Digital implementations of deep feature extractors are intrinsically informative

Max Getter
Chair for Geometry and Analysis
RWTH Aachen University
getter@mathga.rwth-aachen.de

***Abstract*—Rapid information (energy) propagation in deep feature extractors is crucial to balance computational complexity versus expressiveness as a representation of the input. We prove an upper bound for the speed of energy propagation in a unified framework that covers different neural network models, both over Euclidean and non-Euclidean domains. Additional structural information about the signal domain can be used to explicitly determine or improve the rate of decay. To illustrate this, we show global exponential energy decay for a range of 1) feature extractors with discrete-domain input signals, and 2) convolutional neural networks (CNNs) via scattering over locally compact abelian (LCA) groups.**

***Index Terms*—Deep learning, representation learning, feature learning, scattering transform, information retention, energy propagation, neural networks.**

I. Introduction

Deep feature extractors (feature maps) are supposed to produce meaningful representations of the input data for a machine learning task. It is generally hard to predict if a specific feature is relevant to the task at hand [1]. To assess the power of a feature extractor, it is hence necessary to rely on more generic (i.e., task-unspecific) criteria. These include the ability of the feature extractor to encode abstract concepts—resulting in (approximate) invariance to local changes of the input—and its ability to disentangle the factors that are responsible for variation within the data [1]–[4]. Both criteria are related by the key component of information retention, which leaves feature engineers with the challenging task of balancing the discriminative power of the feature extractor (i.e., the ability to disentangle the driving factors for variance within the data) and the potential loss of information (as a result of transforming the input). In this contribution, we focus on the premise that deep feature extractors should contain most of the (relevant) information about their input signals, which is expressed by the aggregate energy content of the features.

A. Prior work

Information retention is a key component of successful feature extraction. It has hence already been addressed by prior works—mainly in the context of scattering CNNs, both over Euclidean domains [5]–[10] and graphs [11], [12]. While the question of energy decay seems to be reasonably settled for scattering CNNs over Euclidean domains, the literature is rather sparse on energy decay in more general feature extractors, in particular, scattering CNNs over non-Euclidean domains.

B. Our contributions

Motivated by the construction of scattering CNNs [5], [13], [14], we provide a unifying framework to study the properties of general deep feature extractors acting on signals defined over arbitrary measure spaces. The framework is particularly well suited to quantitatively analyze their stability and information propagation.

The existing results regarding energy decay are conceptually closely related to each other. We exploit this insight to establish a generic rate for energy decay in general deep feature extractors—building on [11, Proposition 3.3], where the authors show that energy decay is exponential with increasing network depth for certain graph CNNs. As a corollary, we obtain exponential energy decay for general deep feature extractors if the underlying measure space does not contain sets of arbitrarily small positive measure.

We extend ideas from [5], [9] to LCA group scattering and combine them with sumset-estimates based on [15]. This yields exponential energy decay essentially if the frequency supports of the filters are uniformly bounded. As a byproduct, we thereby validate the experimental observations from [16], [17] regarding the energy distribution in scattering CNNs. More generally, we conclude that digital implementations of deep feature extractors are intrinsically informative.

II. Information retention in general feature extractors

A. A general model for feature extraction

For every $\ell \in \mathbb{N}_0$, let $\mathcal{H}^{(\ell)} := L^2(\mathcal{M}^{(\ell)})$ be the Lebesgue-space[1] of square-integrable complex-valued functions over a measure space $\mathcal{M}^{(\ell)} := (M^{(\ell)}, \mathfrak{M}^{(\ell)}, \mu^{(\ell)})$. Set $\mathcal{H} := \mathcal{H}^{(0)}$. For a broad range of both finite- or infinite-depth neural networks, the modules for the ℓ-th layer, $\ell \in \mathbb{N}$, of such a neural network can be described by an at most countable family of bounded linear operators $\mathcal{L}^{(\ell)} \subseteq \mathcal{B}(\mathcal{H}^{(\ell-1)}, \mathcal{H}^{(\ell)})$, which forward the information from layer $\ell - 1$ to layer ℓ,

The author acknowledges funding by the Deutsche Forschungsgemeinschaft (German Research Foundation)—Project number 442047500 through the Collaborative Research Center "Sparsity and Singular Structures" (SFB 1481).

[1]If there is no ambiguity, we omit the Hilbert-space index to denote the corresponding inner product $\langle \cdot, \cdot \rangle$ or norm $\| \cdot \|$.

a nonexpansive[2] map $\sigma^{(\ell)} : \mathcal{H}^{(\ell)} \to \mathcal{H}^{(\ell)}$ with $\sigma^{(\ell)}(0) = 0$, which allows to introduce additional complexity to the model, and a bounded linear operator $A^{(\ell)} \in \mathcal{B}(\mathcal{H}^{(\ell)})$, which allows to generate an output at depth ℓ. Likewise, $A^{(0)} \in \mathcal{B}(\mathcal{H})$ allows to produce a first output at layer depth zero. For any input signal $f \in \mathcal{H}$, the neural network associated with these modules generates the output

$$Sf := \left\{ S[p]f \ : \ \ell \geq 0, \ p \in \mathcal{P}^{(\ell)} \right\},$$

where $\mathcal{P}^{(\ell)} := \mathcal{L}^{(1)} \times \cdots \times \mathcal{L}^{(\ell)}$ for $\ell \in \mathbb{N}$, $\mathcal{P}^{(0)} := \{p_e\}$ contains the unique empty path p_e of length $\ell = 0$ with associated operator $S[p_e]f := A^{(0)}f$, and for any path $p = (L^{(1)}, \ldots, L^{(\ell)}) \in \mathcal{P}^{(\ell)}$ of length $\ell \in \mathbb{N}$,

$$S[p]f := A^{(\ell)}U[p]f, \quad U[p]f := \sigma^{(\ell)}L^{(\ell)} \ldots \sigma^{(1)}L^{(1)}f.$$

B. Energy conservation

The following condition is key to many desirable properties of neural networks employed as feature extractors, and is always assumed below.

Assumption II.1. *For every $\ell \in \mathbb{N}$ and for all $h \in \mathcal{H}^{(\ell-1)}$,*

$$\left\| A^{(\ell-1)}h \right\|^2 + \sum_{L \in \mathcal{L}^{(\ell)}} \|Lh\|^2 \leq \|h\|^2 .$$

Let us briefly comment on the flexibility of the model. Both finite- and infinite-depth neural networks can be modeled by the framework, for if a network consists only of finitely many layers $\ell = 0, \ldots, D$, one can simply take, for all $\ell > D$, $\mathcal{L}^{(\ell)} = \emptyset$ and $A^{(\ell)} = \sigma^{(\ell)} = \mathrm{id}_{\mathcal{H}^{(D)}}$.

The action of the ℓ-th layer of a feedforward neural network [18] is typically described by $f^{(\ell)} = \sigma(W^{(\ell)} f^{(\ell-1)})$, where $W^{(\ell)} \in \mathbb{C}^{d^{(\ell)} \times d^{(\ell-1)}}$ are the (learned) weights including a potential bias term, σ is a pointwise nonlinearity, and $f^{(\ell)}$ is the output of the ℓ-th layer. Clearly, this fits into our framework, by setting $\mathcal{H} = \mathbb{C}^{d^{(0)}}$, $\mathcal{H}^{(\ell)} = \mathbb{C}^{d^{(\ell)}}$, $\mathcal{L}^{(\ell)} = \{W^{(\ell)}\}$, $\sigma^{(\ell)} = \sigma$, for $\ell = 1, \ldots, D$, as well as $A^{(\ell)} = 0_{d^{(\ell)} \times d^{(\ell)}}$ for $\ell = 0, \ldots, D-1$, and $A^{(D)} = \mathrm{id}_{\mathbb{C}^{d^{(D)}}}$. Assumption II.1 then boils down to the requirement that the singular values of $W^{(\ell)}$ are ≤ 1. Linear feedforward neural networks [19] are a special instance thereof, taking pointwise identities for σ.

Section III studies scattering CNNs in the above framework. The framework description is the same for more general CNNs (learned or not). Pooling can be incorporated by adjusting either the nonlinearities or the linear operators $\mathcal{L}^{(\ell)}$.

The next proposition, which collects several properties of S that are based on straightforward generalizations of [13], [14], makes apparent why Assumption II.1 is natural for deep feature extractors.

Proposition II.2. *We have, for every $f \in \mathcal{H}$ and every $N \in \mathbb{N}$,*

$$\sum_{p \in \mathcal{P}^{(N)}} \|S[p]f\|^2 + W_{N+1}(f) \leq W_N(f), \tag{1}$$

where $W_N(f) := \sum_{p \in \mathcal{P}^{(N)}} \|U[p]f\|^2$. Thus,

$$\sum_{n=0}^{N-1} \sum_{p \in \mathcal{P}^{(n)}} \|S[p]f\|^2 + W_N(f) \leq \|f\|^2 . \tag{2}$$

Further, $S : \mathcal{H} \to \bigoplus_{\ell=0}^{\infty} \mathcal{H}^{(\ell)}$ is nonexpansive with $S(0) = 0$, hence norm-decreasing.

Remark II.3. *Assumption II.1 can be relaxed by replacing the right-hand side (RHS) of the inequality with the weaker condition $\leq B^{(\ell)} \|h\|^2$ for a constant $B^{(\ell)} > 1$. Such a relaxation would be necessary to model skip connections (as required in, e.g., residual neural networks). However, upper bounds on $W_N(f)$ are more meaningful if $B^{(\ell)} \leq 1$. Assuming $B^{(\ell)} \leq 1$ also entails that the resulting feature extractor S is nonexpansive, which provably guarantees other desired properties of S, such as its stability with respect to small deformations of the input [5]. Likewise, for scattering CNNs it is common to require an additional lower frame condition of the type $c^{(\ell)} \|h\|^2 \leq \left\|A^{(\ell-1)}h\right\|^2 + \sum_{L \in \mathcal{L}^{(\ell)}} \left\|\sigma^{(\ell)}Lh\right\|^2$, $c^{(\ell)} > 0$ independent of $h \in \mathcal{H}^{(\ell-1)}$. In these cases, energy conservation holds in the sense of $c\|f\|^2 \leq \|Sf\|^2 \leq \|f\|^2$, where $c = \prod_{\ell=1}^{\infty} c^{(\ell)} \geq 0$, cf. [7].*

Starting from (1), a telescoping argument yields

$$\sum_{\ell=N}^{\infty} \sum_{p \in \mathcal{P}^{(\ell)}} \|S[p]f\|^2 \leq \sum_{\ell=N}^{\infty} (W_\ell(f) - W_{\ell+1}(f)) \leq W_N(f).$$

The infinite sum $\sum_{\ell=N}^{\infty} \sum_{p \in \mathcal{P}^{(\ell)}} \|S[p]f\|^2$ provides a natural measure to quantify the information contained in the layers of depth $\geq N$. This quantity is bounded above by $W_N(f)$. Assuming that Sf is nontrivial, truncating the (possibly infinitely deep) feature extractor after the first N layers looses at most $W_N(f)/\|Sf\|^2$ percent of the total energy that is distributed across the entire network, suggesting that a fast decay of this quantity is desirable (see also [7]). Rapid energy decay across layers is also provably closely related to the stability of scattering CNNs with respect to pertubations of the input [5, Corollary 2.15], [13, Theorem 5].

On the basis of Assumption II.1, we derive a generic upper bound for $W_N(f)$, for arbitrary $f \in \mathcal{H}$ and $N \in \mathbb{N}$. Let us start by outlining the general proof strategy, the starting point of which is similar to the one used in previous works on energy propagation in scattering CNNs, e.g. [9]. By Assumption II.1,

$$\begin{aligned} W_{N+1}(f) &= \sum_{p \in \mathcal{P}^{(N)}} \sum_{L \in \mathcal{L}^{(N+1)}} \left\| \sigma^{(N+1)} L U[p] f \right\|^2 \\ &\leq \sum_{p \in \mathcal{P}^{(N)}} \left(\|U[p]f\|^2 - \left\| A^{(N)} U[p] f \right\|^2 \right). \end{aligned} \tag{3}$$

We introduce

$$\begin{aligned} \iota_N :&= \inf_{h \in \mathcal{H}, p \in \mathcal{P}^{(N)} : U[p]h \neq 0} \frac{\left\|A^{(N)}U[p]h\right\|^2}{\|U[p]h\|^2} \\ &= \inf_{g \in \mathcal{R}^{(N)} : \|g\| = 1} \left\| A^{(N)} g \right\|^2 , \end{aligned} \tag{4}$$

[2]That is, 1-Lipschitz, for all $f, g \in \mathcal{H}^{(\ell)}$, $\left\|\sigma^{(\ell)}f - \sigma^{(\ell)}g\right\| \leq \|f - g\|$.

where $\mathcal{R}^{(N)} := \bigcup_{p\in\mathcal{P}^{(N)}} \mathcal{R}(U[p]) \subseteq \mathcal{H}^{(N)}$, $\mathcal{R}$ referring to the range of the operators. Using this to estimate (3), we obtain

$$\begin{aligned} W_{N+1}(f) &\leq \sum_{p\in\mathcal{P}^{(N)}} (1-\iota_N)\,\|U[p]f\|^2 \\ &= W_N(f)\cdot(1-\iota_N) \\ &\leq W_1(f)\cdot\prod_{\ell=1}^{N}(1-\iota_\ell) \\ &\leq \left(\|f\|^2 - \left\|A^{(0)}f\right\|^2\right)\cdot\prod_{\ell=1}^{N}(1-\iota_\ell). \end{aligned} \tag{5}$$

Describing the range $\mathcal{R}(U[p])$ of the operator $U[p]$ for any path p of length ≥ 2 is generally a hard problem, and so is the explicit calculation of ι_N.

C. A generic rate for energy decay

We reduce the question of energy decay to an optimization problem involving the point spectrum of the operators ${A^{(\ell)}}^* A^{(\ell)}$, $\ell \in \mathbb{N}$. To this end, let

$$E^{(\ell)} := \left\{(\lambda,\eta)\in[0,1]\times\mathcal{H}^{(\ell)} \,\middle|\, {A^{(\ell)}}^*A^{(\ell)}\eta = \lambda\eta, \|\eta\| = 1\right\}.$$

Suprema (and infima) over empty sets are to be read as zero.

Theorem II.4. *We have, for every $f \in \mathcal{H}$ and every $N \in \mathbb{N}$,*

$$W_N(f) \leq \left(\|f\|^2 - \left\|A^{(0)}f\right\|^2\right)\cdot\prod_{\ell=1}^{N-1}\left(1 - C^{(\ell)}\right),$$

where

$$C^{(\ell)} := \inf_{g\in\mathcal{R}^{(\ell)}:\|g\|=1}\ \sup_{(\lambda,\eta)\in E^{(\ell)}} |\langle g,\eta\rangle|^2\cdot\lambda \geq 0. \tag{6}$$

Proof. The proof is given in Appendix IV-A. □

Remark II.5. *Note that $\lambda, C^{(\ell)} \in [0,1]$, by Assumption II.1.*

While the case $E^{(\ell)} = \emptyset$ may occur in the very general case, we have $E^{(\ell)} \neq \emptyset$ in many practically relevant settings. In fact, the setting is inspired by that of scattering CNNs over general measure spaces [13], where $A^{(\ell)} = H^t$ is the diffusion operator defined in [13, Equ. (8),(10)] for some $t > 0$, so $E^{(\ell)} = \{(g(\lambda_k)^{2t}, \varphi_k) \mid k \in I\}$ in the notation of [13].

Finally, we note that there are feature extractors for which the generic bound obtained by this approach is sharp, cf. [11].

Theorem II.4 may—at first glance—seem like an artificial reduction to another difficult problem, namely bounding the constant $C^{(\ell)}$ from below. However, at least for some types of specific feature extractors (respectively, their underlying signal spaces), it is actually easy to do so, as we demonstrate below. Indeed, Theorem II.4 generalizes the ideas of [11, Proposition 3.3] that led the authors to conclude exponential energy decay for certain graph CNNs. Likewise, we conclude exponential energy decay for general feature extractors if the measure space does not contain sets of arbitrarily small positive measure, which recovers [11, Proposition 3.3] as a special case of our next corollary[3].

[3]It can also be seen to recover [12, Theorem 3.4] with a minor adjustment in our setting.

Corollary II.6. *Assume that $C_{\mathcal{M}^{(\ell)}} := \inf_m\ \mu^{(\ell)}(m) > 0$, where the infimum is taken over all sets $m \in \mathfrak{M}^{(\ell)}$ for which $\mu^{(\ell)}(m) > 0$. Suppose further that $\sigma^{(\ell)} f \geq 0$ ($\mu^{(\ell)}$-a.e.) for all $f \in \mathcal{H}^{(\ell)}$ and that there exists a pair $(\lambda,\eta) \in E^{(\ell)}$ such that $\lambda^{1/2}\eta \geq C_{A^{(\ell)}}^{1/2}$ holds ($\mu^{(\ell)}$-a.e.) for a constant $C_{A^{(\ell)}} > 0$. Then, we have*

$$C^{(\ell)} \geq C_{\mathcal{M}^{(\ell)}} C_{A^{(\ell)}} > 0.$$

In particular, if this holds for the first $\ell = 1,\dots,N$ layers of the neural network, then we have, for all $f \in \mathcal{H}$,

$$W_N(f) \leq \left(\|f\|^2 - \left\|A^{(0)}f\right\|^2\right)\cdot\prod_{\ell=1}^{N-1}\left(1 - C_{\mathcal{M}^{(\ell)}}C_{A^{(\ell)}}\right).$$

Proof. The proof is given in Appendix IV-B. □

Remark II.7. *The assumptions of Corollary II.6 imply that $\mu^{(\ell)}(M^{(\ell)}) < \infty$. Also note that the derived lower bound is invariant under rescaling of the measure, which is reasonable as this leaves the structure of the feature extractor invariant.*

Let us now continue the discussion from above regarding the concreteness and applicability of our results. First, note that the constraint on the measure space is satisfied for the counting measure, in which case $C_{\mathcal{M}^{(\ell)}} = 1$. This may even be regarded as the most practically relevant measure space, having a feature engineer in mind who works with whatever feature extractor on a digital computer. Second, there is a whole class of feature extractors, for which there exists a constant eigenfunction, $\eta \equiv C_{A^{(\ell)}}^{1/2} = \mu^{(\ell)}(M^{(\ell)})^{-1/2}$. Indeed, for scattering CNNs defined via diffusion operators as in [13, Equ. (8)], the eigenfunction φ_0 of $\mathcal{L}$ (in the notation of [13]) is often constant [13, Remark 3], and we have $(1,\varphi_0) \in E^{(\ell)}$ in the context of the above corollary; see also [13, Theorems 2, 5] and [5, Section 5.1], which are based on similar assumptions.

Finally, the last two arguments are precisely the reason why Corollary II.6 qualitatively[4] recovers [11, Proposition 3.3], also implying exponential energy decay for arbitrary input signals of the graph scattering CNNs described in [11].

III. Information retention in scattering CNNs over LCA groups

A. LCA group scattering

In this section, we demonstrate how structural information about the signal domain can be exploited to derive explicit estimates for the energy propagation in LCA group scattering CNNs. The tools for the following results rely on the properties of LCA groups and their Fourier analysis; a comprehensive overview of which can be found in, e.g., [20].

Let G be an LCA group, and let $\widehat{G}$ be its Pontryagin dual. Fix a Haar measure μ_G on G. For $f \in L^1(G)\cap L^2(G)$, we define the Fourier transform of f by

$$\mathcal{F}(\xi) := \widehat{f}(\xi) := \int_G f(x)\overline{\xi(x)}\,\mathrm{d}\mu_G(x), \quad \xi \in \widehat{G}.$$

[4]There is a difference in the resulting basis of the exponential decay: [11] guarantees $C^{(\ell)} \geq 2/\#G$, while our result only implies the slightly weaker bound $C^{(\ell)} \geq 1/\#G$. This gap ultimately stems from a low-pass condition, which is assumed in [11, Equ. (11)], while our Corollary II.6 does not require an analog thereof.

In the following, we always use the unique Haar measure $\mu_{\widehat{G}}$ on $\widehat{G}$ such that the Fourier transform extends unitarily to an operator $\mathcal{F} : L^2(G) \to L^2(\widehat{G})$.

Throughout this section, let $\Psi \cup \{\chi\} \subseteq L^1(G) \cap L^2(G)$ be a semi-discrete Parseval frame, i.e., an at most countable family of convolution filters that satisfy

$$\forall h \in L^2(G): \ \|h * \chi\|^2 + \sum_{\psi\in\Psi} \|h * \psi\|^2 = \|h\|^2, \quad (7)$$

which is equivalent to the Littlewood-Paley condition,

$$|\widehat{\chi}(\xi)|^2 + \sum_{\psi\in\Psi} |\widehat{\psi}(\xi)|^2 = 1, \quad \text{a.e. } \xi \in \widehat{G}. \quad (8)$$

For $g \in L^1(G) \cap L^2(G)$, let $\mathcal{C}_g \in \mathcal{B}(L^2(G))$ be the convolution operator $f \mapsto f * g$. The filters $\Psi \cup \{\chi\}$ induce a scattering CNN [5], [16], with each layer sharing the same architecture,

$$\mathcal{H}^{(\ell)} := L^2(G), \ \mathcal{L}^{(\ell)} := \{\mathcal{C}_\psi \mid \psi \in \Psi\}, \ A^{(\ell-1)} := \mathcal{C}_\chi,$$

and the pointwise modulus $\sigma^{(\ell)}(f) := |f|$ for $f \in L^2(G)$ as its nonlinearity. By (7), Assumption II.1 and hence (2) hold with equality. By Proposition II.2, the scattering transform associated with these filters is a well-defined, norm-decreasing (nonlinear) operator $S : L^2(G) \to \ell^2\left(L^2(G)\right)$.

B. A setting-specific rate for energy decay

The results in this section are based on scattering-specific ideas from [5], [9], [11], which we combine with analogs for sumset-estimates based on ideas from [15], [21, Section 2.4].

The modulus shifts part of the frequency content of a signal towards a neighborhood of $\mathbb{1}_G \in \widehat{G}$. It is hence natural to assume that the output-generating filter χ satisfies a low-pass condition (which, by (8), is equivalent to a high-pass condition on the filters Ψ) with the intention that this makes the scattering CNN more informative [5], [7], [9].

Assumption III.1. *Let $\Gamma_\Psi := \widehat{G} \setminus \bigcup_{\psi\in\Psi}\{\xi \in \widehat{G} \mid \widehat{\psi}(\xi) \neq 0\}$. We assume that Γ_Ψ is a neighborhood of $\mathbb{1}_G \in \widehat{G}$.*

The next theorem confirms the intuition that the extracted features tend to carry more energy when the frequency gap Γ_Ψ is large and the Fourier supports of Ψ are uniformly small. Part of this stems from the low-pass filter χ capturing more signal information, though the full picture is more nuanced.

Theorem III.2. *Let G be compact. If there exists $S \in \mathbb{N}$ with $\#\operatorname{supp}(\widehat{\psi}) \leq S$ ($\psi \in \Psi$), then, for all $f \in L^2(G)$, $N \in \mathbb{N}$,*

$$W_N(f) \leq \left(\|f\|^2 - \|f * \chi\|^2\right) \cdot (1 - 1/S)^{N-1}.$$

If G is finite, one can choose $S \leq \#G - \#\Gamma_\Psi < \infty$.

Proof. The proof is given in Appendix IV-C. □

Remark III.3. *The proof of Theorem III.2 makes use of the (accessible) Fourier analysis for abelian (compact) groups, whose analogs are more involving in the nonabelian case. If G is finite nonabelian, Corollary II.6 still ensures exponential energy decay, albeit with a slightly worse basis compared with the abelian case. In fact, under the low-pass condition $|\int \chi \, \mathrm{d}\mu_G|^2 = 1$, taking $(\lambda, \eta) = (1, \mu_G(G)^{-1/2} \cdot \mathbb{1}_G)$ in Corollary II.6 yields $\chi * \chi^* * \eta = \eta \geq \mu_G(G)^{-1/2}$. Since $\mu_G(\{e_G\})/\mu_G(G) = 1/\#G$, we have, for all $f \in \mathbb{C}^G$, $N \in \mathbb{N}$,*

$$W_N(f) \leq \left(\|f\|^2 - \|f * \chi\|^2\right) \cdot (1 - 1/\#G)^{N-1}.$$

Theorem III.4. *Let G be any LCA group. There exists an open neighborhood $\Gamma = \Gamma^{-1} \subseteq \widehat{G}$ of $\mathbb{1}_G \in \widehat{G}$ with $\Gamma^8 \subseteq \Gamma_\Psi$. For any such Γ we have, for all $f \in L^2(G)$, $N \in \mathbb{N}$,*

$$W_N(f) \leq \left(\|f\|^2 - \|f * \chi\|^2\right) \cdot \alpha(\Psi, \Gamma)^{N-1}, \quad (9)$$

where

$$\alpha(\Psi, \Gamma) = 1 - \frac{\mu_{\widehat{G}}\left(\Gamma^2\right)^2}{\mathcal{N}(\Psi, \Gamma^2) \cdot \mu_{\widehat{G}}\left(\Gamma^4\right)^2} \in [0, 1],$$

and $\mathcal{N}(\Psi, \Gamma^2)$ denotes the smallest number $n \in \mathbb{N} \cup \{\infty\}$ with the property that for any $\psi \in \Psi$ there exist $\xi_1, \ldots, \xi_n \in \widehat{G}$ such that $\operatorname{supp}(\widehat{\psi}) \subseteq \bigcup_{k=1}^n \xi_k \Gamma^2$. Furthermore, it holds that

$$\mathcal{N}(\Psi, \Gamma^2) \leq \sup_{\psi\in\Psi} \left\lfloor \frac{\mu_{\widehat{G}}(\Gamma \cdot \operatorname{supp}(\widehat{\psi}))}{\mu_{\widehat{G}}\left(\Gamma\right)} \right\rfloor. \quad (10)$$

Thus, $\alpha(\Psi, \Gamma) < 1$ if $\sup_{\psi\in\Psi} \mu_{\widehat{G}}(\Gamma \cdot \operatorname{supp}(\widehat{\psi})) < \infty$, i.e., we then have exponential energy decay globally on $L^2(G)$. In particular, if G is discrete, exponential energy decay holds globally on $L^2(G)$.

Proof. The proof is given in Appendix IV-D. □

Remark III.5. *Theorem III.2, Remark III.3, and Theorem III.4 corroborate the experimental reports (e.g., [16], [17]) that the energy decays exponentially with increasing network depth in digital implementations of scattering CNNs, which is even reinforced by their measure-theoretic analog, Corollary II.6. However, the fact that the digital implementation of any feature extractor necessarily deals with discrete finite signals seems to play a significant role for its energy distribution. In fact, the recent paper [10] shows that the energy decay can even be arbitrarily slow in certain scattering CNNs over the Euclidean domain $\mathbb{R}^d$. Note that this is no contradiction to the findings in this paper, even when considering the digitial implementations of scattering CNNs as suitable discretizations of their Euclidean-domain counterparts: The basis of the exponential decay guaranteed by, e.g., Theorem III.2, converges to 1 if $\sup_{\psi\in\Psi} \#\operatorname{supp}(\widehat{\psi})$ increases at the order of $\#G$ as the size of the group tends to infinity.*

Remark III.6. *Theorem III.4 generalizes [9, Proposition 3.3] (arbitrary LCA groups, explicit bound for the decay basis).*

Our results contribute to the understanding of the energy distribution in deep feature extractors and, as a by-product, partially, positively answer the question of the applicability of a stability result for scattering CNNs posed in [13, Remark 7].

ACKNOWLEDGMENT

The author would like to extend his heartfelt thanks to Hartmut Führ for insightful discussions and careful reading of the manuscript.

References

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 8, pp. 1798–1828, 2013.

[2] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," science, vol. 313, no. 5786, pp. 504–507, 2006.

[3] S. Mallat, "Understanding deep convolutional networks," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2065, p. 20150203, 2016.

[4] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in 2007 IEEE conference on computer vision and pattern recognition. IEEE, 2007, pp. 1–8.

[5] S. Mallat, "Group invariant scattering," Communications on Pure and Applied Mathematics, vol. 65, no. 10, pp. 1331–1398, 2012.

[6] I. Waldspurger, "Exponential decay of scattering coefficients," in 2017 international conference on sampling theory and applications (SampTA). IEEE, 2017, pp. 143–146.

[7] T. Wiatowski, P. Grohs, and H. Bölcskei, "Energy propagation in deep convolutional neural networks," IEEE Transactions on Information Theory, vol. 64, no. 7, pp. 4819–4842, 2017.

[8] ——, "Topology reduction in deep convolutional feature extraction networks," in Wavelets and Sparsity XVII, vol. 10394. SPIE, 2017, pp. 269–280.

[9] W. Czaja and W. Li, "Analysis of time-frequency scattering transforms," Applied and Computational Harmonic Analysis, vol. 47, no. 1, pp. 149–171, 2019.

[10] H. Führ and M. Getter, "Energy propagation in scattering convolution networks can be arbitrarily slow," arXiv preprint arXiv:2406.05121, 2024.

[11] D. Zou and G. Lerman, "Graph convolutional neural networks via scattering," Appl. Comput. Harmon. Anal., vol. 49, no. 3, pp. 1046–1074, 2020.

[12] M. Perlmutter, A. Tong, F. Gao, G. Wolf, and M. Hirn, "Understanding graph neural networks with generalized geometric scattering transforms," SIAM Journal on Mathematics of Data Science, vol. 5, no. 4, pp. 873–898, 2023.

[13] J. Chew, M. Hirn, S. Krishnaswamy, D. Needell, M. Perlmutter, H. Steach, S. Viswanath, and H.-T. Wu, "Geometric scattering on measure spaces," Applied and Computational Harmonic Analysis, p. 101635, 2024.

[14] T. Wiatowski and H. Bölcskei, "A mathematical theory of deep convolutional neural networks for feature extraction," IEEE Transactions on Information Theory, vol. 64, no. 3, pp. 1845–1866, 2017.

[15] I. Ruzsa, "An analog of freiman's theorem in groups," Astérisque, vol. 258, no. 199, pp. 323–326, 1999.

[16] J. Bruna and S. Mallat, "Invariant scattering convolution networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1872–1886, 2013.

[17] J. Andén and S. Mallat, "Multiscale scattering for audio classification." in ISMIR. Miami, Florida, 2011, pp. 657–662.

[18] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," nature, vol. 323, no. 6088, pp. 533–536, 1986.

[19] B. Bah, H. Rauhut, U. Terstiege, and M. Westdickenberg, "Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers," Information and Inference: A Journal of the IMA, vol. 11, no. 1, pp. 307–353, 2022.

[20] W. Rudin, Fourier analysis on groups. Courier Dover Publications, 2017.

[21] T. Tao and V. H. Vu, Additive combinatorics. Cambridge University Press, 2006, vol. 105.

IV. Appendix

A. Proof of Theorem II.4

By (5), it just remains to prove that $\iota_\ell \geq C^{(\ell)}$, $\ell \in \mathbb{N}$. This is trivial if $E^{(\ell)} = \emptyset$ or $\mathcal{R}^{(\ell)} = \{0\}$, recalling that these cases are to be understood as $C^{(\ell)} = 0$. For the other case, choose any $g \in \mathcal{R}^{(\ell)}$ with $\|g\| = 1$ and $(\lambda, \eta) \in E^{(\ell)}$. Since η is normalized, there exists an orthonormal basis $\{\eta\} \dot\cup \Upsilon$ of $\mathcal{H}^{(\ell)} \supseteq \mathcal{R}^{(\ell)}$, which allows us to write

$$A^{(\ell)} g = \langle g, \eta \rangle A^{(\ell)} \eta + \sum_{\upsilon \in \Upsilon} \langle g, \upsilon \rangle A^{(\ell)} \upsilon,$$

where the series converges unconditionally in $\mathcal{H}^{(\ell)}$. Since $(\lambda, \eta) \in E^{(\ell)}$, we find that

$$\left\langle A^{(\ell)} \eta, A^{(\ell)} \upsilon \right\rangle = \left\langle {A^{(\ell)}}^* A^{(\ell)} \eta, \upsilon \right\rangle = \langle \lambda \eta, \upsilon \rangle = 0. \quad (11)$$

Replacing υ by η in (11) also shows that $\left\| A^{(\ell)} \eta \right\|^2 = \lambda$. Consequently, the Pythagorean theorem gives

$$\begin{aligned} \left\| A^{(\ell)} g \right\|^2 &= \left\| \langle g, \eta \rangle A^{(\ell)} \eta \right\|^2 + \left\| \sum_{\upsilon \in \Upsilon} \langle g, \upsilon \rangle A^{(\ell)} \upsilon \right\|^2 \\ &\geq |\langle g, \eta \rangle|^2 \cdot \lambda. \end{aligned} \quad (12)$$

Taking the supremum over all pairs $(\lambda, \eta) \in E^{(\ell)}$ on both sides of (12) and inserting this into (4) concludes the proof.

B. Proof of Corollary II.6

Any $g \in \mathcal{R}^{(\ell)} \subseteq \mathcal{R}(\sigma^{(\ell)})$ satisfies $g \geq 0$ ($\mu^{(\ell)}$-a.e.) by assumption on $\sigma^{(\ell)}$. Thus,

$$\begin{aligned} |\langle g, \eta \rangle|^2 \cdot \lambda &= \left(\int g \lambda^{1/2} \eta \, \mathrm{d}\mu^{(\ell)} \right)^2 \\ &\geq C_{A^{(\ell)}} \cdot \|g\|^2_{L^1(\mathcal{M}^{(\ell)})} \\ &\geq C_{A^{(\ell)}} \cdot C_{\mathcal{M}^{(\ell)}} \cdot \|g\|^2_{L^2(\mathcal{M}^{(\ell)})}, \end{aligned} \quad (13)$$

where (13) is due to the embedding $L^1(\mathcal{M}^{(\ell)}) \hookrightarrow L^2(\mathcal{M}^{(\ell)})$, which itself is a well-known consequence from the assumption that $C_{\mathcal{M}^{(\ell)}} > 0$. Comparing with (6), we conclude

$$C^{(\ell)} \geq C_{A^{(\ell)}} \cdot C_{\mathcal{M}^{(\ell)}}.$$

The upper bound for $W_N(f)$ now follows from Theorem II.4.

C. Proof of Theorem III.2

Any $g \in \mathcal{R}^{(N)}$ can be written as $g = |f * \psi|$ for some $f \in L^2(G)$, $\psi \in \Psi$. Thus,

$$\iota_N \geq \inf_{f \in L^2(G), \psi \in \Psi : \|f * \psi\| = 1} \||f * \psi| * \chi\|^2.$$

By (5), it suffices to show that, for all $f \in L^2(G)$, $0 \neq \psi \in \Psi$,

$$\||f * \psi| * \chi\|^2 \geq (\# \operatorname{supp}(\widehat{\psi}))^{-1} \cdot \|f * \psi\|^2. \quad (14)$$

The Littlewood-Paley condition (8) entails that $|\widehat{\chi}| \equiv 1$ holds on Γ_Ψ. In particular, $|\widehat{\chi}| \geq \delta_{\mathbb{1}_G}$ holds $\mu_{\widehat{G}}$-a.e., and hence, by Parseval's theorem and the convolution theorem,

$$\||f * \psi| * \chi\|^2 = \|\mathcal{F}(|f * \psi|) \cdot \widehat{\chi}\|^2 \geq \|\mathcal{F}(|f * \psi|) \cdot \delta_{\mathbb{1}_G}\|^2. \quad (15)$$

The RHS of (15) equals

$$\mu_{\widehat{G}}(\{\mathbb{1}_G\}) \cdot |\mathcal{F}(|f * \psi|)(\mathbb{1}_G)|^2 = \mu_{\widehat{G}}(\{\mathbb{1}_G\}) \cdot \|f * \psi\|_1^2,$$

and we have $\mathcal{F}(f * \psi) \equiv 0$ on $\widehat{G} \setminus \operatorname{supp}(\widehat{\psi})$, which entails that

$$\|f * \psi\|_1^2 \geq \|\mathcal{F}(f * \psi)\|_\infty^2 \geq \frac{\|\mathcal{F}(f * \psi)\|^2}{\mu_{\widehat{G}}(\{\mathbb{1}_G\}) \cdot \# \operatorname{supp}(\widehat{\psi})}.$$

Finally, applying Parseval's theorem again, and putting all together, we conclude (14); hence, the proof of the theorem.

D. Proof of Theorem III.4

Similar to the Proof of Theorem III.2, it suffices to show that, for all $f \in L^2(G)$, $\psi \in \Psi$,

$$\||f * \psi| * \chi\|^2 \geq \frac{\mu_{\widehat{G}}(\Gamma^2)^2}{\mathcal{N}(\Psi, \Gamma^2) \cdot \mu_{\widehat{G}}(\Gamma^4)^2} \cdot \|f * \psi\|^2. \quad (16)$$

Let $\widehat{\phi} := (\mathbb{1}_{\Gamma^4} * \mathbb{1}_{\Gamma^4}) / \mu_{\widehat{G}}(\Gamma^4)$. This defines an auxiliary function $\phi \in L^2(G)$, which satisfies both $\phi \geq 0$ and $|\widehat{\chi}| \geq \widehat{\phi}$. By the symmetry of Γ, we further have

$$\widehat{\phi}\big|_{\Gamma^2} \geq \frac{\mu_{\widehat{G}}(\Gamma^2)}{\mu_{\widehat{G}}(\Gamma^4)}. \quad (17)$$

Let $n := \mathcal{N}(\Psi, \Gamma^2)$. There are $\xi_1, \ldots, \xi_n \in \widehat{G}$ such that $\operatorname{supp}(\widehat{\psi}) \subseteq \bigcup_{k=1}^n \xi_k \Gamma^2$. By Parseval's theorem and the convolution theorem,

$$\||f * \psi| * \chi\|^2 = \|\mathcal{F}(|f * \psi|) \cdot \widehat{\chi}\|^2 \geq \left\| \mathcal{F}(|f * \psi|) \cdot \widehat{\phi} \right\|^2. \quad (18)$$

Employing (17) and the nonnegativity of ϕ, we obtain

$$\begin{aligned} (18) &= \||f * \psi| * |\xi_k \cdot \phi|\|^2 \\ &\geq \|f * \psi * (\xi_k \cdot \phi)\|^2 \\ &\geq \int_{\xi_k \Gamma^2} |\widehat{f}(\xi)|^2 \cdot |\widehat{\psi}(\xi)|^2 \cdot |\widehat{\phi}(\xi_k^{-1} \cdot \xi)|^2 \, \mathrm{d}\mu_{\widehat{G}}(\xi) \\ &\geq \left(\frac{\mu_{\widehat{G}}(\Gamma^2)}{\mu_{\widehat{G}}(\Gamma^4)} \right)^2 \int_{\xi_k \Gamma^2} |\widehat{f}(\xi)|^2 \cdot |\widehat{\psi}(\xi)|^2 \, \mathrm{d}\mu_{\widehat{G}}(\xi). \end{aligned} \quad (19)$$

Summing over k, and combining (18) and (19) yields

$$\begin{aligned} & n \cdot \||f * \psi| * \chi\|^2 \\ &\geq \left(\frac{\mu_{\widehat{G}}(\Gamma^2)}{\mu_{\widehat{G}}(\Gamma^4)} \right)^2 \int_{\bigcup_{k=1}^n \xi_k \Gamma^2} |\widehat{f}(\xi)|^2 \cdot |\widehat{\psi}(\xi)|^2 \, \mathrm{d}\mu_{\widehat{G}}(\xi) \\ &\geq \left(\frac{\mu_{\widehat{G}}(\Gamma^2)}{\mu_{\widehat{G}}(\Gamma^4)} \right)^2 \int_{\operatorname{supp}(\widehat{\psi})} |\widehat{f}(\xi)|^2 \cdot |\widehat{\psi}(\xi)|^2 \, \mathrm{d}\mu_{\widehat{G}}(\xi). \end{aligned}$$

Since the last integral equals $\|f * \psi\|^2$, rearranging terms concludes the proof of (16).

Finally, (10) follows from a straightforward generalization of Ruzsa's covering lemma for sumsets (cf. [15] and [21, Lemma 2.14]) to LCA groups.

The existence of Γ and the conclusion for G discrete (i.e., $\widehat{G}$ compact) follow from the standard theory for LCA groups.