analysis*. New York, NY, USA: Tata McGraw-Hill Educ., 2006.

[55] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[56] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[57] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[58] K. Bredies and H. K. Pikkarainen, "Inverse problems in spaces of measures," *ESAIM: Control, Optimisation Calculus Variations*, vol. 19, no. 1, pp. 190–218, 2013.

[59] E. J. Candès and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Commun. Pure Appl. Math.*, vol. 67, no. 6, pp. 906–956, 2014.

[60] V. Duval and G. Peyré, "Exact support recovery for sparse spikes deconvolution," *Found. Comput. Math.*, vol. 15, no. 5, pp. 1315–1355, 2015.

[61] S. Aziznejad and M. Unser, "Multikernel regression with sparsity constraint," *SIAM J. Math. Data Sci.*, vol. 3, no. 1, pp. 201–224, 2021.

[62] F. Bach, "Breaking the curse of dimensionality with convex neural networks," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 629–681, 2017.

[63] M. Unser, J. Fageot, and J. P. Ward, "Splines are universal solutions of linear inverse problems with generalized TV regularization," *SIAM Rev.*, vol. 59, no. 4, pp. 769–793, 2017.

[64] T. Debarre, S. Aziznejad, and M. Unser, "Hybrid-spline dictionaries for continuous-domain inverse problems," *IEEE Trans. Signal Process.*, vol. 67, no. 22, pp. 5824–5836, Nov. 2019.
[65] A. Flinth and P. Weiss, "Exact solutions of infinite dimensional total-variation regularized problems," *Inf. Inference: A J. IMA*, vol. 8, no. 3, pp. 407–443, 2019.
[66] K. Bredies and M. Carioni, "Sparsity of solutions for variational inverse problems with finite-dimensional data," *Calculus Variations Partial Differ. Equ.*, vol. 59, no. 1, pp. 1–26, 2020.
[67] Q. Denoyelle, V. Duval, G. Peyré, and E. Soubies, "The sliding Frank-Wolfe algorithm and its application to super-resolution microscopy," *Inverse Problems*, vol. 36, no. 1, 2019, Art. no. 014001.
[68] T. Debarre, J. Fageot, H. Gupta, and M. Unser, "B-Spline-based exact discretization of continuous-domain inverse problems with generalized TV regularization," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4457–4470, Jul. 2019.
[69] M. Simeoni, "Functional penalised basis pursuit on spheres," *Appl. Comput. Harmon. Anal.*, vol. 53, pp. 1–53, 2021.
[70] N. Weaver, *Lipschitz Algebras*. Singapore: World Scientific, 2018, pp. 1–34.
[71] M. Unser and J. Fageot, "Native Banach spaces for splines and variational inverse problems," 2019, *arXiv:1904.10818*.
[72] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1–4, pp. 259–268, 1992.
[73] C. De Boor, *A Practical Guide to Splines*, New York, NY, USA: Springer-Verlag, 1978.
[74] M. Unser, "Splines: A perfect fit for signal and image processing," *IEEE Signal Process. Mag.*, vol. 16, no. 6, pp. 22–38, Nov. 1999.
[75] P. Bohra, J. Campos, H. Gupta, S. Aziznejad, and M. Unser, "Learning activation functions in deep (spline) neural networks," *IEEE Open J. Signal Process.*, vol. 68, pp. 4688–4699, Nov. 2020.
[76] S. Aziznejad, H. Gupta, J. Campos, and M. Unser, "Deep neural networks with trainable activations and controlled Lipschitz constant," *IEEE Trans. Signal Process.*, vol. 68, pp. 4688–4699, Aug. 2020.
[77] M. Unser, "A unifying representer theorem for inverse problems and machine learning," *Found. Comput. Math.*, vol. 21, no. 4, pp. 941–960, 2021.
[78] S. Boyd, N. Parikh, and E. Chu, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.
[79] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, 2014.
[80] M. Unser and S. Aziznejad, "Convex optimization in sums of Banach spaces," *Appl. Comput. Harmon. Anal.*, vol. 56, pp. 1–25, 2022.
[81] A. Kurdila and M. Zabarankin, *Convex Functional Analysis*. Birkhäuser, Basel, Switzerland: Springer, 2006.
[82] H. Gupta, J. Fageot, and M. Unser, "Continuous-domain solutions of linear inverse problems with Tikhonov versus generalized TV regularization," *IEEE Trans. Signal Process.*, vol. 66, no. 17, pp. 4670–4684, Sep. 2018.
[83] R. J. Tibshirani, "The lasso problem and uniqueness," *Electron. J. Statist.*, vol. 7, pp. 1456–1490, 2013.