Journal of Machine Learning Research 20 (2019) 1-30 Submitted 6/18; Revised 5/19; Published 7/19

A Representer Theorem for Deep Neural Networks

Michael Unser MICHAEL.UNSER@EPFL.CH
Biomedical Imaging Group,
École polytechnique fédérale de Lausanne (EPFL),
CH-1015 Lausanne, Switzerland

Editor: Lorenzo Rosasco

Abstract

We propose to optimize the activation functions of a deep neural network by adding a corresponding functional regularization to the cost function. We justify the use of a second-order total-variation criterion. This allows us to derive a general representer theorem for deep neural networks that makes a direct connection with splines and sparsity. Specifically, we show that the optimal network configuration can be achieved with activation functions that are nonuniform linear splines with adaptive knots. The bottom line is that the action of each neuron is encoded by a spline whose parameters (including the number of knots) are optimized during the training procedure. The scheme results in a computational structure that is compatible with existing deep-ReLU, parametric ReLU, APL (adaptive piecewise-linear) and MaxOut architectures. It also suggests novel optimization challenges and makes an explicit link with ℓ_1 minimization and sparsity-promoting techniques.

Keywords: splines, regularization, sparsity, learning, deep neural networks, activation functions

1. Introduction

The basic regression problem in machine learning is to find a parametric representation of a function $f : \mathbb{R}^N \to \mathbb{R}$ given a set of data points $(\boldsymbol{x}_m, y_m) \in \mathbb{R}^{N+1}$ such that $f(\boldsymbol{x}_m)$ is close to y_m for $m = 1, \dots, M$ in an appropriate sense (Bishop, 2006). Classically, there are two steps involved. The first is the *design*, which can be abstracted in the choice of a given parametric class of functions $\boldsymbol{x} \mapsto f(\boldsymbol{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ encodes the parameters. For instance, $f(\boldsymbol{x}|\boldsymbol{\theta})$ could be a neural network with weights $\boldsymbol{\theta}$. The second is the *training*, which basically amounts to an interpolation/approximation problem where the chosen model is fit to the data. In practice, the optimal parameter $\boldsymbol{\theta}_0$ is determined via the functional minimization

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}} \sum_{m=1}^{M} E\big(y_m, f(\boldsymbol{x}_m|\boldsymbol{\theta})\big), \tag{1}$$

where $E : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is a convex error function that quantifies the discrepancy of the fit to the data. A classical choice is $E\big(y_m, f(\boldsymbol{x}_m|\boldsymbol{\theta})\big) = |y_m - f(\boldsymbol{x}_m|\boldsymbol{\theta})|^2$, which yields the least-squares solution.

The most delicate step is the design, because it has to deal with two conflicting requirements. First is the desire for universality, meaning that the parametric model $f(\boldsymbol{x}|\boldsymbol{\theta})$ should be flexible enough to allow for the faithful representation of a large class of functions—ideally, the complete family of continuous functions $\mathbb{R}^N \to \mathbb{R}$, as the dimensionality of $\boldsymbol{\theta}$

goes to infinity. Second is the quest for parsimony, meaning that the model should have a small number of parameters, which is expected to lead to an increase in robustness and trustworthiness.

This work aims at unifying the design of neural networks based on variational principles inspired by kernel methods. To set up the stage, we now briefly review two relevant approaches to supervised learning.

1.1. Kernel Methods

A kernel estimator is a linear model with adjustable parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_M) \in \mathbb{R}^M$ and predefined data centers $\boldsymbol{x}_1, \dots, \boldsymbol{x}_M \in \mathbb{R}^N$ of the form

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{m=1}^{M} \theta_m h(\boldsymbol{x}, \boldsymbol{x}_m), \tag{2}$$

where $\boldsymbol{x} \in \mathbb{R}^N$ is the input variable of the model and where $h : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ is a positive-definite kernel, a preferred choice being the Gaussian kernel $h(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{e}^{-\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2/\sigma^2}$ (Hofmann et al., 2008; Alvarez et al., 2012). This expansion is at the heart of the whole class of kernels methods, including radial-basis functions and support-vector machines (Schölkopf et al., 1997; Vapnik, 2013; Schölkopf and Smola, 2002).

The elegance of kernel estimators lies in that they can be justified based on regularization theory (Poggio and Girosi, 1990; Evgeniou et al., 2000; Poggio and Smale, 2003). The incentive there is to remove some of the arbitrariness of model selection by formulating the learning task as a global minimization problem that takes care of the design and training jointly. The property that makes such an integrated approach feasible is that any Hilbert space $\mathcal{H}$ of continuous functions on $\mathbb{R}^N$ has a unique *reproducing kernel* $h_{\mathcal{H}} : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ such that (*i*) $h_{\mathcal{H}}(\cdot, \boldsymbol{x}_m) \in \mathcal{H}$; and (*ii*) $\langle f, h_{\mathcal{H}}(\cdot, \boldsymbol{x}_m)\rangle_{\mathcal{H}} = f(\boldsymbol{x}_m)$ for any $\boldsymbol{x}_m \in \mathbb{R}^N$ and $f \in \mathcal{H}$ (Aronszajn, 1950). The idea, then, is to formulate the "regularized" version of Problem (1) as

$$f_{\mathrm{RKHS}} = \arg\min_{f \in \mathcal{H}} \left(\sum_{m=1}^{M} E\big(y_m, f(\boldsymbol{x}_m)\big) + \lambda \|f\|_{\mathcal{H}}^2 \right), \tag{3}$$

where the second term penalizes solutions with a large $\|\cdot\|_{\mathcal{H}}$-norm and $\lambda \in \mathbb{R}^+$ is an adjustable tradeoff factor. Under the assumption that the loss function E is convex, the representer theorem (Kimeldorf and Wahba, 1971; Schölkopf et al., 2001; Schölkopf and Smola, 2002) states that the solution of (3) exists, is unique, and is such that $f_{\mathrm{RKHS}} \in \mathrm{span}\{h_{\mathcal{H}}(\cdot, \boldsymbol{x}_m)\}_{m=1}^{M}$. This ultimately results in the same linear expansion as (2). The argument also applies the other way round since any positive-definite kernel h specifies a unique reproducing-kernel Hilbert space (RKHS) $\mathcal{H}$, which then provides the regularization functional $\|f\|_{\mathcal{H}}^2$ in (3) that is matched to the kernel estimator specified by (2).

The other remarkable feature of kernel expansions is their *universality*, under mild conditions on h (Micchelli et al., 2006). In other words, one has the guarantee that the generic linear model of (2) can reproduce *any* continuous function $f : \mathbb{R}^N \to \mathbb{R}$ to a desired degree of accuracy by including sufficiently many centers, with the error vanishing as $M \to \infty$. Moreover, because of the tight connection between kernels, RKHS, and splines (de Boor and

Lynch, 1966; Micchelli, 1986; Wahba, 1990), one can invoke standard results in approximation theory to obtain quantitative estimates of the approximation error of smooth functions as a function of M and of the widest gap between data centers (Wendland, 2005). Finally, there is a well-known link between kernel methods derived from regularization theory and neural networks, albeit "shallow" ones that involve a single nonlinear layer (Poggio and Girosi, 1990).

1.2. Deep Neural Networks

While kernel methods have been a major (and winning) player in machine learning since the mid '90s, they have been recently outperformed by deep neural networks (DNNs) in many real-world applications such as image classification (Krizhevsky et al., 2012), speech recognition (Hinton et al., 2012), and image segmentation (Ronneberger et al., 2015).

The leading idea of deep learning is to build more powerful learning architectures via the stacking/composition of simpler entities (see the review papers by LeCun, Bengio and Hinton (LeCun et al., 2015), Schmidhuber (Schmidhuber, 2015), and Goodfellow's recent textbook (Goodfellow et al., 2016) for more detailed explanations). In this work, we focus on the popular class of feedforward networks that involve a layered composition of affine transformations (linear weights) and pointwise nonlinearities. The deep structure of such a network is specified by its *node descriptor* $(N_0, N_1, \dots, N_L)$, where L is the total number of layers (depth of the network) and N_ℓ is the number of neurons at the ℓth layer. The action of a (scalar) neuron (or node) indexed by (n, ℓ) is described by $\sigma(\mathbf{w}_{n,\ell}^T \boldsymbol{x} - b_{n,\ell})$ where $\boldsymbol{x} \in \mathbb{R}^{N_{\ell-1}}$ denotes the multivariate input of the neuron, $\sigma : \mathbb{R} \to \mathbb{R}$ is a predefined activation function such as a sigmoid or a ReLU (rectified linear unit), $\mathbf{w}_{n,\ell} \in \mathbb{R}^{N_{\ell-1}}$ a set of linear weights, and $b_{n,\ell} \in \mathbb{R}$ an additive bias. The outputs of layer ℓ are then fed as the inputs of layer $(\ell + 1)$ for $\ell = 1, \dots, L - 1$.

To obtain a global description, we group the neurons within a given layer ℓ and specify two corresponding vector-valued maps.

1. Linear step $\boldsymbol{f}_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$ (affine transformation)

$$\boldsymbol{f}_\ell : \boldsymbol{x} \mapsto \boldsymbol{f}_\ell(\boldsymbol{x}) = \mathbf{W}_\ell \boldsymbol{x} - \mathbf{b}_\ell \tag{4}$$

   with weighting matrix $\mathbf{W}_\ell = [\mathbf{w}_{1,\ell} \cdots \mathbf{w}_{N_\ell,\ell}]^T \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and bias vector $\mathbf{b}_\ell = (b_{1,\ell}, \dots, b_{N_\ell,\ell}) \in \mathbb{R}^{N_\ell}$.

2. Nonlinear step $\boldsymbol{\sigma}_\ell : \mathbb{R}^{N_\ell} \to \mathbb{R}^{N_\ell}$ (activation functions)

$$\boldsymbol{\sigma}_\ell : \boldsymbol{x} = (x_1, \dots, x_{N_\ell}) \mapsto \boldsymbol{\sigma}_\ell(\boldsymbol{x}) = \big(\sigma_{1,\ell}(x_1), \dots, \sigma_{N_\ell,\ell}(x_{N_\ell})\big) \tag{5}$$

   with the possibility to adapt the scalar activation functions $\sigma_{n,\ell}$ on a per-node basis.

This allows us to describe the overall action of the full L-layer deep network by

$$\mathbf{f}_{\text{deep}}(\boldsymbol{x}) = (\boldsymbol{\sigma}_L \circ \boldsymbol{f}_L \circ \boldsymbol{\sigma}_{L-1} \circ \cdots \circ \boldsymbol{\sigma}_2 \circ \boldsymbol{f}_2 \circ \boldsymbol{\sigma}_1 \circ \boldsymbol{f}_1)(\boldsymbol{x}), \tag{6}$$

which makes its compositional structure explicit. The design step therefore consists in fixing the architecture of the deep neural net: One must specify $(N_0, N_1, \dots, N_L)$ together with the activation functions $\boldsymbol{\sigma}_\ell : \mathbb{R}^{N_\ell} \to \mathbb{R}^{N_\ell}$. The activations are traditionally chosen to be not only the same for all neurons within a layer, but also the same across layers. This results in a computational structure with adjustable parameters $\boldsymbol{\theta} = (\mathbf{W}_1, \dots, \mathbf{W}_L, \mathbf{b}_1, \dots, \mathbf{b}_L)$ (weights of the linear steps). These are then set during training via the minimization of (1), which is achieved by stochastic gradient descent with efficient error backpropagation (Rumelhart et al., 1986).

While researchers have considered a variety of possible activation functions, such as the traditional sigmoid, a preferred choice that has emerged over the years is the rectified linear unit $\mathrm{ReLU}(x) = \max(x, 0)$ (Glorot et al., 2011). The reasons that support this choice are multiple. The initial motivation was to promote sparsity (in the sense of a decrease in the number of active units), capitalizing on the property that ReLU acts as a gate and works well in combination with ℓ_1-regularization (Glorot et al., 2011). Second is the empirical observation that the training of very deep networks is much faster if the hidden layers are composed of ReLU activation functions (LeCun et al., 2015). Last but not least is the connection between deep ReLU networks and splines—to be further developed in this paper.

A key observation is that a deep ReLU network implements a multivariate input-output relation that is continuous and piecewise-linear (CPWL) (Montufar et al., 2014). This remarkable property is due to the ReLU itself being a linear spline, which has prompted Poggio et al. to interpret deep neural networks as hierarchical splines (Poggio et al., 2015). Moreover, it has been shown that any CPWL function admits a deep ReLU implementation (Wang and Sun, 2005; Arora et al., 2016), which is quite significant since the CPWL family has universal-approximation properties.

The ability of splines to effectively represent arbitrary (univariate) functions (de Boor, 1978; Schumaker, 1981; Unser, 1999) has also been exploited at the more local level of a neuron/node in a network. Several authors have proposed to use spline-related parametric models to optimize the shape of neural activation units. Existing designs include B-spline receptive fields (Lane et al., 1991), Catmul-Rom splines (Vecci et al., 1998), cubic spline activations (Guarnieri et al., 1999), adaptive piecewise-linear (APL) units (Agostinelli et al., 2015), and smooth piecewise-polynomial functions (Hou et al., 2017).

1.3. Road Map

Our purpose in this paper is to strengthen the connection between splines and multilayer ReLU networks even further. To that end, we formulate the design of a deep neural network globally within the context of regularization theory, in direct analogy with the variational formulation of kernel estimators given by (3). The critical aspect, of course, is the selection of an appropriate regularization functional which, for reasons that will be exposed next, will take us outside of the traditional realm of RKHS.

Having set the deep architecture of the neural network, we then formulate the training as a global optimization task whose outcome is a combined set of optimal neuronal activation functions and linear weights. The foundational role of the representer theorem (Theorem 4) is that it will provide us with the parametric representation of the optimal activations, which can then be leveraged to obtain a numerical implementation that is compatible with current architectures; in particular, the popular deep RELU networks.

2. From Deep Neural Networks to Deep Splines

Given the generic structure of a deep neural network, we are interested in investigating the possibility of optimizing the shape of the activation function(s) on a node-by-node basis. We now show how this can be achieved within the context of infinite-dimensional regularization theory.

2.1. Choice of Regularization Functional

For practical relevance, the scheme should favor simple solutions such as identity or linear scaling. This will retain the possibility of performing a classical linear regression. It is also crucial that the activation function σ be differentiable to be compatible with the chain rule when the backpropagation algorithm is used to train the network. Lastly, we want to promote activation functions that are locally linear (such as the ReLU) since these appear to work best in practice. If the two aforementioned constraints are satisfied, then the activation function is CPWL. As this property is conserved through (multivariate) composition, it implies that the resulting map $\boldsymbol{f}_{\text{deep}} : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ is CPWL as well, which is highly desirable for applications (Strang, 2018). Hence, an idealized solution would be a function $\sigma : \mathbb{R} \to \mathbb{R}$ whose second derivative vanishes almost everywhere.

As measure of sparsity, we use the "total-variation" norm $\|\cdot\|_{\mathcal{M}}$ associated with the Banach space

$$\mathcal{M}(\mathbb{R}) = \{f \in \mathcal{S}'(\mathbb{R}) : \|f\|_{\mathcal{M}} \stackrel{\triangle}{=} \sup_{\varphi \in \mathcal{S}(\mathbb{R}) : \|\varphi\|_\infty \le 1} \langle f, \varphi \rangle < \infty\}, \tag{7}$$

where $\mathcal{S}'(\mathbb{R})$ is Schwartz' space of tempered distributions, which is the continuous dual of $\mathcal{S}(\mathbb{R})$, the space of smooth and rapidly decreasing test functions on $\mathbb{R}$. Note that our definition of $\|\cdot\|_{\mathcal{M}}$ (by duality) is equivalent to the notion of total variation used in measure theory (Rudin, 1987). The critical point for us is that the latter is a slight extension of the L_1-norm: The basic property is that $\|f\|_{L_1} \stackrel{\triangle}{=} \int_{\mathbb{R}} |f(x)| \mathrm{d}x = \|f\|_{\mathcal{M}}$ for any $f \in L_1(\mathbb{R})$, which implies that $L_1(\mathbb{R}) \subseteq \mathcal{M}(\mathbb{R})$. However, the shifted Dirac distribution $\delta(\cdot - x_m) \notin L_1(\mathbb{R})$ for any shift $x_m \in \mathbb{R}$, while $\delta(\cdot - x_m) \in \mathcal{M}(\mathbb{R})$ with $\|\delta(\cdot - x_m)\|_{\mathcal{M}} = 1$, which shows that the space $\mathcal{M}(\mathbb{R})$ is (slightly) larger than $L_1(\mathbb{R})$.

To favor neuronal activation functions $\sigma : \mathbb{R} \to \mathbb{R}$ with "sparse" second derivatives, we shall therefore impose a bound on their second total variation, defined as

$$\mathrm{TV}^{(2)}(\sigma) \stackrel{\triangle}{=} \|\mathrm{D}^2 \sigma\|_{\mathcal{M}} = \sup_{\varphi \in \mathcal{S}(\mathbb{R}) : \|\varphi\|_\infty \le 1} \langle \mathrm{D}^2 \sigma, \varphi \rangle \tag{8}$$

where $\mathrm{D}^2 = \frac{\mathrm{d}^2}{\mathrm{d}x^2}$ is the second-derivative operator. The connection with ReLU is that $\mathrm{D}^2\mathrm{ReLU} = \delta$, which confirms that the ReLU activation function is intrinsically sparse with $\mathrm{TV}^{(2)}(\mathrm{ReLU}) = 1$.

Since our formulation involves a joint optimization of all network components, it is important to decouple the effect of the various stages. The only operation that is common to linear transformations and pointwise nonlinearities is a linear scaling, which is therefore transferable from one level to the next. Since most regularization schemes are scale-sensitive, it is essential to prevent such a transfer. We achieve this by restricting the class of admissible weight vectors $\mathbf{w}_{n,\ell}$ acting on a given node indexed by (n,ℓ) to those that have a unit norm. In other words, we shall normalize the scale of all linear modules with the introduction of the new variable $\mathbf{u}_{n,\ell} = \mathbf{w}_{n,\ell}/\|\mathbf{w}_{n,\ell}\|$.

2.2. Topological Properties of the Search Space

The formal definition of our native space (i.e., the space over which the optimization is performed) is

$$\mathrm{BV}^{(2)}(\mathbb{R}) = \{f : \mathbb{R} \to \mathbb{R} \ \text{ s.t. } \ \|\mathrm{D}^2 f\|_{\mathcal{M}} < \infty\}, \tag{9}$$

which is the class of functions with bounded second total variation.

To provide the proper mathematical context for our derivations, we also need to specify a suitable Banach topology for $\mathrm{BV}^{(2)}(\mathbb{R})$. In particular, doing so allows us to show that $\mathrm{BV}^{(2)}(\mathbb{R})$ has universal approximation properties. An important property in this respect is that $\mathrm{BV}^{(2)}(\mathbb{R})$ is a subspace of

$$C_{\mathrm{b},1}(\mathbb{R}) = \{f : \mathbb{R} \to \mathbb{R} \text{ s.t. } f \text{ is continuous and } \|f\|_{\infty,1} = \sup_{x\in\mathbb{R}} \ (1+|x|)^{-1}|f(x)| < \infty\}, \tag{10}$$

which is the space of continuous function with at most linear growth. The key features of $\mathrm{BV}^{(2)}(\mathbb{R})$ that constitute the theoretical backbone of our formulation can then be summarized as follows.

Theorem 1 *The space* $\mathrm{BV}^{(2)}(\mathbb{R})$ *equipped with the norm*

$$\|f\|_{\mathrm{BV}^{(2)}} \overset{\triangle}{=} \|\mathrm{D}^2 f\|_{\mathcal{M}} + \sqrt{|f(0)|^2 + |f(1)-f(0)|^2} \tag{11}$$

has the following properties:

1. *It is a Banach space.*
2. *The Dirac sampling functional* $\delta(\cdot - x_m) : f \mapsto f(x_m)$ *is weak* continuous on* $\mathrm{BV}^{(2)}(\mathbb{R})$ *for any* $x_m \in \mathbb{R}$*. Moreover, it satisfies the continuity bound*

$$|f(x_m)| = |\langle \delta(\cdot - x_m), f\rangle| \le (1+|x_m|)\ \|f\|_{\mathrm{BV}^{(2)}},$$

   for any $f \in \mathrm{BV}^{(2)}(\mathbb{R})$.
3. *Continuous embeddings:* $\mathcal{S}(\mathbb{R}) \subseteq \mathrm{BV}^{(2)}(\mathbb{R}) \subseteq C_{\mathrm{b},1}(\mathbb{R})$.

The reader is referred to Appendix A for the explanation of the concept of weak* continuity, and to Appendices B and C for the details of the proof. Except for Property 1, which is a corollary of (Unser et al., 2017, Theorem 5), the results in Theorem 1 are new. Property 2 is crucial for the proof of Lemma 2 which, as we shall see, is fundamental to our argumentation. Property 3 is very relevant as well since it implies that $\mathrm{BV}^{(2)}(\mathbb{R})$ is rich enough to reproduce any continuous function—in the present case, any neuronal activation—with an arbitrary degree of precision. This can be deduced by transitivity from the universal approximation properties of Schwartz' space $\mathcal{S}(\mathbb{R})$, which even happens to be dense in $\mathcal{S}'(\mathbb{R})$ (Schwartz, 1966).

2.3. Supporting Optimality Results

As preparation for our representer theorem, we now present our lemma on the $\mathrm{TV}^{(2)}$-optimality of piecewise-linear interpolation. This enabling result is deduced from the general spline theory presented in (Unser et al., 2017), as detailed in Appendix D. We then provide arguments to disqualify the use of the more conventional Sobolev-type regularization.

The subproblem of interest is to search for the optimal interpolant of a series of data points within the native space $\mathrm{BV}^{(2)}(\mathbb{R})$.

Lemma 2 ($\mathrm{TV}^{(2)}$**-optimality of piecewise-linear interpolants)** *Consider a series of scalar data points* $(x_m, y_m), m = 1, \dots, M$ *with* $M > 2$ *and* $x_1 \neq x_2$. *Then, under the hypothesis of feasibility (i.e.,* $y_{m_1} = y_{m_2}$ *whenever* $x_{m_1} = x_{m_2}$*), the extremal points of the interpolation problem*

$$\arg \min_{f \in \mathrm{BV}^{(2)}(\mathbb{R})} \|\mathrm{D}^2 f\|_{\mathcal{M}} \quad s.t. \quad f(x_m) = y_m, m = 1, \dots, M$$

are nonuniform splines of degree 1 *(a.k.a. piecewise-linear functions) with no more than* $(M - 2)$ *adaptive knots.*

The proof, together with the relevant background in functional analysis, is given in the appendix. The feasibility hypothesis in Lemma 2 is not restrictive since a function returns a single value for each input point. We are aware of two antecedents to Lemma 2 (i.e., (Fisher and Jerome, 1975, Corollary 2.2), (Mammen and van de Geer, 1997, Proposition 1)); these earlier results, however, are not in a form suitable to our purpose because they restrict the domain of f to a finite interval. Moreover, in both cases, the authors implicitly assume the weak* continuity of the sampling functionals, which is not obvious a priori. In fact, it is the groundwork for establishing this property—namely, the characterization of the predual space $C_{0,\mathrm{D}^2}(\mathbb{R})$ in Theorem 11—that is the most involved part of our derivation (see Appendix B). The statement in Lemma 2 is also more precise because it yields the full solution set (as the convex hull of the extremal points) and gives a stronger bound on the maximum number of knots.

Lemma 2 implies that there exists an optimal interpolator, not necessarily unique, whose generic parametric form is given by

$$f_{\mathrm{spline}}(x) = b_1 + b_2 x + \sum_{k=1}^{K} a_k (x - \tau_k)_+, \tag{12}$$

where $(x)_+ \overset{\triangle}{=} \max(x, 0) = \mathrm{ReLU}(x)$, with the caveat that the intrinsic spline descriptors, given by the (minimal) number K of knots and the knot locations $\tau_1, \dots, \tau_K \in \mathbb{R}$ are not known beforehand. This means that these descriptors need to be optimized jointly with the expansion coefficients $\boldsymbol{b} = (b_1, b_2) \in \mathbb{R}^2$ and $\boldsymbol{a} = (a_1, \dots, a_K) \in \mathbb{R}^K$. Ultimately, this translates into a solution that has a polygonal graph with breakpoints $f_{\mathrm{spline}}(\tau_k), k = 1, \dots, K$ and that perfectly interpolates the data points otherwise, as shown in Figure 1.

Since $\mathrm{TV}^{(2)}$-regularization penalizes the variations of the derivative, it will naturally produce (sparse) solutions with a small number of knots. This means that an optimal spline will typically have fewer knots than there are data points, while the list $\{\tau_1, \dots, \tau_K\}$

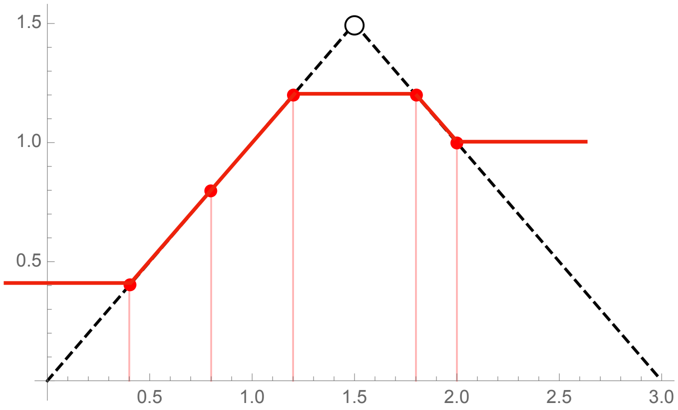

Figure 1: Conventional (solid line) vs. sparse (dashed line) piecewise-linear interpolants. The 5 data points are shown as dots; they coincide with the knots of the conventional interpolant. The sparse solution, by contrast, has a single knot at $\tau_1 = 1.5$ (circle), an argument value that is found in none of the data points.

of its knots, with $K < M$, may not necessarily be a subset of $\{x_1, \dots, x_M\}$, as illustrated in Figure 1. This push towards model simplification (Occam's razor) is highly desirable. It distinguishes this formulation of splines from the more conventional one, which, in the case of interpolation, simply tells us "to connect the dots" with $K = M$ and $\tau_m = x_m$ for $m = 1, \dots, M$ (see the solid-line curve in Figure 1).

It is well known that the classical linear interpolator is the solution of the variational problem in Proposition 3, which we like to see as the precursor of RKHS kernel methods (Prenter, 1975; Wahba, 1990).

Proposition 3 (Sobolev optimality of piecewise-linear interpolation) *Let the native space be the first-order Sobolev space* $H^1(\mathbb{R}) = \{f : \mathbb{R} \to \mathbb{R} \,\big|\, \|\mathrm{D}f\|_{L_2}^2 + |f(0)|^2 < \infty\}$. *Given a series of distinct data points* $(x_m, y_m), m = 1, \dots, M$, *the interpolation problem*

$$\arg \min_{f \in H^1(\mathbb{R})} \int_{\mathbb{R}} |\mathrm{D}f(x)|^2 \mathrm{d}x \;\; s.t. \quad f(x_m) = y_m, \;\; m = 1, \dots, M$$

has a unique piecewise-linear solution that can be written as

$$s_2(x) = b_1 + \sum_{m=1}^{M} a_m (x - x_m)_+. \tag{13}$$

While the result is elegant and translates into a straightforward implementation, the scheme can be cumbersome for large data sets because the number of parameters in (13) increases with the number of data points. The other limitation is that the use of $\|\mathrm{D}f\|_{L_2}$-regularization disqualifies the simple linear solution $f(x) = ax$, which has an infinite cost.

As one may expect, there are also direct extensions of Lemma 2 and Proposition 3 for regularized least-squares approximations. Moreover, the distinction between the two types of solutions—smoothing splines (Schoenberg, 1964) vs. adaptive regression splines (Mammen and van de Geer, 1997)—is even more striking[1] for noisy-data fitting applications, which brings us back to our initial goal: the design and training of neural networks.

2.4. Representer Theorem for Deep Neural Networks

Our aim is to determine the optimal activation functions for a deep neural network in a task-dependent fashion. This problem is inherently ill-posed because activations are infinite-dimensional entities while we only have access to finite data. As in the case of interpolation, we resolve the ambiguity by imposing an appropriate form of regularization. Having singled out $\mathrm{TV}^{(2)}$ as the most favorable choice, we now proceed with the enunciation of our representer theorem for deep neural networks. We have purposefully stated the optimization problem in a generic form that is compatible with the current practice in DNN. Specifically, the cost function in (15) includes a standard data term that penalizes data misfit plus a regularization to constrain the values of the linear weights of the network (e.g., $R_\ell(\mathbf{U}_\ell) = \|\mathbf{U}_\ell\|_{\mathrm{F}}^2$ in the case of the popular weight-decay penalty). The novelty is the additional optimization over the neuronal activations $\sigma_{n,\ell}$ and the insertion of the $\mathrm{TV}^{(2)}$ term to regularize their shape.

Theorem 4 (**$\mathrm{TV}^{(2)}$-optimality of deep spline networks**) *Let the L-layer feedforward neural network $\mathbf{f} : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ with node descriptor $(N_0, N_1, \dots, N_L)$ take the form*

$$\boldsymbol{x} \mapsto \mathbf{f}(\boldsymbol{x}) = (\boldsymbol{\sigma}_L \circ \boldsymbol{\ell}_L \circ \boldsymbol{\sigma}_{L-1} \circ \cdots \circ \boldsymbol{\ell}_2 \circ \boldsymbol{\sigma}_1 \circ \boldsymbol{\ell}_1)(\boldsymbol{x}), \tag{14}$$

which is an alternating composition of the normalized linear transformations $\boldsymbol{\ell}_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}, \boldsymbol{x} \mapsto \mathbf{U}_\ell \boldsymbol{x}$ with linear weights $\mathbf{U}_\ell = [\mathbf{u}_{1,\ell} \ \cdots \ \mathbf{u}_{N_\ell,\ell}]^T \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ such that $\|\mathbf{u}_{n,\ell}\| = 1$ and the nonlinear activations $\boldsymbol{\sigma}_\ell : \mathbb{R}^{N_\ell} \to \mathbb{R}^{N_\ell}, \boldsymbol{x} \mapsto \big(\sigma_{1,\ell}(x_1), \dots, \sigma_{N_\ell,\ell}(x_n)\big)$ with $\sigma_{1,\ell}, \dots, \sigma_{N_\ell,\ell} \in \mathrm{BV}^{(2)}(\mathbb{R})$. Given a series of data points $(\boldsymbol{x}_m, \boldsymbol{y}_m)_{m=1}^M$, we then define the training problem

$$\arg \min_{(\mathbf{U}_\ell),(\sigma_{n,\ell} \in \mathrm{BV}^{(2)}(\mathbb{R}))} \left(\sum_{m=1}^{M} E\big(\boldsymbol{y}_m, \mathbf{f}(\boldsymbol{x}_m)\big) + \mu \sum_{\ell=1}^{N} R_\ell(\mathbf{U}_\ell) + \lambda \sum_{\ell=1,}^{L} \sum_{n=1}^{N_\ell} \mathrm{TV}^{(2)}(\sigma_{n,\ell}) \right), \tag{15}$$

where $E : \mathbb{R}^{N_L} \times \mathbb{R}^{N_L} \to \mathbb{R}^+$ is an arbitrary convex error function such that $E(\boldsymbol{y}, \boldsymbol{y}) = 0$ for any $\boldsymbol{y} \in \mathbb{R}^{N_\ell}$, $R_\ell : \mathbb{R}^{N_\ell \times N_\ell} \to \mathbb{R}^+$ is some arbitrary convex cost that favors certain types of linear transformations, and $\lambda, \mu \in \mathbb{R}^+$ are two adjustable regularization parameters. If the solution of (15) *exists, then it is achieved by a deep spline network with individual activations of the form*

$$\sigma_{n,\ell}(x) = b_{1,n,\ell} + b_{2,n,\ell} x + \sum_{k=1}^{K_{n,\ell}} a_{k,n,\ell} (x - \tau_{k,n,\ell})_+, \tag{16}$$

[1] In the least-squares setting, one can adjust the strength of $\mathrm{TV}^{(2)}$-regularization to control the number of knots and thereby produce solutions with $K \ll M$.

with adaptive parameters $K_{n,\ell} \leq (M-2)$, $\tau_{1,n,\ell}, \dots, \tau_{K_{n,\ell},n,\ell} \in \mathbb{R}$, *and* $b_{1,n,\ell}, b_{2,n,\ell}, a_{1,n,\ell}, \dots, a_{K_{n,\ell},n,\ell} \in \mathbb{R}$.

Proof Let the function $\tilde{\mathbf{f}} : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ be a (not necessarily unique) solution of the problem summarized by (15). This solution is described by (14) with some optimal choice of transformation matrices $\tilde{\mathbf{U}}_\ell$ and pointwise nonlinearities $\tilde{\sigma}_{n,\ell} : \mathbb{R} \to \mathbb{R}$ for $\ell = 1, \dots, L$ and $n = 1, \dots, N_\ell$.

As we apply $\tilde{\mathbf{f}}$ to the data point $\boldsymbol{x} = \boldsymbol{x}_m$ and progressively move through the layers of the network, we generate a series of vectors $\boldsymbol{z}_{m,\ell} \in \mathbb{R}^{N_\ell}$, according to a recursive definition as follows.

- Initialization (input of the network): $\tilde{\boldsymbol{y}}_{m,0} = \boldsymbol{x}_m$.
- Recursive update: For $\ell = 1, \dots, L$, calculate

$$\boldsymbol{z}_{m,\ell} = (z_{1,m,\ell}, \dots, z_{N_\ell,m,\ell}) = \tilde{\mathbf{U}}_\ell \tilde{\boldsymbol{y}}_{m,\ell-1} \tag{17}$$

and construct $\tilde{\boldsymbol{y}}_{m,\ell} = (\tilde{y}_{1,m,\ell}, \dots, \tilde{y}_{N_\ell,m,\ell}) \in \mathbb{R}^{N_\ell}$ with

$$\tilde{y}_{n,m,\ell} = \tilde{\sigma}_{n,\ell}(z_{n,m,\ell}) \quad n = 1, \dots, N_\ell. \tag{18}$$

At the output level, we get $\tilde{\mathbf{f}}(\boldsymbol{x}_m) = \tilde{\boldsymbol{y}}_{m,L}$ for $m = 1, \dots, M$, which are the values that determine the data-fidelity part of the criterion associated with the optimal network and represented by the term $\sum_{m=1}^{M} E\big(\boldsymbol{y}_m, \mathbf{f}(\boldsymbol{x}_m)\big)$ in (15). Likewise, the specification of the optimal linear transforms $\tilde{\mathbf{U}}_1, \dots, \tilde{\mathbf{U}}_L$ fixes the regularization cost $\sum_{\ell=1}^{L} R_\ell(\mathbf{U}_\ell)$. Having set these quantities, we concentrate on the final element of the problem: the characterization of the "optimal" activations $\tilde{\sigma}_{n,\ell} : \mathbb{R} \to \mathbb{R}$ in-between the locations $z_{n,m,\ell}$ associated with the "auxiliary" data points $\tilde{y}_{n,m,\ell} = \tilde{\sigma}_{n,\ell}(z_{n,m,\ell})$, $m = 1, \dots, M$. The key is to recognize that we can now consider the various activation functions individually because the variation of $\tilde{\sigma}_{n,\ell}$ in-between data points is entirely controlled by $\mathrm{TV}^{(2)}(\tilde{\sigma}_{n,\ell})$ without any influence on the other terms of the cost functional. Since the solution $\tilde{\mathbf{f}}$ achieves the global optimum, we have that

$$\tilde{\sigma}_{n,\ell} \in \arg \min_{f \in \mathrm{BV}^{(2)}(\mathbb{R})} \|\mathrm{D}^2 f\|_{\mathcal{M}} \quad \text{s.t.} \quad f(z_{n,m,\ell}) = \tilde{y}_{n,m,\ell}, \quad m = 1, \dots, M,$$

where the "auxiliary" data pairs $(z_{n,m,\ell}, \tilde{y}_{n,m,\ell})$ are specified by (18). After this reformulation, we can apply Lemma 2, which proves that, at each node (n, ℓ), the minimum is achieved by a nonuniform spline with a number $K_{n,\ell}$ of knots smaller than the number of data points.

Since the hypothesis of feasibility is implicit in the construction, there is only one case not covered by Lemma 2: the singular scenario where all the auxiliary data points associated to a node are equal. Fortunately, this does not break the argument because such a configuration calls for a (zero-cost) solution of the form $b_1 + b_2 x$ (which is a special case of (12) with $K = 0$), except for the twist that there are now infinitely many possibilities with $b_1 + b_2 z_1 = \tilde{y}_1$. ■

This result translates into a computational structure where each node of the network $\big($with fixed index $(n, \ell)\big)$ is characterized by

- its number $0 \leq K = K_{n,\ell}$ of knots (ideally, much smaller than M);
- the location $\{\tau_{k,n,\ell}\}_{k=1}^{K_{n,\ell}}$ of these knots (equivalent to ReLU biases);
- the expansion coefficients $b_{1,n,\ell}, b_{2,n,\ell}, a_{1,n,\ell}, \dots, a_{K,n,\ell}$, also written as $\mathbf{b} = (b_1, b_2) \in \mathbb{R}^2$ and $\mathbf{a} = (a_1, \dots, a_K) \in \mathbb{R}^K$ to avoid notational overload.

The fundamental point is that these parameters (including the number of knots) are data-dependent and adjusted automatically through the minimization of (15). All this takes place during training.

3. Interpretation and Discussion

Theorem 4 tells us that we can configure a neural network optimally by restricting our attention to piecewise-linear activation functions $\sigma_{n,\ell}$, or *spline activations*, for short. In effect, this means that the "infinite-dimensional" minimization problem specified by (15) can be converted into a tractable finite-dimensional problem where, for each node (n, ℓ), the parameters to be optimized are the number $K_{n,\ell}$ of knots, the locations $\{\tau_{k,n,\ell}\}_{k=1}^{K_{n,\ell}}$ of the spline knots, and the linear weights $b_{1,n,\ell}, b_{2,n,\ell}, a_{1,n,\ell}, \dots, a_{K_{n,\ell},n,\ell} \in \mathbb{R}$. The enabling property is going to be (21), which converts the continuous-domain regularization into a discrete ℓ_1-norm. This is consistent with the expectation that bounding the second-order total-variation favors solutions with sparse second derivatives—i.e., linear splines with the fewest possible number of knots. The idea is that ℓ_1-minimization promotes a reduction of the number of active coefficients $a_{k,n,\ell}$ (Donoho, 2006; Unser et al., 2016).

The other important feature is that the knots are adaptive and that they can be learned during training using the standard backpropagation algorithm. What is required is the derivative of the activation functions. It is given by

$$\sigma'_{n,\ell}(x) = b_{2,n,\ell} + \sum_{k=1}^{K_{n,\ell}} a_{k,n,\ell} \mathbb{1}_{[\tau_{k,n,\ell},+\infty)}(x), \tag{19}$$

where $\mathbb{1}_{[\tau,+\infty)}(x)$ is an indicator function that is zero for $x < \tau$ and 1 otherwise. These derivatives are piecewise-constant splines with jumps of height $a_{k,n,\ell}$ at the knot locations $\tau_{k,n,\ell}$. By differentiating (19) once more, we get that

$$\sigma''_{n,\ell}(x) = \sum_{k=1}^{K_{n,\ell}} a_{k,n,\ell} \delta(x - \tau_{k,n,\ell}), \tag{20}$$

where δ is the Dirac distribution. Owing to the property that $\|\delta(\cdot - \tau_{k,n,\ell})\|_{\mathcal{M}} = 1$, we then readily deduce that

$$\mathrm{TV}^{(2)}\{\sigma_{n,\ell}\} = \|\sigma''_{n,\ell}\|_{\mathcal{M}} = \sum_{k=1}^{K_{n,\ell}} |a_{k,n,\ell}| = \|\mathbf{a}_{n,\ell}\|_1, \tag{21}$$

which converts the continuous-domain regularization into a more familiar minimum ℓ_1-norm constraint on the underlying expansion coefficients.

3.1. Link with Existing Techniques

What is even more interesting, from a practical point of view, is that the corresponding system translates into a deep ReLU network modulo a slight modification of the standard architecture described by (6). Indeed, the primary basis functions in (16) are shifted ReLUs, so that each spline activation $\sigma_{n,\ell}$ can be realized by way of a simple one-layer ReLU subnetwork with the spline knots being encoded in the biases. In particular, when the only active coefficients is $a_{n,\ell} = a_{1,n,\ell}$ (i.e., $b_{1,n,\ell} = 0$, $b_{2,n,\ell} = 0$, and $K_{n,\ell} = 1$), we have a perfect equivalence with the classical deep ReLU structure described by (6) with $\sigma_{n,\ell}(x) = (x)_+$. The enabling property is that

$$(\mathbf{w}_{n,\ell}^T \boldsymbol{x} - z_{n,\ell})_+ = (a_{n,\ell}\mathbf{u}_{n,\ell}^T \boldsymbol{x} - z_{n,\ell})_+ = a_{n,\ell}(\mathbf{u}_{n,\ell}^T \boldsymbol{x} - \tau_{n,\ell})_+,$$

with $\mathbf{u}_{n,\ell} = \mathbf{w}_{n,\ell}/\|\mathbf{w}_{n,\ell}\|$, $a_{n,\ell} = \|\mathbf{w}_{n,\ell}\|$, and $\tau_{n,\ell} = z_{n,\ell}/a_{n,\ell}$. Concretely, this means that, for every layer ℓ, we can absorb the single ReLU coefficients $a_{n,\ell}, n = 1, \dots, N_\ell$ into the prior linear transformation and consider unnormalized transformations $\mathbf{W}_\ell = [\mathbf{w}_{1,\ell} \ \dots \ \mathbf{w}_{N_\ell,\ell}]^T$ $\big($as in (4)$\big)$ rather than the normalized ones of Theorem 4 with $\mathbf{u}_{n,\ell} = \mathbf{w}_{n,\ell}/\|\mathbf{w}_{n,\ell}\|$.

Theorem 4 then suggests that the next step in complexity is to add the linear term $b_{1,n,\ell}+b_{2,n,\ell}x$ to each node, since its regularization cost vanishes. Interestingly, the suggested configuration—one ReLU plus an adjustable linear term per neuron—is equivalent to the parametric ReLU model (PReLU) of He et al. (2015) which has been found to systematically outperform the baseline ReLU configuration in real-world applications. The other design extreme is to let $\lambda \to \infty$, in which case the whole network collapses, leading to an affine mapping of the form $\mathbf{f}(\boldsymbol{x}) = \mathbf{W}\boldsymbol{x} - \mathbf{b}$ with $\mathbf{W} \in \mathbb{R}^{N_L \times N_0}$ and $\mathbf{b} \in \mathbb{R}^{N_L}$. More generally, the framework provides us with the possibility of controlling the number of knots (and, hence, the complexity of the network) through the simple adjustment of the regularization parameter λ, with the number of knots increasing as $\lambda \to 0$.

Among the various attempts in the literature to optimize the shape of the activation functions in deep neural networks, there is one scheme that is remarkably close to the optimal solution suggested by our theorem: the APL (adaptive piecewise-linear activation) framework of Agostinelli et al. (2015) in which each neuron is represented as a linear combination of shifted ReLUs, with the parameter being determined during training. The only difference is that their number of ReLUs is fixed a priori and that their model does not include the linear term $b_1 + b_2 x$. While the formulation of Agostinelli et al.'s does not involve any explicit regularization, they found in their experiments that is was helpful to add some mild ℓ_2 penalty on the ReLU coefficients (in contrast to the sparsity-promoting ℓ_1-penalty that results from our theorem) to avoid numerical instability. The good news in support of our theorem is that they report substantial improvement (9.4% and 7.5% relative-error decrease, respectively) on state-of-the-art CNN (with fixed RELU activations) on the CIFAR-10 and CIFAR-100 classification benchmarks.

A characteristic property of deep spline networks, to be considered here as a superset of the traditional deep ReLU networks, is that they produce an input-output relation that is continuous and piecewise-linear (CPWL) in the following sense: the corresponding function $\mathbf{f}$ is continuous $\mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$; its domain $\mathbb{R}^{N_0} = \bigcup_{k=1}^{K} P_k$ can be partitioned into a finite set of non-overlapping convex polytopes P_k over which it is affine (Tarela and Martinez, 1999; Wang and Sun, 2005). More precisely, $\mathbf{f}(\boldsymbol{x}) = \boldsymbol{f}_k(\boldsymbol{x})$ for all $\boldsymbol{x} \in P_k$ where $\boldsymbol{f}_k : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$

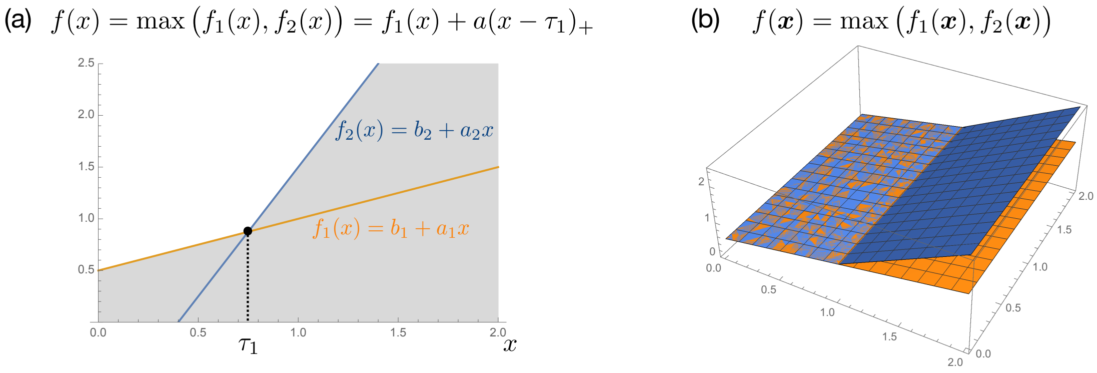

Figure 2: Deep-spline implementation of MaxOut with $N = 2$. (a) In one dimension, the spline parameters are $a = (a_2 - a_1)$ and $\tau_1 = (b_2 - b_1)/(a_2 - a_1)$. (b) In two or more dimensions, $f(\boldsymbol{x}) = b_1 + \boldsymbol{a}_1^T\boldsymbol{x} + a(\boldsymbol{u}^T\boldsymbol{x} - \tau_1)_+$, where $\boldsymbol{u}$ is the unit vector perpendicular to the hinge.

has the same parametric form as in (4). This simply follows from the observation that $\boldsymbol{\sigma}_\ell = (\sigma_{1,\ell}, \dots, \sigma_{N_\ell,\ell})$, with $\sigma_{n,\ell}$ as specified by (16), is CPWL and that the CPWL property is conserved through functional composition. In fact, the CPWL property for $N = 1$ is equivalent to the function being a nonuniform spline of degree 1.

Another powerful architecture that is known to generate CPWL functions is the MaxOut network (Goodfellow et al., 2013). There, the nonlinear steps $\boldsymbol{\sigma}_\ell$ in (6) are replaced by max-pooling operations. It turns out that these operations are also expressible in terms of deep splines, as illustrated in Figure 2 for the simple case where the maximum is taken over two inputs. Interestingly, this conversion requires the use of the linear term which is absent in conventional ReLU networks. This reinforces the argument made by Goodfellow et al. concerning the capability of MaxOut to learn activation functions.

An attractive feature that is offered by the deep-spline parameterization is the possibility of suppressing a network layer or, rather, of merging two adjacent ones when the optimal solution is such that $K_{n,\ell} = 0$ for ℓ fixed and $n = 1, \dots, N_\ell$. This is a property that results from the presence of the linear component and has not been exploited so far.

3.2. Generalizations

The optimality result in Theorem 4 holds for a remarkably broad family of cost functions, which should cover all cases of practical interest. The first condition is that the data term, as its name suggest, be solely dependent upon $\boldsymbol{y}_m$ and $\boldsymbol{f}(\boldsymbol{x}_m)$. The second is that the regularization of the weights (the part that constrains the linear steps) and the regularization of the individual activation functions be decoupled from each others. Obvious generalizations of the result include

- cases where the optimization in (15) is performed over a subset of the components while other network elements such as critical linear weights, activation functions[2], or

[2] A prominent example is the use of the softmax function (Bishop, 2006; Goodfellow et al., 2016) to convert the output of a neural network into a set of pseudo-probabilities.

even pooling operators are fixed beforehand (in particular, this includes the important subclass of networks that are not fully connected);

- configurations such as those found in convolutional networks where some (tunable) activation functions are shared among multiple nodes;
- generalized forms of regularization where $\mathrm{TV}^{(2)}(\sigma_{n,\ell})$ is substituted by $\psi\big(\mathrm{TV}^{(2)}(\sigma_{n,\ell})\big)$, where $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ is any monotonically increasing function.

While the first two scenarios require a slight reformulation of the optimization problem, it is still possible to invoke the same kind of "interpolation" argument as in the proof of Theorem 4. The third generalization is obvious since the (constrained) miminization of $\mathrm{TV}^{(2)}(\sigma_{n,\ell})$ is equivalent to the minimization of $\psi\big(\mathrm{TV}^{(2)}(\sigma_{n,\ell})\big)$.

The statement in Theorem 4 refers to the global optimum of (15), which is often hard to reach in practice because the underlying problem is highly non-convex. It turns out that the argument of the proof is also applicable to local minima and/or saddle points of the cost functional.

By relying on the supporting mathematics in Appendix D for general spline-admissible operators L, it is possible to revisit the proof of Theorem 4 to determine the parametric form of the optimal activations for higher-order versions of TV regularization; i.e., $\mathrm{TV}^{(n)}(\sigma) = \|\mathrm{D}^n\sigma\|_{\mathcal{M}}$. This yields optimal activations that are nonuniform polynomial splines of degree $n > 2$. While such solutions have a higher order of differentiability, they are less favorable globally because the underlying spline property is not retained through composition, meaning that the larger the number of layers, the larger the polynomal degree of the "polytopes" of the resulting network. By contrast, the CPWL property of the linear splines in (16) is preserved through composition, so that the resulting deep spline DNN can be also be interpreted as a flat (or shallow) multidimensional piecewise-linear spline. The other way of inducing CPWL activations is through the quadratic Sobolev 1 regularization of Proposition 3. However, this solution has two shortcomings: (i) its inability to represent the identity, which would result in an infinite cost, and (ii) its lack of sparsity.

3.3. Comparison with Kernel Methods

We like to contrast the result in Theorem 4 with the classical representer theorem of machine learning (Schölkopf et al., 2001). The commonality is that both theorems provide a parametric representation of the solution in the form of a linear "kernel" expansion. The primary distinction is that the classical representer theorem is restricted to "shallow" networks with $L = 1$. Yet, there is another difference even more crucial for our purpose: the fact that the knots τ_k in (16) are *adaptive* and few ($K \ll M$), while the centers $\boldsymbol{x}_m$ in (2) are *fixed* and as numerous as there are data points in the training set. In addition, the ReLU function $(x)_+$ is not a kernel in the traditional sense of the term because it is not positive-definite. We note, however, that it can be substituted by another equivalent spline generator $|x|$, which is conditionally positive-definite (Micchelli, 1986; Wendland, 2005). Again, the property that makes this feasible is the presence of the linear term $b_{1,n,\ell} + b_{2,n,\ell}x$.

There is also a conceptual similarity between the result of Theorem 4 and a recent representer theorem for deep kernel networks (Bohn et al., 2018) that results in a solution that is a composition of L multivalued kernel estimators of the classical RKHS form given by (2). Again, the two main differences with the present framework are: (i) each layer of the deep kernel network is a multivariate nonlinear map, which does not necessarily allow for affine transformations (e.g., linear regressions) and (ii) the kernel expansion in each layer requires as many basis functions as there are training data; this amounts to a total of $L \times M$ linear parameters. This can rapidly become prohibitive, not to mention the complexity of the underlying (non-convex) optimization task. The first shortcoming can easily be fixed by inserting intermediate affine transformations, in direct analogy with the type of architecture covered by Theorem 4. The second limitation is more fundamental and can probably only be removed by adopting some kind of generalized TV regularization in the spirit of Unser et al. (2017); in short, this calls for an extension of Theorem 4 to multivariate activations, which is currently work in progress.

3.4. Towards a Practical Implementation

While the solution of Theorem 4 is conceptually appealing, it can be expected to be harder to implement than fixed-kernel/RKHS methods since the optimization is not only over the linear weights $\boldsymbol{a}_{n,\ell}$ and $\boldsymbol{b}_{n,\ell}$, but also over the number and position of the corresponding spline knots. There is also always a risk that an increase in the number of degrees of freedom may compromise the generalization ability of the resulting network, which means that the method will need to be carefully tested and validated on real data. A possible strategy for making the optimization easier is to constrain the ReLU units to lie on a grid, in the spirit of Gupta et al. (2018), and then to rely on standard iterative ℓ_1-norm minimization techniques to produce a sparse solution (Donoho, 2006; Foucart and Rauhut, 2013; Unser et al., 2016). Such a scheme may still require some explicit knot-deletion step, either as post-processing or during the training iterations, to effectively trim down the number of parameters. A potential difficulty is that minimum $\mathrm{TV}^{(2)}$ interpolants are typically non-unique, because the underlying regularization is semi-convex. This means that the solution found by an iterative algorithm, assuming that the minimum of the regularization energy is achieved, is not necessarily the sparsest one within the (convex) solution set. Designing an algorithm that can effectively deal with this issue will be a very valuable contribution to the field.

4. Conclusion

The main contribution of this work is to provide the theoretical foundations for an integrated approach to neural networks where a subpart of the design—the optimal shaping of activations—can be transferred to the training part of the process and formulated as a global optimization problem. It also deepens the connection between splines and multilayer ReLU networks, as a pleasing side product. While the concept seems promising and includes the theoretical possibility of suppressing unnecessary layers, it raises a number of issues that can only be answered through extensive experimentation with real data. There are already strong indications in the literature (e.g., the improved performance of PReLU) of the practical usefulness of the linear-activation component that is suggested by the theory and not present in traditional ReLU systems. The task ahead is to demonstrate the capability of more complex spline activations to improve upon the state-of-the-art. (Except for a potential risk of over-parameterization, deep-spline networks should perform at least as well as deep ReLU, PReLU, or APL networks since the latter constitute a subset of the former.)

We expect the greatest challenge for training a deep spline network to be the proper optimization of the number of knots at each neuron, given that the solution with the fewest parameters is the most desirable. In short, we are still in need of a practical and efficient solution for training a deep neural network with fully adaptable activations that globally produces a continuous and piecewise-linear input-output relation; in other words, a DNN that implements an adaptive multidimensional linear spline.

Appendices

The proof of Lemma 2 is based on some foundational results in (Unser et al., 2017) that rely on compactness arguments requiring the weak* topology. We therefore start with a brief review of the relevant notions from functional analysis (Appendix A). We then specify the topology of $\mathrm{BV}^{(2)}(\mathbb{R})$ in Appendix B and precisely delineate its predual space in Theorem 11. This latter is the key to the proof of Theorem 1 and Lemma 2 that are presented in Appendix C and D, respectively.

Appendix A. Background on Continuity and Weak* Continuity

Definition 5 *Let $v : u \mapsto \langle v, u\rangle$ be a linear functional on a Banach space $\mathcal{U}$ equipped with the norm $\|\cdot\|_{\mathcal{U}}$. Then, $v : \mathcal{U} \to \mathbb{R}$ is said to be continuous if $\lim_{n\to\infty}\langle v, u_n\rangle = \langle v, u\rangle$ for any sequence (u_n) in $\mathcal{U}$ such that $\lim_{n\to\infty}\|u_n - u\|_{\mathcal{U}} = 0$.*

We recall that $\mathcal{U}'$ (the continuous dual of $\mathcal{U}$) is the vector space that is formed of the linear functionals that are continuous on $\mathcal{U}$; it is a Banach space equipped with the dual norm

$$\|v\|_{\mathcal{U}'} = \sup_{\varphi\in\mathcal{U}:\|u\|_{\mathcal{X}}\le 1} \langle v, u\rangle.$$

By following up the above property, one specifies the space of linear functionals that are continuous on $\mathcal{U}'$, which yields the Banach space $\mathcal{U}''$. A standard result in functional analysis is that $\mathcal{U}$ is continuously embedded in its bidual $\mathcal{U}''$, which is indicated as $\mathcal{U} \hookrightarrow \mathcal{U}''$, with the two spaces being isometrically isomorphic (i.e., $\mathcal{U} = \mathcal{U}''$) if and only if $\mathcal{U}$ is reflexive (Rudin, 1991). In other words, the construction of the bidual $\mathcal{U}''$ gets us back to the initial space in the reflexive case only.

A primary case of interest for this paper is $\mathcal{U}' = \mathcal{M}(\mathbb{R})$, which is not reflexive. For such a scenario, the proper way to deduce the *predual* space $\mathcal{U}$ from $\mathcal{V} = \mathcal{U}'$ is through the identification of the linear functionals that are weak*-continuous on $\mathcal{V}$.

Definition 6 (weak* topology) *A sequence $(v_n)_{n=1}^{\infty}$ in $\mathcal{V} = \mathcal{U}'$ is said to converge to v in the weak* topology if $\lim_{n\to\infty}\langle v_n - v, u\rangle = 0$ for all $u \in \mathcal{U}$.*

Definition 7 (weak* continuity) *A linear functional $u : \mathcal{U}' \to \mathbb{R}$ is said to be weak*-continuous if $\lim_{n\to\infty}\langle u, v_n\rangle = \langle u, v\rangle$ for any sequence (v_n) that converges to v in the weak* topology.*

Proposition 8 (see (Reed and Simon, 1980, Theorem IV.20, p. 114)) *The only weak* continuous linear functionals on $\mathcal{U}'$ are the elements of $\mathcal{U}$.*

The main point is that, despite the qualifier "weak", the functional property of weak* continuity is actually more restrictive than continuity.

In practice, it is relatively straightforward to establish the continuity of $u : \mathcal{V} \to \mathbb{R}$ since the property is equivalent to the existence of a constant $C > 0$ such

$$|\langle u, v\rangle| \leq C\|v\|_{\mathcal{V}}$$

for all $v \in \mathcal{V}$, which also yields $\|u\|_{\mathcal{V}'} \leq C < \infty$. By contrast, proving that $v : \mathcal{V} \to \mathbb{R}$ is weak*-continuous in the non-reflexive scenario requires the precise characterization of the predual of $\mathcal{V}$, which is typically more demanding mathematically. For instance, the property that $\mathcal{M}(\mathbb{R}) = \big(C_0(\mathbb{R})\big)'$ is a fundamental result in measure theory known as the Riesz-Markov theorem (Rudin, 1987). For example, the functionals $\varphi \mapsto \langle 1, \varphi\rangle$ and $\varphi \mapsto \langle \mathbb{1}_{[0,1]}, \varphi\rangle$ are continuous on $\mathcal{M}(\mathbb{R}) = \big(C_0(\mathbb{R})\big)'$ because the "generalized" functions 1 and $\mathbb{1}_{[0,1]}$ are bounded in the sup-norm. However, they both fail to be weak*-continuous. Indeed, $1 \notin C_0(\mathbb{R})$ because the unit-valued function does not decay at infinity, and $\mathbb{1}_{[0,1]} \notin C_0(\mathbb{R})$ because it is not continuous everywhere. In the latter example, we may recover weak* continuity by considering a smoothed version of the indicator function.

These considerations are central to the proof of Lemma 2 because it requires the weak* continuity of the sampling functional $\delta(\cdot - x_m)$. While the sampling operation is continuous on $\mathrm{BV}^{(2)}(\mathbb{R})$, it is not necessarily weak*-continuous; at least, not in the canonical topology that is proposed in (Unser et al., 2017, p. 780) (e.g., polynomial-spline example with $N_0 = 2$ and $\boldsymbol{\phi} = (\delta, -\delta')$). This is the reason why we need to revisit the construction of our native space, as detailed in Appendix B, and establish a new operational criterion for testing weak* continuity (Theorem 11).

Appendix B. Banach Structure of $\mathrm{BV}^{(2)}(\mathbb{R})$ and of Its Predual Space

While the definition of $\mathrm{BV}^{(2)}(\mathbb{R})$ given in (9) is convenient for expository purposes, it is not directly usable for mathematical analysis because the functional $\|\mathrm{D}^2 f\|_{\mathcal{M}}$ is only a semi-norm. We therefore rely on (Unser et al., 2017, Theorem 5) with $\mathrm{L} = \mathrm{D}^2$ to characterize $\mathrm{BV}^{(2)}(\mathbb{R})$ as a Banach space, with the choice of the biorthogonal system $(\boldsymbol{\phi}, \boldsymbol{p})$ for $\mathcal{N}_{\mathrm{D}^2}$ still being left open.

Proposition 9 (Banach structure of $\mathrm{BV}^{(2)}(\mathbb{R})$) *Let $(\boldsymbol{\phi}, \boldsymbol{p})$ be a biorthogonal system for $\mathcal{N}_{\mathrm{D}^2} = \mathrm{span}\{1, x\}$. Then, $\mathrm{BV}^{(2)}(\mathbb{R})$ equipped with the norm*

$$\|f\| = \|\mathrm{D}^2 f\|_{\mathcal{M}} + \sqrt{|\langle \phi_1, f\rangle|^2 + |\langle \phi_2, f\rangle|^2} \tag{22}$$

is a (non-reflexive) Banach space. Moreover, every $f \in \mathrm{BV}^{(2)}(\mathbb{R})$ has the unique direct-sum decomposition

$$f = \mathrm{G}_{\boldsymbol{\phi}}\{w\} + p, \tag{23}$$

where $w = \mathrm{D}^2 f \in \mathcal{M}(\mathbb{R})$, $p = \sum_{n=1}^{2} \langle f, \phi_n \rangle p_n \in \mathcal{N}_{\mathrm{D}^2}$, *and* $\mathrm{G}_{\boldsymbol{\phi}} : w \mapsto \int_{\mathbb{R}} g_{\boldsymbol{\phi}}(\cdot, y) w(y) \mathrm{d}y$, *with*

$$g_{\boldsymbol{\phi}}(x, y) = (x - y)_+ - \sum_{n=1}^{2} p_n(x) \langle \phi_n, (\cdot - y)_+ \rangle. \tag{24}$$

Central to our formulation is the unique operator $\mathrm{G}_{\boldsymbol{\phi}} : \mathcal{M}(\mathbb{R}) \to \mathrm{BV}^{(2)}(\mathbb{R})$ such that

$$\mathrm{D}^2 \mathrm{G}_{\boldsymbol{\phi}}\{w\} = w \qquad \text{(right-inverse property)} \tag{25}$$

$$\langle \phi_1, \mathrm{G}_{\boldsymbol{\phi}}\{w\} \rangle = 0, \quad \langle \phi_2, \mathrm{G}_{\boldsymbol{\phi}}\{w\} \rangle = 0 \qquad \text{(boundary conditions)} \tag{26}$$

for all $w \in \mathcal{M}(\mathbb{R})$. Specifically, (26) ensures the orthogonality of the two components of the direct-sum decomposition of f in (23), while (25) and the biorthogonality of $(\boldsymbol{\phi}, \boldsymbol{p})$ guarantees its unicity.

Since $\mathrm{BV}^{(2)}(\mathbb{R})$ is non-reflexive, the characterization of its predual is required for testing the hypothesis of weak* continuity. To that end, we first recall that the predual of $\mathcal{M}(\mathbb{R}) = \big(C_0(\mathbb{R})\big)'$ is the space $C_0(\mathbb{R})$ of continuous functions that vanish at infinity equipped with the sup-norm (Rudin, 1987). Moreover, since $\mathcal{S}(\mathbb{R})$ (Schwartz' space of smooth and rapidly decaying functions) is dense in $C_0(\mathbb{R})$ (Schwartz, 1966), the latter can also be described as the completion of $\mathcal{S}(\mathbb{R})$ equipped with the sup-norm, in conformity with the definition of $\mathcal{M}(\mathbb{R})$ given by (7).

We now present an explicit construction and characterization of the predual of $\mathrm{BV}^{(2)}(\mathbb{R})$. This description is consistent with an earlier theorem of ours (Unser et al., 2017, Theorem 6) applicable to general spline spaces; however, it contributes two novel elements: (i) the operational criterion for space membership provided by the first property, and (ii) the construction of the predual space $C_{\mathrm{D}^2,\boldsymbol{\phi}}(\mathbb{R})$ via the completion of $\mathcal{S}(\mathbb{R})$, which requires additional hypotheses on $\boldsymbol{\phi}$.

Definition 10 *Let* $\boldsymbol{p} = (p_1, p_2)$ *be a basis of* $\mathcal{N}_{\mathrm{D}^2} = \mathrm{span}\{1, x\}$ *and* $\boldsymbol{\phi} = (\phi_1, \phi_2)$ *a complementary set of (generalized) functions whose Fourier transforms are denoted by* $\widehat{\phi}_1, \widehat{\phi}_2$. *Then, the system* $(\boldsymbol{p}, \boldsymbol{\phi})$ *is said to be admissible for* D^2 *if*

1. *the basis functions are biorthogonal; i.e.,* $\langle \phi_m, p_n \rangle = \delta_{m,n}$, $(m, n = 1, 2)$
2. $\widehat{\phi}_1, \widehat{\phi}_2 \in L_{1,2}(\mathbb{R}) = \{f : \mathbb{R} \mapsto \mathbb{R} \ \big|\ \int_{\mathbb{R}} (1 + |\omega|)^{-2} |f(\omega)| \mathrm{d}\omega < \infty\}$, *with the two functions being continuously differentiable twice at* $\omega = 0$.

Theorem 11 (Predual of native space) *Let* $(\boldsymbol{\phi}, \boldsymbol{p})$ *be an admissible system in the sense of Definition 10. Then, the function space*

$$C_{\mathrm{D}^2,\boldsymbol{\phi}}(\mathbb{R}) = \{g = \mathrm{D}^2 v + a_1 \phi_1 + a_2 \phi_2 : v \in C_0(\mathbb{R}), \mathbf{a} = (a_1, a_2) \in \mathbb{R}^2\} \tag{27}$$

has the following properties:

1. *every* $g \in C_{\mathrm{D}^2,\boldsymbol{\phi}}(\mathbb{R})$ *has a unique direct-sum representation as in* (27) *with* $v = \mathrm{G}_{\boldsymbol{\phi}}^*\{g\}$, $a_1 = \langle p_1, g \rangle$, *and* $a_2 = \langle p_2, g \rangle$, *where* $\mathrm{G}_{\boldsymbol{\phi}}^*$ *is the adjoint of* $\mathrm{G}_{\boldsymbol{\phi}}$ *specified by* (24);

2. *$C_{\mathrm{D}^2,\boldsymbol{\phi}}(\mathbb{R})$ is a (non-reflexive) Banach space equipped with the norm*

$$\|g\|_{C_{\mathrm{D}^2,\boldsymbol{\phi}}} \overset{\triangle}{=} \max(\|\mathrm{G}^*_{\boldsymbol{\phi}} g\|_\infty, \|\boldsymbol{p}(g)\|_2) = \max(\|v\|_\infty, \|\boldsymbol{a}\|_2); \tag{28}$$

3. *$C_{\mathrm{D}^2,\boldsymbol{\phi}}(\mathbb{R})$ is the predual of $\mathrm{BV}^{(2)}(\mathbb{R})$ in Proposition 9; i.e., $\mathrm{BV}^{(2)}(\mathbb{R}) = \big(C_{\mathrm{D}^2,\boldsymbol{\phi}}(\mathbb{R})\big)'$;*

4. *$C_{\mathrm{D}^2,\boldsymbol{\phi}}(\mathbb{R})$ is the completion of $\mathcal{S}(\mathbb{R})$ equipped with the $\|\cdot\|_{C_{\mathrm{D}^2,\boldsymbol{\phi}}}$-norm.*

Proof : The main idea is that the construction expressed by (27) is the direct sum of the two linear spaces $\mathcal{U}$ and $\mathcal{N}_{\boldsymbol{\phi}}$ whose Banach topology and completion properties are revealed next.

(i) Topology of the Space $\mathcal{N}_{\boldsymbol{\phi}} = \mathrm{span}\{\phi_1, \phi_2\}$ and of Its Dual
The space $\mathcal{N}_{\boldsymbol{\phi}}$ collects the two last components of g in (27) and is equipped with the discrete ℓ_2-norm $\|\phi\|_{\mathcal{N}_{\boldsymbol{\phi}}} = \|\mathbf{a}\|_2$ with $\mathbf{a} = \boldsymbol{p}(\phi) = (\langle p_1, \phi\rangle, \langle p_2, \phi\rangle)$. We also specify the projection operator $C_{\mathrm{D}^2,\boldsymbol{\phi}}(\mathbb{R}) \to \mathcal{N}_{\boldsymbol{\phi}}$:

$$\mathrm{Proj}_{\mathcal{N}_{\boldsymbol{\phi}}}\{g\} = \langle p_1, g\rangle \phi_1 + \langle p_2, g\rangle \phi_2.$$

The complementary space is $\mathcal{N}_{\boldsymbol{p}} = \mathrm{span}\{p_1, p_2\}$ equipped with the norm $\|p\|_{\mathcal{N}_{\boldsymbol{p}}} = \|\boldsymbol{p}(p)\|_2 = \|\mathbf{b}\|_2$ with $\mathbf{b} = \boldsymbol{\phi}(p) = (\langle \phi_1, p\rangle, \langle \phi_2, p\rangle)$. Thanks to the biorthogonality of $\boldsymbol{\phi}$ and $\boldsymbol{p}$, for all $p = b_1 p_1 + b_2 p_2 \in \mathcal{N}_{\boldsymbol{p}}$, we have that

$$\|p\|_{\mathcal{N}'_{\boldsymbol{\phi}}} = \sup_{\phi \in \mathcal{N}_{\boldsymbol{\phi}} : \|\phi\|_{\mathcal{N}_{\boldsymbol{\phi}}} \le 1} \langle \phi, p\rangle = \sup_{\mathbf{a} \in \mathbb{R}^2 : \|\mathbf{a}\|_2 \le 1} \mathbf{a}^T \mathbf{b} = \|\mathbf{b}\|_2 = \|p\|_{\mathcal{N}_{\boldsymbol{p}}},$$

which shows that $\mathcal{N}_{\boldsymbol{p}} = \mathcal{N}'_{\boldsymbol{\phi}}$ is the continuous dual of $\mathcal{N}'_{\boldsymbol{\phi}}$.

(ii) Range of the Operator $\mathrm{G}^*_{\boldsymbol{\phi}}$
To derive the required properties, we restrict the domain of $\mathrm{G}^*_{\boldsymbol{\phi}}$ to the subspace

$$\mathcal{S}_{\boldsymbol{p}^\perp}(\mathbb{R}) = \big\{\psi \in \mathcal{S}(\mathbb{R}) : \langle p_1, \psi\rangle = 0, \langle p_2, \psi\rangle = 0\big\} \subset \mathcal{S}(\mathbb{R}).$$

By using the explicit form (24) of the kernel of $\mathrm{G}_{\boldsymbol{\phi}}$, we find that, for any $\psi \in \mathcal{S}_{\boldsymbol{p}^\perp}(\mathbb{R})$,

$$\begin{aligned}\mathrm{G}^*_{\boldsymbol{\phi}}\{\psi\}(x) &= \int_{\mathbb{R}} \big((y-x)_+ - q_1(x)p_1(y) - q_2(x)p_2(y)\big)\psi(y)\mathrm{d}y \\ &= \int_{\mathbb{R}} (y-x)_+ \psi(y)\mathrm{d}y - q_1(x) \underbrace{\langle p_1, \psi\rangle}_{0} - q_2(x) \underbrace{\langle p_2, \psi\rangle}_{0} \\ &= \mathrm{D}^{-2*}\{\psi\}(x),\end{aligned} \tag{29}$$

where $q_n(y) = \langle \phi_n, (\cdot - y)_+\rangle$ for $n = 1, 2$, and D^{-2*} is the 2-fold (adjoint) integration operator whose frequency response is $\big(-1/\omega^2 - \mathrm{i}\pi\delta'(\omega)\big) = \mathcal{F}\{(-x)_+\}(\omega)$. Based on (29), we then show that

$$\forall \psi \in \mathcal{S}_{\boldsymbol{p}^\perp}(\mathbb{R}) : \quad \mathrm{G}^*_{\boldsymbol{\phi}}\{\psi\} \in C_0(\mathbb{R}), \tag{30}$$

which, as we shall see, implies the boundedness of $\mathrm{G}^*_{\boldsymbol{\phi}} : \mathcal{S}_{\boldsymbol{p}^\perp}(\mathbb{R}) \to C_0(\mathbb{R}) \hookrightarrow \mathcal{S}'(\mathbb{R})$. Property (30) is established by examining the Fourier transform[3] of $f = \mathrm{D}^{-2*}\{\psi\}$ given by

$$\hat{f}(\omega) = -\widehat{\psi}(\omega)/\omega^2 - \mathrm{i}\pi\big(\widehat{\psi}(0)\delta'(\omega) - \widehat{\psi}^{(1)}(0)\delta(\omega)\big) \tag{31}$$

with $\widehat{\psi} = \mathcal{F}\{\psi\} \in \mathcal{S}(\mathbb{R})$. Since $\boldsymbol{p}(\psi) = \boldsymbol{0} \Leftrightarrow \widehat{\psi}(0) = \widehat{\psi}^{(1)}(0) = 0$, we first simplify (31) to $\hat{f}(\omega) = \big(-\widehat{\psi}(\omega)/\omega^2\big)$ and then invoke a Taylor-series argument to deduce the continuity of $\hat{f}(\omega)$ at $\omega = 0$. This, together with the boundedness and rapid decay of $\widehat{\psi}(\omega)$, implies that $\hat{f} \in L_1(\mathbb{R})$. The announced result—the continuity, boundedness, and decay of $f(x)$ at infinity—then follows from the Riemann-Lebesgue lemma.

(iii) The Banach Topology of $\mathcal{U}$
The definition of $\mathcal{U}$, which corresponds to the first component in (27), is

$$\mathcal{U} = \{f = \mathrm{D}^2 v : v \in C_0(\mathbb{R})\},$$

equipped with the norm $\|\mathrm{D}^2 v\|_{\mathcal{U}} = \|v\|_\infty$, which establishes an isometric isomorphism with $C_0(\mathbb{R})$. Our intent now is to prove that $\|f\|_{\mathcal{U}} = \|\mathrm{G}^*_{\boldsymbol{\phi}} f\|_\infty$ for all $f \in \mathcal{U}$, which is equivalent to showing that $\mathrm{G}^*_{\boldsymbol{\phi}}$ is the inverse of $\mathrm{D}^2 : C_0(\mathbb{R}) \to \mathcal{U}$.

We shall achieve this through an extension process that builds upon the properties of the operator $\mathrm{G}^*_{\boldsymbol{\phi}}$ established in Step (*ii*). We start by considering the semi-norm $\psi \mapsto \|\psi\|_{\tilde{\mathcal{U}}} \overset{\triangle}{=} \|\mathrm{G}^*_{\boldsymbol{\phi}}\psi\|_\infty$, which is well-defined over $\mathcal{S}_{\boldsymbol{p}^\perp}(\mathbb{R}) \subset \mathcal{S}(\mathbb{R})$. Since $\mathrm{G}^*_{\boldsymbol{\phi}}\psi = \mathrm{D}^{-2*}\psi$ for all $\psi \in \mathcal{S}_{\boldsymbol{p}^\perp}(\mathbb{R})$ and $\mathrm{D}^2\mathrm{D}^{-2*}\varphi = \varphi$ any $\varphi \in \mathcal{S}(\mathbb{R})$, we have that $\|\mathrm{G}^*_{\boldsymbol{\phi}}\psi\|_\infty = 0 \Leftrightarrow \psi = 0$, which shows that $\|\cdot\|_{\tilde{\mathcal{U}}}$ is a norm over $\mathcal{S}_{\boldsymbol{p}^\perp}(\mathbb{R})$, as expected. This allows us to rephrase the inclusion property from Step (*ii*) as: $\mathrm{G}^*_{\boldsymbol{\phi}}$ isometrically maps $(\mathcal{S}_{\boldsymbol{p}^\perp}(\mathbb{R}), \|\cdot\|_{\tilde{\mathcal{U}}})$ to the Banach space $(C_0(\mathbb{R}), \|\cdot\|_\infty)$, which is the form suitable for the bounded linear transformation (BLT) extension theorem.

Theorem 12 (Reed and Simon (1980, Theorem I.7, p. 9)) *Let* G *be a bounded linear transformation from a normed space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ to a complete normed space $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$. Then,* G *has a unique extension to a bounded linear transformation (with the same bound) from the completion of $\mathcal{X}$ to $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$.*

Consequently, the restricted operator from Step (*ii*) uniquely extends to an isometry $\mathrm{G}^*_{\boldsymbol{\phi}} : \tilde{\mathcal{U}} \to C_0(\mathbb{R})$ where the Banach space $\tilde{\mathcal{U}}$ is the completion of $\mathcal{S}_{\boldsymbol{p}^\perp}(\mathbb{R})$ in the $\|\cdot\|_{\tilde{\mathcal{U}}}$-norm. The final element is that $\mathrm{D}^2\mathrm{G}^*_{\boldsymbol{\phi}}\psi = \mathrm{D}^2\mathrm{D}^{-2*}\psi = \psi$ for all $\mathcal{S}_{\boldsymbol{p}^\perp}(\mathbb{R}) \subseteq \tilde{\mathcal{U}}$, which indicates that D^2 is the inverse of $\mathrm{G}^*_{\boldsymbol{\phi}}$ on $\mathcal{S}_{\boldsymbol{p}^\perp}(\mathbb{R})$. Since the latter is a dense subset of $\tilde{\mathcal{U}}$, we can extend the property to the entire space, which ultimately proves that $\mathcal{U} = \tilde{\mathcal{U}}$.

(iv) The space $C_{\mathrm{D}^2,\boldsymbol{\phi}}(\mathbb{R}) = \mathcal{U} \oplus \mathcal{N}_{\boldsymbol{\phi}}$
The inclusion $g \in C_{\mathrm{D}^2,\boldsymbol{\phi}}(\mathbb{R})$ is equivalent to $g = f + \phi$ where $f = \mathrm{D}^2 v$ with $v \in C_0(\mathbb{R})$ and $\phi = a_1\phi_1 + a_2\phi_2$. The components (f, ϕ) are retrieved as $f = \mathrm{Proj}_{\mathcal{U}}\{g\} = \mathrm{D}^2\mathrm{G}^*_{\boldsymbol{\phi}} g$ and $\phi = \mathrm{Proj}_{\mathcal{N}_{\boldsymbol{\phi}}}\{g\}$. The conditions $\mathrm{G}^*_{\boldsymbol{\phi}}\phi = 0$ and $\mathrm{Proj}_{\mathcal{N}_{\boldsymbol{\phi}}}\{\mathrm{D}^2 v\} = 0$ for all $\phi \in \mathcal{N}_{\boldsymbol{\phi}}$ and $v \in C_0(\mathbb{R})$ ensure that $\mathcal{U} \cap \mathcal{N}_{\boldsymbol{\phi}} = \{0\}$ so that the sum is direct. The other relevant identity

[3] We use the product rule $\psi(\cdot)\delta' = \psi(0)\delta' - \psi'(0)\delta$, which follows from the definition of the distribution $\delta' : \varphi \mapsto \langle \delta', \varphi \rangle = -\varphi'(0)$.

from Step (*iii*) is $f = \mathrm{D}^2\mathrm{G}^*_{\boldsymbol{\phi}} f$ for all $f \in \mathcal{U}$. Consequently, $C_{\mathrm{D}^2,\boldsymbol{\phi}}(\mathbb{R}) = \mathcal{U} \oplus \mathcal{N}_{\boldsymbol{\phi}}$ is a Banach space when equipped with $\|(f,\phi)\|_\infty$, which is the composite norm given by (28).

(v) The Identification of $\mathrm{BV}^{(2)}(\mathbb{R}) = \big(C_{\mathrm{D}^2,\boldsymbol{\phi}}(\mathbb{R})\big)'$
First, we identify the norm of $\mathcal{U}'$ by applying a standard duality argument:

$$\begin{aligned}\|u^*\|_{\mathcal{U}'} &= \sup_{u\in\mathcal{U}:\|u\|_{\mathcal{U}}\le 1} \langle u^*, u\rangle = \sup_{v\in C_0(\mathbb{R}):\|v\|_\infty\le 1} \langle u^*, \mathrm{D}^2 v\rangle \\ &= \sup_{v\in\mathcal{S}(\mathbb{R}):\|v\|_\infty\le 1} \langle \mathrm{D}^2 u^*, v\rangle = \|\mathrm{D}^2 u^*\|_{\mathcal{M}}\end{aligned}$$

where we have used the identity $u = \mathrm{D}^2 v$ with $v \in C_0(\mathbb{R})$ and the denseness of $\mathcal{S}(\mathbb{R})$ in $C_0(\mathbb{R})$. The dual of $C_{\mathrm{D}^2,\boldsymbol{\phi}}(\mathbb{R})$ in Step (*iv*) is then given by $\mathcal{U}' \oplus \mathcal{N}'_{\boldsymbol{\phi}} = \mathcal{U}' \oplus \mathcal{N}_{\boldsymbol{p}}$ equipped with the composite norm $\|(u^*, p)\|_1 = \|u^*\|_{\mathcal{U}'} + \|\boldsymbol{\phi}(p)\|_2 = \|\mathrm{D}^2 f\|_{\mathcal{M}} + \|\boldsymbol{\phi}(f)\|_2 = \|f\|_{\mathrm{BV}^{(2)}}$, which is the dual norm of $\|(f,\phi)\|_\infty = \max(\|f\|_{\mathcal{U}}, \|\boldsymbol{p}(\phi)\|_2)$.

(vi) The Space $C_{\mathrm{D}^2,\boldsymbol{\phi}}(\mathbb{R})$ *Is The Completion of* $\mathcal{S}(\mathbb{R})$ *in the* $\|\cdot\|_{C_{\mathrm{D}^2,\boldsymbol{\phi}}}$*-Norm*
The idea is to amend the extension technique of Step (*iii*) by selecting a second biorthogonal system $(\boldsymbol{\varphi}, \boldsymbol{p})$ such that $\mathcal{N}_{\boldsymbol{\varphi}} = \mathrm{span}\{\varphi_1, \varphi_2\} \subset \mathcal{S}(\mathbb{R})$. This yields the direct-sum decomposition of $\varphi = \tilde{\psi} + \tilde{\phi} \in \mathcal{S}(\mathbb{R})$ with $\tilde{\phi} = \mathrm{Proj}_{\mathcal{N}_{\boldsymbol{\varphi}}}\{\varphi\} \in \mathcal{N}_{\boldsymbol{\varphi}}$ and $\tilde{\psi} = (\varphi - \tilde{\phi}) \in \mathcal{S}_{\boldsymbol{p}^\perp}(\mathbb{R})$. While we already know that $\mathrm{G}^*_{\boldsymbol{\phi}}\tilde{\psi} \in C_0(\mathbb{R})$, the delicate point is to make sure that the same holds true for $\mathrm{G}^*_{\boldsymbol{\phi}}\tilde{\phi}$. Since $\tilde{\phi} \in \mathrm{span}\{\varphi_1, \varphi_2\}$, the last requirement is equivalent to

$$\mathrm{G}^*_{\boldsymbol{\phi}}\{\varphi_n\} = \mathrm{D}^{-2*}(\mathrm{Id} - \mathrm{Proj}_{\mathcal{N}_{\boldsymbol{\phi}}})\{\varphi_n\} = \mathrm{D}^{-2*}\{\varphi_n - \phi_n\} \in C_0(\mathbb{R}) \tag{32}$$

for $n = 1, 2$. With the same arguments as in Step (*ii*) (Riemann-Lebesgue lemma), we ensure that (32) is met by imposing the Fourier-domain condition

$$\frac{\widehat{\phi}_n(\omega) - \widehat{\varphi}_n(\omega)}{\omega^2} \in L_1(\mathbb{R}), \tag{33}$$

which results from the second hypothesis in Definition 10. In effect, the role of $\widehat{\varphi}_n \in \mathcal{S}(\mathbb{R})$ in (33) is to temper the singularity of $1/\omega^2$ at the origin, in reason of the condition $\boldsymbol{p}(\varphi_n - \phi_n) = \boldsymbol{0}$, which induces a second-order zero in the numerator. This correction does not impact integrability otherwise because of the rapid decay of $\widehat{\varphi}_n$.

Having established that $\mathrm{G}^*_{\boldsymbol{\phi}}\{\tilde{\psi} + \tilde{\phi}\} \in C_0(\mathbb{R})$, we can now check that

$$\|\varphi\|_{C_{\mathrm{D}^2,\boldsymbol{\phi}}} = \max(\|\mathrm{G}^*_{\boldsymbol{\phi}}\{\tilde{\psi} + \tilde{\phi}\}\|_\infty, \|\boldsymbol{p}(\tilde{\phi})\|_2) = 0 \Leftrightarrow (\tilde{\psi}, \tilde{\phi}) = (0, 0) \Leftrightarrow \varphi = 0,$$

which proves that $\|\cdot\|_{C_{\mathrm{D}^2,\boldsymbol{\phi}}}$ is a valid norm over $\mathcal{S}(\mathbb{R}) = \mathcal{S}_{\boldsymbol{p}^\perp}(\mathbb{R}) \oplus \mathcal{N}_{\boldsymbol{\varphi}}$. We then deduce the desired completion result from the BLT theorem by observing that $\mathrm{G}^*_{\boldsymbol{\phi}} : (\tilde{\psi}, \tilde{\phi}) \mapsto \mathrm{G}^*_{\boldsymbol{\phi}}\{\tilde{\psi} + \tilde{\phi}\}$ is bounded from $(\mathcal{S}(\mathbb{R}), \|\cdot\|_{C_{\mathrm{D}^2,\boldsymbol{\phi}}})$ to $(C_0(\mathbb{R}), \|\cdot\|_\infty)$. The boundedness of the operator simply follows from the inequality

$$\|\mathrm{G}^*_{\boldsymbol{\phi}}\varphi\|_\infty \le \|\varphi\|_{C_{\mathrm{D}^2,\boldsymbol{\phi}}} = \max(\|\mathrm{G}^*_{\boldsymbol{\phi}}\varphi\|_\infty, \|\boldsymbol{p}(\varphi)\|_2) < \infty$$

for any $\varphi \in \mathcal{S}(\mathbb{R})$. ∎

By considering the dual form of Property 4 in Theorem 11 (which is a new result, to the best of our knowledge), we obtain an alternative, self-contained definition of our native space as

$$\mathrm{BV}^{(2)}(\mathbb{R}) = \{f \in \mathcal{S}'(\mathbb{R}) : \sup_{\varphi \in \mathcal{S}(\mathbb{R}) : \|\varphi\|_{C_{\mathrm{D}^2,\boldsymbol{\phi}}} \le 1} \langle f, \varphi \rangle < \infty\}, \tag{34}$$

which is the direct analog of (7).

Another important observation is that the "canonical" choice $\boldsymbol{\phi} = (\delta, -\delta')$ from (Unser et al., 2017) does not fulfill the second condition in Definition 10 (it actually fails by a tiny margin because $(-\mathrm{i}\omega)$ is only in $L_{1,2+\epsilon}(\mathbb{R})$ for any $\epsilon > 0$). This means that Property 4 does not apply to that particular case, even though the underlying native spaces are hardly distinguishable as sets. The only significant difference is in the specification of the corresponding weak* topology which, as we shall see, is essential to the proof of Lemma 2.

Appendix C. Proof of Theorem 1

(i) Explicit Banach Topology and Identification of the Right Inverse of D^2

The first statement in Theorem 1 results for the application of Proposition 9 with our specific choice of biorthogonal system $\boldsymbol{\phi} = (\phi_1, \phi_2) = \big(\delta, -\delta + \delta(\cdot - 1)\big)$ and $\boldsymbol{p} = (p_1, p_2)$, with $p_1(x) = 1$ and $p_2(x) = x$. This system satisfies the admissibility conditions in Definition 10. In particular, we have that $\langle p_1, \phi_1\rangle = p_1(0) = 1$, $\langle p_1, \phi_2\rangle = -p_1(0) + p_1(1) = 0$, $\langle p_2, \phi_1\rangle = p_2(0) = 0$ and $\langle p_2, \phi_2\rangle = -p_2(0) + p_2(1) = 1$ (biorthogonality). By inserting these functionals in (22), we obtain the formula of the norm for $\mathrm{BV}^{(2)}(\mathbb{R})$ in (11).

The corresponding expression (24) of the kernel of $\mathrm{G}_{\boldsymbol{\phi}}$ is

$$g_{\boldsymbol{\phi}}(x, y) = (x - y)_+ - (1 - x)(-y)_+ - x(1 - y)_+, \tag{35}$$

whose overall behavior is illustrated in Figure 3. We observe that the functions $y \mapsto g_{\boldsymbol{\phi}}(x, y)$ are continuous, triangle-shaped B-splines with the following characteristics:

- for $x \le 0$: $y \mapsto g_{\boldsymbol{\phi}}(x, y)$ is supported in $[x, 1]$ and takes its maximum at $y = 0$;
- for $x \in (0, 1)$: $y \mapsto g_{\boldsymbol{\phi}}(x, y)$ is supported in $[0, 1]$ and takes its extremum at $y = x$;
- for $x \ge 1$: $y \mapsto g_{\boldsymbol{\phi}}(x, y)$ is supported in $[0, x]$ and takes its maximum at $y = 1$.

(ii) Weak Continuity of Sampling Functionals*

The key here is that $\mathrm{G}^*_{\boldsymbol{\phi}}\{\delta(\cdot - x_m)\}(y) = g_{\boldsymbol{\phi}}(x_m, y)$ where $g_{\boldsymbol{\phi}}(x_m, \cdot)$, as defined by (35), is continuous, bounded and compactly supported, and, hence, vanishing at $\pm\infty$. Consequently, $\delta(\cdot - x_m) = \mathrm{D}^2 v + a_1\phi_1 + a_2\phi_2$ with $v = g_{\boldsymbol{\phi}}(x_m, \cdot) \in C_0(\mathbb{R})$, $a_1 = \langle 1, \delta(\cdot - x_m)\rangle = 1$, and $a_2 = \langle x, \delta(\cdot - x_m)\rangle = x_m$ in accordance with (27) in Theorem 11, which proves that $\delta(\cdot - x_m) \in C_{\mathrm{D}^2,\boldsymbol{\phi}}(\mathbb{R})$. This establishes its weak* continuity on $\big(C_{\mathrm{D}^2,\boldsymbol{\phi}}(\mathbb{R})\big)'$ (by Proposition 8).

Based on the observation that $\|g_{\boldsymbol{\phi}}(x_m, \cdot)\|_\infty \le |x_m|$, we then easily estimate the norm of $\delta(\cdot - x_m)$ as

$$\begin{aligned}\|\delta(\cdot - x_m)\|'_{\mathrm{BV}^{(2)}} &= \max\big(\sup_{y \in \mathbb{R}} |g_{\boldsymbol{\phi}}(x_m, y)|, \|\boldsymbol{a}\|_2\big)\\ &\le (1 + |x_m|) < \infty.\end{aligned}$$

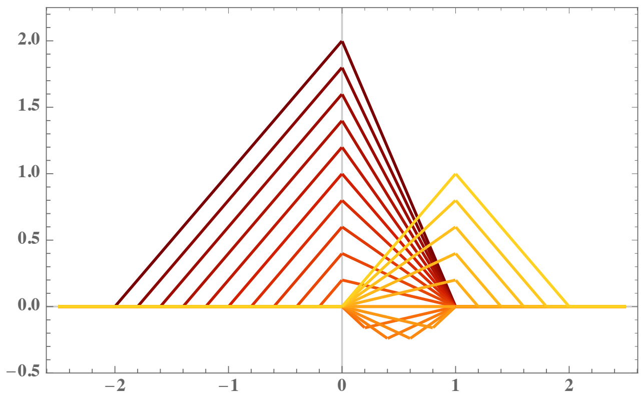

Figure 3: Graphs of the function $y \mapsto g_{\boldsymbol{\phi}}(x,y)$ for a series of values of $x=-2,\dots,2$ with step of 0.25. This illustrates the property that $g_{\boldsymbol{\phi}}(x,\cdot) \in C_0(\mathbb{R})$ for any $x \in \mathbb{R}$, which is critical to the proof of Lemma 2. By contrast, the canonical solution of (Unser et al., 2017) with $\boldsymbol{\phi}=(\delta,-\delta')$ would have resulted in a series of piecewise-linear functions with a discontinuous drop to 0 at $x=0$.

Finally, we recall that the property that two Banach spaces $\mathcal{U}$ and $\mathcal{U}'$ form a dual pair implies that $|\langle u,u'\rangle| \le \|u\|_{\mathcal{U}}\|u'\|_{\mathcal{U}'}$ for any $u \in \mathcal{U}$ and $u' \in \mathcal{U}'$. Taking $\mathcal{U}=C_{\mathrm{D}^2,\boldsymbol{\phi}}(\mathbb{R})$ and $u=\delta(\cdot - x_m)$ allows us to translate the above norm estimate into the announced continuity bound.

(iii) Embedding Properties

Property 4 in Theorem 11 tells us that $\mathcal{S}(\mathbb{R}) \hookrightarrow C_{\mathrm{D}^2,\boldsymbol{\phi}}(\mathbb{R})$ with the embedding being dense. This, together with the observation that $C_{\mathrm{D}^2,\boldsymbol{\phi}}(\mathbb{R}) \hookrightarrow \mathcal{S}'(\mathbb{R}^d)$, implies that $\mathcal{S}(\mathbb{R}) \hookrightarrow \mathrm{BV}^{(2)}(\mathbb{R}) \hookrightarrow \mathcal{S}'(\mathbb{R})$ (by duality) with the outer embedding being dense since $\mathcal{S}(\mathbb{R})$ is itself dense in $\mathcal{S}'(\mathbb{R})$. As for the embedding in $C_{\mathrm{b},1}(\mathbb{R})$, we first invoke Proposition 14 below, which states that $\mathrm{BV}^{(2)}(\mathbb{R}) \hookrightarrow L_{\infty,1}(\mathbb{R})$. Since $C_{\mathrm{b},1}(\mathbb{R})$ is isometrically embedded in $L_{\infty,1}(\mathbb{R})$ and the members of $\mathrm{BV}^{(2)}(\mathbb{R})$ are continuous (because they are differentiable), we then readily deduce that $\mathcal{S}(\mathbb{R}) \hookrightarrow \mathrm{BV}^{(2)}(\mathbb{R}) \hookrightarrow C_{\mathrm{b},1}(\mathbb{R})$.

We note that the proof of the embedding $\mathrm{BV}^{(2)}(\mathbb{R}) \hookrightarrow C_{\mathrm{b},1}(\mathbb{R})$ does not depend upon the specific choice of biorthogonal system $(\boldsymbol{\phi},\boldsymbol{p})$, as long as the admissibility condition in Definition 10 holds. By contrast, the proof of the weak* continuity of $\delta(\cdot - x_0)$ in Item (*ii*) is specific. While it is sufficient for our purpose, we have recent evidence (work in progress) that the argument is extendable to a broad class of equivalent biorthogonal systems, which makes us believe that there is actually a tight connection between Properties 2 and 3 in Theorem 1. Figure 3 and the argumentation in Step (*ii*) are enlightening in that respect because they make us appreciate the regularization effect of the second component in (24), which counterbalances the growth of the primary term $g(x,y)=(x-y)_+$ (the kernel of the shift-invariant inverse D^{-2*}). The other constraint is that the ϕ_n need to be regular enough (e.g., no worse than a Dirac) to preserve continuity.

Appendix D. Proof of Lemma 2

Proof The lemma is deduced from (Unser et al., 2017, Theorem 4): an abstract optimality result for generalized spline interpolation that holds for an extended class of admissible regularization operators L and for arbitrary linear functionals ($\nu_m : f \mapsto \langle \nu_m, f\rangle$), subject to the weak*-continuity requirement. The relevant version of the result for functions $f : \mathbb{R} \to \mathbb{R}$ is restated here in the explicit form of Theorem 15.

The maximal polynomial rate of growth (n_0) of functions is controlled via their inclusion in the space

$$L_{\infty,n_0}(\mathbb{R}) = \{f : \mathbb{R} \to \mathbb{R} \ \text{ s.t. } \|f\|_{\infty,n_0} \overset{\triangle}{=} \operatorname*{ess\,sup}_{x\in\mathbb{R}}(1 + |x|)^{-n_0}|f(x)| < \infty\}.$$

Definition 13 (Spline-admissible operator) *A linear operator* $\mathrm{L} : \mathcal{M}_{\mathrm{L}}(\mathbb{R}) \to \mathcal{M}(\mathbb{R})$*, where* $\mathcal{M}_{\mathrm{L}}(\mathbb{R}) \supseteq \mathcal{S}(\mathbb{R})$ *is an appropriate subspace of* $\mathcal{S}'(\mathbb{R})$*, is called* spline-admissible *if*

1. *it is shift-invariant;*

2. *there exists a function* $\rho_{\mathrm{L}} : \mathbb{R} \to \mathbb{R}$ *of slow growth (the Green's function of* L*) such that* $\mathrm{L}\rho_{\mathrm{L}} = \delta$*, where* δ *is the Dirac impulse. The rate of polynomial growth of* ρ_{L} *is* $n_0 = \inf\{n \in \mathbb{N} : \rho_{\mathrm{L}} \in L_{\infty,n}(\mathbb{R})\}$.

3. *the (growth-restricted) null space of* L*,*

$$\mathcal{N}_{\mathrm{L}} = \{q \in L_{\infty,n_0}(\mathbb{R}) : \mathrm{L}q = 0\},$$

has the finite dimension $N_0 \geq 0$.

The native space of L, $\mathcal{M}_{\mathrm{L}}(\mathbb{R})$, is then identified as

$$\mathcal{M}_{\mathrm{L}}(\mathbb{R}) = \{f \in L_{\infty,n_0}(\mathbb{R}) : \|\mathrm{L}f\|_{\mathcal{M}} < \infty\}. \tag{36}$$

In addition, it is assumed that $\mathcal{M}_{\mathrm{L}}(\mathbb{R})$ is equipped with an appropriate Banach topology which gives a concrete meaning to the underlying notion of (weak*-) continuity.

As expected, the operator $\mathrm{L} = \mathrm{D}^2$ is spline-admissible: Its causal Green's function is $\rho_{\mathrm{D}^2}(x) = (x)_+$ (ReLU), which exhibits the algebraic rate of growth $n_0 = 1$, while its null space $\mathcal{N}_{\mathrm{D}^2} = \mathrm{span}\{p_1, p_2\}$ with $p_1(x) = 1$ and $p_2(x) = x$ is finite-dimensional with $N_0 = 2$. These are precisely the basis functions associated with L that appear in (12).

We now show that the slow-growth condition with $n_0 = 1$ is implicit in the specification of $\mathrm{BV}^{(2)}(\mathbb{R})$ given by (9) and/or Proposition 9 so that our definition of the native space is consistent with (36), with $\mathrm{BV}^{(2)}(\mathbb{R}) = \mathcal{M}_{\mathrm{D}^2}(\mathbb{R})$.

Proposition 14 *With the choice of topology specified in Theorem 1,* $\mathrm{BV}^{(2)}(\mathbb{R}) \hookrightarrow L_{\infty,1}(\mathbb{R})$*, while*

$$f \in \mathrm{BV}^{(2)}(\mathbb{R}) \quad \Leftrightarrow \quad \mathrm{TV}^{(2)}(f) = \sup_{\|\varphi\|_\infty \leq 1 : \varphi \in \mathcal{S}(\mathbb{R})} \langle f, \mathrm{D}^2\varphi\rangle = \|\mathrm{D}^2 f\|_{\mathcal{M}} < \infty.$$

Proof The key is the bound $\|g_{\boldsymbol{\phi}}(x,\cdot)\|_\infty \le |x|$ for any $x \in \mathbb{R}$ (see Figure 3 and accompanying explanations), which implies that

$$C_{\boldsymbol{\phi}} = \operatorname*{ess\,sup}_{x,y\in\mathbb{R}} (1+|x|)^{-1}|g_{\boldsymbol{\phi}}(x,y)| < \infty.$$

This ensures the continuity of the operator $\mathrm{G}_{\boldsymbol{\phi}} : \mathcal{M}(\mathbb{R}) \to L_{\infty,1}(\mathbb{R})$ with $\|\mathrm{G}_{\boldsymbol{\phi}}\| = C_{\boldsymbol{\phi}}$ by (Unser et al., 2017, Theorem 3). Next, we use the property that any $f \in \mathrm{BV}^{(2)}(\mathbb{R})$ admits the unique decomposition $f = \mathrm{G}_{\boldsymbol{\phi}} w + p$ with $w = \mathrm{L}f \in \mathcal{M}(\mathbb{R})$ and $p = \sum_{n=1}^{2} \langle \phi_n, f\rangle p_n \in \mathcal{N}_{\boldsymbol{p}}$, so that

$$\begin{aligned}
\|f\|_{\infty,1} &\le \|\mathrm{G}_{\boldsymbol{\phi}} w\|_{\infty,1} + \|p\|_{\infty,1}\\
&\le C_{\boldsymbol{\phi}}\|w\|_{\mathcal{M}} + \sum_{n=1}^{2} |\langle \phi_n, f\rangle|\, \|p_n\|_{\infty,1}\\
&\le C_{\boldsymbol{\phi}}\|\mathrm{L}f\|_{\mathcal{M}} + \|\boldsymbol{p}(f)\|_2 \sum_{n=1}^{2} \|p_n\|_{\infty,1}\\
&\le \left(C_{\boldsymbol{\phi}} + \sum_{n=1}^{2} \|p_n\|_{\infty,1} \right) \|f\|_{\mathrm{BV}^{(2)}},
\end{aligned}$$

which proves that $\mathrm{BV}^{(2)}(\mathbb{R})$ is continuously embedded in $L_{\infty,1}(\mathbb{R})$. The reason for using the dual definition of the $\mathrm{TV}^{(2)}$ semi-norm in the last statement of the proposition is that the formula remains valid for any $f \in \mathcal{S}'(\mathbb{R})$ with $\mathrm{TV}^{(2)}(f) = \infty \Leftrightarrow f \notin \mathrm{BV}^{(2)}(\mathbb{R})$. Likewise, $\mathrm{TV}^{(2)}(f) = 0 \Leftrightarrow f \in \mathcal{N}_{\boldsymbol{p}}$. ■

Theorem 15 (Generalized spline interpolant) *Let us assume that the following conditions are met:*

1. *The operator* $\mathrm{L} : \mathcal{M}_{\mathrm{L}}(\mathbb{R}) \to \mathcal{M}(\mathbb{R})$ *is spline-admissible in the sense of Definition 13.*
2. *The linear measurement operator* $\boldsymbol{\nu} : f \mapsto \boldsymbol{\nu}(f) = \big(\langle \nu_1, f\rangle, \dots, \langle \nu_M, f\rangle\big)$ *maps* $\mathcal{M}_{\mathrm{L}}(\mathbb{R}^d) \to \mathbb{R}^M$ *and is weak*-continuous on* $\mathcal{M}_{\mathrm{L}}(\mathbb{R}^d) = \big(C_{\mathrm{L}}(\mathbb{R}^d)\big)'$.
3. *The recovery problem is well-posed over the null space of* L: $\boldsymbol{\nu}(q_1) = \boldsymbol{\nu}(q_2) \Leftrightarrow q_1 = q_2$, *for any* $q_1, q_2 \in \mathcal{N}_{\mathrm{L}}$.

Then, the extremal points of the (feasible) generalized interpolation problem

$$\beta = \min_{f \in \mathcal{M}_{\mathrm{L}}(\mathbb{R})} \|\mathrm{L}f\|_{\mathcal{M}} \quad \text{s.t.} \quad \boldsymbol{\nu}(f) = \mathbf{y} \tag{37}$$

are necessarily nonuniform L*-splines of the form*

$$s(x) = \sum_{n=1}^{N_0} b_n p_n(x) + \sum_{k=1}^{K} a_k \rho_{\mathrm{L}}(x - \tau_k) \tag{38}$$

with parameters $\mathbf{b} = (b_1, \dots, b_{N_0}) \in \mathbb{R}^{N_0}$*,* $K \le (M - N_0)$ *(effective number of knots),* $\{\tau_k\}_{k=1}^K$ *with* $\tau_k \in \mathbb{R}$*, and* $\mathbf{a} = (a_1, \dots, a_K) \in \mathbb{R}^K$*. Here,* $\{p_n\}_{n=1}^{N_0}$ *is a basis of* $\mathcal{N}_{\mathrm{L}}$ *and* $\mathrm{L}\rho_{\mathrm{L}} = \delta$ *so that* $\beta = \|\mathrm{L}s\|_{\mathcal{M}} = \sum_{k=1}^K |a_k| = \|\mathbf{a}\|_1$*. The full solution set of* (37) *is the weak*-closed convex hull of those extremal points.*

Hence, we only need to ensure that the underlying mathematical hypotheses are met for the spline-admissible operator $\mathrm{L} = \mathrm{D}^2$ and $\nu_m = \delta(\cdot - x_m)$.

- Weak* continuity of sampling functionals. This is taken care by the second statement in Theorem 1, the proof of which is quite involved because $\mathrm{BV}^{(2)}(\mathbb{R})$ is non-reflexive.
- Well-posedness of reconstruction for $f \in \mathcal{N}_{\mathrm{D}^2} = \mathrm{span}\{1, x\}$. It is well-known that the classical linear-regression problem

$$\mathbf{b} = \arg\min_{b_1, b_2} \sum_{m=1}^{M} |y_m - (b_1 + b_2 x_m)|^2$$

  is well-posed and has a unique solution if and only if $S = \{x_m\}_{m=1}^M$ contains at least two distinct points, say, $x_1 \neq x_2$, which takes care of the final hypothesis in Theorem 15.

■

Acknowlegdments

The research was partially supported by the Swiss National Science Foundation under Grant 200020-184646. The author is thankful to Julien Fageot, Shayan Aziznejad, Anais Badoual, Kyong Hwan Jin, and Harshit Gupta for helpful interactions.

References

Forest Agostinelli, Matthew Hoffman, Peter Sadowski, and Pierre Baldi. Learning activation functions to improve deep neural networks. In *Proc. Int. Conf. Learn. Representations, arXiv:1412.6830*, 2015.

Mauricio A. Alvarez, Lorenzo Rosasco, and Neil D. Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. *preprint arXiv:1611.01491*, 2016.

Christopher M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

Bastian Bohn, Michael Griebel, and Christian Rieger. A representer theorem for deep kernel learning. *arXiv:1709.10441v3*, 2018.

Carl de Boor. *A Practical Guide to Splines.* Springer-Verlag, New York, 1978.

Carl de Boor and Robert E. Lynch. On splines and their minimum properties. *Journal of Mathematics and Mechanics*, 15(6):953–969, 1966.

David L. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006.

Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1):1–50, 2000.

S. D. Fisher and J. W. Jerome. Spline solutions to L_1 extremal problems in one and several variables. *Journal of Approximation Theory*, 13(1):73–83, 1975.

Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing.* Springer, 2013.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2011.

Ian J. Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. *Proceedings of Machine Learning Research*, 28(3):1319–1327, 2013.

Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*, volume 1. MIT press Cambridge, 2016.

Stefano Guarnieri, Francesco Piazza, and Aurelio Uncini. Multilayer feedforward networks with adaptive spline activation function. *IEEE Transactions on Neural Networks*, 10(3): 672–683, 1999.

Harshit Gupta, Julien Fageot, and Michael Unser. Continuous-domain solutions of linear inverse problems with Tikhonov *versus* generalized TV regularization. *IEEE Transactions on Signal Processing*, 66(17):4670–4684, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-Rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

Thomas Hofmann, Bernhart Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008.

Le Hou, Dimitris Samaras, Tahsin Kurc, Yi Gao, and Joel Saltz. ConvNets with smooth adaptive activation functions for regression. In *Artificial Intelligence and Statistics*, pages 430–439, 2017.

George Kimeldorf and Grace Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

Stephen H. Lane, Marshall Flax, David Handelman, and Jack Gelfand. Multi-layer perceptrons with B-spline receptive field functions. In *Advances in Neural Information Processing Systems*, pages 684–692, 1991.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015.

Enno Mammen and Sara van de Geer. Locally adaptive regression splines. *Annals of Statistics*, 25(1):387–413, 1997.

Charles A. Micchelli. Interpolation of scattered data: Distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2(1):11–22, 1986.

Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.

Guido F. Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2924–2932, 2014.

Tomaso Poggio and Federico Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247(4945):978–982, 1990.

Tomaso Poggio and Steve Smale. The mathematics of learning: Dealing with data. *Notices of the AMS*, 50(5):537–544, 2003.

Tomaso Poggio, Lorenzo Rosasco, Amnon Shashua, Nadav Cohen, and Fabio Anselmi. Notes on hierarchical splines, DCLNs and i-theory. Technical report, Center for Brains, Minds and Machines (CBMM), 2015.

P.M. Prenter. *Splines and Variational Methods.* Wiley, New York, 1975.

Michael Reed and Barry Simon. *Methods of Modern Mathematical Physics. Vol. 1: Functional Analysis.* Academic Press, 1980.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 9351, pages 234–241. Springer, LNCS, 2015.

Walter Rudin. *Real and Complex Analysis.* McGraw-Hill, New York, 3rd edition, 1987.

Walter Rudin. *Functional Analysis.* McGraw-Hill, New York, 2nd edition, 1991. McGraw-Hill Series in Higher Mathematics.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533, 1986.

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

I.J. Schoenberg. Spline functions and the problem of graduation. *Proceedings of the National Academy of Sciences*, 52(4):947–950, 1964.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT press, 2002.

Bernhard Schölkopf, Kah-Kay Sung, Chris J. C. Burges, Federico Girosi, Partha Niyogi, Tomaso Poggio, and Vladimir Vapnik. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing*, 45 (11):2758–2765, 1997.

Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In David Helmbold and Bob Williamson, editors, *Computational Learning Theory*, pages 416–426. Springer Berlin Heidelberg, 2001.

Larry L. Schumaker. *Spline Functions: Basic Theory.* Wiley, New York, 1981.

Laurent Schwartz. *Théorie des Distributions.* Hermann, Paris, 1966.

Gilbert Strang. The functions of deep learning. *SIAM News*, 51(10):1,4, 2018.

J.M. Tarela and M.V. Martinez. Region configurations for realizability of lattice piecewise-linear models. *Mathematical and Computer Modelling*, 30(11):17–27, 1999.

Michael Unser. Splines: A perfect fit for signal and image processing. *IEEE Signal Processing Magazine*, 16(6):22–38, 1999.

Michael Unser, Julien Fageot, and Harshit Gupta. Representer theorems for sparsity-promoting ℓ_1 regularization. *IEEE Transactions on Information Theory*, 62(9):5167–5180, 2016.

Michael Unser, Julien Fageot, and John Paul Ward. Splines are universal solutions of linear inverse problems with generalized-TV regularization. *SIAM Review*, 59(4):769–793, 2017.

Vladimir Vapnik. *The Nature of Statistical Learning Theory.* Springer Science & Business Media, 2013.

Lorenzo Vecci, Francesco Piazza, and Aurelio Uncini. Learning and approximation capabilities of adaptive spline activation function neural networks. *Neural Networks*, 11(2): 259–270, 1998.

Grace Wahba. *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics, Philadelphia, PA, 1990.

Shuning Wang and Xusheng Sun. Generalization of hinging hyperplanes. *IEEE Transactions on Information Theory*, 51(12):4425–4431, 2005.

Holger Wendland. *Scattered Data Approximations.* Cambridge University Press, 2005.