Topology Reduction in Deep Convolutional Feature Extraction Networks

Thomas Wiatowski[a], Philipp Grohs[b], and Helmut Bölcskei[a]

[a]Department of Information Technology and Electrical Engineering, ETH Zurich, Sternwartstrasse 7, 8092 Zurich, Switzerland
[b]Faculty of Mathematics, University of Vienna, Oskar Morgenstern Platz 1, 1090 Vienna, Austria

ABSTRACT

Deep convolutional neural networks (CNNs) used in practice employ potentially hundreds of layers and 10,000s of nodes. Such network sizes entail significant computational complexity due to the large number of convolutions that need to be carried out; in addition, a large number of parameters needs to be learned and stored. Very deep and wide CNNs may therefore not be well suited to applications operating under severe resource constraints as is the case, e.g., in low-power embedded and mobile platforms. This paper aims at understanding the impact of CNN topology, specifically depth and width, on the network's feature extraction capabilities. We address this question for the class of scattering networks that employ either Weyl-Heisenberg filters or wavelets, the modulus non-linearity, and no pooling. The exponential feature map energy decay results in Wiatowski et al., 2017, are generalized to $\mathcal{O}(a^{-N})$, where an *arbitrary* decay factor $a > 1$ can be realized through suitable choice of the Weyl-Heisenberg prototype function or the mother wavelet. We then show how networks of fixed (possibly small) depth N can be designed to guarantee that $((1-\varepsilon)\cdot 100)\%$ of the input signal's energy are contained in the feature vector. Based on the notion of operationally significant nodes, we characterize, partly rigorously and partly heuristically, the topology-reducing effects of (effectively) band-limited input signals, band-limited filters, and feature map symmetries. Finally, for networks based on Weyl-Heisenberg filters, we determine the prototype function bandwidth that minimizes—for fixed network depth N—the average number of operationally significant nodes per layer.

Keywords: Machine learning, deep convolutional neural networks, scattering networks, feature extraction, wavelets, Weyl-Heisenberg frames

1. INTRODUCTION

Feature extraction based on deep convolutional neural networks (CNNs) has been applied with significant success in a wide range of practical machine learning tasks [1]. 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 and 10,000s of nodes [2, 3] (e.g., the CNN in [2] has a depth of 152 with an average number of 472 nodes per layer). Such network sizes entail formidable computational challenges, both in the training phase due to the large number of parameters to be learned (e.g., the CNN in [3] has 144 million parameters), and in operating the network due to the large number of convolutions that need to be carried out (e.g., the CNN in [2] entails 11.3 billion FLOPS to pass a single image through the network). Moreover, storing the learned network parameters requires large amounts of memory. Very deep and wide CNNs may therefore not be suited to applications operating under strong resource constraints as is the case, e.g., in low-power embedded and mobile platforms [4]. It is hence important to understand the impact of CNN topology, specifically depth and width, on the network's feature extraction capabilities.

Further author information: (Send correspondence to T.W.)
T.W.: E-mail: withomas@nari.ee.ethz.ch, Telephone: +41 44 63 22804
P.G.: E-mail: philipp.grohs@univie.ac.at, Telephone: +43 1 4277 55741
H.B.: E-mail: boelcskei@nari.ee.ethz.ch, Telephone: +41 44 63 23433

We address this question for the class of scattering networks as introduced in [5] and extended in [6]. Scattering network-based feature extractors were shown to yield classification performance competitive with the state-of-the-art on various data sets [7–9]. Moreover, a mathematical theory exists, which allows to establish formally that such feature extractors are—under certain technical conditions—horizontally [5] or vertically [6] translation-invariant, energy-conserving [10–12], deformation-stable in the sense of [5] or exhibit limited sensitivity to deformations on input signal classes such as band-limited functions [6], cartoon functions [13], and Lipschitz functions [13].

Estimates of the number N of layers (i.e., the network depth) needed to have $((1-\varepsilon)\cdot 100)\%$ of the input signal energy be contained in the feature vector—obtained by aggregating filtered versions of the propagated signals (a.k.a. feature maps)—were recently obtained in [12]. The results in [12] apply to scattering networks employing the modulus non-linearity, no pooling, and general filters that are analytic, constitute Parseval frames [14], and are allowed to be different in different network layers. The main findings of [12] state that the feature map energy decays at least as fast as i) $\mathcal{O}(N^{-\alpha})$, for an explicitly specified $\alpha > 0$, for general filters, ii) $\mathcal{O}((3/2)^{-N})$ for broad families of Weyl-Heisenberg (WH) filters, and iii) $\mathcal{O}((5/3)^{-N})$ for broad families of wavelet filters.

Contributions. For scattering networks that employ the modulus non-linearity and no pooling, we generalize the exponential energy decay results in [12] to $\mathcal{O}(a^{-N})$, where an *arbitrary* decay factor $a > 1$ can be realized by suitable choice of the WH prototype function or the mother wavelet. We then show how networks of fixed (possibly small) depth N can be designed to guarantee that $((1-\varepsilon)\cdot 100)\%$ of the input signal's energy are contained in the feature vector. Based on the notion of operationally significant nodes, we characterize, partly rigorously and partly heuristically, the topology-reducing effects of (effectively) band-limited input signals, band-limited filters, and feature map symmetries. The results we obtain suggest a classification into shallow, single-layer, constant-width, expanding-width, depth-pruned, and extremely narrow scattering networks. Finally, for networks based on WH filters, we determine the prototype function bandwidth that minimizes—for a fixed network depth N—the average number of operationally significant nodes per layer.

2. CNN-BASED FEATURE EXTRACTORS

For the general notation employed in this paper, we refer to [12, Section 1]. We set the stage by briefly reviewing the basics of scattering network-based feature extractors. The presentation follows closely that in [12, Section 2]. Throughout the paper, we focus on the 1-D case and employ the module sequence

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

i.e., each network layer is associated with (i) the same collection of filters $\Psi = \{\chi\} \cup \{g_\lambda\}_{\lambda\in\Lambda} \subseteq L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, where χ, referred to as output-generating filter, and the g_λ, indexed by a countable set Λ, satisfy the Parseval frame condition [14]

$$\|f * \chi\|_2^2 + \sum_{\lambda\in\Lambda} \|f * g_\lambda\|^2 = \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}), \tag{1}$$

(ii) the modulus non-linearity $|\cdot| : L^2(\mathbb{R}) \to L^2(\mathbb{R})$, $|f|(x) := |f(x)|$, and (iii) no pooling, which, in the terminology of [6], corresponds to pooling through the identity operator with pooling factor equal to one. Associated with the module $(\Psi, |\cdot|, \mathrm{Id})$, the operator $U[\lambda]$ defined in [6, Eq. 12] particularizes to

$$U[\lambda]f = \big|f * g_\lambda\big|. \tag{2}$$

We extend (2) to paths on index sets

$$q = (\lambda_1, \lambda_2, \ldots, \lambda_n) \in \underbrace{\Lambda \times \Lambda \times \cdots \times \Lambda}_{n \text{ times}} =: \Lambda^n, \quad n \in \mathbb{N},$$

according to $U[q]f = U[(\lambda_1, \lambda_2, \ldots, \lambda_n)]f := U[\lambda_n] \cdots U[\lambda_2]U[\lambda_1]f$, where, for the empty path $e := \emptyset$, we set $\Lambda^0 := \{e\}$ and $U[e]f := f$, for $f \in L^2(\mathbb{R})$. The signals $U[q]f$ are often referred to as feature maps in the deep

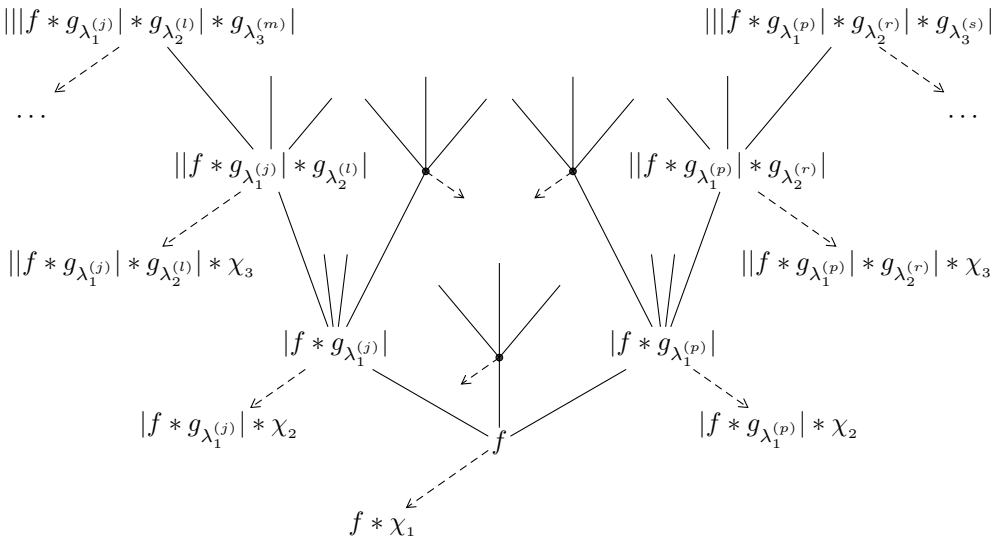

Figure 1: Network architecture underlying the feature extractor (3). The index $\lambda_n^{(k)}$ corresponds to the k-th filter $g_{\lambda^{(k)}}$ of the collection Ψ_n associated with the n-th network layer. The function χ_{n+1} is the output-generating filter of the n-th network layer. The root of the network corresponds to $n = 0$.

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 [6, Def. 3]

$$\Phi_\Omega(f) := \bigcup_{n=0}^{\infty} \Phi_\Omega^n(f), \tag{3}$$

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* Φ_Ω was shown in [6, 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 [6, 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 [6, Theorem 2], cartoon functions [13, Theorem 1], and Lipschitz functions [13, Corollary 1]. More recently, it was shown in [12, Theorem 1] that Φ_Ω is energy-conserving in the sense of the energy contained in the feature vector $\Phi_\Omega(f)$ being proportional to that of the corresponding input signal f.

3. FEATURE MAP ENERGY DECAY

The total energy contained in the feature maps in the n-th network layer is given by

$$W_n(f) := \sum_{q\in\Lambda^n} \|U[q]f\|_2^2, \quad f \in L^2(\mathbb{R}).$$

Our goal is to construct WH and wavelet filters that realize exponential energy decay according to $W_n(f) = \mathcal{O}(a^{-n})$, with arbitrary $a > 1$. In particular, we want to tune the decay factor a by adjusting a single parameter, which will be seen to determine the WH prototype function or the mother wavelet bandwidth. This will be accomplished through the following constructions:

*Throughout, we refer to Φ_Ω as *feature extractor* and to $\Phi_\Omega(f)$ as *feature vector* corresponding to the input signal f.

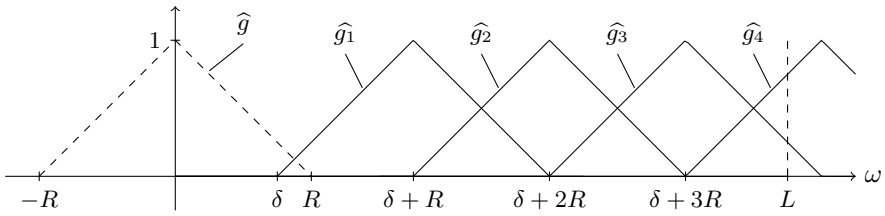

Figure 2: Illustration of the Fourier transforms of the WH filters g_k on the frequency band $[0, L]$. The Fourier transform $\widehat{g}$ of the prototype function g is supported on the interval $[-R, R]$.

i) *WH filters*: For fixed $R > 0$, $\delta \geq \frac{R}{2}$, let the functions $g, \phi \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ satisfy the Littlewood-Paley condition

$$|\widehat{\phi}(\omega)|^2 + \sum_{k=1}^{\infty} |\widehat{g}(\omega - (Rk + \delta))|^2 = 1, \quad a.e.\ \omega \geq 0,$$

with $\mathrm{supp}(\widehat{g}) = [-R, R]$, $\widehat{g}(-\omega) = \widehat{g}(\omega)$, and $\widehat{g}$ real-valued. Moreover, let $g_k(x) := e^{2\pi i(Rk+\delta)x} g(x)$, $k \geq 1$, $g_k(x) := e^{-2\pi i(R|k|+\delta)x} g(x)$, $k \leq -1$, and set $\chi(x) := \phi(x)$, $x \in \mathbb{R}$. The Fourier transforms $\widehat{g_k}$ and $\widehat{g}$ are illustrated in Figure 2.

ii) *Wavelets*: For fixed $r > 1$, let the mother and father wavelets $\psi, \phi \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ satisfy the Littlewood-Paley condition [15]

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

with $\mathrm{supp}(\widehat{\psi}) = [r^{-1}, r]$ and $\widehat{\psi}$ real-valued. Moreover, let $g_j(x) := r^j \psi(r^j x)$, $j \geq 1$, $g_j(x) := r^{|j|} \psi(-r^{|j|} x)$, $j \leq -1$, and let the output-generating filter be $\chi(x) := \phi(x)$, $x \in \mathbb{R}$. The Fourier transforms of the wavelets g_j and the mother wavelet ψ are illustrated in Figure 3.

The conditions we impose can be satisfied by constructing g, ϕ in i) from a function whose Fourier transform is a 1-D B-spline [16, Section 1], and ψ, ϕ in ii) from, e.g., the analytic Meyer wavelet [17, Section 3.3.5]. We emphasize that both the WH and the wavelet filters satisfy—by construction—the analyticity and highpass condition [12, Assumption 1] as well as the symmetry property

$$\widehat{g_\lambda}(-\omega) = \widehat{g_{-\lambda}}(\omega), \quad \forall\, \lambda \in \mathbb{Z}\backslash\{0\},\ \forall\, \omega \in \mathbb{R}, \tag{4}$$

which will turn out (in Section 5) to be key in reducing the number of "operationally relevant nodes", a notion defined in (15) below. We refer to the intervals $[-\delta, \delta]$ and $[-1, 1]$ as "spectral gaps" left by the WH and wavelet filters, respectively, as we have $\mathrm{supp}\,(\widehat{g_k}\,) \cap [-\delta, \delta] = \emptyset$, for all $k \in \mathbb{Z}\backslash\{0\}$, in the WH case, and $\mathrm{supp}\,(\widehat{g_j}\,) \cap [-1, 1] = \emptyset$, for all $j \in \mathbb{Z}\backslash\{0\}$, in the wavelet case.

The results in this paper apply to input signals that belong to the class of Sobolev functions $H^s(\mathbb{R}) = \big\{ f \in L^2(\mathbb{R}) \;\big|\; \int_{\mathbb{R}} (1 + |\omega|^2)^s |\widehat{f}(\omega)|^2 \mathrm{d}\omega < \infty \big\}$, $s \geq 0$, where the parameter s acts as a smoothness index. The following spectral decay behavior of Sobolev functions is of relevance throughout the paper. For every $f \in H^s(\mathbb{R})$, there exist $\beta, \mu, L > 0$ (all depending on f) such that

$$|\widehat{f}(\omega)| \leq \mu (1 + |\omega|^2)^{-(\frac{s}{2} + \frac{1}{4} + \frac{\beta}{4})}, \quad \text{a.e. } |\omega| \geq L, \tag{5}$$

where $L > 0$ plays the role of an effective bandwidth of f (see, e.g., [19, Sec. 6.2.1]). Moreover, Sobolev functions encompass a wide range of practically relevant signal classes such as square-integrable functions $L^2(\mathbb{R}) = H^0(\mathbb{R})$, strictly (L)-band-limited functions $L^2_L(\mathbb{R}) \subseteq H^s(\mathbb{R})$, for all $L > 0$ and all $s \geq 0$, and cartoon functions [20] $\mathcal{C}^K_{\mathrm{CART}} \subseteq H^s(\mathbb{R})$, for all $K > 0$ and all $s \in (0, \frac{1}{2})$ (see [12, Lemma 1]). We note that cartoon functions are widely used in the mathematical signal processing literature [7, 12, 13, 21] as a model for natural images such as, e.g., images of handwritten digits [22].

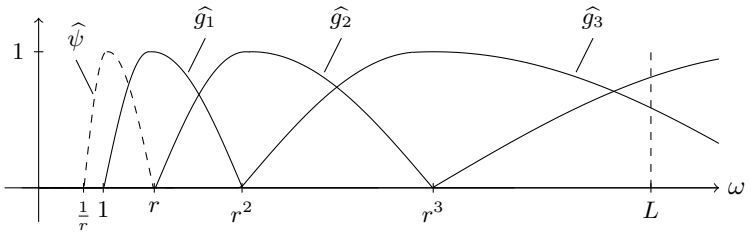

Figure 3: Illustration of the Fourier transforms of the wavelet filters g_j on the frequency band $[0, L]$. The Fourier transform $\widehat{\psi}$ of the mother wavelet ψ is supported on the interval $[r^{-1}, r]$.

Our first main result is the following.

THEOREM 3.1. *For the WH case, let $R > 0$, $\delta \geq \frac{R}{2}$, and set*

$$a = \frac{1}{2} + \frac{\delta}{R}. \tag{6}$$

For the wavelet case, let $r > 1$ and set

$$a = \frac{r^2 + 1}{r^2 - 1}. \tag{7}$$

Then, in both cases, we have

$$W_n(f) = \mathcal{O}\Big(a^{-\frac{n(2s+\beta)}{(2s+\beta+1)}}\Big), \qquad \forall f \in H^s(\mathbb{R}), \tag{8}$$

where the parameter $\beta > 0$ depends on f according to (5).

Proof. The proof is structurally very similar to that of [12, Theorem 2] and will hence not be presented in detail. In a nutshell, the new elements needed to establish (8) are, for the WH case, to replace the so-called modulation weights $\nu_k := Rk + \frac{2}{3}R$, $k \geq 1$, $\nu_k := -\nu_{|k|}$, $k \leq -1$, defined in [12, Eq. 117] by $\nu_k := Rk + \delta - \frac{R^2}{R+2\delta}$, $k \geq 1$, $\nu_k := -\nu_{|k|}$, $k \leq -1$, and similarly, in the wavelet case, to replace the modulation weights $\nu_j := \frac{4}{5}2^j$, $j \geq 1$, $\nu_j := -\frac{4}{5}2^{|j|}$, $j \leq -1$, defined in [12, Eq. 102] by the r-dependent modulation weights $\nu_j := \frac{2r}{r^2+1}r^j$, $j \geq 1$, $\nu_j := -\frac{2r}{r^2+1}r^{|j|}$, $j \leq -1$. The rest of the proof follows closely that of [12, Theorem 2]. □

The identities (6) and (7) show that the filter constructions we propose, indeed, allow to tune the decay factor a through a single parameter, namely R in the WH case and r in the wavelet case. Reducing R, r results in faster energy decay (see also Figure 4). Particularizing (6) to $R = \delta$ and (7) to $r = 2$ recovers the decay factors $a = \frac{3}{2}$ and $a = \frac{5}{2}$, respectively, established in [12, Theorem 2]. Finally, we refer the reader to [12, Section 3] and references therein for an overview of previous work on the decay rate of $W_n(f)$.

4. DEPTH-CONSTRAINED SCATTERING NETWORKS

We now turn to the design of scattering networks of fixed (possibly small) depth N that capture most of the input signal's features. This will be formalized by seeking WH and wavelet filters that, for given $\varepsilon > 0$ and given depth $N \in \mathbb{N}$, result in feature extractors satisfying[†]

$$(1 - \varepsilon)\|f\|_2^2 \leq \sum_{n=0}^{N} |||\Phi_\Omega^n(f)|||^2 \leq \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}). \tag{9}$$

The lower bound in (9) guarantees that at least $((1 - \varepsilon) \cdot 100)\%$ of the input signal energy are contained in the feature vector $\{\Phi_\Omega^n(f)\}_{n=0}^N$ generated in the first N network layers. We note that establishing the upper bound in (9) does not pose any significant difficulties as it follows straight from the results in [6, Appendix E]. The lower bound in (9) implies a trivial kernel for the feature extractor Φ_Ω and thereby ensures that the only signal f that is mapped to the all-zeros feature vector is $f = 0$. We emphasize that the energy decay results in Theorem 3.1 pertain to the feature maps $U[q]f$, whereas energy conservation according to (9) applies to the feature vector $\{\Phi_\Omega^n(f)\}_{n=0}^N$.

[†]The feature space norm is defined as $|||\Phi_\Omega^n(f)|||^2 := \sum_{q \in \Lambda^n} \|(U[q]f) * \chi_{n+1}\|_2^2$.

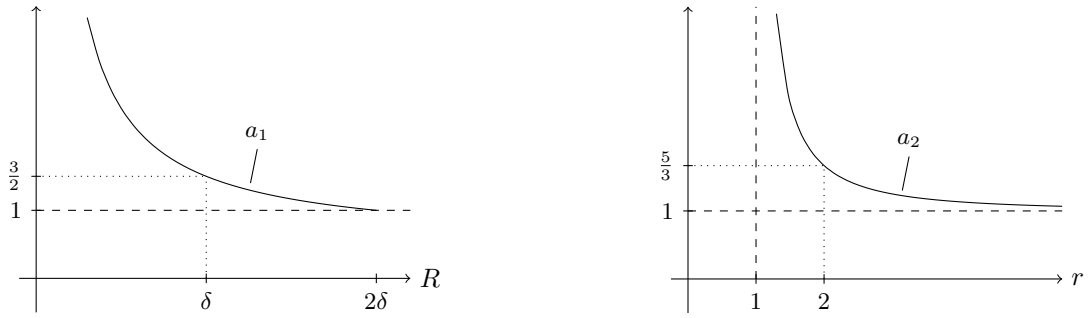

Figure 4: Illustration of the functions $a_1(R) := \frac{1}{2} + \frac{\delta}{R}$, for $R \leq 2\delta$, (left plot) and $a_2(r) := \frac{r^2+1}{r^2-1}$, for $r > 1$, (right plot).

The next result explains how to choose R in the WH and r in the wavelet case so as to satisfy (9). In particular, we shall see that for every (possibly small) $\varepsilon > 0$ and every $N \in \mathbb{N}$, say $\varepsilon = 0.01$ and $N = 1$, there exist $R > 0$ and $r > 1$ such that (9) holds.

THEOREM 4.1. *For the WH case, let $R > 0$, $\delta \geq \frac{R}{2}$. For the wavelet case, let $r > 1$ and $\delta = 1$. Moreover, let $l > 1$, take $f \in H^s(\mathbb{R})\backslash\{0\}$, fix $\varepsilon \in (0,1)$ and $N \in \mathbb{N}$, and define*

$$\kappa := \left(\frac{\tau}{(1-(1-\varepsilon)^{\frac{1}{4l}})\delta} \right)^{1/N},$$

where

$$\tau := \max\left\{ \delta, L, \left(\frac{2\mu^2}{(2s+\beta)(1-\sqrt{1-\varepsilon})\|f\|_2^2} \right)^{\frac{1}{2s+\beta}} \right\}.$$

Here, the parameters $\beta, \mu, L > 0$ depend on f in the sense of (5). If, in the WH case,

$$0 < R \leq \sqrt{\frac{\delta}{\kappa - \frac{1}{2}}}, \tag{10}$$

or, in the wavelet case,

$$1 < r \leq \sqrt{\frac{\kappa+1}{\kappa-1}}, \tag{11}$$

then (9) holds.

Proof. Let a be the decay factor in (7) or (6). Then, it follows from [12, Corollary 2] that

$$a \geq \left(\frac{\tau}{(1-(1-\varepsilon)^{\frac{1}{4l}})\delta} \right)^{1/N} = \kappa \tag{12}$$

is sufficient for (9) to hold. In the WH case, we have $a = \frac{1}{2} + \frac{\delta}{R}$, $\delta \geq \frac{R}{2}$, which, when combined with (12), yields

$$\frac{1}{2} + \frac{\delta}{R} \geq \kappa. \tag{13}$$

Rearranging terms in (13) establishes (10). Next, in the wavelet case, we have $a = \frac{r^2+1}{r^2-1}$, $r > 1$, which, when combined with (12), yields

$$\frac{r^2+1}{r^2-1} \geq \kappa. \tag{14}$$

Rearranging terms in (14) establishes (11) and thereby completes the proof. □

5. NUMBER OF OPERATIONALLY SIGNIFICANT NODES

While the results presented thus far were of mathematically formal nature, in the present section, we shall allow ourselves to argue on a less formal level. The energy decay and conservation results established so far assume an infinite number of filters in the module $(\Psi, |\cdot|, \mathrm{Id})$, and hence an infinite number of nodes in each network layer. Formally, this is a consequence of the filters $\{g_\lambda\}_{\lambda\in\Lambda}$ depending on an index set Λ with $\mathrm{card}(\Lambda) = \infty$, which, in turn, is needed to satisfy the frame condition (1) for all $f \in L^2(\mathbb{R})$. However, as Sobolev functions exhibit spectral decay according to (5), one can consider them effectively band-limited, with L in (5) acting as the effective bandwidth. The error induced by ignoring the spectral components outside the effective spectral support set $\mathrm{esupp}(\widehat{f}) = [-L, L]$ can be bounded according to

$$\begin{aligned}\|f - f_L\|_2^2 = \|\widehat{f} - \widehat{f_L}\|_2^2 = \int_{\mathbb{R}} |\widehat{f}(\omega) - \widehat{f_L}(\omega)|^2 \mathrm{d}\omega = \int_{|\omega|>L} |\widehat{f}(\omega)|^2 \mathrm{d}\omega \leq \mu^2 \int_{|\omega|>L} (1+|\omega|^2)^{-(s+\frac{1}{2}+\frac{\beta}{2})} \mathrm{d}\omega \\ = 2\mu^2 \int_L^\infty (1+r^2)^{-(s+\frac{1}{2}+\frac{\beta}{2})} \mathrm{d}r \leq 2\mu^2 \int_L^\infty r^{-(2s+1+\beta)} \mathrm{d}r = \frac{2\mu^2}{(2s+\beta)L^{2s+\beta}},\end{aligned}$$

where $\widehat{f_L} := \widehat{f} \cdot \mathbb{1}_{[-L,L]}$. This error decreases with increasing smoothness index s. As the function $f * g_\lambda$, with $f \in H^s(\mathbb{R})$ and strictly band-limited $g_\lambda \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, is strictly band-limited, and the modulus non-linearity results in (roughly) a doubling of bandwidth, as heuristically argued below, we allow ourselves to assume in the following that all feature maps are effectively band-limited. Consequently, it is sensible to ask how many nodes are actually needed in the n-th network layer to capture the feature map energy contained in[‡] $\mathrm{esupp}(\widehat{U[q]f})$, $q \in \Lambda^{n-1}$. We formalize this question by defining the number of *operationally significant nodes* in the n-th network layer as

$$\Xi(n) := \mathrm{card}\Big(\Big\{U[q]f \;\Big|\; q \in \Lambda_{\mathrm{sig}}^n\Big\}\Big), \quad n \geq 0, \tag{15}$$

where the set Λ_{sig}^n is defined (recursively) according to $\Lambda_{\mathrm{sig}}^0 := \Lambda^0$, $\Lambda_{\mathrm{sig}}^1 := \big\{\lambda \in \Lambda \;\big|\; \mathrm{esupp}(\widehat{f}) \cap \mathrm{supp}(\widehat{g_\lambda}) \neq \emptyset\big\}$, and

$$\Lambda_{\mathrm{sig}}^n := \Big\{(q,\lambda) \;\Big|\; q \in \Lambda_{\mathrm{sig}}^{n-1} \;\text{ and }\; \lambda \in \Lambda \;\text{ with }\; \mathrm{esupp}(\widehat{U[q]f}) \cap \mathrm{supp}(\widehat{g_\lambda}) \neq \emptyset\Big\}, \quad n \geq 2. \tag{16}$$

For the root of the network, i.e., $n = 0$, we have $\Xi(0) = 1$, owing to $U[q]f = U[e]f = f$. The definition of $\Xi(n)$ accounts for a topology reduction, relative to the full tree in Figure 1, caused by i) feature map symmetries (see (19) below) and reflected by counting the number of distinct[§] feature maps $U[q]f$ in (15) only[¶], and ii) width pruning owing to the effective band-limitation of the input signal f (and hence the effective band-limitation of the feature maps $U[q]f$) and the strict band-limitation of the filters g_λ. Note that the specific spectral structure of f, e.g., a multi-band structure, can lead to further topology reduction. As this effect is, however, somewhat artificial, it will consistently be ignored in the remainder of the paper. We honor the dependence of $\Xi(n)$ on i) the filters in Ψ (and their parameters δ, r, and R) and ii) the effective bandwidth L of the input signal f through the notation $\Xi_{\mathrm{WH}}(n, R, \delta, L)$ and $\Xi_{\mathrm{wav}}(n, r, L)$.

Next, our goal is to determine $\Xi(n)$, for $n \geq 1$. Starting with the WH case and $n = 1$, we have $\mathrm{esupp}(\widehat{f}) \cap \mathrm{supp}(\widehat{g_k}) \neq \emptyset$ if $L > \delta$ (which prevents $\mathrm{esupp}(\widehat{f}) = [-L, L]$ from being fully contained in the spectral gap $[-\delta, \delta]$ left by the filters $\{g_k\}_{k\in\mathbb{Z}\setminus\{0\}}$, see Figure 5, top row) and $|k| \leq \lceil (L-\delta)R^{-1} \rceil$ (see Figure 5, bottom row). This yields

$$\Lambda_{\mathrm{WH,\,sig}}^1 = \big\{k \in \mathbb{Z}\setminus\{0\} \;\big|\; |k| \leq \lceil (L-\delta)R^{-1} \rceil\big\}, \qquad \text{if } L > \delta,$$

[‡]We remark that the notation $\mathrm{esupp}(\widehat{U[q]f})$, for $q \in \Lambda^n$, with $n \geq 1$, is used to denote the "effective spectral support" of the feature map $U[q]f$ in a somewhat casual sense, i.e., not necessarily strictly pertaining to the interval $[-L, L]$ based on the Sobolev bandwidth L according to (5).

[§]We recall that the cardinality of a set equals the number of *distinct* elements in the set, e.g., $\mathrm{card}(\{a, a, b\}) = 2$, for $a \neq b$.

[¶]We emphasize that the location (in the full tree in Figure 1) of identical feature maps (not counted in $\Xi(n)$) is uniquely determined. In practice, it therefore suffices to, indeed, compute these features only once and arrange identical copies accordingly in $\Phi_\Omega(f)$.

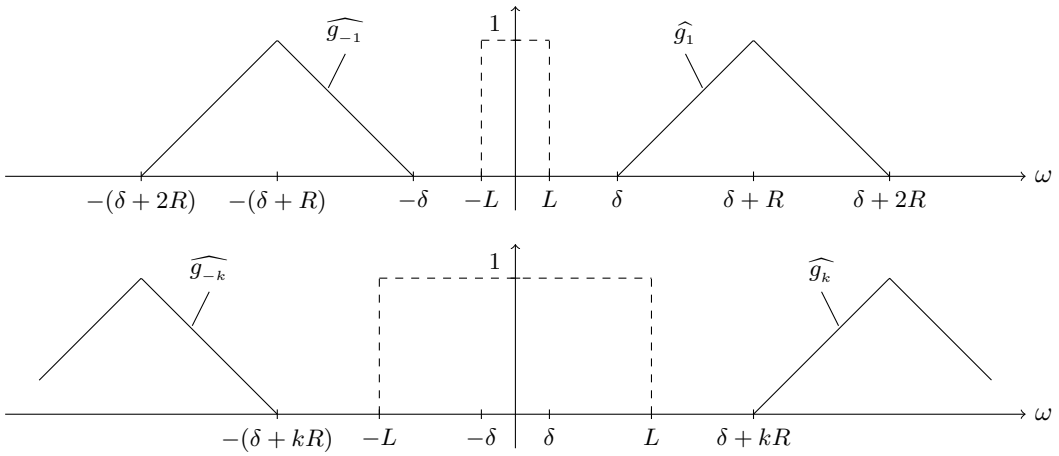

Figure 5: Top row: If $L \leq \delta$, then $\operatorname{esupp}(\widehat{f}) = [-L, L]$ is fully contained in the spectral gap $[-\delta, \delta]$ left by the filters $\{g_k\}_{k \in \mathbb{Z}\setminus\{0\}}$. Bottom row: If $L > \delta$ and $|k| > \lceil (L-\delta)R^{-1} \rceil$, the spectral supports of the filters g_k do not overlap with $\operatorname{esupp}(\widehat{f}) = [-L, L]$.

and $\Lambda^1_{\mathrm{WH}} = \emptyset$, if $L \leq \delta$, which, in turn, implies

$$\Xi_{\mathrm{WH}}(1, R, \delta, L) = 2 \cdot \lceil (L-\delta)R^{-1} \rceil, \qquad \text{if } L > \delta,$$

and $\Xi_{\mathrm{WH}}(1, R, \delta, L) = 0$, if $L \leq \delta$. Next, determining $\Xi(2)$ requires, by (16), studying the spectral characteristics of the feature maps $U[k]f = |f * g_k|$. We note that, owing to the modulus non-linearity, characterizing the effective spectral support of $U[k]f = |f * g_k|$ is non-trivial. We can, however, take a cue from the behavior of the *squared* modulus non-linearity, i.e.,

$$W[k]f := |f * g_k|^2 = (f * g_k) \cdot (\overline{f * g_k}),$$

and note that $\widehat{W[k]f}$ is simply the auto-correlation of $\widehat{f} \cdot \widehat{g_k}$ (see the second row in Figure 6). The squared modulus non-linearity therefore doubles the spectral support of $f * g_k$ and "demodulates" in the sense of the spectrum $\widehat{W[k]f}$ being located symmetrically around the origin, both irrespectively of the spectral location of $f * g_k$. The key observation is now that the modulus non-linearity behaves very similarly, as illustrated in Figure 6, third row. In the following, we shall therefore allow ourselves to work with $\operatorname{esupp}(\widehat{U[k]f}) \subseteq [-2R, 2R]$, for all $k \in \mathbb{Z}\setminus\{0\}$. We hasten to add that this statement is based solely on numerical evidence and we do not have a corresponding formal result. It is interesting to observe that the sigmoid, rectified linear unit, and hyperbolic tangent non-linearities, all exhibit very different behavior in this regard (see Figure 6, bottom row, for an illustration for the rectified linear unit). By induction over n, one can show that, for all $n \geq 2$, we have

$$\Lambda^n_{\mathrm{WH,\,sig}} = \big\{ (q,k) \;\big|\; q \in \Lambda^{n-1}_{\mathrm{WH,\,sig}} \text{ and } |k| \leq \lceil 2 - \delta R^{-1} \rceil \big\}, \qquad \text{if } 2R, L > \delta, \tag{17}$$

and $\Lambda^n_{\mathrm{WH,\,sig}} = 0$, else, which implies

$$\Xi_{\mathrm{WH}}(n, R, \delta, L) = \begin{cases} 1, & \text{if } n = 0, \\ 2 \cdot \lceil \frac{L-\delta}{R} \rceil, & \text{if } n = 1 \text{ and } L > \delta, \\ 2 \cdot \lceil \frac{L-\delta}{R} \rceil \lceil 2 - \frac{\delta}{R} \rceil^{n-1}, & \text{if } n \geq 2 \text{ and } 2R, L > \delta, \end{cases} \tag{18}$$

and $\Xi_{\mathrm{WH}}(n, R, \delta, L) = 0$, else. We remark that (18) follows from (17) upon noting that, from the second network layer onwards, either $U[(k_1, \ldots, k_{n-1}, -k_n)]f$ or $U[(k_1, \ldots, k_{n-1}, k_n)]f$ only contribute to $\Xi(n)$ (which,

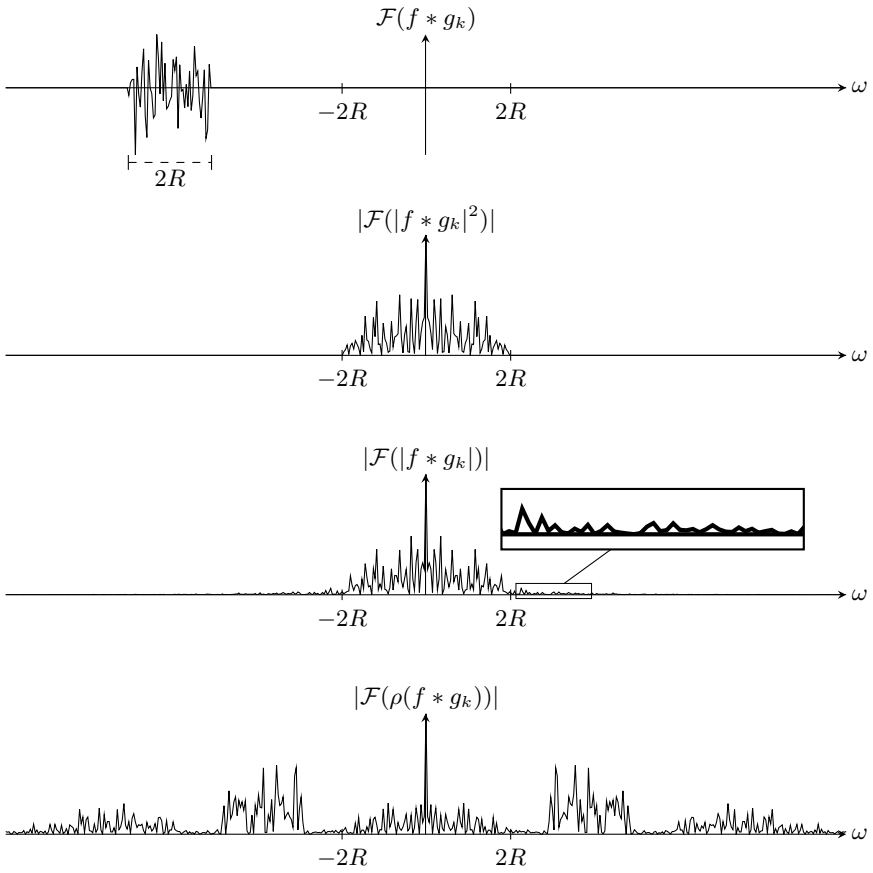

Figure 6: Illustration of the demodulation and bandwidth doubling effect of the squared modulus (second row) and the modulus non-linearities (third row). The WH filters have prototype function g with $\mathrm{supp}(\widehat{g}) = [-R, R]$. The rectified linear unit non-linearity (bottom row) is defined as $\rho(z) := \max\{0, \mathrm{Real}(z)\} + \max\{0, \mathrm{Imag}(z)\}$, $z \in \mathbb{C}$.

as explained above, only counts the number of *distinct* feature maps). This follows from the symmetry relation

$$\begin{aligned} U[(k_1, \ldots, k_{n-1}, -k_n)]f = U[-k_n]U[(k_1, \ldots, k_{n-1})]f &= \big| \underbrace{(U[(k_1, \ldots, k_{n-1})]f)}_{\text{real-valued}} * g_{-k_n} \big| \\ &= \big|(U[(k_1, \ldots, k_{n-1})]f) * g_{k_n}\big| = U[(k_1, \ldots, k_n)]f, \quad \forall n \geq 2, \end{aligned} \tag{19}$$

where the first equality in (19) is by the following.

LEMMA 5.1. *Let $f \in L^2(\mathbb{R})$ be real-valued and g_λ either a WH or a wavelet filter as defined in Section 3. Then, we have*

$$U[-\lambda]f = |f * g_{-\lambda}| = |f * g_\lambda| = U[\lambda]f, \ \forall \lambda \in \Lambda.$$

Proof. The proof follows from basic Fourier calculus and by exploiting the symmetry property (4). □

For the wavelet case, arguments similar to those leading to (17) yield, for all $n \geq 1$,

$$\Lambda^n_{\mathrm{wav,\,sig}} = \big\{(q, j) \;\big|\; q \in \Lambda^{n-1}_{\mathrm{wav,\,sig}} \text{ and } |j| \leq \lceil \log_r(L^{(n-1)}) \rceil \big\}, \qquad \text{if } L^{(n-1)} > 1, \tag{20}$$

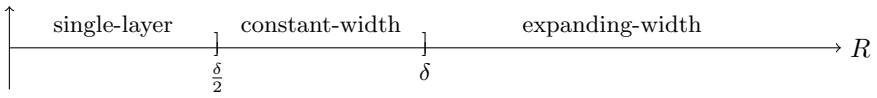

Figure 7: Transition between network topologies (as induced by operationally significant nodes) as a function of R.

and $\Lambda^n_{\text{wav, sig}} = \emptyset$, if $L^{(n-1)} \leq 1$, where $L^{(n)} := L(r^2-1)^n$, for $n \geq 0$. This implies

$$\Xi_{\text{wav}}(n,r,L) = \begin{cases} 1, & \text{if } n = 0, \\ 2 \cdot \lceil \log_r(L) \rceil, & \text{if } n = 1 \text{ and } L > 1, \\ \mathcal{O}\big(\log_r^n(L) + 2^n(n-1)!\big), & \text{if } n \geq 2 \text{ and } L > 1 \text{ and } r > \sqrt{2}, \\ 2 \cdot \lceil \log_r(L) \rceil^n, & \text{if } n \geq 2 \text{ and } L > 1 \text{ and } r = \sqrt{2}, \\ \mathcal{O}\big(\log_r^n(L)\big), & \text{if } M > n \geq 2 \text{ and } L > 1 \text{ and } r < \sqrt{2}, \end{cases} \tag{21}$$

and $\Xi_{\text{wav}}(n,r,L) = 0$, else, where $M := 1 + \log_{r^2-1}(L)$. We can see that the parameter $r > 1$ of the mother wavelet crucially impacts the index sets (20) and thereby the number of operationally significant nodes (21). Specifically, r determines whether the effective bandwidths $L^{(n)} = L(r^2-1)^n$ of the feature maps increase, decrease, or remain constant as the layer index n increases. For $r > \sqrt{2}$ we have bandwidth expansion, for $r < \sqrt{2}$ bandwidth contraction, and for $r = \sqrt{2}$ the effective bandwidths of the feature maps $U[q]f$ remain constant in n, i.e., $L^{(n)} = L$, for all $n \in \mathbb{N}$.

6. NETWORK TOPOLOGY INDUCED BY OPERATIONALLY SIGNIFICANT NODES

The scattering network architecture defined in Section 2 has a tree topology with an infinite number of nodes per layer. The analysis in the previous section revealed, however, that the number $\Xi(n)$ of operationally significant nodes is finite in every network layer $n = 0, \ldots, N$. The goal of this section is to determine and characterize the network topology and the corresponding feature vector $\{\Phi^n_\Omega(f)\}_{n=0}^N$ induced by the operationally significant nodes. For WH filters, we distinguish between the following cases:

i) *Shallow feature extraction*: If $L \leq \delta$ (i.e., the effective spectral support of the input signal f is fully contained in the spectral gap $[-\delta, \delta]$, see Figure 5, top row), then $\Xi_{\text{WH}}(0,R,\delta,L) = 1$ and $\Xi_{\text{WH}}(n,R,\delta,L) = 0$, $n \geq 1$. The feature vector $\{\Phi^n_\Omega(f)\}_{n=0}^N$ consists of a single element, namely $f * \chi$, which is simply the output at the root of the network.

ii) *Single-layer network*: If $L > \delta$ and $2R \leq \delta$ (i.e., the effective spectral support of all feature maps is fully contained in the spectral gap $[-\delta, \delta]$), then $\Xi_{\text{WH}}(0,R,\delta,L) = 1$, $\Xi_{\text{WH}}(1,R,\delta,L) = 2 \cdot \lceil \frac{L-\delta}{R} \rceil$, and $\Xi_{\text{WH}}(n,R,\delta,L) = 0$, $n \geq 2$, which renders the network to have a single layer only. The corresponding feature vector is given by $\{\Phi^n_\Omega(f)\}_{n=0}^N = \{f * \chi\} \cup \{|f * g_k| * \chi\}_{|k| \leq \lceil (L-\delta)R^{-1} \rceil}$.

iii) *Constant-width network*: If $L > \delta$ and $R \leq \delta < 2R$ (i.e., only the spectral supports of the filters g_k, $k \in \{-1, 1\}$, overlap with the interval $[-2R, 2R]$), then the number of operationally significant nodes $\Xi_{\text{WH}}(n,R,\delta,L) = 2 \cdot \lceil \frac{L-\delta}{R} \rceil$, $n \geq 1$, is constant in n (for $n = 0$, we have $\Xi_{\text{WH}}(0,R,\delta,L) = 1$). In this constant-width network, every network layer $n \geq 1$ contributes with $2 \cdot \lceil \frac{L-\delta}{R} \rceil$ elements to the feature vector.

iv) *Expanding-width network*: If $L > \delta$ and $\delta < R$ (i.e., at least four filters g_k overlap with the interval $[-2R, 2R]$), then $\Xi_{\text{WH}}(n,R,\delta,L) = 2 \cdot \lceil \frac{L-\delta}{R} \rceil \lceil 2 - \frac{\delta}{R} \rceil^{n-1}$, $n \geq 1$, which renders the network expanding width (for $n = 0$, we have $\Xi_{\text{WH}}(0,R,\delta,L) = 1$).

We note that for $L > \delta$, it is the bandwidth R of the WH prototype function g that determines the transition between the network topologies above, see Figure 7.

We next turn to wavelet filters with the following cases of interest:

i) *Depth-pruned network*: If $L > \max\{1, r\}$, $r < \sqrt{2}$, and $N > M = 1 + \log_{r^2-1}(L)$, (i.e., the effective bandwidths $L^{(n)} = L(r^2-1)^n$ of the feature maps are decreasing in n and are eventually smaller than 1 and hence contained in the spectral gap $[-1, 1]$), then we have $\{\Phi_\Omega^n(f)\}_{n=0}^N = \{\Phi_\Omega^n(f)\}_{n=0}^M$. This means that from the M-th layer onwards, there are no more non-zero signals to be propagated to deeper layers.

ii) *Extremely-narrow network*: If $1 < L \leq r = \sqrt{2}$ (i.e., the effective bandwidths $L^{(n)} = L$ of the feature maps are constant in n, with $n \geq 1$, and overlap with the spectral supports of g_k, $k \in \{-1, 1\}$, only), then the number of operationally significant nodes $\Xi_{\text{wav}}(n, r, L) = 2$, $n \geq 1$, is constant in n (for $n = 0$, we have $\Xi_{\text{wav}}(0, r, L) = 1$). Every network layer $n \geq 1$ contributes with two elements to the feature vector.

7. MINIMIZING THE AVERAGE NUMBER OF OPERATIONALLY SIGNIFICANT NODES PER LAYER

The purpose of this section is to analyze the impact of the feature map energy decay rate on the average number of operationally significant nodes per layer. For simplicity of exposition, throughout this chapter, we focus on the WH case. We take the parameters N, δ, and L to be fixed and assume i) that the effective bandwidth L of the input signal satisfies $L > \delta$ (which guarantees that we are not in the (trivial) shallow feature extraction situation, see Section 6) and ii) that the network depth satisfies $N \geq 3$.

We first recall that, thanks to (6), the (exponential) decay factor a can be tuned through the parameter R. Specifically, reducing the bandwidth R of the WH prototype function g implies faster (guaranteed) energy decay (see also Figure 4). Increasing R implies slower (guaranteed) energy decay with R eventually violating the condition $R \geq 2\delta$ needed for validity of the statement in Theorem 3.1. In the following, we determine the optimal value R^* in the exponential-decay regime $R \in (0, 2\delta)$ of $W_n(f)$ that minimizes the average number of operationally significant nodes per layer given by

$$\Theta_{\text{WH}}(N, R, \delta, L) := \frac{1}{N} \sum_{n=1}^{N} \Xi_{\text{WH}}(n, R, \delta, L). \tag{22}$$

In order to minimize the expression in (22) over the interval $(0, 2\delta)$, we distinguish between three cases:

i) If $R \in (\delta, 2\delta)$, then we are in the situation of an expanding-width network, and we have

$$\Theta_{\text{WH}}(N, R, \delta, L) = 2 \cdot \left\lceil \frac{L-\delta}{R} \right\rceil \frac{1}{N} \sum_{n=0}^{N-1} \left[\underbrace{2 - \frac{\delta}{R}}_{\in (1, \frac{3}{2})} \right]^n = 2 \cdot \left\lceil \frac{L-\delta}{R} \right\rceil \frac{1}{N} \sum_{n=0}^{N-1} 2^n = 2 \cdot \left\lceil \frac{L-\delta}{R} \right\rceil \frac{1}{N} (2^N - 1). \tag{23}$$

ii) For $R \in (\frac{\delta}{2}, \delta]$, we have a constant-width network and

$$\Theta_{\text{WH}}(N, R, \delta, L) = 2 \cdot \left\lceil \frac{L-\delta}{R} \right\rceil \frac{1}{N} \sum_{n=0}^{N-1} \left[\underbrace{2 - \frac{\delta}{R}}_{\in (0, 1]} \right]^n = 2 \cdot \left\lceil \frac{L-\delta}{R} \right\rceil. \tag{24}$$

iii) If $R \in (0, \frac{\delta}{2}]$, we get a single-layer network, and $\Theta_{\text{WH}}(N, R, \delta, L) = 2 \cdot \lceil \frac{L-\delta}{R} \rceil$.

Next, we note that the function $R \mapsto \lceil \frac{L-\delta}{R} \rceil$, $R \in (0, 2\delta)$, is monotonically decreasing in R, which allows us to conclude that, owing to ii) and iii), $R^* \notin (0, \frac{\delta}{2}]$. Moreover, thanks to the monotonicity of the mapping $R \mapsto \lceil \frac{L-\delta}{R} \rceil$, $R \in (0, 2\delta)$, it is sufficient to evaluate the expression (23) for $R = 2\delta$ and (24) for $R = \delta$ and to determine which of the resulting two values is smaller. Specifically, we have

$$\Theta_{\text{WH}}(N, 2\delta, \delta, L) = 2 \cdot \left\lceil \frac{L}{2\delta} - \frac{1}{2} \right\rceil \frac{1}{N} (2^N - 1) = 2 \cdot \left\lceil \frac{1}{2} \left(\frac{L}{\delta} - 1 \right) \right\rceil \frac{1}{N} (2^N - 1) \geq \left\lceil \frac{L}{\delta} - 1 \right\rceil \frac{1}{N} (2^N - 1) \tag{25}$$

$$> 2 \cdot \left\lceil \frac{L}{\delta} - 1 \right\rceil = \Theta_{\text{WH}}(N, \delta, \delta, L), \tag{26}$$

where in (25) we used $\lceil \frac{x}{2} \rceil \geq \frac{1}{2} \lceil x \rceil$, $x \geq 0$, and (26) is thanks to $N \geq 3$, which, in turn, is by assumption. This implies $R^* = \delta$ and renders the network constant-width.

REFERENCES

[1] Goodfellow, I., Bengio, Y., and Courville, A., [*Deep Learning*], MIT Press (2016).

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

[3] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," *Proc. of International Conference on Learning Representations (ICLR)* (2014).

[4] Lane, N. D. and Georgiev, P., "Can deep learning revolutionize mobile sensing?," *Proc. of International Workshop on Mobile Computing Systems and Applications*, 117–122 (2015).

[5] Mallat, S., "Group invariant scattering," *Comm. Pure Appl. Math.* **65**(10), 1331–1398 (2012).

[6] Wiatowski, T. and Bölcskei, H., "A mathematical theory of deep convolutional neural networks for feature extraction," *arXiv:1512.06293* (2015).

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

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

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

[10] Waldspurger, I., *Wavelet transform modulus: Phase retrieval and scattering*, PhD thesis, École Normale Supérieure Paris (2015).

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

[12] Wiatowski, T., Grohs, P., and Bölcskei, H., "Energy propagation in deep convolutional neural networks," *arXiv:1704.03636* (2017).

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

[14] Ali, S. T., Antoine, J. P., and Gazeau, J. P., "Continuous frames in Hilbert spaces," *Annals of Physics* **222**(1), 1–37 (1993).

[15] Frazier, M., Jawerth, B., and Weiss, G., [*Littlewood-Paley Theory and the Study of Function Spaces*], American Mathematical Society (1991).

[16] Gröchening, K., Janssen, A. J. E. M., Kaiblinger, N., and Pfander, G. E., "Note on B-splines, wavelet scaling functions, and Gabor frames," *IEEE Trans. Inf. Theory* **49**(12), 3318–3320 (2003).

[17] Daubechies, I., [*Ten Lectures on Wavelets*], SIAM (1992).

[18] Wendland, H., [*Scattered Data Approximation*], Cambridge University Press (2004).

[19] Grafakos, L., [*Modern Fourier Analysis*], Springer (2009, second edition).

[20] Donoho, D. L., "Sparse components of images and optimal atomic decompositions," *Constructive Approximation* **17**(3), 353–382 (2001).

[21] Grohs, P., Keiper, S., Kutyniok, G., and Schäfer, M., "Cartoon approximation with α-curvelets," *J. Fourier Anal. Appl.*, 1–59 (2015).

[22] LeCun, Y. and Cortes, C., "The MNIST database of handwritten digits," (1998). `http://yann.lecun.com/exdb/mnist`.