SIAM J. MATH. DATA SCI.
Vol. 5, No. 2, pp. 422–445

Measuring Complexity of Learning Schemes Using Hessian-Schatten Total Variation[*]

Shayan Aziznejad[†], Joaquim Campos[†], and Michael Unser[†]

Abstract. In this paper, we introduce the Hessian-Schatten total variation (HTV)—a novel seminorm that quantifies the total "rugosity" of multivariate functions. Our motivation for defining HTV is to assess the complexity of supervised-learning schemes. We start by specifying the adequate matrix-valued Banach spaces that are equipped with suitable classes of mixed norms. We then show that the HTV is invariant to rotations, scalings, and translations. Additionally, its minimum value is achieved for linear mappings, which supports the common intuition that linear regression is the least complex learning model. Next, we present closed-form expressions of the HTV for two general classes of functions. The first one is the class of Sobolev functions with a certain degree of regularity, for which we show that the HTV coincides with the Hessian-Schatten seminorm that is sometimes used as a regularizer for image reconstruction. The second one is the class of continuous and piecewise-linear (CPWL) functions. In this case, we show that the HTV reflects the total change in slopes between linear regions that have a common facet. Hence, it can be viewed as a convex relaxation (ℓ_1-type) of the number of linear regions (ℓ_0-type) of CPWL mappings. Finally, we illustrate the use of our proposed seminorm.

Key words. Hessian operator, Schatten norm, total variation, continuous and piecewise-linear functions, supervised learning

MSC codes. 26B30, 46E27, 68T05, 68Q32

DOI. 10.1137/22M147517X

1. Introduction.

Given the sequence $(\boldsymbol{x}_m, y_m) \in \mathbb{R}^d \times \mathbb{R}$, $m = 1, \dots, M$, of data points, the goal of supervised learning is to construct a mapping $f : \mathbb{R}^d \to \mathbb{R}$ that adequately explains the data, i.e., $f(\boldsymbol{x}_m) \approx y_m$, while avoiding the problem of overfitting [24, 26, 65]. This is often formulated as a minimization problem of the form

$$\min_{f \in \mathcal{F}} \left(\sum_{m=1}^{M} E\left(f(\boldsymbol{x}_m), y_m\right) + \lambda \mathcal{R}(f) \right), \tag{1.1}$$

where $\mathcal{F}$ is the search space, $E : \mathbb{R} \times \mathbb{R}$ is a loss function that quantifies data discrepancy, and $\mathcal{R} : \mathcal{F} \to \mathbb{R}$ is a functional that enforces regularization. The regularization parameter $\lambda > 0$ adjusts the contribution of the two terms. A classical example is learning over reproducing-kernel Hilbert spaces (RKHS), where $\mathcal{F} = \mathcal{H}(\mathbb{R}^d)$ is an RKHS and $\mathcal{R}(f) = \|f\|_{\mathcal{H}}^2$ [50, 51].

[*] Received by the editors January 21, 2022; accepted for publication (in revised form) January 23, 2023; published electronically June 1, 2023.

https://doi.org/10.1137/22M147517X

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

[†] Biomedical Imaging Group, EPFL, 1015 Lausanne, Switzerland (sh.aziznejad@gmail.com, joaquim.campos@epfl.ch, michael.unser@epfl.ch).

The key result in this framework is the kernel representer theorem that provides a parametric form for the learned mapping [30, 58]. This foundational result is at the heart of many kernel-based schemes, such as support-vector machines [20, 59, 61]. Moreover, there has been an interesting lines of work regarding the statistical optimality of kernel-based methods [13, 39, 52, 62]. A central element in these analyses is that the regularization functional $\mathcal{R}(\cdot)$ (in this case, the underlying Hilbertian norm) directly controls the complexity of the learned mapping [6, section 2.4].

Although kernel methods are supported by a sound theory, they have been outperformed by deep neural networks (DNNs) in various areas of application [23, 34]. DNN-based methods are the current state of the art in several image processing tasks, such as inverse problems [29], image classification [32], and image segmentation [55]. Unlike kernel methods, DNNs have intricate nonlinear structures and the reason for their outstanding performance is not yet fully understood [6]. A possible approach to the comparison of DNNs is to quantify the "complexity" of the learned mapping. For example, neural networks with rectified linear units (ReLU), $\mathrm{ReLU}(x) = \max(x, 0)$ [22], are known to produce continuous and piecewise-linear (CPWL) mappings. Consequently, the number of linear regions of the input-output mapping has been proposed as a measure of complexity in this case [25, 49]. While this is an interesting metric to study, it has two limitations. The first is that this quantity is only defined for CPWL functions and, consequently, only applicable to ReLU neural networks. This prevents one from building a framework that would include neural networks with more modern activation functions [14, 27, 40, 53]. The second limitation is that this measure is not robust, in the sense that the input-output mapping might have many small linear regions around the training data points and still be able to generalize well. For example, the learned mapping can be the superposition of two CPWL functions: a dominant one with a few significant linear regions and a much smaller residual one with many tiny regions which enables the mapping to interpolate the training data points exactly. This phenomenon, which is called "benign overfitting" [5, 38] and is observed in overparametrized models, cannot be reflected in the aforementioned complexity measure.

In this paper, we introduce a novel family of seminorms—the Hessian-Schatten total variation (HTV)—and we propose its use as a way to quantify the complexity of learning schemes. Our definition of the HTV is based on a second-order extension of the space of functions with bounded variation [1]. We show that the HTV seminorm satisfies the following desirable properties:

1. It assigns the zero value for linear regression, which is the simplest learning scheme.
2. It is invariant (up to a multiplicative factor) to simple transformations (such as linear isometries and scaling) over the input domain.
3. It is defined for both smooth and CPWL functions. Hence, it is applicable to a broad class of learning schemes, including ReLU neural networks and radial-basis functions.
4. It favors CPWL functions with a small number of linear regions, thus promoting a simpler (and, hence, more interpretable) representation of the data (Occam's razor principle).

We provide closed-form formulas for the HTV of both Sobolev and CPWL functions. For Sobolev functions, the HTV coincides with the Hessian-Schatten seminorm, which is often used as a regularization term in linear inverse problems [35, 37]. For CPWL functions, the HTV is a convex relaxation of the number of linear regions. This is analogous to the classical ℓ_0 penalty in the field of compressed sensing, where it is often replaced by its convex proxy, the ℓ_1 norm, to ensure tractability [18, 19].

1.1. Related works.

The Banach space of functions with bounded Hessian was originally introduced by Demengel [17]. This space, together with its associated seminorm, the second-order total variation (TV), has been used in various domains of image processing [7, 28, 31]. Intuitively, the second-order TV computes the Frobenius norm of the Hessian and integrates it over the whole domain. More recently, Bredies et al. have extended this notion to higher-order derivatives by defining the total generalized variation and introduced its use as a regularization functional for solving ill-posed inverse problems [9, 11]. We refer the reader to [10] for a complete review of the higher-order TV-based methods for solving inverse problems. In our work, we extend the second-order total variation to cover all Schatten matrix norms with arbitrary parameter $p \in [1, +\infty]$. In addition to this extension, we introduce the use of the HTV seminorm as a complexity measure in analyzing supervised learning schemes. Particularly, we were motivated by this incentive to study the CPWL family and ReLU neural networks in more depth.

Another relevant functional is the Radon-domain total-variation seminorm [43, 46, 47, 48], which fulfils many of the desirable properties of a good complexity measure.

1.2. Outline.

The paper is organized as follows: We start section 2 with some mathematical preliminaries that are essential for this paper. In section 3, we introduce the HTV seminorm and prove its desirable properties. We then compute the HTV of two general classes of functions (Sobolev and CPWL) in section 4. Finally, we illustrate the practical aspects of our proposed seminorm with examples in section 5.

2. Preliminaries.

Throughout the paper, we denote the input domain by $\Omega \subseteq \mathbb{R}^d$, and we assume Ω to be an open ball of radius $R > 0$, with the convention that the case $R = +\infty$ corresponds to $\Omega = \mathbb{R}^d$.

2.1. Schatten matrix norms.

For any $p \in [1, +\infty]$, the Schatten-p norm of a real-valued matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ is defined as

$$\|\mathbf{A}\|_{S_p} \triangleq \begin{cases} \left(\sum_{i=1}^{d} |\sigma_i(\mathbf{A})|^p\right)^{\frac{1}{p}}, & 1 \leq p < +\infty, \\ \max_i |\sigma_i(\mathbf{A})|, & p = +\infty, \end{cases} \tag{2.1}$$

where $(\sigma_1(\mathbf{A}), \dots, \sigma_d(\mathbf{A}))$ are the singular values of $\mathbf{A}$ [8]. It is known that the dual of the Schatten-p norm is the Schatten-q norm, where $q \in [1, \infty]$ is the Hölder conjugate of p such that $\frac{1}{p} + \frac{1}{q} = 1$. This result stems from a variant of the Hölder inequality for Schatten norms. It states that

$$\langle \mathbf{A}, \mathbf{B} \rangle \triangleq \mathrm{Tr}\left(\mathbf{A}^T \mathbf{B}\right) \leq \|\mathbf{A}\|_{S_p} \|\mathbf{B}\|_{S_q} \tag{2.2}$$

for any pair of matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d \times d}$ (see [36] for a simple proof).

2.2. Total-variation norm. Schwartz's space of infinitely differentiable and compactly supported test functions $\varphi : \Omega \to \mathbb{R}$ is denoted by $\mathcal{D}(\Omega)$. Its continuous dual $\mathcal{D}'(\Omega)$ is the space of distributions [60]. The Banach space $\mathcal{C}_0(\Omega)$ is the completion of $\mathcal{D}(\Omega)$ with respect to the L_∞ norm $\|f\|_{L_\infty} \overset{\Delta}{=} \sup_{\boldsymbol{x}\in\Omega} |f(\boldsymbol{x})|$. The bottom line is that the space $\mathcal{C}_0(\Omega)$ is formed of continuous functions $f : \Omega \to \mathbb{R}$ that vanish at infinity. The Riesz–Markov theorem states that the dual of $\mathcal{C}_0(\Omega)$ is the space $\mathcal{M}(\Omega) = (\mathcal{C}_0(\Omega))'$ of bounded Radon measures equipped with the total-variation norm [56]

$$\|w\|_{\mathcal{M}} \overset{\Delta}{=} \sup_{\varphi\in\mathcal{D}(\Omega)\setminus\{0\}} \frac{\langle w, \varphi\rangle}{\|\varphi\|_\infty}. \tag{2.3}$$

The space $\mathcal{M}(\Omega)$ is a superset of the space $L_1(\Omega)$ of absolutely integrable measurable functions with $\|f\|_{\mathcal{M}} = \|f\|_{L_1}$ for any $f \in L_1(\Omega)$. Moreover, it contains shifted Dirac impulses with $\|\delta(\cdot - \boldsymbol{x}_0)\|_{\mathcal{M}} = 1$ for any $\boldsymbol{x}_0 \in \Omega$. The latter can be generalized to any distribution of the form $w_{\boldsymbol{a}} = \sum_{n\in\mathbb{Z}} a_n \delta(\cdot - \boldsymbol{x}_n)$ with $\|w_{\boldsymbol{a}}\|_{\mathcal{M}} = \|\boldsymbol{a}\|_{\ell_1}$ for any $\boldsymbol{a} = (a_n) \in \ell_1(\mathbb{Z})$ and any sequence of distinct locations $(\boldsymbol{x}_n) \subseteq \Omega$.

2.3. Matrix-valued Banach spaces. In this work, we are interested in the matrix-valued extension of the spaces defined in section 2.2. We denote by $\mathcal{C}_0(\Omega; \mathbb{R}^{d\times d})$ the space of continuous matrix-valued functions $\mathbf{F} : \Omega \to \mathbb{R}^{d\times d}$ that vanish at infinity so that $\lim_{\|\boldsymbol{x}\|\to\infty} \|\mathbf{F}(\boldsymbol{x})\| = 0$ whenever the domain is unbounded. (Note that this definition does not depend on the choice of the norms, because they are all equivalent in finite-dimensional vector spaces.) Any matrix-valued function $\mathbf{F} : \Omega \to \mathbb{R}^{d\times d}$ has the unique representation

$$\mathbf{F} = [f_{i,j}] = \begin{pmatrix} f_{1,1} & \cdots & f_{1,d} \\ \vdots & \ddots & \vdots \\ f_{d,1} & \cdots & f_{d,d} \end{pmatrix}, \tag{2.4}$$

where each entry $f_{i,j} : \Omega \to \mathbb{R}$ is a scalar-valued function for $i, j = 1, \dots, d$. In this representation, the space $\mathcal{C}_0(\Omega; \mathbb{R}^{d\times d})$ is the collection of matrix-valued functions of the form (2.4) with $f_{i,j} \in \mathcal{C}_0(\Omega)$.

Definition 2.1. *Let $q \in [1, +\infty]$. For any $\mathbf{F} \in \mathcal{C}_0(\Omega; \mathbb{R}^{d\times d})$, the L_∞-S_q mixed norm is defined as*

$$\|\mathbf{F}\|_{L_\infty, S_q} \overset{\Delta}{=} \left\| \begin{pmatrix} \|f_{1,1}\|_{L_\infty} & \cdots & \|f_{1,d}\|_{L_\infty} \\ \vdots & \ddots & \vdots \\ \|f_{d,1}\|_{L_\infty} & \cdots & \|f_{d,d}\|_{L_\infty} \end{pmatrix} \right\|_{S_q}. \tag{2.5}$$

Remark 1. In Definition 2.1, the S_q-norm appears as the outer norm. We remain faithful to this convention throughout the paper and always denote mixed norms in order of appearance, where the first is the inner-norm and the second the outer-norm.

Following [63], we deduce that $(\mathcal{C}_0(\Omega; \mathbb{R}^{d\times d}), \|\cdot\|_{L_\infty, S_q})$ is a bona fide Banach space, whose dual is $(\mathcal{M}(\Omega, \mathbb{R}^{d\times d}), \|\cdot\|_{\mathcal{M}, S_p})$, where $\mathcal{M}(\Omega; \mathbb{R}^{d\times d})$ is the collection of matrix-valued Radon measures of the form

$$\mathbf{W} = [w_{i,j}] = \begin{pmatrix} w_{1,1} & \cdots & w_{1,d} \\ \vdots & \ddots & \vdots \\ w_{d,1} & \cdots & w_{d,d} \end{pmatrix}, \quad w_{i,j} \in \mathcal{M}(\Omega) \quad \forall i,j = 1,\dots,d, \tag{2.6}$$

and the mixed $\mathcal{M}$-S_p norm is defined as

$$\|\mathbf{W}\|_{\mathcal{M},S_p} \triangleq \left\| \begin{pmatrix} \|w_{1,1}\|_{\mathcal{M}} & \cdots & \|w_{1,d}\|_{\mathcal{M}} \\ \vdots & \ddots & \vdots \\ \|w_{d,1}\|_{\mathcal{M}} & \cdots & \|w_{d,d}\|_{\mathcal{M}} \end{pmatrix} \right\|_{S_p}. \tag{2.7}$$

The duality product $\langle \cdot, \cdot \rangle : \mathcal{M}(\Omega; \mathbb{R}^{d\times d}) \times \mathcal{C}_0(\Omega; \mathbb{R}^{d\times d}) \to \mathbb{R}$ is then defined as

$$\langle \mathbf{W}, \mathbf{F} \rangle \triangleq \sum_{i=1}^{d} \sum_{j=1}^{d} \langle w_{i,j}, f_{i,j} \rangle. \tag{2.8}$$

Finally, we denote by $L_1(\Omega; \mathbb{R}^{d\times d})$, $\mathcal{D}(\Omega; \mathbb{R}^{d\times d})$, and $\mathcal{D}'(\Omega; \mathbb{R}^{d\times d})$ the matrix-valued generalizations of the spaces $L_1(\Omega)$, $\mathcal{D}(\Omega)$, and $\mathcal{D}'(\Omega)$, respectively. These spaces are equipped with the natural direct-product topology. Consequently, we are allowed to use the notation $\mathcal{D}'(\Omega; \mathbb{R}^{d\times d})$, because the latter space is indeed the topological dual of $\mathcal{D}(\Omega; \mathbb{R}^{d\times d})$.

2.4. Generalized Hessian operator.

The operators $\frac{\partial^2}{\partial x_i \partial x_j} : \mathcal{D}'(\Omega) \to \mathcal{D}'(\Omega)$ are viewed as second-order weak partial derivatives. More precisely, for any $i,j = 1,\dots,d$ and any $w \in \mathcal{D}'(\Omega)$, the distribution $\frac{\partial^2}{\partial x_i \partial x_j}\{w\} \in \mathcal{D}'(\Omega)$ is defined as

$$\left\langle \frac{\partial^2}{\partial x_i \partial x_j} w, \varphi \right\rangle = \left\langle w, \frac{\partial^2}{\partial x_i \partial x_j} \varphi \right\rangle$$

for all test functions $\varphi \in \mathcal{D}(\Omega)$. This leads to the following definition of the generalized Hessian operator over the space of distributions.

Definition 2.2. *The Hessian operator* $\mathrm{H} : \mathcal{D}'(\Omega) \to \mathcal{D}'(\Omega; \mathbb{R}^{d\times d})$ *is defined as*

$$\mathrm{H}\{f\} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{pmatrix}. \tag{2.9}$$

3. The Hessian-Schatten total variation.

In order to properly define the HTV seminorm, we start by introducing a novel class of mixed norms over $\mathcal{C}_0(\Omega; \mathbb{R}^{d\times d})$.

Definition 3.1. *Let* $q \in [1, +\infty]$. *For any* $\mathbf{F} \in \mathcal{C}_0(\Omega; \mathbb{R}^{d\times d})$, *the mixed* S_q-L_∞ *norm is defined as*

$$\|\mathbf{F}\|_{S_q, L_\infty} = \sup_{\boldsymbol{x} \in \Omega} \|\mathbf{F}(\boldsymbol{x})\|_{S_q}. \tag{3.1}$$

In section 2.3, we highlighted that the dual norm of the L_∞-S_q mixed norm is $\mathcal{M}$-S_p, which is defined over matrix-valued Radon measures. In Definition 3.1, we switched the order of application of the individual norms; however, the two norms induce the same topology over the space $\mathcal{C}_0(\Omega; \mathbb{R}^{d\times d})$.

Theorem 3.2. *Regarding the mixed norms defined in Definitions 2.1 and 3.1*

1. *The functional* $\mathbf{F} \mapsto \|\mathbf{F}\|_{S_q,L_\infty}$ *is a well-defined (finite) norm over* $\mathcal{C}_0(\Omega;\mathbb{R}^{d\times d})$.
2. *The* L_∞-S_q *and the* S_q-L_∞ *mixed norms are equivalent, in the sense that there exist positive constants* $A, B > 0$ *such that, for all* $\mathbf{F} \in \mathcal{C}_0(\Omega;\mathbb{R}^{d\times d})$, *we have that*
$$A\|\mathbf{F}\|_{S_q,L_\infty} \le \|\mathbf{F}\|_{L_\infty,S_q} \le B\|\mathbf{F}\|_{S_q,L_\infty}. \tag{3.2}$$
3. *The normed space* $(\mathcal{C}_0(\Omega;\mathbb{R}^{d\times d}), \|\cdot\|_{S_q,L_\infty})$ *is a bona fide Banach space.*

The proof is available in Appendix A. Using the outcomes of Theorem 3.2 and, in particular, item 3, we are now ready to introduce the S_p-$\mathcal{M}$ mixed norm defined over the space of matrix-valued Radon measures.

Definition 3.3. *For any matrix-valued Radon measure* $\mathbf{W} \in \mathcal{M}(\Omega,\mathbb{R}^{d\times d})$, *the* S_p-$\mathcal{M}$ *mixed norm is defined as*
$$\|\mathbf{W}\|_{S_p,\mathcal{M}} \triangleq \sup\left\{\langle \mathbf{W},\mathbf{F}\rangle : \mathbf{F} \in \mathcal{C}_0(\Omega;\mathbb{R}^{d\times d}), \|\mathbf{F}\|_{S_q,L_\infty} = 1\right\}. \tag{3.3}$$

Intuitively, the S_p-$\mathcal{M}$ norm of a matrix-valued function $\mathbf{F} : \Omega \to \mathbb{R}^{d\times d}$ is equal to the total-variation norm of the function $\boldsymbol{x} \mapsto \|\mathbf{F}(\boldsymbol{x})\|_{S_q}$. However, this intuition cannot directly lead to a general definition, because the space $\mathcal{M}(\Omega;\mathbb{R}^{d\times d})$ contains elements that do not have a pointwise definition. We are therefore forced to define this norm by duality, as opposed to the $\mathcal{M}$-S_p norm given in (2.7).

We also remark that, due to the dense embedding $\mathcal{D}(\Omega;\mathbb{R}^{d\times d}) \hookrightarrow \mathcal{C}_0(\Omega;\mathbb{R}^{d\times d})$, one can alternatively express the S_p-$\mathcal{M}$ norm as
$$\|\mathbf{W}\|_{S_p,\mathcal{M}} = \sup\left\{\langle \mathbf{W},\mathbf{F}\rangle : \mathbf{F} \in \mathcal{D}(\Omega;\mathbb{R}^{d\times d}), \|\mathbf{F}\|_{S_q,L_\infty} = 1\right\}, \tag{3.4}$$
which is well-defined for all matrix-valued distributions. However, the only elements of $\mathcal{D}'(\Omega;\mathbb{R}^{d\times d})$ of a finite S_p-$\mathcal{M}$ norm are precisely the matrix-valued finite Radon measures. In other words, $\mathcal{M}(\Omega,\mathbb{R}^{d\times d})$ is the largest subspace of $\mathcal{D}'(\Omega;\mathbb{R}^{d\times d})$ with a finite S_p-$\mathcal{M}$ norm.

In what follows, we strengthen the intuition behind the S_p-$\mathcal{M}$ norm by computing it for two general classes of functions/distributions in $\mathcal{M}(\Omega,\mathbb{R}^{d\times d})$ that are particularly important in our framework: the absolutely integrable matrix-valued functions and the Dirac fence distributions.

Definition 3.4. *For any nonzero matrix* $\mathbf{A} \in \mathbb{R}^{d\times d}$, *any convex compact set* $C \subset \mathbb{R}^{d_1}$ *with* $d_1 < d$, *and any measurable transformation* $\mathbf{T} : \mathbb{R}^{d_1} \to \mathbb{R}^{d-d_1}$ *(not necessarily linear) such that* $C \times \mathbf{T}(C) \subseteq \Omega$, *we define the corresponding Dirac fence* $\mathbf{D} \in \mathcal{M}(\Omega;\mathbb{R}^{d\times d})$ *as*
$$\mathbf{D}(\boldsymbol{x}_1,\boldsymbol{x}_2) = \mathbf{A}\mathbb{1}_{\boldsymbol{x}_1\in C}\,\delta(\boldsymbol{x}_2 - \mathbf{T}\boldsymbol{x}_1\}), \quad \boldsymbol{x}_1 \in \mathbb{R}^{d_1}, \boldsymbol{x}_2 \in \mathbb{R}^{d-d_1}, (\boldsymbol{x}_1,\boldsymbol{x}_2) \in \Omega. \tag{3.5}$$

Dirac fence distributions are natural generalizations of *the Dirac impulse* to nonlinear (and bounded) manifolds [44]. More precisely, for any test function $\mathbf{F} \in \mathcal{C}_0(\Omega;\mathbb{R}^{d\times d})$ and any Dirac fence $\mathbf{D}$ of the form (3.5), we have that
$$\langle \mathbf{D},\mathbf{F}\rangle = \int_C \mathrm{Tr}\left(\mathbf{A}^T\mathbf{F}(\boldsymbol{x}_1,\mathbf{T}\boldsymbol{x}_1)\right)\mathrm{d}\boldsymbol{x}_1 \in \mathbb{R}. \tag{3.6}$$

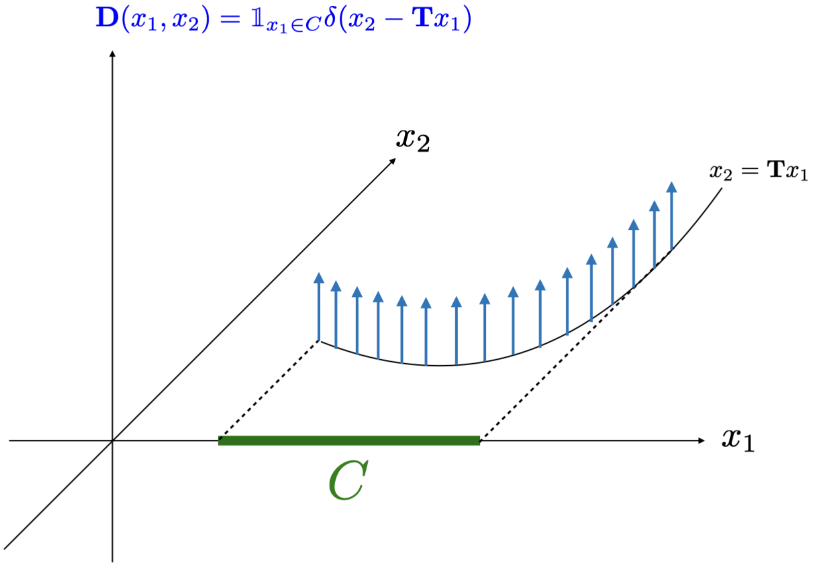

$\mathbf{D}(x_1, x_2) = \mathbb{1}_{x_1 \in C}\delta(x_2 - \mathbf{T}x_1)$
x_2
$x_2 = \mathbf{T}x_1$
x_1
C

Figure 1. *Illustration of a Dirac fence with $d = 2$ and $d_1 = 1$.*

Intuitively, this corresponds to considering a "continuum" of low-dimensional Dirac impulses on the d_1-dimensional compact manifold $C \times \mathbf{T}(C)$ that is embedded in Ω, as illustrated in Figure 1.

Theorem 3.5. *Let $p \in [1, +\infty)$. Then the following hold:*

1. *For any matrix-valued function $\mathbf{W} \in L_1(\Omega, \mathbb{R}^{d\times d}) \subseteq \mathcal{M}(\Omega; \mathbb{R}^{d\times d})$, we have that*

$$\|\mathbf{W}\|_{S_p,\mathcal{M}} = \| \, \|\mathbf{W}(\cdot)\|_{S_p} \, \|_{L_1} = \int_\Omega \left(\sum_{i=1}^{d} |\sigma_i\left(\mathbf{W}(\boldsymbol{x})\right)|^p \right)^{\frac{1}{p}} \mathrm{d}\boldsymbol{x}. \tag{3.7}$$

2. *For any Dirac fence distribution $\mathbf{D}$ of the form (3.5), we have that*

$$\|\mathbf{D}\|_{S_p,\mathcal{M}} = \|\mathbf{A}\|_{S_p} \mathrm{Leb}(C), \tag{3.8}$$

   where $\mathrm{Leb}(C)$ denotes the Lebesgue measure of $C \subseteq \mathbb{R}^{d_1}$.

3. *Consider two Dirac fences $\mathbf{D}_1$ and $\mathbf{D}_2$ of the form*

$$\mathbf{D}_i(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathbf{A}_i \mathbb{1}_{\boldsymbol{x}_1 \in C_i} \delta(\boldsymbol{x}_2 - \mathbf{T}_i \boldsymbol{x}_1\}), \quad i = 1, 2,$$

   and assume that the "intersection" of the two fences is of measure zero, in the sense that $C_0 = \{\boldsymbol{x}_1 \in C_1 \cap C_2 : \mathbf{T}_1 \boldsymbol{x}_1 = \mathbf{T}_2 \boldsymbol{x}_1\}$ is a subset of $\mathbb{R}^{d_1}$ whose Lebesgue measure is zero. Then, we have that

$$\|\mathbf{D}_1 + \mathbf{D}_2\|_{S_p,\mathcal{M}} = \|\mathbf{D}_1\|_{S_p,\mathcal{M}} + \|\mathbf{D}_2\|_{S_p,\mathcal{M}}. \tag{3.9}$$

The proof can be found in Appendix B. We are now ready to define the HTV seminorm.

Definition 3.6. *Let $p \in [1,+\infty]$. The Hessian-Schatten total variation of any $f \in \mathcal{D}'(\Omega)$ is defined as*

$$\mathrm{HTV}_p(f) = \|\mathrm{H}\{f\}\|_{S_p,\mathcal{M}} = \sup\left\{\langle \mathrm{H}\{f\},\mathbf{F}\rangle : \mathbf{F} \in \mathcal{D}(\Omega;\mathbb{R}^{d\times d}), \|\mathbf{F}\|_{S_q,L_\infty} = 1\right\}, \tag{3.10}$$

where $q \in [1,+\infty]$ is the Hölder conjugate of p with $\frac{1}{p}+\frac{1}{q}=1$.

In some cases, the HTV seminorm yields the same value for all choices of $p \in [1,+\infty]$. For instance, in dimension $d=1$, the HTV coincides with the second-order total-variation (TV-2) seminorm, $\mathrm{TV}^{(2)}(f) = \|\mathrm{D}^2\{f\}\|_{\mathcal{M}}$, where D denotes the weak derivative operator $\mathcal{D}'(\mathbb{R}) \to \mathcal{D}'(\mathbb{R})$. Another interesting example (see section 4.2) is the class of CPWL functions. In these cases, we can remove the subscript p and denote the seminorm by $\mathrm{HTV}(f)$ for brevity.

We now prove some desirable properties of the HTV functional. The proofs can be found in Appendix C.

Theorem 3.7. *The HTV seminorm satisfies the following properties:*

1. Null space*: A distribution has a vanishing HTV if and only if it can be identified as an affine function. In other words, we have that*

$$\mathcal{N}_{\mathrm{HTV}_p}(\Omega) = \left\{f \in \mathcal{D}'(\Omega) : \mathrm{HTV}_p(f) = 0\right\} = \left\{\boldsymbol{x} \mapsto \boldsymbol{a}^T\boldsymbol{x} + b : \boldsymbol{a} \in \mathbb{R}^d, b \in \mathbb{R}\right\}.$$

2. Invariance*: Let $\Omega = \mathbb{R}^d$. For any $f \in \mathcal{D}'(\mathbb{R}^d)$, we have that*

$$\begin{aligned}
&\mathrm{HTV}_p(f(\cdot - \boldsymbol{x}_0)) = \mathrm{HTV}_p(f) && \forall \boldsymbol{x}_0 \in \mathbb{R}^d,\\
&\mathrm{HTV}_p(f(\alpha\cdot)) = |\alpha|^{2-d}\mathrm{HTV}_p(f) && \forall \alpha \in \mathbb{R},\\
&\mathrm{HTV}_p(f(\mathbf{U}\cdot)) = \mathrm{HTV}_p(f) && \forall \mathbf{U} \in \mathbb{R}^{d\times d} : \textit{Orthonormal.}
\end{aligned}$$

4. Closed-form expressions for the HTV of special functions.

Although Definition 3.6 introduces a formal way to compute the HTV of a given element $f \in \mathcal{D}'(\Omega)$, it is still very abstract and not practical. This is the reason why we now provide closed-form expressions for the HTV of two general classes of functions.

4.1. Sobolev functions.

Let $W_1^2(\Omega)$ be the Sobolev space of absolutely integrable and twice-differentiable functions $f : \Omega \to \mathbb{R}$ whose first- and second-order partial derivatives are in $L_1(\Omega)$. We note that, for compact domains Ω, this space contains the input-output relation of neural networks with activation functions that are twice-differentiable almost everywhere (e.g., sigmoid [15], Swish [53], Mish [40], GeLU [27]).

Proposition 4.1 (Sobolev compatibility). *Let $p \in [1,+\infty]$. Then, for any Sobolev function $f \in W_1^2(\Omega)$, we have that*

$$\mathrm{HTV}_p(f) = \|\mathrm{H}\{f\}\|_{S_p,L_1} = \int_\Omega \|\mathrm{H}\{f\}(\boldsymbol{x})\|_{S_p}\mathrm{d}\boldsymbol{x}.$$

Proof. This is a consequence of Theorem 3.5 since, for any $f \in W_1^2(\Omega)$, the matrix-valued function $\mathrm{H}\{f\} : \Omega \to \mathbb{R}^{d\times d} : \boldsymbol{x} \mapsto \mathrm{H}\{f\}(\boldsymbol{x})$ is measurable and is in $L_1(\Omega;\mathbb{R}^{d\times d})$. ■

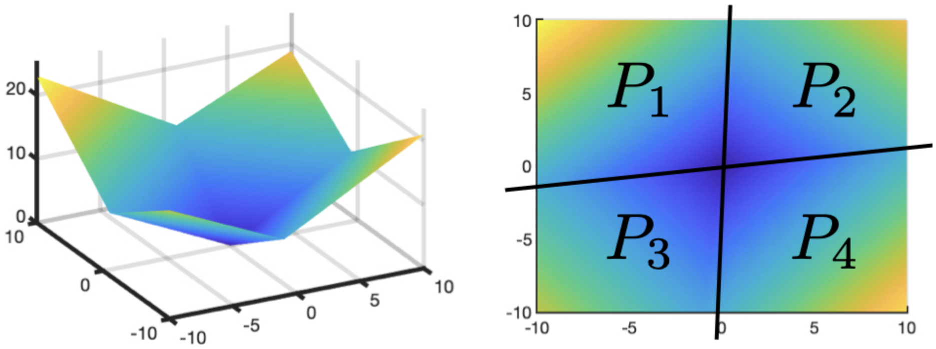


Figure 2. *Illustration of a CPWL function $f:\mathbb{R}^2\to\mathbb{R}$. Left: 3D view. Right: 2D partitioning.*

Remark 2. In Proposition 4.1, we only need the second-order partial derivatives to be in $L_1(\Omega)$. Following Poincare's inequality, this is equivalent to the Sobolev criteria for bounded domains. However, in the case of $\Omega=\mathbb{R}^d$, our results are applicable to Beppo–Levi spaces [33].

Interestingly, Proposition 4.1 demonstrates that our introduced seminorm is a generalization of the Hessian-Schatten regularization that has been used in inverse problems and image reconstruction [35, 37].

4.2. Continuous and piecewise-linear mappings. A function $f:\Omega\to\mathbb{R}$ is said to be continuous and piecewise linear if the following hold:

1. It is continuous.
2. There exists a finite partitioning $\Omega=P_1\sqcup P_2\sqcup\cdots\sqcup P_N$ such that, for any $n=1,\dots,N$, P_n is a polytope with the property that the restricted function $f|_{P_n}$ is an affine mapping of the form $f|_{P_n}(\boldsymbol{x})=\boldsymbol{a}_n^T\boldsymbol{x}+b_n$ for all $\boldsymbol{x}\in P_n$.

An example of a CPWL function is shown in Figure 2. Let us highlight that there is an intimate link between CPWL functions and ReLU neural networks. Indeed, it has been shown that the input-output relation of any feed forward ReLU neural network is a CPWL function [41, 49]. Moreover, any CPWL function can be represented *exactly* by some ReLU neural network whose depth is at most $(\lceil\log_2(d+1)\rceil+1)$ [2, Theorem 2.1].

Theorem 4.2. *Let $f:\Omega\to\mathbb{R}$ be the CPWL function described above. For any $p\in[1,+\infty]$, the corresponding HTV of f is given as*

$$\mathrm{HTV}_p(f)=\frac{1}{2}\sum_{n=1}^{N}\sum_{k\in\mathrm{adj}_n}\|\boldsymbol{a}_n-\boldsymbol{a}_k\|_2H^{d-1}(P_n\cap P_k), \tag{4.1}$$

where adj_n is the set of indices $k\in\{1,\dots,N\}$ such that P_n and P_k are neighbors and H^{d-1} denotes the $(d-1)$-dimensional Hausdorff measure.

The proof of Theorem 4.2 is provided in Appendix D. We conclude from (4.1) that the HTV seminorm accounts for the change of (directional) slope in all the junctions in the partitioning.

Specifically, the HTV of a CPWL function is proportional to a weighted ℓ_1 penalty on the vector of slope changes, where the weights are proportional to the volume of the intersection region. This can be seen as a convex relaxation of the number of linear regions. The latter has the disadvantage that it is unable to differentiate between small and large changes of slope. Another noteworthy observation is the invariance of the HTV of CPWL functions to the value of $p \in [1, +\infty]$, which is unlike the case of Sobolev functions in Proposition 4.1. This is due to the extreme sparsity of the Hessian of CPWL functions. In fact, the Hessian matrix is zero everywhere except at the borders of linear regions. There, it is a Dirac fence weighted by a rank-1 matrix. The invariance then follows from the observation that the Schatten-p norms collapse to a single value in rank-1 matrices (i.e., their only nonzero singular value).

It is known from the literature on low-rank matrix recovery that among the Schatten norms, only the case $p = 1$ is relevant for obtaining "sparse" elements [54]. By analogy, we conjecture that the only member of the HTV family that favors CPWL functions is HTV_1.

To demonstrate the applicability of Theorem 4.2, we now provide a closed-form expression for the HTV of a 2-layer neural network $f : \Omega \to \mathbb{R}$ with ReLU activation functions and skip connections. Let us recall that the input-output mapping of such networks admits the representation

$$f(\boldsymbol{x}) = c_0 + \boldsymbol{c}_1^T \boldsymbol{x} + \sum_{n=1}^{N} v_n \mathrm{ReLU}\left(\boldsymbol{w}_n^T \boldsymbol{x} - b_n\right) \tag{4.2}$$

for some $N \in \mathbb{N}$, $\boldsymbol{w}_n \in \mathbb{R}^d$, and $v_n, b_n \in \mathbb{R}$ for $n = 1, \dots, N$.

Proposition 4.3. *Let $\Omega = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 \leq R\}$ be the d-dimensional Euclidean ball of radius $R > 0$. Then, the HTV of any 2-layer neural network $f : \Omega \to \mathbb{R}$ of the form (4.2) can be computed as*

$$\mathrm{HTV}(f) = \omega_{d-1}(R) \sum_{n=1}^{N} \gamma_n |v_n| \|\boldsymbol{w}_n\|_2, \tag{4.3}$$

with the weight coefficients $\gamma_n = \left(1 - \frac{b_n^2}{\|\boldsymbol{w}_n\|_2^2 R^2}\right)_+^{\frac{d-1}{2}} \in [0, 1]$ for $n = 1, \dots, N$, where $\omega_{d-1}(R)$ denotes the volume of the $(d-1)$-dimensional sphere of radius R.

Proof. We shall proceed by induction on N. For $N = 0$, f is an affine mapping whose HTV is zero. Assuming that the $\mathrm{HTV}(f)$ is given by (4.3), we then consider the function

$$g = f + v_{N+1} \mathrm{ReLU}\left(\boldsymbol{w}_{N+1}^T \cdot - b_{N+1}\right) \tag{4.4}$$

for some arbitrary $v_{N+1}, b_{N+1} \in \mathbb{R}$, and $\boldsymbol{w}_{N+1} \in \mathbb{R}^d$. Next, we invoke the positive homogeneity of ReLU to obtain

$$g = f + \tilde{v} \mathrm{ReLU}\left(\boldsymbol{u}^T \cdot - \tilde{b}\right), \tag{4.5}$$

where $\tilde{v} = v_{N+1} \|\boldsymbol{w}_{N+1}\|_2$, $\boldsymbol{u} = \boldsymbol{w} / \|\boldsymbol{w}_{N+1}\|_2$, and $\tilde{b} = b_n / \|\boldsymbol{w}_{N+1}\|_2$. There are two cases:

Case I: $|\tilde{b}| \geq R$. There, $(g - f)$ is an affine mapping over Ω and, hence, we have that $\mathrm{HTV}(f) = \mathrm{HTV}(g)$.

Case II: $|\tilde{b}| < R$. In this case, the ridge function $\mathrm{ReLU}(\boldsymbol{u}^T \cdot - \tilde{b})$ splits some of the linear regions of f in two. Let us denote one of these regions by P whose gradient vector is $\boldsymbol{a}$. This region is now divided in two: $P = P_1 \sqcup P_2$ with gradient vectors $\boldsymbol{a}_1 = \boldsymbol{a}$ and $\boldsymbol{a}_2 = \boldsymbol{a} + \tilde{v}\boldsymbol{u}$, respectively. Hence, the contribution of the ridge function in P is equal to $H^{d-1}(P_1 \cap P_2)|\tilde{v}|$. Summing this up over all relevant regions yields

$$\mathrm{HTV}(g) = \mathrm{HTV}(f) + |\tilde{v}| H^{d-1}(E), \qquad E = \left\{ \boldsymbol{x} \in \Omega : \boldsymbol{u}^T \boldsymbol{x} = \tilde{b} \right\}. \tag{4.6}$$

Since the domain Ω is rotation-invariant, we can assume without loss of generality that $\boldsymbol{u} = \mathbf{e}_1$. Hence,

$$\begin{aligned} H^{d-1}(E) &= H^{d-1}\left(\{\boldsymbol{x} \in \mathbb{R}^d : x_1 = \tilde{b}, \|\boldsymbol{x}\|_2 \le R\}\right) \\ &= \mathrm{Vol}\left(\{\boldsymbol{x} \in \mathbb{R}^{d-1} : \|\boldsymbol{x}\|_2^2 \le R^2 - \tilde{b}^2\}\right) \\ &= \left(1 - \frac{\tilde{b}^2}{R^2}\right)^{\frac{d-1}{2}} \omega_{d-1}(R). \end{aligned}$$

Combining this with (4.6) yields the announced bound. ■

Proposition 4.3 indicates that, for shallow neural networks, the HTV is a weighted analogue of the path-norm regularizer [42], with the weights of each path depending on the corresponding bias. This makes sense for bounded domains, because the effective influence of each ridge (i.e., the way it changes the shape of the overall function) is inversely proportional to its off-set. In particular, if the ridge is far out (Case I in the proof), then it simply induces an affine term over Ω.

Let us also highlight that for large values of R, the weight coefficients tend to 1, i.e., $\lim_{R\to+\infty} \gamma_n = 1$. This means that for large domains, the HTV has an effect similar to the path-norm regularizer, which also means that it is closely linked to weight decay regularization [45]. Moreover, the path-norm regularizer is known to be equal to the Radon-domain total variation of the input-output mapping [46]. Specifically, for shallow neural networks $f : \mathbb{R}^d \to \mathbb{R}$, we have that

$$\mathcal{R}\mathrm{TV}^{(2)}(f) = \lim_{R\to+\infty} \frac{\mathrm{HTV}(f|_\Omega)}{\omega_{d-1}(R)}. \tag{4.7}$$

Although we have only established this connection for shallow neural networks, we believe the two measures of complexity are linked for deeper architectures as well. This requires understanding the effect of function composition (at least, in the CPWL family) in the computation of the HTV. The question, however, is very challenging and constitutes an interesting direction for future research.

5. Illustrations of usage.

In this section, we illustrate the behavior of the HTV seminorm in different scenarios. The associated codes are available online[1]. In our first example, we consider the problem of learning one-dimensional mappings from noisy data. In this example, we compare five different learning schemes:

[1] https://github.com/joaquimcampos/HTV-Learn

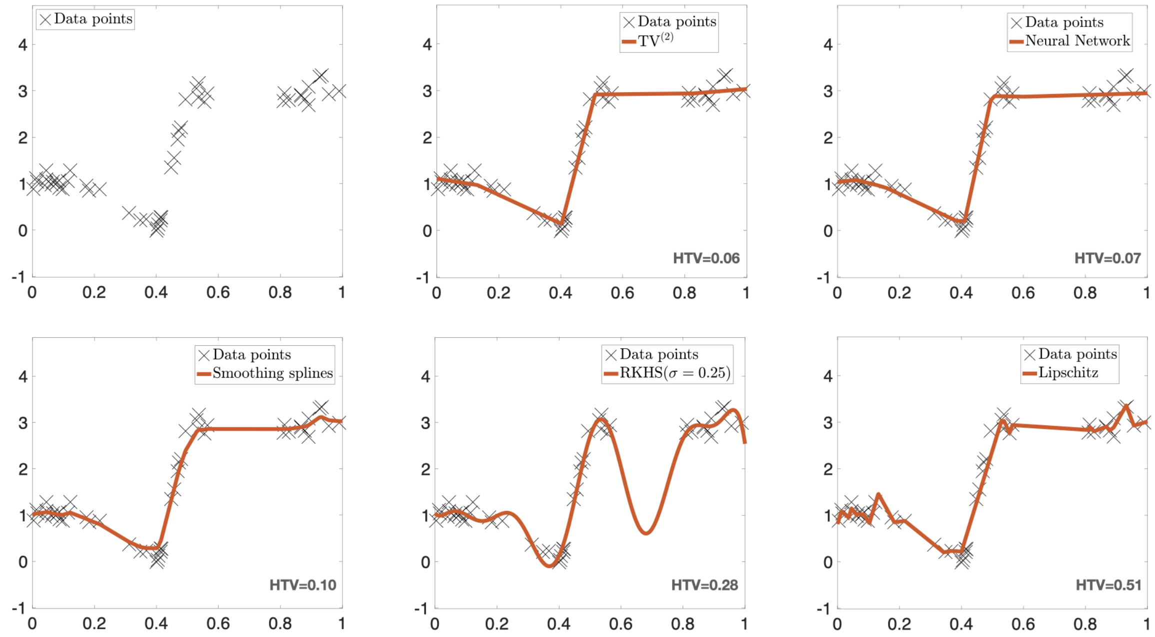


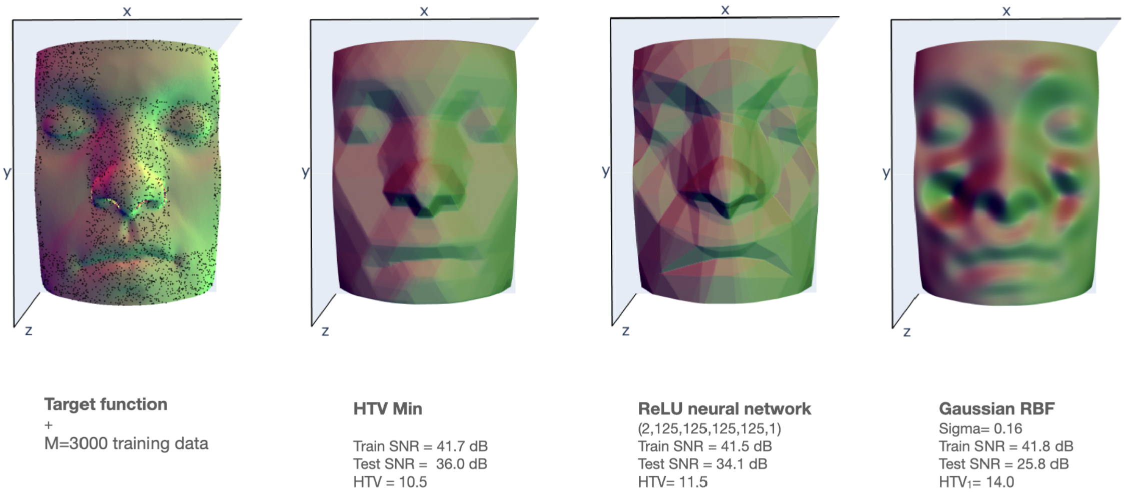
Figure 3. *Comparison of five different learning schemes in the* 1*D setting.*

1. A ReLU neural network with three hidden layers, each layer consisting of 10 neurons;
2. CPWL learning using TV-2 regularization [16];
3. CPWL learning using Lipschitz regularization [3];
4. CPWL learning using the L_2 norm of the first derivative as the regularization term (smoothing spline);
5. RKHS learning with a Gaussian reproducing kernel $k(x,y) = \exp(-(x-y)^2/(2\sigma^2))$ whose width is $\sigma = 1/4$.

We set the hyperparameters of each method such that they all have a similar training loss. The learned mappings are depicted in Figure 3, where we have also indicated their corresponding HTV value. As can be seen, the models that have a lower HTV are simpler and visually more satisfactory. Moreover, we observe that the neural network produces a CPWL mapping with complexity similar to the one produced by the TV-2 regularization scheme, which is expected to yield the mapping with the smallest HTV. This is in line with the recent results in deep-learning theory that indicate the existence of certain implicit regularizations in the learning of neural networks [57].

Next, we consider a 2D learning example where we take $M = 3000$ samples from a 2D height map obtained from a facial dataset[2]. Note that there are gaps in the training data, which makes the fitting problem more challenging. In this case, we compare three different learning schemes:

[2] https://www.turbosquid.com/3d-models/3d-male-head-model-1357522



Target function
+
M=3000 training data

HTV Min
Train SNR = 41.7 dB
Test SNR = 36.0 dB
HTV = 10.5

ReLU neural network
(2,125,125,125,125,1)
Train SNR = 41.5 dB
Test SNR = 34.1 dB
HTV= 11.5

Gaussian RBF
Sigma= 0.16
Train SNR = 41.8 dB
Test SNR = 25.8 dB
HTV_1= 14.0

Figure 4. *Learning of a 2D height map of a face from its nonuniform samples.*

1. A ReLU neural network with four hidden layers, each layer consisting of 125 hidden neurons (the neural network has been explicitly regularized using the weight decay scheme with the parameter $\mu = 1\text{e}-6$);
2. RKHS learning with a Gaussian radial-basis function whose width is $\sigma = 0.196$ and regularization parameter is $\lambda = 1\text{e}-3$;
3. the framework of learning 2D functions with HTV regularization with the parameter $\lambda = 1\text{e}-2$ [12].

We tune the hyperparameters of each framework (including the width of the neural network) to have a similar training error. The results are depicted in Figure 4. Similarly to the previous case, this example illustrates the property that the HTV favors simple and intuitive models that are visually more adequate.

Finally, we study the role of hyperparameters in the complexity of the final learned mapping in the 2D dataset. To that end, we plot in Figure 5 the HTV_p (for three different values of p) of the mapping learned by the RBF scheme versus the regularization parameter λ and the kernel width σ. As expected, sharper kernels and lower values of λ correspond to a higher HTV in the output.

Analogously, we have plotted in Figure 6 the HTV of the neural network used in Figure 4 versus the weight decay parameter μ. The result suggests that the two metrics follow the same trend.

6. Conclusion.

In this paper, we have introduced the Hessian-Schatten total-variation (HTV) seminorm and proposed its use as a complexity measure for the study of learning schemes. Our notion of complexity is very general and can be applied to different scenarios. We have proven that the HTV enjoys the properties that are expected of a good complexity measure, such as invariance to simple transformations and zero penalization of linear regressors. We then computed the HTV of two general classes of functions. In each case, we derived

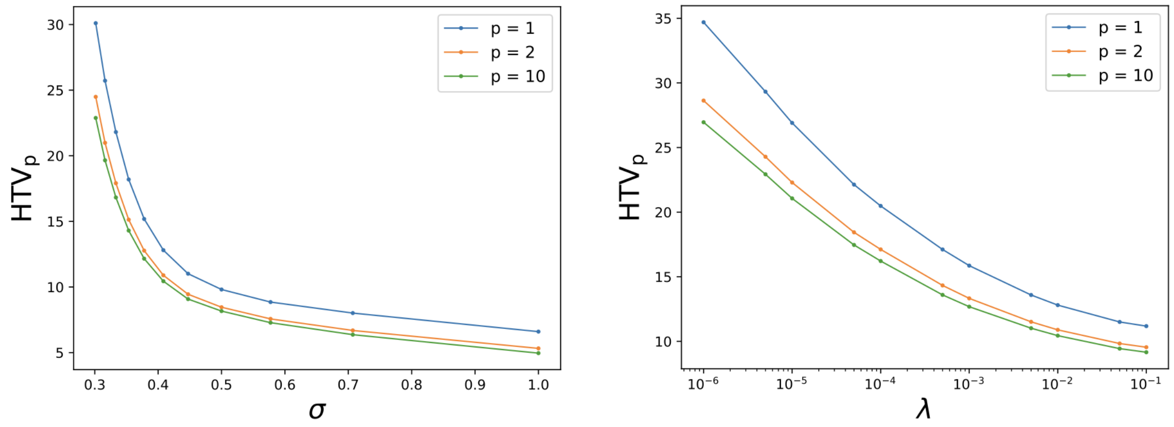


Figure 5. *The HTV of the learned kernel estimator versus the regularization weight λ (left) and the kernel width σ (right) in the 2D face example.*

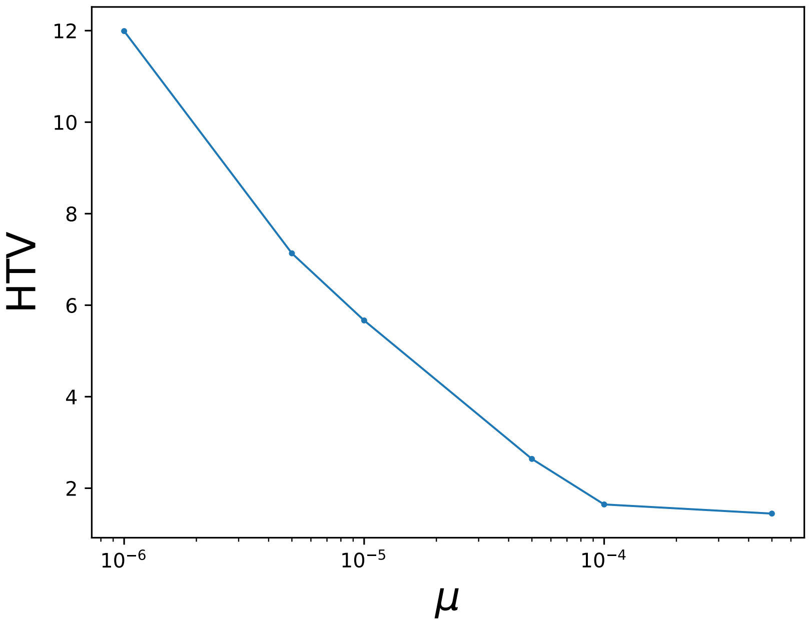


Figure 6. *The HTV of the learned neural network versus the weight decay parameter μ in the 2D face example.*

simple formulas for the HTV that allowed us to interpret its underlying behavior. Finally, we have provided some illustrative examples of usage for the comparison of learning algorithms. Future research directions could be to use this notion of complexity to study learning schemes, in particular their generalization power.

Appendix A. Proof of Theorem 3.2.

Proof. It is known that all norms are equivalent in finite-dimensional vector spaces. Consequently, there exist positive constants $c_1, c_2 > 0$ such that

$$\forall \mathbf{A} = [a_{i,j}] \in \mathbb{R}^{d\times d}, \quad c_1 \|\mathbf{A}\|_{\text{sum}} \leq \|\mathbf{A}\|_{S_q} \leq c_2 \|\mathbf{A}\|_{\text{sum}},$$

where $\|\mathbf{A}\|_{\text{sum}} = \sum_{i=1}^{d} \sum_{j=1}^{d} |a_{i,j}|$. This immediately yields that

$$c_1 \sum_{i=1}^{d} \sum_{j=1}^{d} \|f_{i,j}\|_{L_\infty} \leq \|\mathbf{F}\|_{L_\infty, S_q} \leq c_2 \sum_{i=1}^{d} \sum_{j=1}^{d} \|f_{i,j}\|_{L_\infty}, \tag{A.1}$$

as well as that

$$c_1 \sup_{\boldsymbol{x}\in\Omega} \|\mathbf{F}(\boldsymbol{x})\|_{\text{sum}} \leq \|\mathbf{F}\|_{S_q, L_\infty} \leq c_2 \sup_{\boldsymbol{x}\in\Omega} \|\mathbf{F}(\boldsymbol{x})\|_{\text{sum}}, \tag{A.2}$$

for all $\mathbf{F} \in \mathcal{C}_0(\Omega; \mathbb{R}^{d\times d})$. On the one hand, we have that

$$\sup_{\boldsymbol{x}\in\Omega} \|\mathbf{F}(\boldsymbol{x})\|_{\text{sum}} = \sup_{\boldsymbol{x}\in\Omega} \left(\sum_{i,j=1}^{d} |f_{i,j}(\boldsymbol{x})| \right) \leq \sum_{i,j=1}^{d} \sup_{\boldsymbol{x}\in\Omega} |f_{i,j}(\boldsymbol{x})| = \sum_{i,j=1}^{d} \|f_{i,j}\|_{L_\infty}. \tag{A.3}$$

Combining (A.1), (A.2), and (A.3), we then deduce that

$$\|\mathbf{F}\|_{S_q, L_\infty} \leq c_2 \sup_{\boldsymbol{x}\in\Omega} \|\mathbf{F}(\boldsymbol{x})\|_{\text{sum}} \leq c_2 \sum_{i,j=1}^{d} \|f_{i,j}\|_{L_\infty} \leq \frac{c_2}{c_1} \|\mathbf{F}\|_{L_\infty, S_q}. \tag{A.4}$$

On the other hand, using $\|\mathbf{F}(\boldsymbol{x})\|_{\text{sum}} \geq |f_{i,j}(\boldsymbol{x})|$ for all $i, j = 1, \dots, d$, we obtain that

$$\|f_{i,j}\|_{L_\infty} = \sup_{\boldsymbol{x}\in\Omega} |f_{i,j}(\boldsymbol{x})| \leq \sup_{\boldsymbol{x}\in\Omega} \|\mathbf{F}(\boldsymbol{x})\|_{\text{sum}} \quad \forall i, j = 1, \dots, d.$$

Summing over all $i, j = 1, \dots, d$ then gives that

$$\sum_{i=1}^{d} \sum_{j=1}^{d} \|f_{i,j}\|_{L_\infty} \leq d^2 \sup_{\boldsymbol{x}\in\Omega} \|\mathbf{F}(\boldsymbol{x})\|_{\text{sum}}. \tag{A.5}$$

Combining (A.1), (A.2), and (A.5), we obtain that

$$\|\mathbf{F}\|_{L_\infty, S_q} \leq c_2 \sum_{i=1}^{d} \sum_{j=1}^{d} \|f_{i,j}\|_{L_\infty} \leq c_2 d^2 \sup_{\boldsymbol{x}\in\Omega} \|\mathbf{F}(\boldsymbol{x})\|_{\text{sum}} \leq \frac{c_2}{c_1} d^2 \|\mathbf{F}\|_{S_q, L_\infty}. \tag{A.6}$$

Finally, the inequalities (A.4) and (A.6) yield (3.2) with $A = \frac{c_1}{c_2}$ and $B = \frac{c_2}{c_1} d^2$ (item 2). Further, it guarantees that the functional $\mathbf{F} \mapsto \|\mathbf{F}\|_{S_q, L_\infty}$ is well-defined (finite) for all $\mathbf{F} \in \mathcal{C}_0(\Omega, \mathbb{R}^{d\times d})$. It is then easy to verify the remaining norm properties (positivity, homogeneity, and the triangle inequality) of $\|\cdot\|_{S_q, L_\infty}$ (item 1). As for item 3, we note that the norm equivalence implies that both norms induce the same topology over $\mathcal{C}_0(\Omega; \mathbb{R}^{d\times d})$. Hence, $(\mathcal{C}_0(\Omega; \mathbb{R}^{d\times d}), \|\cdot\|_{S_q, L_\infty})$ is a bona fide Banach space. ■

Appendix B. Proof of Theorem 3.5.

Proof. Item 1. We first show that the right-hand side of (3.7) is well-defined and admits a finite value. First, note that $\|\mathbf{W}(\cdot)\|_{S_p}$ is the composition of the measurable function $\mathbf{W}:\Omega\to\mathbb{R}^{d\times d}$ and the Schatten-p norm $\|\cdot\|_{S_p}:\mathbb{R}^{d\times d}\to\mathbb{R}$ that is continuous and, consequently, measurable. This implies that $\|\mathbf{W}(\cdot)\|_{S_p}$ is also a measurable function and, hence, its L_1 norm is well-defined. The last step is to show that the L_1 norm is finite. From the norm-equivalence property of finite-dimensional vector spaces, we deduce the existence of $b>0$ such that, for any $\mathbf{A}=[a_{i,j}]\in\mathbb{R}^{d\times d}$, we have that

$$\|\mathbf{A}\|_{S_p}\le b\|\mathbf{A}\|_{\mathrm{sum}}, \tag{B.1}$$

where $\|\mathbf{A}\|_{\mathrm{sum}}=\sum_{i=1}^d\sum_{j=1}^d|a_{i,j}|$. This implies that

$$\|\,\|\mathbf{W}(\cdot)\|_{S_p}\,\|_{L_1}=\int_\Omega\|\mathbf{W}(\boldsymbol{x})\|_{S_p}\,\mathrm{d}\boldsymbol{x}\le b\int_\Omega\|\mathbf{W}(\boldsymbol{x})\|_{\mathrm{sum}}\,\mathrm{d}\boldsymbol{x}\overset{\text{(i)}}{=}b\sum_{i=1}^d\sum_{j=1}^d\|w_{i,j}\|_{L_1}<+\infty,$$

where we have used Fubini's theorem to deduce (i). Now, one readily verifies that

$$\langle\mathbf{W},\mathbf{F}\rangle=\sum_{i,j=1}^d\langle w_{i,j},f_{i,j}\rangle=\sum_{i,j=1}^d\int_\Omega w_{i,j}(\boldsymbol{x})f_{i,j}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=\int_\Omega\left(\sum_{i,j=1}^d w_{i,j}(\boldsymbol{x})f_{i,j}(\boldsymbol{x})\right)\mathrm{d}\boldsymbol{x},$$

$$\int_\Omega\left|\sum_{i,j=1}^d w_{i,j}(\boldsymbol{x})f_{i,j}(\boldsymbol{x})\right|\mathrm{d}\boldsymbol{x}\overset{\text{(i)}}{\le}\int_\Omega\|\mathbf{W}(\boldsymbol{x})\|_{S_p}\|\mathbf{F}(\boldsymbol{x})\|_{S_q}\mathrm{d}\boldsymbol{x}\overset{\text{(ii)}}{\le}\|\,\|\mathbf{W}(\cdot)\|_{S_p}\,\|_{L_1}\|\mathbf{F}\|_{S_q,L_\infty},$$

where we have used the Hölder inequality for Schatten norms (see (2.2)) in (i) and the one for L_p norms in (ii). We conclude that

$$\|\mathbf{W}\|_{S_p,\mathcal{M}}\le\|\,\|\mathbf{W}(\cdot)\|_{S_p}\,\|_{L_1}. \tag{B.2}$$

To show the equality, we need to prove that, for any $\epsilon>0$, there exists an element $\mathbf{F}_\epsilon\in\mathcal{C}_0(\Omega;\mathbb{R}^{d\times d})$ with $\|\mathbf{F}_\epsilon\|_{S_q,L_\infty}=1$ such that

$$\langle\mathbf{W},\mathbf{F}_\epsilon\rangle\ge\|\,\|\mathbf{W}(\cdot)\|_{S_p}\,\|_{L_1}-\epsilon. \tag{B.3}$$

Consider the function $\mathbf{F}:\Omega\to\mathbb{R}^{d\times d}$ with

$$\mathbf{F}(\boldsymbol{x})=\begin{cases}\frac{\mathrm{J}_{S_p,\mathrm{rank}}(\mathbf{W}(\boldsymbol{x}))}{\|\mathbf{W}(\boldsymbol{x})\|_{S_p}}, & \mathbf{W}(\boldsymbol{x})\neq\mathbf{0},\\ 0 & \text{otherwise,}\end{cases} \tag{B.4}$$

where $\mathrm{J}_{S_p,\mathrm{rank}}:\mathbb{R}^{d\times d}\to\mathbb{R}$ is the sparse duality mapping that maps $\mathbf{A}\in\mathbb{R}^{d\times d}$ to its minimum rank (S_p,S_q)-conjugate (see [4] for the definition and the proof of well-definedness)[3]. We first note that $\mathbf{F}$ is a measurable function. Indeed, from [4], we know that $\mathrm{J}_{S_p,\mathrm{rank}}$ is a measurable mapping over $\mathbb{R}^{d\times d}$. Hence, its composition with the measurable function $\mathbf{W}$ is

[3] This function coincides with the usual duality mapping for $p\in(1,+\infty)$, and the rank constraint is only needed for the special case $p=1$.

also measurable. Moreover, norms are continuous (and, so, Borel-measurable) functionals. Therefore, we have that $\mathbf{F}(\boldsymbol{x}) = \mathbb{1}_{\mathbf{W}\neq\mathbf{0}}\frac{\mathrm{J}_{S_p,\mathrm{rank}}(\mathbf{W}(\boldsymbol{x}))}{\|\mathbf{W}(\boldsymbol{x})\|_{S_p}}$ is also Borel-measurable. Knowing the measurability of $\mathbf{F}$, we observe that

$$\text{(B.5)} \qquad \int_\Omega \mathrm{Tr}(\mathbf{W}^T(\boldsymbol{x})\mathbf{F}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} = \int_\Omega \|\mathbf{W}(\boldsymbol{x})\|_{S_p}\mathrm{d}\boldsymbol{x} = \|\,\|\mathbf{W}(\cdot)\|_{S_p}\|_{L_1}.$$

We also note that $\|\mathbf{F}\|_{S_q,L_\infty} = 1$. The final step is to use Lusin's theorem (see [21, Theorem 7.10]) to find an ϵ-approximation $\mathbf{F}_\epsilon \in \mathcal{C}_0(\Omega;\mathbb{R}^{d\times d})$ of $\mathbf{F}$ on the unit S_q-L_∞ ball so that

$$\text{(B.6)} \qquad \left|\int_\Omega \mathrm{Tr}(\mathbf{W}^T(\boldsymbol{x})\mathbf{F}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} - \int_\Omega \mathrm{Tr}(\mathbf{W}^T(\boldsymbol{x})\mathbf{F}_\epsilon(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}\right| \leq \epsilon.$$

Now, combining (B.6) with (B.5), we deduce (B.3), which completes the proof.

Item 2. We first recall that the application of a distribution $\mathbf{D}$ of the form (3.5) to any element $\mathbf{F} \in \mathcal{C}_0(\Omega;\mathbb{R}^{d\times d})$ can be computed as

$$\text{(B.7)} \qquad \langle \mathbf{D}, \mathbf{F}\rangle = \int_C \mathrm{Tr}\left(\mathbf{A}^T\mathbf{F}(\boldsymbol{x}, \mathrm{T}\boldsymbol{x})\right)\mathrm{d}\boldsymbol{x}.$$

Using Hölder's inequality, for any $\mathbf{F} \in \mathcal{C}_0(\Omega;\mathbb{R}^{d\times d})$ with $\|\mathbf{F}\|_{S_q,L_\infty} = 1$, we obtain that

$$\begin{aligned}\int_C \mathrm{Tr}\left(\mathbf{A}^T\mathbf{F}(\boldsymbol{x}, \mathrm{T}\boldsymbol{x})\right)\mathrm{d}\boldsymbol{x} &\leq \int_C \|\mathbf{A}\|_{S_p}\|\mathbf{F}(\boldsymbol{x}, \mathrm{T}\boldsymbol{x})\|_{S_q}\mathrm{d}\boldsymbol{x}\\ &\leq \|\mathbf{A}\|_{S_p}\int_C 1\mathrm{d}\boldsymbol{x} = \|\mathbf{A}\|_{S_1}\mathrm{Leb}(C),\end{aligned}$$

which implies that $\|\mathbf{D}\|_{S_p,\mathcal{M}} \leq \|\mathbf{A}\|_{S_p}\mathrm{Leb}(C)$. To verify the equality, we consider an element $\mathbf{F} \in \mathcal{C}_0(\Omega;\mathbb{R}^{d\times d})$ whose restriction on C is the constant matrix $\mathbf{A}^* = \|\mathbf{A}\|_{S_p}^{-1}\mathrm{J}_{S_p,\mathrm{rank}}(\mathbf{A})$.

Item 3. Following the assumption that $\mathrm{Leb}(C_0) = 0$, for any $\epsilon > 0$, there exists a measurable set $E \subseteq \mathbb{R}^{d_1}$ with $\mathrm{Leb}(E) = \epsilon/2$ such that $C_0 \subseteq E$. From the construction, we deduce that the sets $C_1\backslash E$ and $C_2\backslash E$ are separable; hence, there exists a function $\mathbf{F}_\epsilon \in \mathcal{C}_0(\Omega;\mathbb{R}^{d\times d})$ with $\|\mathbf{F}_\epsilon\|_{S_q,L_\infty} = 1$ such that

$$\mathbf{F}_\epsilon(\boldsymbol{x}_1, \mathbf{T}_i\boldsymbol{x}_1) = \mathbf{A}_i^* \quad \forall \boldsymbol{x}_1 \in C_i\backslash C_0, i = 1, 2,$$

where $\mathbf{A}_i^* = \|\mathbf{A}_i\|_{S_p}^{-1}\mathrm{J}_{S_p,\mathrm{rank}}(\mathbf{A}_i)$, $i = 1, 2$. This implies that, for $i = 1, 2$, we have

$$\begin{aligned}\langle \mathbf{D}_i, \mathbf{F}_\epsilon\rangle &= \int_{C_i} \mathrm{Tr}\left(\mathbf{A}_i^T\mathbf{F}_\epsilon(\boldsymbol{x}_1, \mathbf{T}_i\boldsymbol{x}_1)\right)\mathrm{d}\boldsymbol{x}_1\\ &= \int_{C_0} \mathrm{Tr}\left(\mathbf{A}_i^T\mathbf{F}_\epsilon(\boldsymbol{x}_1, \mathbf{T}_i\boldsymbol{x}_1)\right)\mathrm{d}\boldsymbol{x}_1 + \int_{C_i\backslash C_0} \mathrm{Tr}\left(\mathbf{A}_i^T\mathbf{A}_i^*\right)\mathrm{d}\boldsymbol{x}_1\\ &\geq -\mathrm{Leb}(C_0)\|\mathbf{A}_i\|_{S_p} + \mathrm{Leb}(C_i\backslash C_0)\|\mathbf{A}_i\|_{S_p}\\ &\geq \|\mathbf{A}_i\|_{S_p}(\mathrm{Leb}(C_i) - \epsilon).\end{aligned}$$

Hence, for any $\epsilon > 0$, we have that

$$\begin{aligned}\|\mathbf{D}_1 + \mathbf{D}_2\|_{S_p,\mathcal{M}} &\geq \langle \mathbf{D}_1 + \mathbf{D}_2, \mathbf{F}_\epsilon\rangle\\ &\geq \|\mathbf{A}_1\|_{S_p}\mathrm{Leb}(C_1) + \|\mathbf{A}_2\|_{S_p}\mathrm{Leb}(C_2) - \epsilon(\|\mathbf{A}_1\|_{S_p} + \|\mathbf{A}_2\|_{S_p}).\end{aligned}$$

By letting $\epsilon \to 0$, we deduce that $\|\mathbf{D}_1 + \mathbf{D}_2\|_{S_p,\mathcal{M}} \geq \|\mathbf{D}_1\|_{S_p,\mathcal{M}} + \|\mathbf{D}_2\|_{S_p,\mathcal{M}}$, which, together with the triangle inequality, yields the announced equality. ■

Appendix C. Proof of Theorem 3.7.

Proof. Item 1. Starting from $\mathrm{H}\{f\} = \mathbf{0}$, we deduce that $\frac{\partial^2 f}{\partial x_i^2} = 0$ for $i = 1, \dots, d$. Following Proposition 6.1 in [64], we deduce that the null space of $\frac{\partial^2}{\partial x_1^2}$ can only contain (multivariate) polynomials. Using this, we infer that any p in the null space of $\frac{\partial^2}{\partial x_1^2}$ is of the form $p(\boldsymbol{x}) = a_1 x_1 + q_1(\boldsymbol{x})$ for some $a_1 \in \mathbb{R}$ and some multivariate polynomial q_1 that does not depend on x_1. Finally, one verifies by induction that $q_1(\boldsymbol{x}) = \sum_{i=2}^d a_i x_i + q_0(\boldsymbol{x})$, where q_0 is a multivariate polynomial that does not depend on any of its variables and so is constant, i.e., $q_0(\boldsymbol{x}) = b$ for some $b \in \mathbb{R}$. We conclude the proof by remarking that any affine mapping is indeed in the null space of H.

Item 2. By invoking that $\mathrm{H}\{f(\cdot - \boldsymbol{x}_0)\} = \mathrm{H}\{f\}(\cdot - \boldsymbol{x}_0)$, we immediately deduce that

$$\begin{aligned}\mathrm{HTV}\left(f(\cdot - \boldsymbol{x}_0)\right) &= \sup\left\{\langle \mathrm{H}\{f\}(\cdot - \boldsymbol{x}_0), \mathbf{F}\rangle : \mathbf{F} \in \mathcal{D}(\mathbb{R}^d; \mathbb{R}^{d\times d}), \|\mathbf{F}\|_{S_q, L_\infty} = 1\right\}\\ &= \sup\left\{\langle \mathrm{H}\{f\}, \mathbf{F}(\cdot + \boldsymbol{x}_0)\rangle : \mathbf{F} \in \mathcal{D}(\mathbb{R}^d; \mathbb{R}^{d\times d}), \|\mathbf{F}\|_{S_q, L_\infty} = 1\right\}\\ &= \mathrm{HTV}(f).\end{aligned}$$

Similarly, following the chain rule, we obtain that $\mathrm{H}\{f(\alpha\cdot)\} = \alpha^2 \mathrm{H}\{f\}(\alpha\cdot)$. This yields that

$$\begin{aligned}\mathrm{HTV}\left(f(\alpha\cdot)\right) &= \alpha^2 \sup\left\{\langle \mathrm{H}\{f\}(\alpha\cdot), \mathbf{F}\rangle : \mathbf{F} \in \mathcal{D}(\mathbb{R}^d; \mathbb{R}^{d\times d}), \|\mathbf{F}\|_{S_q, L_\infty} = 1\right\}\\ &= \alpha^2 \sup\left\{\langle \mathrm{H}\{f\}, \alpha^{-d}\mathbf{F}(\alpha^{-1}\cdot)\rangle : \mathbf{F} \in \mathcal{D}(\mathbb{R}^d; \mathbb{R}^{d\times d}), \|\mathbf{F}\|_{S_q, L_\infty} = 1\right\}\\ &= |\alpha|^{2-d} \sup\left\{\langle \mathrm{H}\{f\}, \mathbf{F}(\cdot)\rangle : \mathbf{F} \in \mathcal{D}(\mathbb{R}^d; \mathbb{R}^{d\times d}), \|\mathbf{F}\|_{S_q, L_\infty} = 1\right\}\\ &= |\alpha|^{2-d}\mathrm{HTV}(f).\end{aligned}$$

As for the last invariance property, we use the formula for the Hessian of a rotated function

$$\mathrm{H}\{f(\mathbf{U}\cdot)\} = \mathbf{U}^T \mathrm{H}\{f\}(\mathbf{U}\cdot)\mathbf{U}.$$

This implies that

$$\begin{aligned}\mathrm{HTV}\left(f(\mathbf{U}\cdot)\right) &= \sup\left\{\langle \mathbf{U}^T \mathrm{H}\{f\}(\mathbf{U}\cdot)\mathbf{U}, \mathbf{F}\rangle : \mathbf{F} \in \mathcal{D}(\mathbb{R}^d; \mathbb{R}^{d\times d}), \|\mathbf{F}\|_{S_q, L_\infty} = 1\right\}\\ &= \sup\left\{\langle \mathrm{H}\{f\}(\mathbf{U}\cdot), \mathbf{U}\mathbf{F}(\cdot)\mathbf{U}^T\rangle : \mathbf{F} \in \mathcal{D}(\mathbb{R}^d; \mathbb{R}^{d\times d}), \|\mathbf{F}\|_{S_q, L_\infty} = 1\right\}\\ &= \sup\left\{\langle \mathrm{H}\{f\}, \mathbf{U}\mathbf{F}(\mathbf{U}^T\cdot)\mathbf{U}^T\rangle : \mathbf{F} \in \mathcal{D}(\mathbb{R}^d; \mathbb{R}^{d\times d}), \|\mathbf{F}\|_{S_q, L_\infty} = 1\right\}\\ &= \mathrm{HTV}(f),\end{aligned}$$

where the last equality follows from the invariance of Schatten norms under orthogonal transformations (as exploited, for example, in [35, 37]). ■

Appendix D. Proof of Theorem 4.2. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a CPWL function with linear regions $P_n \subseteq \Omega$ and affine parameters $\boldsymbol{a}_n \in \mathbb{R}^d$ and $b_n \in \mathbb{R}$ for $n = 1, \dots, N$. We first compute the gradient of f.

Lemma D.1. *The gradient of a CPWL function $f : \Omega \to \mathbb{R}$ as described above can be expressed as*

$$\boldsymbol{\nabla} f(\boldsymbol{x}) = \sum_{n=1}^{N} \boldsymbol{a}_n \mathbb{1}_{P_n}(\boldsymbol{x}) \tag{D.1}$$

for almost every $\boldsymbol{x} \in \Omega$.

Proof. The interior of P_n is denoted by U_n, with $n = 1, \dots, N$. We then note that $\Omega \backslash (\bigcup_{n=1}^{N} U_n)$ is a set of measure zero. Hence, it is sufficient to show that $\boldsymbol{\nabla} f(\boldsymbol{x}) = \boldsymbol{a}_n$ for any $\boldsymbol{x}_0 = (x_{0,1}, \dots, x_{0,d}) \in U_n$. We define the functions $g_i : \mathbb{R} \to \mathbb{R}$ as

$$g_i(x) = f(x_{0,1}, \dots, x_{0,i-1}, x, x_{0,i+1}, \dots, x_{0,d}).$$

Following the definition of CPWL mappings, g_i is a linear spline (i.e., a 1D continuous and piecewise-linear function). Hence, it is locally linear and can be expressed as $g_i(x) = a_{n,i}x + (\sum_{j \neq i} a_{n,j} x_{0,j} + b)$ in an open neighborhood of $x_{0,i}$. Moreover, it is clear that $a_{n,i} = g_i'(x_{0,i}) = \frac{\partial f}{\partial x_i}(\boldsymbol{x}_0)$. Hence,

$$\boldsymbol{\nabla} f(\boldsymbol{x}_0) = \left(\frac{\partial f}{\partial x_1}(\boldsymbol{x}_0), \dots, \frac{\partial f}{\partial x_d}(\boldsymbol{x}_0) \right) = (a_{n,1}, \dots, a_{n,d}) = \boldsymbol{a}_n.$$ ■

Proof of Theorem **4.2**. We start by introducing some notions that are required in the proof. For each $n = 1, \dots, N$ and $k \in \text{adj}_n$, we denote the intersection of P_n and P_k by $L_{n,k} = P_n \cap P_k$, which is itself a convex polytope with codimension $(d-1)$, in the sense that it lies on a hyperplane $H_{n,k} = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{u}_{n,k}^T \boldsymbol{x} + \beta_{n,k} = 0\}$ for some normal vector $\boldsymbol{u}_{n,k} = (u_{n,k,i}) \in \mathbb{R}^d$ with $\|\boldsymbol{u}_{n,k}\|_2 = 1$ and some shift value $\beta_{n,k} \in \mathbb{R}$. We adopt the convention that $\boldsymbol{u}_{n,k}$ refers to the outward normal vector, so that $\boldsymbol{u}_{n,k}^T \boldsymbol{x} + \beta_{n,k} \leq 0$ for all $\boldsymbol{x} \in P_n$. We divide the proof in four steps:

Step 1: Transformation to the general position. First, without any loss of generality, we assume that all entries of $\boldsymbol{u}_{n,k}$ for all $n = 1, \dots, N$ and $k \in \text{adj}_n$ are nonzero. Consider a unitary matrix $\mathbf{V} \in \mathbb{R}^{d \times d}$ such that $[\mathbf{V}\boldsymbol{u}_{n,k}]_i \neq 0$ for all $n = 1, \dots, N$, $k \in \text{adj}_n$, and $i = 1, \dots, d$. We remark that the function $g = f(\mathbf{V}\cdot)$ is CPWL with linear regions $\tilde{P}_n = \mathbf{V}^T P_n$ and affine parameters $\tilde{\boldsymbol{a}}_n = \mathbf{V}^T \boldsymbol{a}_n$ and $\tilde{b}_n = b_n$ for $n = 1, \dots, N$. Now, if (4.1) holds for g, then we can invoke the invariance properties of the HTV (see Theorem 3.7) to deduce that

$$\begin{aligned}
\text{HTV}_p(f) &= \text{HTV}_p(g) \\
&= \frac{1}{2} \sum_{n=1}^{N} \sum_{k \in \text{adj}_n} \|\tilde{\boldsymbol{a}}_n - \tilde{\boldsymbol{a}}_k\|_2 H^{d-1}(\tilde{P}_n \cap \tilde{P}_k) \\
&= \frac{1}{2} \sum_{n=1}^{N} \sum_{k \in \text{adj}_n} \|\mathbf{V}^T(\boldsymbol{a}_n - \boldsymbol{a}_k)\|_2 H^{d-1}(\mathbf{V}^T(P_n \cap P_k)) \\
&= \frac{1}{2} \sum_{n=1}^{N} \sum_{k \in \text{adj}_n} \|\boldsymbol{a}_n - \boldsymbol{a}_k\|_2 H^{d-1}(P_n \cap P_k),
\end{aligned}$$

where the last equality is due to the invariance of the Hausdorff measure and the ℓ_2 norm to orthonormal transformations.

Step 2: Calculation of the Hessian distribution. From now on, we assume that all entries of $u_{n,k}$ are nonzero, with

$$u_{n,k,i} = 0, \qquad n = 0, \dots, N, \quad k \in \mathrm{adj}_n, \quad i = 1, \dots, d.$$

This allows us to view $H_{n,k}$ as the graph of the affine mapping $T_{n,k} : \mathbb{R}^{d-1} \to \mathbb{R}$, with

$$T_{n,k}(x_1, \dots, x_{d-1}) = \beta_{n,k} - \frac{\sum_{i=1}^{d-1} u_{n,k,i} x_i}{u_{n,k,d}},$$

and to define $C_{n,k} = \{\boldsymbol{x} \in \mathbb{R}^{d-1} : (\boldsymbol{x}, T_{n,k}\boldsymbol{x}) \in L_{n,k}\} \subseteq \Omega$ as the preimage of $L_{n,k}$ over $T_{n,k}$. We also remark that, due to this affine projection, the $(d-1)$-dimensional Hausdorff measure of $L_{n,k}$ and the Lebesgue measure of $C_{n,k}$ are related by the coefficient $u_{n,k,d}$. Indeed, we have that $H^{d-1}(L_{n,k}) = \frac{\mathrm{Leb}(C_{n,k})}{|u_{n,k,d}|}$. Using these notions, we now compute the matrix-valued distribution $\mathrm{H}\{f\} \in \mathcal{M}(\Omega; \mathbb{R}^{d \times d})$. We first note that, for all $n = 0, \dots, N$ and $i = 1, \dots, d$, we have that

$$\frac{\partial \mathbb{1}_{P_n}}{\partial x_i}(\boldsymbol{x}) = \sum_{k \in \mathrm{adj}_n} -\mathrm{sgn}(u_{n,k,i})\delta\left(x_i + \frac{\sum_{j \neq i} u_{n,k,j} x_j + \beta_{n,k}}{u_{n,k,i}}\right) \mathbb{1}_{L_{n,k}}(\boldsymbol{x}). \tag{D.2}$$

Using the relation $\delta(\alpha \cdot) = |\alpha|^{-1}\delta(\cdot)$ for all $\alpha \in \mathbb{R}$, we obtain that

$$\frac{\partial \mathbb{1}_{P_n}}{\partial x_i}(\boldsymbol{x}) = \sum_{k \in \mathrm{adj}_n} \frac{-u_{n,k,i}}{|u_{n,k,d}|} \delta\left(x_d - T_{n,k}\boldsymbol{x}_1\right) \mathbb{1}_{L_{n,k}}(\boldsymbol{x}), \tag{D.3}$$

where $\boldsymbol{x}_1 = (x_1, \dots, x_{d-1}) \in \mathbb{R}^{d-1}$. Following the definition of $C_{n,k}$, we immediately get that

$$\delta\left(x_d - T_{n,k}\boldsymbol{x}_1\right) \mathbb{1}_{L_{n,k}}(\boldsymbol{x}) = \delta\left(x_d - T_{n,k}\boldsymbol{x}_2\right) \mathbb{1}_{C_{n,k}}(\boldsymbol{x}_1), \tag{D.4}$$

which leads to

$$\frac{\partial \mathbb{1}_{P_n}}{\partial x_i}(\boldsymbol{x}) = \sum_{k \in \mathrm{adj}_n} \frac{-u_{n,k,i}}{|u_{n,k,d}|} \delta\left(x_d - T_{n,k}\boldsymbol{x}_1\right) \mathbb{1}_{C_{n,k}}(\boldsymbol{x}_1). \tag{D.5}$$

Combining (D.5) with Lemma D.1, we then deduce that

$$\begin{aligned}\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x}) &= \sum_{n=1}^{N} a_{n,j} \frac{\partial \mathbb{1}_{P_n}}{\partial x_i}(\boldsymbol{x}) \\ &= \sum_{n=1}^{N} a_{n,j} \sum_{k \in \mathrm{adj}_n} \frac{-u_{n,k,i}}{|u_{n,k,d}|} \delta\left(x_d - T_{n,k}\boldsymbol{x}_1\right) \mathbb{1}_{C_{n,k}}(\boldsymbol{x}_1).\end{aligned}$$

Now, since $L_{n,k} = P_n \cap P_k$ and $\boldsymbol{u}_{n,k} = (-\boldsymbol{u}_{k,n})$, we can rewrite the second-order partial derivatives as

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x}) = \frac{1}{2} \sum_{n=1}^{N} \sum_{k \in \mathrm{adj}_n} (a_{k,j} - a_{n,j}) \frac{u_{n,k,i}}{|u_{n,k,d}|} \delta\left(x_d - T_{n,k}\boldsymbol{x}_1\right) \mathbb{1}_{C_{n,k}}(\boldsymbol{x}_1).$$

Putting it in matrix form, we conclude that the Hessian is a sum of disjoint Dirac fences, as in

$$\begin{aligned}\mathrm{H}\{f\}(\boldsymbol{x}) &= \left[\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{x})\right]\\ &= \frac{1}{2}\sum_{n=1}^{N}\sum_{k\in \mathrm{adj}_n}\left[(a_{k,j}-a_{n,j})\frac{u_{n,k,i}}{|u_{n,k,d}|}\right]\delta\left(x_d - T_{n,k}\boldsymbol{x}_1\right)\mathbb{1}_{C_{n,k}}(\boldsymbol{x}_1). \qquad \text{(D.6)}\end{aligned}$$

Step 3: Computation of the HTV. By invoking item 3 of Theorem 3.5, we deduce that

$$\begin{aligned}\mathrm{HTV}_p(f) &= \|\mathrm{H}\{f\}\|_{S_p,\mathcal{M}}\\ &= \frac{1}{2}\sum_{n=1}^{N}\sum_{k\in \mathrm{adj}_n}\left\|\left[(a_{k,j}-a_{n,j})\frac{u_{n,k,i}}{|u_{n,k,d}|}\right]\delta\left(x_d - T_{n,k}\boldsymbol{x}_1\right)\mathbb{1}_{C_{n,k}}(\boldsymbol{x}_1)\right\|_{S_p,\mathcal{M}}\\ &= \frac{1}{2}\sum_{n=1}^{N}\sum_{k\in \mathrm{adj}_n}\left\|\left[(a_{k,j}-a_{n,j})\frac{u_{n,k,i}}{|u_{n,k,d}|}\right]\right\|_{S_p}\mathrm{Leb}(C_{n,k}), \qquad \text{(D.7)}\end{aligned}$$

where the last equality results from item 2 of Theorem 3.5.
Finally, we use the continuity of f to deduce that, for any pair of points $\boldsymbol{p}_1, \boldsymbol{p}_2 \in H_{n,k}$, we have that

$$\boldsymbol{a}_n^T\boldsymbol{p}_i + b_n = \boldsymbol{a}_k^T\boldsymbol{p}_i + b_k, \qquad i = 1, 2.$$

Subtracting the above equalities for $i = 1$ and $i = 2$, we obtain that

$$\boldsymbol{a}_n^T(\boldsymbol{p}_1 - \boldsymbol{p}_2) = \boldsymbol{a}_k^T(\boldsymbol{p}_1 - \boldsymbol{p}_2).$$

However, $(\boldsymbol{p}_1 - \boldsymbol{p}_2)$ is orthogonal to $\boldsymbol{u}_{n,k}$. Hence, the vector $(\boldsymbol{a}_k - \boldsymbol{a}_n)$ points in the direction of $\boldsymbol{u}_{n,k}$. This implies that the matrix

$$\left[(a_{k,j}-a_{n,j})\frac{u_{n,k,i}}{|u_{n,k,d}|}\right] = |u_{n,k,d}|^{-1}\boldsymbol{u}_{n,k}(\boldsymbol{a}_k - \boldsymbol{a}_n)^T = \frac{\|\boldsymbol{a}_k - \boldsymbol{a}_n\|_2}{|u_{n,k,d}|}\boldsymbol{u}_{n,k}\boldsymbol{u}_{n,k}^T$$

is rank-1 and symmetric. Hence, for any $p \in [1, +\infty]$, its Schatten-p norm is equal to the absolute value of its trace. The replacement of this in (D.7) and the use of $H^{d-1}(L_{n,k}) = \frac{\mathrm{Leb}(C_{n,k})}{|u_{n,k,d}|}$ yields the announced expression (4.1). ■

REFERENCES

[1] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford Math. Monogr. 254, The Clarendon Press, Oxford University Press, New York, 2000.

[2] R. Arora, A. Basu, P. Mianjy, and A. Mukherjee, *Understanding Deep Neural Networks with Rectified Linear Units*, preprint, https://arxiv.org/abs/1611.01491, 2016.

[3] S. Aziznejad, T. Debarre, and M. Unser, *Sparsest Univariate Learning Models Under Lipschitz Constraint*, preprint, https://arxiv.org/abs/2112.13542, 2021.

[4] S. Aziznejad and M. Unser, *Duality mapping for Schatten matrix norms*, Numer. Funct. Anal. Optim., 42 (2021), pp. 679–695.

[5] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, *Benign overfitting in linear regression*, Proc. Natl. Acad. Sci. USA, 117 (2020), pp. 30063–30070.

[6] P. L. Bartlett, A. Montanari, and A. Rakhlin, *Deep learning: A statistical viewpoint*, Acta Numer., 30 (2021), pp. 87–201.

[7] M. Bergounioux and L. Piffet, *A second-order model for image denoising*, Set-Valued Var. Anal., 18 (2010), pp. 277–306.

[8] R. Bhatia, *Matrix Analysis*, Grad. Texts in Math. 169, Springer, New York, 2013.

[9] K. Bredies and M. Holler, *Regularization of linear inverse problems with total generalized variation*, J. Inverse Ill-Posed Probl., 22 (2014), pp. 871–913.

[10] K. Bredies and M. Holler, *Higher-order total variation approaches and generalisations*, Inverse Problems, 36 (2020), 123001.

[11] K. Bredies, K. Kunisch, and T. Pock, *Total generalized variation*, SIAM J. Imaging Sci., 3 (2010), pp. 492–526, https://doi.org/10.1137/090769521.

[12] J. Campos, S. Aziznejad, and M. Unser, *Learning of continuous and piecewise-linear functions with Hessian total-variation regularization*, IEEE Open J. Signal Process., 3 (2021), pp. 36–48.

[13] A. Caponnetto and E. De Vito, *Optimal rates for the regularized least-squares algorithm*, Found. Comput. Math., 7 (2007), pp. 331–368.

[14] D. A. Clevert, T. Unterthiner, and S. Hochreiter, *Fast and Accurate Deep Network Learning by Exponential Linear Units (elus)*, preprint, https://arxiv.org/abs/1511.07289, 2015.

[15] G. Cybenko, *Approximation by superpositions of a sigmoidal function*, Math. Control Signals Systems, 2 (1989), pp. 303–314.

[16] T. Debarre, Q. Denoyelle, M. Unser, and J. Fageot, *Sparsest piecewise-linear regression of one-dimensional data*, J. Comput. Appl. Math., 406 (2022), 114044.

[17] F. Demengel, *Fonctions à hessien borné*, Ann. Inst. Fourier (Grenoble), 34 (1984), pp. 155–190.

[18] D. L. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.

[19] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*, Cambridge University Press, Cambridge, UK, 2012.

[20] T. Evgeniou, M. Pontil, and T. Poggio, *Regularization networks and support vector machines*, Adv. Comput. Math., 13 (2000), pp. 1–50.

[21] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*, Pure Appl. Math. (New York) 40, John Wiley & Sons, New York, 1999.

[22] X. Glorot, A. Bordes, and Y. Bengio, *Deep sparse rectifier neural networks*, in Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 2011, pp. 315–323.

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

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

[25] B. Hanin and D. Rolnick, *Complexity of Linear Regions in Deep Networks*, preprint, https://arxiv.org/abs/1901.09021, 2019.

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

[27] D. Hendrycks and K. Gimpel, *Gaussian Error Linear Units (gelus)*, preprint, https://arxiv.org/abs/1606.08415, 2016.

[28] W. Hinterberger and O. Scherzer, *Variational methods on the space of functions of bounded Hessian for convexification and denoising*, Computing, 76 (2006), pp. 109–133.

[29] K. Jin, M. McCann, E. Froustey, and M. Unser, *Deep convolutional neural network for inverse problems in imaging*, IEEE Trans. Image Process., 26 (2017), pp. 4509–4522.

[30] G. Kimeldorf and G. Wahba, *Some results on Tchebycheffian spline functions*, J. Math. Anal. Appl., 33 (1971), pp. 82–95.

[31] F. Knoll, K. Bredies, T. Pock, and R. Stollberger, *Second order total generalized variation (TGV) for MRI*, Magn. Reson. Med., 65 (2011), pp. 480–491.

[32] A. Krizhevsky, I. Sutskever, and G. Hinton, *Imagenet classification with deep convolutional neural networks*, in Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2012, pp. 1097–1105.

[33] T. Kurokawa, *Riesz potentials, higher Riesz transforms and Beppo Levi spaces*, Hiroshima Math. J., 18 (1988), pp. 541–597.

[34] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, Nature, 521 (2015), pp. 436–444.

[35] S. Lefkimmiatis, A. Bourquard, and M. Unser, *Hessian-based norm regularization for image restoration with biomedical applications*, IEEE Trans. Image Process., 21 (2012), pp. 983–995.

[36] S. Lefkimmiatis, A. Roussos, P. Maragos, and M. Unser, *Structure tensor total variation*, SIAM J. Imaging Sci., 8 (2015), pp. 1090–1122, https://doi.org/10.1137/14098154X.

[37] S. Lefkimmiatis, J. Ward, and M. Unser, *Hessian Schatten-norm regularization for linear inverse problems*, IEEE Trans. Image Process., 22 (2013), pp. 1873–1888.

[38] Z. Li, Z. H. Zhou, and A. Gretton, *Towards an Understanding of Benign Overfitting in Neural Networks*, preprint, https://arxiv.org/abs/2106.03212, 2021.

[39] S. Mendelson and J. Neeman, *Regularization in kernel learning*, Ann. Statist., 38 (2010), pp. 526–565.

[40] D. Misra, *Mish: A Self Regularized Non-monotonic Neural Activation Function*, preprint, https://arxiv.org/abs/1908.08681, 2019.

[41] G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio, *On the number of linear regions of deep neural networks*, in Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2014, pp. 2924–2932.

[42] B. Neyshabur, R. R. Salakhutdinov, and N. Srebro, *Path-SGD: Path-Normalized Optimization in Deep Neural Networks*, Adv. Neural Inform. Process. Syst. 28, MIT Press, Cambridge, MA, 2015.

[43] G. Ongie, R. Willett, D. Soudry, and N. Srebro, *A function space view of bounded norm infinite width ReLU nets: The multivariate case*, in Proceedings of the Eighth International Conference on Learning Representations (ICLR'20), Addis Ababa, Ethiopia, 2020.

[44] L. Onural, *Impulse functions over curves and surfaces and their applications to diffraction*, J. Math. Anal. Appl., 322 (2006), pp. 18–27.

[45] R. Parhi and R. D. Nowak, *The role of neural network activation functions*, IEEE Signal Process. Lett., 27 (2020), pp. 1779–1783.

[46] R. Parhi and R. D. Nowak, *Banach space representer theorems for neural networks and ridge splines*, J. Mach. Learn. Res., 22 (2021), pp. 1–40.

[47] R. Parhi and R. D. Nowak, *Near-minimax optimal estimation with shallow ReLU neural networks*, IEEE Trans. Inform. Theory, 69 (2023), pp. 1125–1140.

[48] R. Parhi and R. D. Nowak, *What kinds of functions do deep neural networks learn? Insights from variational spline theory*, SIAM J. Math. Data Sci., 4 (2022), pp. 464–489, https://doi.org/10.1137/21M1418642.

[49] R. Pascanu, G. Montufar, and Y. Bengio, *On the Number of Response Regions of Deep Feed Forward Networks with Piece-wise Linear Activations*, preprint, https://arxiv.org/abs/1312.6098, 2013.

[50] T. Poggio and F. Girosi, *Networks for approximation and learning*, Proc. IEEE, 78 (1990), pp. 1481–1497.

[51] T. Poggio and F. Girosi, *Regularization algorithms for learning that are equivalent to multilayer networks*, Science, 247 (1990), pp. 978–982.

[52] A. Rakhlin and X. Zhai, *Consistency of interpolation with Laplace kernels is a high-dimensional phenomenon*, in Proceedings of the Conference on Learning Theory, PMLR, 2019, pp. 2595–2623.

[53] P. Ramachandran, B. Zoph, and Q. V. Le, *Searching for Activation Functions*, preprint, https://arxiv.org/abs/1710.05941, 2017.

[54] B. Recht, M. Fazel, and P. A. Parrilo, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, SIAM Rev., 52 (2010), pp. 471–501, https://doi.org/10.1137/070697835.

[55] O. Ronneberger, P. Fischer, and T. Brox, *U-Net: Convolutional networks for biomedical image segmentation*, in Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2015, pp. 234–241.

[56] W. Rudin, *Real and Complex Analysis*, Tata McGraw-Hill Education, New York, 2006.

[57] P. Savarese, I. Evron, D. Soudry, and N. Srebro, *How Do Infinite Width Bounded Norm Networks Look in Function Space*?, preprint, https://arxiv.org/abs/1902.05040, 2019.

[58] B. Schölkopf, R. Herbrich, and A. J. Smola, *A generalized representer theorem*, in Proceedings of the International Conference on Computational Learning Theory, Springer, Berlin, Heidelberg, 2001, pp. 416–426.

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

[60] L. Schwartz, *Théorie des Distributions*, Vol. 2, Hermann Paris, 1957.

[61] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer, New York, 2008.

[62] I. Steinwart, D. R. Hush, and C. Scovel, *Optimal rates for regularized least squares regression*, in Proceedings of the Conference on Learning Theory, 2009, pp. 79–93.

[63] M. Unser and S. Aziznejad, *Convex optimization in sums of Banach spaces*, Appl. Comput. Harmon. Anal., 56 (2022), pp. 1–25.

[64] M. Unser and P. Tafti, *An Introduction to Sparse Stochastic Processes*, Cambridge University Press, Cambridge, UK, 2014.

[65] G. Wahba, *Spline Models for Observational Data*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 59, SIAM, Philadelphia, 1990, https://doi.org/10.1137/1.9781611970128.