Energy Decay and Conservation in Deep Convolutional Neural Networks

Philipp Grohs[1], Thomas Wiatowski[2], and Helmut Bölcskei[2]
[1]Dept. Math., University of Vienna, Austria
[2]Dept. IT & EE, ETH Zurich, Switzerland
philipp.grohs@univie.ac.at, withomas@nari.ee.ethz.ch, boelcskei@nari.ee.ethz.ch

***Abstract*—Many practical machine learning tasks employ very deep convolutional neural networks. Such large depths pose formidable computational challenges in training and operating the network. It is therefore important to understand how fast the energy contained in the propagated signals (a.k.a. feature maps) decays across layers. In addition, it is desirable that the feature extractor generated by the network be informative in the sense of the only signal mapping to the all-zeros feature vector being the zero input signal. This "trivial null-set" property can be accomplished by asking for "energy conservation" in the sense of the energy in the feature vector being proportional to that of the corresponding input signal. This paper establishes conditions for energy conservation (and thus for a trivial null-set) for a wide class of deep convolutional neural network-based feature extractors and characterizes corresponding feature map energy decay rates. Specifically, we consider general scattering networks employing the modulus non-linearity and we find that under mild analyticity and high-pass conditions on the filters (which encompass, inter alia, various constructions of Weyl-Heisenberg filters, wavelets, ridgelets, (α)-curvelets, and shearlets) the feature map energy decays at least polynomially fast. For broad families of wavelets and Weyl-Heisenberg filters, the guaranteed decay rate is shown to be exponential. Moreover, we provide handy estimates of the number of layers needed to have at least $((1-\varepsilon)\cdot 100)\%$ of the input signal energy be contained in the feature vector.**

I. Introduction

Feature extraction based on deep convolutional neural networks (DCNNs) has been applied with significant success in a wide range of practical machine learning tasks [1], [2]. Many of these applications, such as, e.g., the classification of images in the ImageNet data set, employ very deep networks with potentially hundreds of layers [3]. Such network depths entail formidable computational challenges in the training phase due to the large number of parameters to be learned, and in operating the network due to the large number of convolutions that need to be carried out. It is therefore paramount to understand how fast the energy contained in the signals generated in the individual network layers, a.k.a. feature maps, decays across layers. In addition, it is important that the feature vector (obtained by aggregating filtered versions of the feature maps) be informative in the sense of the only signal mapping to the all-zeros feature vector being the zero input signal. This "trivial null-set" property for the feature extractor can be obtained by asking for the energy in the feature vector being proportional to that of the corresponding input signal, a property we shall refer to as "energy conservation".

Scattering networks as introduced in [4] and extended in [5] constitute an important class of feature extractors based on nodes that implement convolutional transforms with pre-specified or learned filters in each network layer (e.g., wavelets [4], [6], uniform covering filters [7], or general filters [5]), followed by a non-linearity (e.g., the modulus [4], [6], [7], or a general Lipschitz non-linearity [5]), and a pooling operation (e.g., sub-sampling or average-pooling [5]). Scattering network-based feature extractors were shown to yield classification performance competitive with the state-of-the-art on various data sets [8]–[13]. Moreover, a mathematical theory exists, which allows to establish formally that such feature extractors are—under certain technical conditions—horizontally [4] or vertically [5] translation-invariant and deformation-stable in the sense of [4], or exhibit limited sensitivity to deformations in the sense of [5] on input signal classes such as band-limited functions [5], [14], cartoon functions [15], and Lipschitz functions [15].

It was shown recently that the energy in the feature maps generated by scattering networks employing, in every network layer, the same set of (certain) Parseval wavelets [6, Section 5] or "uniform covering" [7] filters (both satisfying analyticity and vanishing moments conditions), the modulus non-linearity, and no pooling, decays at least exponentially fast and "strict" energy conservation (which, in turn, implies a trivial null-set) for the infinite-depth feature vector holds. Specifically, the feature map energy decay was shown to be at least of order $\mathcal{O}(a^{-N})$, for some *unspecified* $a>1$, where N denotes the network depth. We note that d-dimensional uniform covering filters as introduced in [7] are functions whose Fourier transforms' support sets can be covered by a union of finitely many balls. This covering condition is satisfied by, e.g., Weyl-Heisenberg filters [16] with a band-limited prototype function, but fails to hold for multi-scale filters such as wavelets [17], [18], (α)-curvelets [19]–[21], shearlets [22], [23], or ridgelets [24]–[26], see [7, Remark 2.2 (b)].

Contributions. The first main contribution of this paper is a characterization of the feature map energy decay rate in DCNNs employing the modulus non-linearity, no pooling, and *general* filters that constitute a frame [17], [27]–[29], but not necessarily a Parseval frame, and are allowed to be different in different network layers. We find that, under mild analyticity and high-pass conditions on the filters, the energy decay rate is at least polynomial in the network depth, i.e., the decay is

at least of order $\mathcal{O}(N^{-\alpha})$, and we *explicitly* specify the decay exponent $\alpha > 0$. This result encompasses, inter alia, various constructions of Weyl-Heisenberg filters, wavelets, ridgelets, (α)-curvelets, shearlets, and learned filters (of course as long as the learning algorithm imposes the analyticity and high-pass conditions we require). For broad families of wavelets and Weyl-Heisenberg filters, the guaranteed energy decay rate is shown to be exponential in the network depth, i.e., the decay is at least of order $\mathcal{O}(a^{-N})$ with the decay factor given as $a = \frac{5}{3}$ in the wavelet case and $a = \frac{3}{2}$ in the Weyl-Heisenberg case. We hasten to add that our results constitute *guaranteed* decay rates and do not preclude the energy from decaying faster in practice.

Our second main contribution shows that the energy decay results above are compatible with a trivial null-set for finite- and infinite-depth networks. Specifically, this is accomplished by establishing energy proportionality between the feature vector and the underlying input signal with the proportionality constant lower- and upper-bounded by the frame bounds of the filters employed in the different layers.

Finally, for input signals that belong to the class of band-limited functions, our energy decay and conservation results are shown to yield handy estimates of the number of layers needed to have at least $((1-\varepsilon)\cdot 100)\%$ of the input signal energy be contained in the feature vector. For example, in the case of exponential energy decay with $a = \frac{5}{3}$ and for band-limited input signals, only 8 layers are needed to absorb 95% of the input signal's energy.

For proofs of the results in this paper and the general notation used we refer to [30].

II. DCNN-BASED FEATURE EXTRACTORS

Throughout the paper, we use the terminology of [5], consider (unless explicitly stated otherwise) input signals $f \in L^2(\mathbb{R}^d)$, and employ the module-sequence

$$\Omega := \big((\Psi_n, |\cdot|, \mathrm{Id})\big)_{n\in\mathbb{N}}, \tag{1}$$

i.e., each network layer is associated with (i) a collection of filters $\Psi_n := \{\chi_n\} \cup \{g_{\lambda_n}\}_{\lambda_n\in\Lambda_n} \subseteq L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$, where χ_n, referred to as output-generating filter, and the g_{λ_n}, indexed by a countable set Λ_n, satisfy the frame condition [17], [27], [29]

$$A_n\|f\|_2^2 \leq \|f*\chi_n\|_2^2 + \sum_{\lambda_n\in\Lambda_n} \|f*g_{\lambda_n}\|^2 \leq B_n\|f\|_2^2, \tag{2}$$

for all $f \in L^2(\mathbb{R}^d)$, for some $A_n, B_n > 0$, (ii) the modulus non-linearity $|\cdot| : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$, $|f|(x) := |f(x)|$, and (iii) no pooling, which, in the terminology of [5], corresponds to pooling through the identity operator with pooling factor equal to one. Associated with the module $(\Psi_n, |\cdot|, \mathrm{Id})$, the operator $U_n[\lambda_n]$ defined in [5, Eq. 12] particularizes to

$$U_n[\lambda_n]f = \big|f * g_{\lambda_n}\big|. \tag{3}$$

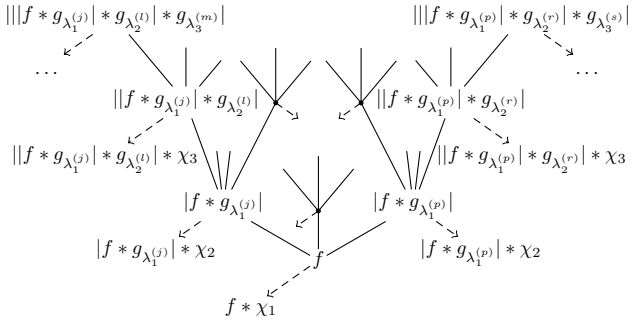

Fig. 1: Network architecture underlying the feature extractor (5). The index $\lambda_n^{(k)}$ corresponds to the filter $g_{\lambda^{(k)}}$ of the collection Ψ_n associated with the n-th network layer. The function χ_{n+1} is the output-generating filter of the n-th network layer. The root of the network corresponds to $n = 0$.

We extend (3) to paths on index sets $q = (\lambda_1, \lambda_2, \ldots, \lambda_n) \in \Lambda_1 \times \Lambda_2 \times \cdots \times \Lambda_n =: \Lambda^n$, $n \in \mathbb{N}$, according to

$$\begin{aligned} U[q]f &= U[(\lambda_1, \lambda_2, ..., \lambda_n)]f \\ &:= U_n[\lambda_n]\cdots U_2[\lambda_2]U_1[\lambda_1]f, \end{aligned} \tag{4}$$

where, for the empty path $e := \emptyset$, we set $\Lambda^0 := \{e\}$ and $U[e]f := f$. The signals $U[q]f$, $q \in \Lambda^n$, associated with the n-th network layer, are often referred to as feature maps in the deep learning literature. The feature vector $\Phi_\Omega(f)$ is obtained by aggregating filtered versions of the feature maps. More formally, $\Phi_\Omega(f)$ is defined as [5, Definition 3]

$$\Phi_\Omega(f) := \bigcup_{n=0}^{\infty} \Phi_\Omega^n(f), \tag{5}$$

where $\Phi_\Omega^n(f) := \{(U[q]f) * \chi_{n+1}\}_{q\in\Lambda^n}$ are the features generated in the n-th network layer, see Figure 1. Here, $n = 0$ corresponds to the root of the network. The function χ_{n+1} is the output-generating filter of the n-th network layer. The feature extractor

$$\Phi_\Omega : L^2(\mathbb{R}^d) \to \big(L^2(\mathbb{R}^d)\big)^{\bigcup_{n=0}^{\infty}\Lambda^n}$$

was shown in [5, Theorem 1] to be vertically translation-invariant, provided although that pooling is employed, with pooling factors $S_n \geq 1$, $n \in \mathbb{N}$, (see [5, Eq. 6] for the definition of the general pooling operator) such that $\lim_{N\to\infty}\prod_{n=1}^{N} S_n = \infty$. Moreover, Φ_Ω exhibits limited sensitivity to certain non-linear deformations on (input) signal classes such as band-limited functions [5, Theorem 2], cartoon functions [15, Theorem 1], and Lipschitz functions [15, Corollary 1].

III. ENERGY DECAY AND ENERGY CONSERVATION

The first central goal of this paper is to understand how fast the energy contained in the feature maps decays across layers. Specifically, we shall study the decay of

$$W_N(f) := \sum_{q\in\Lambda^N} \|U[q]f\|_2^2, \quad f \in L^2(\mathbb{R}^d), \tag{6}$$

as a function of network depth N. Moreover, it is desirable that the infinite-depth feature vector $\Phi_\Omega(f)$ be informative in the sense of the only signal mapping to the all-zeros feature vector being the zero input signal, i.e., Φ_Ω has a trivial null-set

$$\mathcal{N}(\Phi_\Omega) := \{f \in L^2(\mathbb{R}^d) \mid \Phi_\Omega(f) = 0\} \stackrel{!}{=} \{0\}. \tag{7}$$

$\mathcal{N}(\Phi_\Omega) = \{0\}$ can be guaranteed by asking for "energy conservation" in the sense of

$$A_\Omega \|f\|_2^2 \leq |||\Phi_\Omega(f)|||^2 \leq B_\Omega \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d), \tag{8}$$

for some constants $A_\Omega, B_\Omega > 0$ (possibly depending on the module-sequence Ω) and with the feature space norm $|||\Phi_\Omega(f)||| := \big(\sum_{n=0}^\infty |||\Phi_\Omega^n(f)|||^2\big)^{1/2}$, where $|||\Phi_\Omega^n(f)||| := \big(\sum_{q\in\Lambda^n} \|(U[q]f)*\chi_{n+1}\|_2^2\big)^{1/2}$. Indeed, (7) follows from (8) as the upper bound in (8) yields $\{0\} \subseteq \mathcal{N}(\Phi_\Omega)$, and the lower bound implies $\{0\} \supseteq \mathcal{N}(\Phi_\Omega)$. We emphasize that, as Φ_Ω is a non-linear operator (owing to the modulus non-linearities), characterizing its null-set is non-trivial in general. The upper bound in (8) was established in [5, Appendix E]. While the existence of this upper bound is implied by the filters Ψ_n, $n \in \mathbb{N}$, satisfying the frame property (2) [5, Appendix E], perhaps surprisingly, this is not enough to guarantee $A_\Omega > 0$ (see [30, Appendix A] for an example). We refer the reader to Section V for results on the null-set of the *finite-depth* feature extractor $\bigcup_{n=0}^N \Phi_\Omega^n$.

Previous work on the decay rate of $W_N(f)$ in [6, Section 5] shows that for wavelet-based networks (i.e., in every network layer the filters $\Psi = \{\chi\} \cup \{g_\lambda\}_{\lambda\in\Lambda}$ in (1) are taken to be (specific) 1-D wavelets that constitute a Parseval frame, with χ a low-pass filter) there exist $\varepsilon > 0$ and $a > 1$ (both constants unspecified) such that

$$W_N(f) \leq \int_{\mathbb{R}} |\widehat{f}(\omega)|^2 \Big(1 - \Big|\widehat{r_g}\Big(\frac{\omega}{\varepsilon a^{N-1}}\Big)\Big|^2\Big)\mathrm{d}\omega, \tag{9}$$

for real-valued 1-D signals $f \in L^2(\mathbb{R})$ and $N \geq 2$, where $\widehat{r_g}(\omega) := e^{-\omega^2}$. For scattering networks that employ, in every network layer, uniform covering filters $\Psi = \{\chi\} \cup \{g_\lambda\}_{\lambda\in\Lambda} \subseteq L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ forming a Parseval frame (where χ, again, is a low-pass filter), exponential energy decay according to

$$W_N(f) = \mathcal{O}(a^{-N}), \quad \forall f \in L^2(\mathbb{R}^d), \tag{10}$$

for an unspecified $a > 1$, was established in [7, Proposition 3.3]. Moreover, [6, Section 5] and [7, Theorem 3.6 (a)] state—for the respective module-sequences—that (8) holds with $A_\Omega = B_\Omega = 1$ and hence

$$|||\Phi_\Omega(f)|||^2 = \|f\|_2^2. \tag{11}$$

The first main goal of the present paper is to establish i) for d-dimensional complex-valued input signals that $W_N(f)$ decays polynomially according to

$$W_N(f) \leq B_\Omega^N \int_{\mathbb{R}^d} |\widehat{f}(\omega)|^2 \Big(1 - \Big|\widehat{r_l}\Big(\frac{\omega}{N^\alpha}\Big)\Big|^2\Big)\mathrm{d}\omega, \tag{12}$$

for $f \in L^2(\mathbb{R}^d)$ and $N \geq 1$, where $\alpha = 1$, for $d = 1$, and $\alpha = \log_2(\sqrt{d/(d-1/2)})$, for $d \geq 2$, $B_\Omega^N = \prod_{k=1}^N \max\{1, B_k\}$, and $\widehat{r_l} : \mathbb{R}^d \to \mathbb{R}$, $\widehat{r_l}(\omega) = (1 - |\omega|)_+^l$, with $l > \lfloor d/2 \rfloor + 1$, for networks based on general filters $\{\chi_n\} \cup \{g_{\lambda_n}\}_{\lambda_n\in\Lambda_n}$ that satisfy mild analyticity and high-pass conditions and are allowed to be different in different network layers (with the proviso that χ_n, $n \in \mathbb{N}$, is of low-pass nature in a sense to be made precise), and ii) for 1-D complex-valued input signals that (6) decays exponentially according to

$$W_N(f) \leq \int_{\mathbb{R}} |\widehat{f}(\omega)|^2 \Big(1 - \Big|\widehat{r_l}\Big(\frac{\omega}{a^{N-1}}\Big)\Big|^2\Big)\mathrm{d}\omega, \tag{13}$$

for $f \in L^2(\mathbb{R})$ and $N \geq 1$, for networks that are based, in every network layer, on a broad family of wavelets, with the decay factor given explicitly as $a = \frac{5}{3}$, or on a broad family of Weyl-Heisenberg filters [5, Appendix B], with decay factor $a = \frac{3}{2}$. Thanks to the right-hand side (RHS) of (12) and (13) not depending on the specific filters $\{\chi_n\} \cup \{g_{\lambda_n}\}_{\lambda_n\in\Lambda_n}$, we will be able to establish—under smoothness assumptions on the input signal f—universal energy decay results. Specifically, particularizing the RHS expressions in (12) and (13) to Sobolev-class input signals $f \in H^s(\mathbb{R}^d)$, $s > 0$, where

$$H^s(\mathbb{R}^d) = \Big\{f \in L^2(\mathbb{R}^d) \;\Big|\; \|f\|_{H^s} < \infty\Big\},$$

with $\|f\|_{H^s} := \big(\int_{\mathbb{R}^d} (1 + |\omega|^2)^s |\widehat{f}(\omega)|^2 \mathrm{d}\omega\big)^{1/2}$, we show that (12) yields polynomial energy decay according to

$$W_N(f) = \mathcal{O}\big(N^{-\gamma\alpha}\big), \tag{14}$$

and (13) exponential energy decay

$$W_N(f) = \mathcal{O}\big(a^{-\gamma N}\big), \tag{15}$$

where $\gamma := \min\{1, 2s\}$ in both cases.

Our second central goal is to prove energy conservation according to (8) (which, as explained above, implies $\mathcal{N}(\Phi_\Omega) = \{0\}$) for the network configurations corresponding to the energy decay results (12) and (13). Finally, we provide handy estimates of the number of layers needed to have at least $((1-\varepsilon)\cdot 100)\%$ of the input signal energy be contained in the feature vector.

IV. Main results

Throughout the paper, we make the following assumptions on the filters $\{g_{\lambda_n}\}_{\lambda_n\in\Lambda_n}$.

Assumption 1. *The $\{g_{\lambda_n}\}_{\lambda_n\in\Lambda_n}$, $n \in \mathbb{N}$, are analytic in the following sense: For every layer index $n \in \mathbb{N}$, for every $\lambda_n \in \Lambda_n$, there exists an orthant $H_{A_{\lambda_n}} \subseteq \mathbb{R}^d$, with $A_{\lambda_n} \in O(d)$, such that*

$$\mathit{supp}(\widehat{g_{\lambda_n}}) \subseteq H_{A_{\lambda_n}}. \tag{16}$$

Moreover, there exists $\delta > 0$ so that

$$\sum_{\lambda_n\in\Lambda_n} |\widehat{g_{\lambda_n}}(\omega)|^2 = 0, \quad \textit{a.e. } \omega \in B_\delta(0). \tag{17}$$

In the 1-D case, i.e., for $d = 1$, Assumption 1 simply amounts to every filter g_{λ_n} satisfying

$$\text{either} \quad \mathrm{supp}(\widehat{g_{\lambda_n}}) \subseteq (-\infty, -\delta] \quad \text{or} \quad \mathrm{supp}(\widehat{g_{\lambda_n}}) \subseteq [\delta, \infty),$$

which constitutes an "analyticity" and "high-pass" condition. For dimensions $d \geq 2$, Assumption 1 requires that every filter g_{λ_n} be of high-pass nature and have a Fourier transform supported in a (not necessarily canonical) orthant. Since the frame condition (2) is equivalent to the Littlewood-Paley condition [31]

$$A_n \leq |\widehat{\chi_n}(\omega)|^2 + \sum_{\lambda_n \in \Lambda_n} |\widehat{g_{\lambda_n}}(\omega)|^2 \leq B_n, \quad \text{a.e. } \omega \in \mathbb{R}^d, \tag{18}$$

(17) implies low-pass characteristics for χ_n to fill the spectral gap $B_\delta(0)$ left by the filters $\{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$.

The conditions (16) and (17) we impose on the Ψ_n, $n \in \mathbb{N}$, are not overly restrictive as they encompass, inter alia, various constructions of Weyl-Heisenberg filters (e.g., a 1-D B-spline as prototype function [32, Section 1]), wavelets (e.g., analytic Meyer wavelets [17, Section 3.3.5] in 1-D, and Cauchy wavelets [33] in 2-D), and specific constructions of ridgelets [26, Section 2.2], curvelets [20, Section 4.1], α-curvelets [21, Section 3], and shearlets (e.g., cone-adapted shearlets [34, Section 4.3]). We refer the reader to [5, Appendices B and C] for a brief review of some of these filter structures.

We are now ready to state our main result on energy decay and energy conservation.

Theorem 1. *Let Ω be the module-sequence* (1) *with filters $\{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$ satisfying the conditions in Assumption 1, and let $\delta > 0$ be the radius of the spectral gap $B_\delta(0)$ left by the filters $\{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$ according to* (17)*. Furthermore, let $s \geq 0$, $A_\Omega^N := \prod_{k=1}^N \min\{1, A_k\}$, $B_\Omega^N := \prod_{k=1}^N \max\{1, B_k\}$, and*

$$\alpha := \begin{cases} 1, & d = 1, \\ \log_2(\sqrt{d/(d-1/2)}), & d \geq 2. \end{cases} \tag{19}$$

i) We have

$$W_N(f) \leq B_\Omega^N \int_{\mathbb{R}^d} |\widehat{f}(\omega)|^2 \Big(1 - \Big|\widehat{r_l}\Big(\frac{\omega}{N^\alpha \delta}\Big)\Big|^2\Big) \mathrm{d}\omega, \tag{20}$$

for $f \in L^2(\mathbb{R}^d)$ and $N \geq 1$, where $\widehat{r_l} : \mathbb{R}^d \to \mathbb{R}$, $\widehat{r_l}(\omega) := (1 - |\omega|)_+^l$, with $l > \lfloor d/2 \rfloor + 1$.

ii) For every Sobolev function $f \in H^s(\mathbb{R}^d)$, $s > 0$, we have

$$W_N(f) = \mathcal{O}\big(B_\Omega^N N^{-\gamma\alpha}\big), \tag{21}$$

where $\gamma := \min\{1, 2s\}$.

iii) If, in addition to Assumption 1,

$$0 < A_\Omega := \lim_{N\to\infty} A_\Omega^N \leq B_\Omega := \lim_{N\to\infty} B_\Omega^N < \infty, \tag{22}$$

then we have energy conservation according to

$$A_\Omega \|f\|_2^2 \leq |||\Phi_\Omega(f)|||^2 \leq B_\Omega \|f\|_2^2, \tag{23}$$

for all $f \in L^2(\mathbb{R}^d)$.

The strength of the results in Theorem 1 derives itself from the fact that the only condition we need to impose on the filters Ψ_n is Assumption 1, which, as already mentioned, is met by a wide array of filters. Moreover, condition (22) is easily satisfied by normalizing the filters Ψ_n, $n \in \mathbb{N}$, appropriately (see, e.g., [5, Proposition 3]). We note that this normalization, when applied to filters that satisfy Assumption 1, yields filters that still meet Assumption 1.

The identity (21) establishes, upon normalization [5, Proposition 3] of the Ψ_n to get $B_n \leq 1$, $n \in \mathbb{N}$, that the energy decay rate, i.e., the decay rate of $W_N(f)$, is at least polynomial in N. We hasten to add that (20) does not preclude the energy from decaying faster in practice.

The next result shows that, under additional structural assumptions on the filters $\{g_{\lambda_n}\}_{\lambda_n \in \Lambda}$, the guaranteed energy decay rate can be improved from polynomial to exponential. Specifically, we can get exponential energy decay for broad families of wavelets and Weyl-Heisenberg filters. For conceptual reasons, we consider the 1-D case and, for simplicity of exposition, we employ filters that constitute Parseval frames and are identical across network layers.

Theorem 2. *Let $\widehat{r_l} : \mathbb{R} \to \mathbb{R}$, $\widehat{r_l}(\omega) := (1 - |\omega|)_+^l$, with $l > 1$.*

i) Wavelets: Let the mother and father wavelets $\psi, \phi \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ satisfy supp$(\widehat{\psi}) \subseteq [1/2, 2]$ and

$$|\widehat{\phi}(\omega)|^2 + \sum_{j=1}^\infty |\widehat{\psi}(2^{-j}\omega)|^2 = 1, \quad a.e.\ \omega \geq 0. \tag{24}$$

Moreover, let $g_j(x) := 2^j \psi(2^j x)$, for $x \in \mathbb{R}$, $j \geq 1$, and $g_j(x) := 2^{|j|}\psi(-2^{|j|}x)$, for $x \in \mathbb{R}$, $j \leq -1$, and set $\chi(x) := \phi(x)$, for $x \in \mathbb{R}$. Let Ω be the module-sequence (1) *with filters $\Psi = \{\chi\} \cup \{g_j\}_{j \in \mathbb{Z}\setminus\{0\}}$ in every network layer. Then,*

$$W_N(f) \leq \int_{\mathbb{R}} |\widehat{f}(\omega)|^2 \Big(1 - \Big|\widehat{r_l}\Big(\frac{\omega}{(5/3)^{N-1}}\Big)\Big|^2\Big) \mathrm{d}\omega, \tag{25}$$

for $f \in L^2(\mathbb{R})$ and $N \geq 1$. Moreover, for every Sobolev function $f \in H^s(\mathbb{R})$, $s > 0$, we have

$$W_N(f) = \mathcal{O}\big((5/3)^{-\gamma N}\big), \tag{26}$$

where $\gamma := \min\{1, 2s\}$.

ii) Weyl-Heisenberg filters: For $R \in \mathbb{R}$, let the functions $g, \phi \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ satisfy supp$(\widehat{g}) \subseteq [-R, R]$, $\widehat{g}(-\omega) = \widehat{g}(\omega)$, for $\omega \in \mathbb{R}$, and

$$|\widehat{\phi}(\omega)|^2 + \sum_{k=1}^\infty |\widehat{g}(\omega - R(k+1))|^2 = 1, \tag{27}$$

a.e. $\omega \geq 0$. Moreover, let $g_k(x) := e^{2\pi i(k+1)Rx} g(x)$, for $x \in \mathbb{R}$, $k \geq 1$, and $g_k(x) := e^{-2\pi i(|k|+1)Rx} g(x)$, for $x \in \mathbb{R}$, $k \leq -1$, and set $\chi(x) := \phi(x)$, for $x \in \mathbb{R}$. Let Ω be the module-sequence (1) *with filters $\Psi = \{\chi\} \cup \{g_k\}_{k \in \mathbb{Z}\setminus\{0\}}$ in every network layer. Then,*

$$W_N(f) \leq \int_{\mathbb{R}} |\widehat{f}(\omega)|^2 \Big(1 - \Big|\widehat{r_l}\Big(\frac{\omega}{(3/2)^{N-1}R}\Big)\Big|^2\Big) \mathrm{d}\omega, \tag{28}$$

for $f \in L^2(\mathbb{R})$ and $N \geq 1$. Moreover, for every Sobolev function $f \in H^s(\mathbb{R})$, $s > 0$, we have

$$W_N(f) = \mathcal{O}\big((3/2)^{-\gamma N}\big), \tag{29}$$

where $\gamma := \min\{1, 2s\}$.

The conditions we impose on the mother and father wavelet ψ, ϕ in i) are satisfied, e.g., by analytic Meyer wavelets

[17, Section 3.3.5], and those on the prototype function g and low-pass filter ϕ in ii) by B-splines [32, Section 1]. Moreover, as shown in [35, Theorem 3.1], the exponential energy decay results in (26) and (29) can be generalized to $\mathcal{O}(a^{-N})$ with *arbitrary* decay factor $a > 1$ realized through suitable choice of the mother wavelet or the Weyl-Heisenberg prototype function.

For a detailed discussion on the relation between the results in Theorems 1 and 2 and related results in the literature [6], [7], [36], we refer the reader to [30, Section IV].

V. Number of layers needed

DCNNs used in practice employ potentially hundreds of layers [3]. Such network depths entail formidable computational challenges both in training and in operating the network. It is therefore important to understand how many layers are needed to have most of the input signal energy be contained in the feature vector. This will be done by considering Parseval frames in all layers, i.e., frames with frame bounds $A_n = B_n = 1$, $n \in \mathbb{N}$, and by asking for bounds of the form

$$(1-\varepsilon) \leq \frac{\sum_{n=0}^{N} |||\Phi_\Omega^n(f)|||^2}{\|f\|_2^2} \leq 1, \tag{30}$$

i.e., by determining the network depth N guaranteeing that at least $((1-\varepsilon)\cdot 100)\%$ of the input signal energy are captured by the corresponding depth-N feature vector $\bigcup_{n=0}^{N} \Phi_\Omega^n(f)$. Moreover, (30) ensures that the depth-N feature extractor $\bigcup_{n=0}^{N} \Phi_\Omega^n$ exhibits a trivial null-set.

The following results establish handy estimates of the number N of layers needed to guarantee (30).

Corollary 1.

i) Let Ω be the module-sequence (1) *with filters $\{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$ satisfying the conditions in Assumption 1, and let the corresponding frame bounds be $A_n = B_n = 1$, $n \in \mathbb{N}$. Let $\delta > 0$ be the radius of the spectral gap $B_\delta(0)$ left by the filters $\{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$ according to* (17)*. Furthermore, let $l > \lfloor d/2 \rfloor + 1$, $\varepsilon \in (0,1)$, α as defined in* (19)*, and $f \in L^2(\mathbb{R}^d)$ L-band-limited. If*

$$N \geq \left\lceil \left(\frac{L}{(1-(1-\varepsilon)^{\frac{1}{2l}})\delta} \right)^{1/\alpha} - 1 \right\rceil, \tag{31}$$

then (30) *holds.*

ii) Assume that the conditions in Theorem 2 i) and ii) hold. For the wavelet case, let $a = \frac{5}{3}$ and $\delta = 1$ (where δ corresponds to the radius of the spectral gap left by the wavelets $\{g_j\}_{j \in \mathbb{Z}\setminus\{0\}}$). For the Weyl-Heisenberg case, let $a = \frac{3}{2}$ and $\delta = R$ (here, δ corresponds to the radius of the spectral gap left by the Weyl-Heisenberg filters $\{g_k\}_{k \in \mathbb{Z}\setminus\{0\}}$). Moreover, let $l > 1$, $\varepsilon \in (0,1)$, and $f \in L^2(\mathbb{R})$ L-band-limited. If

$$N \geq \left\lceil \log_a \left(\frac{L}{(1-(1-\varepsilon)^{\frac{1}{2l}})\delta} \right) \right\rceil, \tag{32}$$

then (30) *holds in both cases.*

	$(1-\varepsilon)$					
	0.25	0.5	0.75	0.9	0.95	0.99
wavelets	2	3	4	6	8	11
Weyl-Heisenberg filters	2	4	5	8	10	14
general filters	2	3	7	19	39	199

Table I: Number N of layers needed to ensure that $((1-\varepsilon)\cdot 100)\%$ of the input signal energy are contained in the features generated in the first N network layers.

Corollary 1 nicely shows how the description complexity of the signal class under consideration, namely the bandwidth L and the dimension d through the decay exponent α defined in (19) determine the number N of layers needed. Specifically, (31) and (32) show that larger bandwidths L and larger dimension d render the input signal f more "complex", which requires deeper networks to capture most of the energy of f. The dependence of the lower bounds in (31) and (32) on the network properties, through the module-sequence Ω, is through the decay factor $a > 1$ and the radius δ of the spectral gap left by the filters $\{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$.

The following numerical example provides quantitative insights on the influence of the parameter ε on (31) and (32). Specifically, we set $L = 1$, $\delta = 1$, $d = 1$ (which implies $\alpha = 1$, see (19)), $l = 1.0001$, and show in Table I the number N of layers needed according to (31) and (32) for different values of ε. The results show that 95% of the input signal energy are contained in the first 8 layers in the wavelet case and in the first 10 layers in the Weyl-Heisenberg case. We can therefore conclude that in practice a relatively small number of layers is needed to have most of the input signal energy be contained in the feature vector. In contrast, for general filters, where we can guarantee polynomial energy decay only, $N = 39$ layers are needed to absorb 95% of the input signal energy. We hasten to add, however, that (20) simply *guarantees* polynomial energy decay and does not preclude the energy from decaying faster in practice.

References

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

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 770–778.

[4] S. Mallat, "Group invariant scattering," *Comm. Pure Appl. Math.*, vol. 65, no. 10, pp. 1331–1398, 2012.

[5] T. Wiatowski and H. Bölcskei, "A mathematical theory of deep convolutional neural networks for feature extraction," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1845–1866, 2018.

[6] I. Waldspurger, "Wavelet transform modulus: Phase retrieval and scattering," Ph.D. dissertation, École Normale Supérieure, Paris, 2015.

[7] W. Czaja and W. Li, "Analysis of time-frequency scattering transforms," *arXiv:1606.08677*, 2017.

[8] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1872–1886, 2013.

[9] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4114–4128, 2014.

[10] L. Sifre, "Rigid-motion scattering for texture classification," Ph.D. dissertation, Centre de Mathématiques Appliquées, École Polytechnique Paris-Saclay, 2014.

[11] T. Wiatowski, M. Tschannen, A. Stanić, P. Grohs, and H. Bölcskei, "Discrete deep feature extraction: A theory and new architectures," in *Proc. of International Conference on Machine Learning (ICML)*, 2016, pp. 2149–2158.

[12] M. Tschannen, T. Kramer, G. Marti, M. Heinzmann, and T. Wiatowski, "Heart sound classification using deep structured features," in *Proc. of Computing in Cardiology (CinC)*, 2016, pp. 565–568.

[13] M. Tschannen, L. Cavigelli, F. Mentzer, T. Wiatowski, and L. Benini, "Deep structured features for semantic segmentation," *Proc. of European Signal Processing Conference (EUSIPCO)*, pp. 61–65, 2017.

[14] R. Balan, M. Singh, and D. Zou, "Lipschitz properties for deep convolutional networks," *arXiv:1701.05217*, 2017.

[15] P. Grohs, T. Wiatowski, and H. Bölcskei, "Deep convolutional neural networks on cartoon functions," in *Proc. of IEEE International Symposium on Information Theory (ISIT)*, 2016, pp. 1163–1167.

[16] K. Gröchenig, *Foundations of time-frequency analysis*. Birkhäuser, 2001.

[17] I. Daubechies, *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, 1992.

[18] S. Mallat, *A wavelet tour of signal processing: The sparse way*, 3rd ed. Academic Press, 2009.

[19] E. J. Candès and D. L. Donoho, "New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities," *Comm. Pure Appl. Math.*, vol. 57, no. 2, pp. 219–266, 2004.

[20] ——, "Continuous curvelet transform: II. Discretization and frames," *Appl. Comput. Harmon. Anal.*, vol. 19, no. 2, pp. 198–222, 2005.

[21] P. Grohs, S. Keiper, G. Kutyniok, and M. Schäfer, "Cartoon approximation with α-curvelets," *J. Fourier Anal. Appl.*, vol. 22, no. 6, pp. 1235–1293, 2016.

[22] K. Guo, G. Kutyniok, and D. Labate, "Sparse multidimensional representations using anisotropic dilation and shear operators," in *Wavelets and Splines*, G. Chen and M. J. Lai, Eds. Nashboro Press, 2006, pp. 189–201.

[23] G. Kutyniok and D. Labate, Eds., *Shearlets: Multiscale analysis for multivariate data*. Birkhäuser, 2012.

[24] E. J. Candès, "Ridgelets: Theory and applications," Ph.D. dissertation, Stanford University, 1998.

[25] E. J. Candès and D. L. Donoho, "Ridgelets: A key to higher-dimensional intermittency?" *Philos. Trans. R. Soc. London Ser. A*, vol. 357, no. 1760, pp. 2495–2509, 1999.

[26] P. Grohs, "Ridgelet-type frame decompositions for Sobolev spaces related to linear transport," *J. Fourier Anal. Appl.*, vol. 18, no. 2, pp. 309–325, 2012.

[27] S. T. Ali, J. P. Antoine, and J. P. Gazeau, "Continuous frames in Hilbert spaces," *Annals of Physics*, vol. 222, no. 1, pp. 1–37, 1993.

[28] O. Christensen, *An introduction to frames and Riesz bases*. Birkhäuser, 2003.

[29] G. Kaiser, *A friendly guide to wavelets*. Birkhäuser, 1994.

[30] T. Wiatowski, P. Grohs, and H. Bölcskei, "Energy propagation in deep convolutional neural networks," *IEEE Trans. Inf. Theory*, 2018, to appear.

[31] M. Frazier, B. Jawerth, and G. Weiss, *Littlewood-Paley theory and the study of function spaces*. American Mathematical Society, 1991.

[32] K. Gröchenig, A. J. E. M. Janssen, N. Kaiblinger, and G. E. Pfander, "Note on B-Splines, wavelet scaling functions, and Gabor frames," *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3318–3320, 2003.

[33] P. Vandergheynst, "Directional dyadic wavelet transforms: Design and algorithms," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 363–372, 2002.

[34] G. Kutyniok and D. Labate, "Introduction to shearlets," in *Shearlets: Multiscale analysis for multivariate data*, G. Kutyniok and D. Labate, Eds. Birkhäuser, 2012, pp. 1–38.

[35] T. Wiatowski, P. Grohs, and H. Bölcskei, "Topology reduction in deep convolutional feature extraction networks," *Proc. of SPIE (Wavelets and Sparsity XVII)*, vol. 10394, pp. 1 039 418:1–1 039 418:12, 2017.

[36] I. Waldspurger, "Exponential decay of scattering coefficients," *Proc. of International Conference on Sampling Theory and Applications (SampTA)*, pp. 143–146, 2017.