Which Frequencies do CNNs Need? Emergent Bottleneck Structure in Feature Learning

Yuxiao Wen [1] **Arthur Jacot** [1]

Abstract

We describe the emergence of a Convolution Bottleneck (CBN) structure in CNNs, where the network uses its first few layers to transform the input representation into a representation that is supported only along a few frequencies and channels, before using the last few layers to map back to the outputs. We define the CBN rank, which describes the number and type of frequencies that are kept inside the bottleneck, and partially prove that the parameter norm required to represent a function f scales as depth times the CBN rank f. We also show that the parameter norm depends at next order on the regularity of f. We show that any network with almost optimal parameter norm will exhibit a CBN structure in both the weights and - under the assumption that the network is stable under large learning rate - the activations, which motivates the common practice of down-sampling; and we verify that the CBN results still hold with down-sampling. Finally we use the CBN structure to interpret the functions learned by CNNs on a number of tasks.

1. Introduction

Convolutional Neural Networks (CNNs) have played a key role in the success of deep learning (Lecun et al., 1998; Krizhevsky et al., 2012). It seems that the structure of CNNs is particularly well adapted to tasks on natural images. But we still lack a description of this structure, though many theories have been proposed.

The most common explanation, is that some fundamental properties of natural images are encoded in the structure of CNNs, such as translation invariance and locality. These intuitions have motivated special network architectures that encode additional properties such as rotation symmetries (Cohen et al., 2019), or the design of feature maps such as the scattering transform (Mallat, 2012) that encode similar symmetries, upon which more traditional statistical models can then be used.

A CNN at initialization gives rise to features and kernels, either the Neural Network Gaussian Process (NNGP) kernel (Neal, 1996; Cho & Saul, 2009) or the Neural Tangent Kernel (NTK) (Jacot et al., 2018). The symmetries and invariances enforced by the locality, weight-sharing and pooling of CNNs are reflected in the kernels (Bietti & Mairal, 2019; Arora et al., 2019; Mei et al., 2021; Misiakiewicz & Mei, 2022), thus reducing the intrinsic dimension of the task and improving generalization (Mei et al., 2021; Misiakiewicz & Mei, 2022).

While the aforementioned results rely on a connection between fully-connected neural networks (FC-NNs) and kernel methods, other results have shown that the inductive bias coming from the CNN architecture is much more general, and applies to any training method that satisfies some reasonable property such as rotation equivariance (Li et al., 2020b; Xiao & Pennington, 2022; Wang & Wu, 2023).

But even those expertly designed kernel and features fail in general to match the performances of CNNs (Arora et al., 2019; Li et al., 2019). A possible explanation is that feature learning allows CNNs to identify low-dimensional structures in the task during training, thus further reducing the dimensionality of the task, beyond the dimension reduction that is enforced by the CNN architecture. This is supported by the empirical observation that CNNs can learn additional symmetries during training (Petrini et al., 2021).

While there is a large literature of empirical analysis of features learned by CNNs (Karantzas et al., 2022) there remains very little theoretical work outside of linear CNNs (Dai et al., 2021).

The appearance of low-dimensional features and symmetry learning has already been observed in FC-NNs (Jacot, 2023a;b). This paper extends these results to CNNs, showing a very similar bottleneck structure, though with some important differences resulted from the CNN architecture,

[1] Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA. Correspondence to: Yuxiao Wen <yuxiaowen@nyu.edu>, Arthur Jacot <arthur.jacot@nyu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

in particular the translation invariance and pooling.

1.1. Bottleneck Structure in CNNs

Recent papers (Jacot, 2023a;b) have observed a bottleneck structure in L_2-regularized FC-NNs, where the representation learned in the middle layers are low-dimensional, which implies a bias towards learning symmetries.

In this paper, we extend most of the results in (Jacot, 2023a;b) to CNNs. An important distinction is that instead of the FC-NN bottleneck structure which favors learning any type of low-dimensional representations in the middle of the network, CNN favor representations that are supported along a finite number of frequencies, with an additional preference towards lower frequencies due to the existence of pooling:

- We decompose the representation cost $R(f;\Omega,L)$ (Gunasekar et al., 2018b) of CNNs, which describes the implicit bias of CNNs with L_2-regularization, as:

$$R(f;\Omega,L) = LR^{(0)}(f;\Omega) + R^{(1)}(f;\Omega) + o(1).$$

- We conjecture (and partially prove) that the first term $R^{(0)}$ equals the so-called Convolution Bottleneck rank $\mathrm{Rank}_{\mathrm{CBN}}$, which is small for functions f that can be decomposed as first mapping to a representation that is supported along a finite number of frequencies, with a preference for lower frequencies in the presence of pooling, and then mapping back to the outputs (that may be high dimensional and high frequency).
- The second term $R^{(1)}$ plays a complementary role as a measure of regularity that bounds the Jacobian of f.
- We show that under some conditions, almost all weight matrices W_ℓ of the network will have a few large singular values, matching the frequencies that are kept in the CBN-rank decomposition. Also, under the additional assumption that the parameters are stable under reasonable learning rate, one can show that the activations are also supported on the same few frequencies.
- The emergence of this bottleneck structure, where the middle representation of the network are only supported along a few low frequencies, motivates the use of down sampling, as is commonly done in practice. We show that for functions that accept such a low-frequency hidden structure, the $R^{(0)}$ term is unaffected by down-sampling in the middle of the network.

The low-dimensionality and low-frequency of the representations inside the bottleneck makes them highly interpretable. We illustrate this with a set of numerical experiments in Section 6.

2. Preliminaries

In this section, we first formally define the convolution operation in CNNs and related notations to express convolution in the form of matrix multiplication. Then we define the parameterization of the CNNs and their representation cost.

2.1. Convolution in Matrix Form

For any $a, b \in \mathbb{R}^n$, we define the (cyclic) convolution $a * b$ by

$$(a * b)_i \equiv \sum_{j=1}^{n} a_j b_{i-1+j \mod n}, \quad i = 1, \dots, n.$$

The cross-channel convolution typically used in CNNs with input $x \in \mathbb{R}^{n\times c_1}$ and filter $w \in \mathbb{R}^{n\times c_2\times c_1}$ are denoted by $w \circledast x \in \mathbb{R}^{n\times c_2}$ and defined as follows:

$$(w \circledast x)_{:,k} = \sum_{s=1}^{c_1} w_{:,k,s} * x_{:,s}, \quad k = 1, \dots, c_2.$$

Note that $a * b = Ab$ with the circulant matrix

$$A = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \\ a_n & a_1 & \cdots & a_{n-1} \\ \vdots & & & \vdots \\ a_2 & a_3 & \cdots & a_1 \end{bmatrix}.$$

For the cross-channel convolution, we can also define its equivalent matrix representation by $W \in \mathbb{R}^{nc_2\times nc_1}$ with $W_{i,k;1,s} = w_{i,k,s}$ and $W_{i+p,k;j+p,s} = W_{i,k;j,s}$ for $i, j, p \in [n]$ and $k, s \in [c]$, where the addition is taken modulo n. One can verify that for $x \in \mathbb{R}^{n\times c_1}$,

$$(Wx)_{:,k} = \sum_{s=1}^{c_1} w_{:,k,s} * x_{:,s}, \quad \text{for } k \in [c_2].$$

Let $F_n \in \mathbb{C}^{n\times n}$ be the discrete Fourier transform (DFT) matrix in n dimension, i.e. $(F_n)_{i,j} = \frac{1}{\sqrt{n}}\omega_n^{(i-1)(j-1)}$ where $\omega_n = e^{2\pi i/n}$. Note that $F_n a$ gives the DFT coefficients of $a \in \mathbb{R}^n$. Also, $a * b = \sqrt{n}F_n^*\mathrm{diag}(F_n a)F_n b$ where $F_n^* = F_n^{-1}$ denotes the conjugate transpose. With these matrix representations and results mentioned above, we may view convolutions as linear transformations in the Fourier domain and apply standard linear algebra results in the proofs.

2.2. Network Parameterization

In this paper, we consider the following parameterization of CNNs: let $x \in \Omega \subseteq \mathbb{R}^{n\times c_{in}}$ be the input where Ω is a compact subset, n be the input size, and c_{in} the number of input channels. We adopt the index convention that $A_{:,j}$ denotes the j-th column of a matrix A and similarly for vectors and tensors. For an L-layer CNN, for $\ell = 0, 1, \dots, L-1$, the activation $\alpha_\ell(x) \in \mathbb{R}^{n\times c_\ell}$ at the ℓ-th layer is defined recursively by

$$\begin{aligned}\alpha_0(x) &= x\\ \tilde{\alpha}_\ell(x) &= \mathbf{1}b_\ell^T + w_\ell \circledast \alpha_{\ell-1}(x)\\ \alpha_\ell(x)_{:,c} &= \sigma(m * \tilde{\alpha}_\ell(x)_{:,c}), \quad c = 1,\dots,c_\ell\end{aligned}$$

where $w_\ell \in \mathbb{R}^{n\times c_\ell \times c_{\ell-1}}$ are the weight filters, $b_\ell \in \mathbb{R}^{c_\ell}$ the biases, $\mathbf{1} \in \mathbb{R}^n$ the all-one vector, $m \in \mathbb{R}^n$ a user-specified pooling filter applied to each channel, and nonlinearity $\sigma = \mathrm{ReLU}$. The last layer is linear:

$$\alpha_L(x) = \tilde{\alpha}_L(x) = \mathbf{1}b_L^T + w_L \circledast \alpha_{L-1}(x).$$

As remarked in Section 2.1, we write instead

$$\alpha_\ell(x) = \sigma\left(M(W_\ell \alpha_{\ell-1}(x) + \mathbf{1}b_\ell^T)\right)$$

and focus on this matrix representation in the rest of this work. CNNs with this parameterization is naturally translationally equivariant, and discussion on its universality is deferred to Appendix A.

2.3. Representation Cost

The representation cost of a function f is the minimum norm of the parameter θ for a depth-L CNN f_θ to represent it over the input domain:

$$R(f;\Omega,L) = \min_{f_\theta|\Omega = f|\Omega} \|\theta\|^2$$

where the minimum is taken over all possible parameters $\theta = (W_1, b_1, \dots, W_L, b_L)$ with $f_\theta(x) = f(x)\ \forall x \in \Omega$. We let $R(f;\Omega,L) = \infty$ if no such parameter exists. This representation cost describes the natural bias on the optimized CNN representation induced by introducing the L_2 regularization on the parameter θ for arbitrary training cost function $\mathcal{L}$:

$$\min_\theta \mathcal{L}(f_\theta) + \lambda\|\theta\|^2 = \min_{f\in\mathcal{N}_m} \mathcal{L}(f) + \lambda R(f;\Omega,L) \qquad (1)$$

where $\mathcal{N}_m$ denotes the set of all translationally equivariant piecewise linear (TEPL) functions that can be represented by a CNN on Ω with pooling filter m.

Remark 2.1. Another natural definition for the representation cost is using the norm of the convolution filters w_ℓ instead of the matrix representation W_ℓ. This only changes the parameter norm by a constant factor $\|w_\ell\|_F^2 = \frac{1}{n}\|W_\ell\|_F^2$ so that the result presented in this paper can easily be adapted to this other setting. Detailed discussion on adaptation to the filter norm is left in Appendix F.

3. Large Depth Representation Cost

Our goal is to describe the bottleneck structure that appears in deep CNNs trained with L_2-regularization, e.g. Figure 2, where the weight matrices in the middle layers of the network keep only a small number of large singular values corresponding mostly to low frequencies. This bottleneck structure affects the representation cost of large depth networks.

Our intuition, which is supported by our theoretical results, is that this structure emerges because it minimizes the 'cost of representing the identity': For large depths, most of the layers of the network will be dedicated to 'keeping information', i.e. to represent the identity (or an orthogonal transformation) on the data. To represent the identity with a small parameter norm, it is optimal for the pre-activations to be positive, so that the ReLU equals the identity on them, and to be supported along a few low frequencies, because the weight matrix W_ℓ can then be chosen so that MW_ℓ is equals the identity along these frequencies and zero orthogonal to them. More precisely if the image of $\tilde{\alpha}_{\ell,c}$ is positive and only supported along the frequencies $I_c \subset [n]$ for each channel $c = 1,\dots,c_\ell$, there we can choose W_ℓ such that $MW_\ell\sigma(\tilde{\alpha}_\ell(x)) = \tilde{\alpha}_\ell(x)$ and

$$\|W_\ell\|_F^2 = \sum_{c=1}^{c_\ell}\sum_{i\in I_c} \tilde{m}_i^{-2}. \qquad (2)$$

This we call the 'cost of identity' which is a sum over the cost $\tilde{m}_i^{-2}$ of representing each frequency i that we keep. In the absence of pooling $M = Id$ each frequency has the same cost, but for average pooling or other types of low-pass pooling, higher frequencies have a higher cost.

3.1. Convolutional Bottleneck Rank

In the infinite depth limit $L \to \infty$ almost all layers will be dedicated to 'representing the identity', and their parameter norm will be roughly as described in equation 2. It is therefore optimal for the network to map in a few layers from the input representations to a representation supported along a few low frequencies, and then use the last few layers to map back to the outputs. The TEPL functions f for which such a decomposition is possible are eactly those that have a small Convolutional Bottleneck (CBN) rank:

$$\mathrm{Rank}_{\mathrm{CBN}}(f;\Omega) := \inf_{\substack{f = h\circ g\\ g = g_1\oplus\cdots\oplus g_k}} \sum_{c=1}^{k}\sum_{i\in I_c} \tilde{m}_i^{-2}$$

where $\oplus$ denotes channel concatenation, f can be factorized into $g : \mathbb{R}^{n\times c_{in}} \to \mathbb{R}^{n\times k}$ and $h : \mathbb{R}^{n\times k} \to \mathbb{R}^{n\times c_{out}}$ for some $k \in \mathbb{N}$, $I_c \subset [n]$ denotes the supported DFT frequencies of the mapping g_c of g to the c-th channel for $c = 1,\dots,k$, and $\tilde{m} = F_n m$ gives the DFT coefficients of the pooling m.

TEPL functions f with a small CBN rank can be represented by a deep CNN with a small parameter norm:

Theorem 3.1. *For any translationally equivariant function f with finite CBN rank, there is a constant c that depends only on the target function f s.t.*

$$R(f;\Omega,L)\leq L\mathrm{Rank}_{\text{CBN}}(f;\Omega)+c.$$

Proof. We sketch the proof idea of this theorem here. Suppose $f=h\circ g$ attains the bottleneck rank. By Lemma G.1, there is a CNN with depth $L_g=\lfloor\log(nc_{in}+1)\rfloor+2$ ($L_h=\lfloor\log(nk+1)\rfloor+2$ resp.) and parameter θ_g (θ_h resp.) that represents g (h resp.). For $L\geq L_g+L_h$, we can construct the following CNN that represents f: Let the first L_g layers represent g and the last L_h layers represent h. Let the middle $L-L_g-L_h$ layers be identity layers on $g(\Omega)$. The overall parameter norm of this CNN is

$$\|\theta\|^2=\|\theta_g\|^2+\|\theta_h\|^2+(L-L_g-L_h)\sum_{c=1}^{k}\sum_{t\in I_c}\tilde{m}_t^{-2}.$$

□

If we define the limiting representation cost $R^{(0)}(f;\Omega)$ as the limit $\lim_{L\to\infty}\frac{1}{L}R(f;\Omega,L)$, then this implies the upper bound $R^{(0)}(f;\Omega)\leq\mathrm{Rank}_{\text{CBN}}(f;\Omega)$, but we conjecture that the two are actually equal. This conjecture is inspired by the fact that in numerical experiments, there is a similar bottleneck structure as the one in the proof of Theorem 3.1, suggesting that such a structure might be indeed optimal (up $o(L)$ terms).

We give the following theoretical support for our conjecture. First the $R^{(0)}$ shares a number of properties typical of a notion of rank with $\mathrm{Rank}_{\text{CBN}}$, such as

$$R^0(f_2\circ f_1;\Omega)\leq\min\{R^0(f_2;f_1(\Omega)),R^0(f_1;\Omega)\};$$
$$R^0(f_2+f_1;\Omega)\leq R^0(f_2;\Omega)+R^0(f_1;\Omega).$$

These properties as well as others are proven in Appendix B.

Second, and more importantly, we give a lower bound for $R^{(0)}(f;\Omega)$ in terms of the Jacobian $Jf(x)$ at a point x, which matches the upperbound for a large family of TEPL functions f.

This lower bound will be expressed in terms of the following pooling-dependent rank: for any translation equivariant matrix $A\in\mathbb{R}^{n\min\{c_{in},c_{out}\}\times n\min\{c_{in},c_{out}\}}$, define

$$\mathrm{Rank}_m(A)=\sum_{c=1}^{\min\{c_{in},c_{out}\}}\sum_{t=1}^{n}\tilde{m}_t^{-2}\mathbb{1}[s_{c,t}(A)\neq 0]$$

where $s_{c,t}(A)$ denotes the c-th singular value of A along the t-frequency for $c=1,\dots,\min\{c_{in},c_{out}\}$ and $t=1,\dots,n$. Note that in the absence of pooling (i.e. $m=id$), it reduces to the matrix rank $\mathrm{Rank}_{id}(A)=\mathrm{Rank}(A)$.

Theorem 3.2. *For any translationally equivariant function f, let $Jf(x)$ be the Jacobian of f at x. The following pooling-dependent lower bounds hold:*

1. $\frac{1}{\tilde{m}_{\max}^2}\max_{x\in\Omega}\mathrm{Rank}(Jf(x))\leq R^{(0)}(f;\Omega)$

   In particular, when there is no pooling, $\max_{x\in\Omega}\mathrm{Rank}(Jf(x))\leq R^{(0)}(f;\Omega)$.

2. $\max_{x\in\Omega_-}\mathrm{Rank}_m(Jf(x))\leq R^{(0)}(f;\Omega)$

   where the max is taken over the subset $\Omega_-:=\{x\in\Omega\,|\,x_{p,i}=x_{q,i}\ \forall i=1,\dots,c_{in},\forall p,q=1,\dots,n\}$, i.e. all x that are constant along each channel.

If there is a point $x\in\Omega_-$ (or $x\in\Omega$ when there is no pooling) that matches the lower bound in Theorem 3.2 and the upper bound Theorem 3.1, we prove the conjecture that $R^{(0)}=\mathrm{Rank}_{\text{CBN}}$. For example, if the target function f is a linear one-layer CNN and $\exists x\in\Omega_-$ is an interior point in Ω, by matching the upper and the lower bounds, we have

$$R^{(0)}(f;\Omega)=\mathrm{Rank}_{\text{CBN}}(f;\Omega)=\sum_{c=1}^{\min\{c_{in},c_{out}\}}\sum_{t\in I_c}\tilde{m}_t^{-2}$$

where I_c is the DFT frequencies supported by the weight filter W at the c-th output channel (see proof in Appendix B).

3.2. Finite Depth Correction

There are many functions with the same CBN rank, some more complex than others, depending on the complexity of the functions g and h. The $R^{(0)}$-term fails to capture the complexity of g and h as can be seen in the sketch of proof of Theorem 3.1, where the corresponding parameter norms $\|\theta_g\|^2$ and $\|\theta_h\|^2$ have negligible contribution to the parameter norm in contrast to the middle identity layers and do not affect $R^{(0)}$. To capture these subdominant terms, we consider the following correction term:

Definition 3.3. Define the finite depth correction term by

$$R^{(1)}(f;\Omega):=\lim_{L\to\infty}R(f;\Omega,L)-LR^0(f;\Omega).$$

This correction term $R^{(1)}$ serves as a “regularity control” on the learned CNNs:

Proposition 3.4.
1. *For any $x\in\Omega_-$, $R^{(1)}(f;\Omega)\geq 2\sum_{s_{c,t}\neq 0}\tilde{m}_t^{-2}\log(s_{c,t}\tilde{m}_t)$ with $s_{c,t}$ being the (c,t)-th singular values of $Jf(x)$ for $c=1,\dots,\min\{c_{in},c_{out}\}$ and $t=1,\dots,n$.*

2. *For all $x\in\Omega$, if there is no pooling, then $R^{(1)}(f;\Omega)\geq 2\log|Jf(x)|_+$.*

3. *If $R^{(0)}(f\circ g;\Omega)=R^{(0)}(f;g(\Omega))=R^{(0)}(g;\Omega)$, then $R^{(1)}(f\circ g;\Omega)\leq R^{(1)}(f;g(\Omega))+R^{(1)}(g;\Omega)$.*

4. *If $R^{(0)}(f+g;\Omega) = R^{(0)}(f;\Omega) + R^{(0)}(g;\Omega)$, then $R^{(1)}(f+g;\Omega) \leq R^{(1)}(f;\Omega) + R^{(1)}(g;\Omega)$.*

As shown by the first and second point of Proposition 3.4, the finite depth correction $R^{(1)}(f;\Omega)$ controls the regularity of the learned function f by upper bounding the (weighted) sum of the log singular values of the Jacobian. The third statement in Proposition 3.4 indicates that among functions with the same $R^{(0)}$ cost, their "regularity control" satisfies subadditivity.

We can rewrite the L_2-regulartized training objective in Equation 1 approximately in terms of $R^{(0)}$ and $R^{(1)}$:

$$\min_{f \in \mathcal{N}_m} \mathcal{L}(f) + \lambda L R^{(0)}(f;\Omega) + \lambda R^{(1)}(f;\Omega) \tag{3}$$

where the depth L now plays a role of balancing the rank estimation and the regularity control. If our conjecture $R^{(0)}(f;\Omega) = \mathrm{Rank}_{\mathrm{CBN}}(f;\Omega)$ holds, then we may classify the functions $f \in \mathcal{N}_m$ into subsets according to their BN-rank $R^{(0)}(f;\Omega)$, i.e. for each possible combination $I_k \in \mathcal{P}([\min\{c_{in}, c_{out}\}] \times [n])$ we can define

$$\mathcal{N}_{m,k} := \left\{ f \in \mathcal{N}_m \,:\, R^{(0)}(f;\Omega) = \sum_{(c,t) \in I_k} \tilde{m}_t^{-2} \right\}.$$

For fixed depth L and within each $\mathcal{N}_k$, the objective 3 minimizes the loss and the $R^{(1)}$ term that controls the regularity via $\min_{f \in \mathcal{N}_{m,k}} \mathcal{L}(f) + \lambda R^{(1)}(f;\Omega)$ and hence the objective itself becomes

$$\min_{k \in [K]} \left\{ \lambda L \sum_{(c,t) \in I_k} \tilde{m}_t^{-2} + \min_{f \in \mathcal{N}_k} \mathcal{L}(f) + \lambda R^{(1)}(f;\Omega) \right\}.$$

This reformulated objective suggests that for each possible bottleneck rank, indexed by $k \in [K]$, there is a regular minimizer $f_k \in \mathcal{N}_{m,k}$, and the depth L only decides which f_k is the global minimizer by trading off the bottleneck rank term and the inner minimization term (which controls the regularity of f_k). This reformulated objective suggests that as $L \to \infty$, regularized training is biased toward low-rank CNNs whose inner representations concentrate to frequencies where most information is kept (with large $\tilde{m}_t$).

4. Bottleneck Structure in Weights and Pre-Activations

Although we cannot prove the conjecture in its entirety, we are indeed able to show a bottleneck structure in the weights and the pre-activations of CNNs with sufficiently small parameter norms.

Bottleneck Structure In Weights

In the proof of Theorem 3.1, we construct a CNN where the weights in most layers support only a few frequencies. This bottleneck structure in the weights is also observed in the numerical experiments in Section 6. We show that when the parameter norm is sufficiently small, this bottleneck structure is common in the weights:

Theorem 4.1. *Suppose $\exists k > 0$ such that the parameter norm $\|\theta\|^2 \leq kL + c$ is small enough for $k = \max_{x \in \Omega_-} \mathrm{Rank}_m(Jf_\theta(x))$. Let $x_0 \in \mathrm{argmax}_{x \in \Omega_-} \mathrm{Rank}_m Jf_\theta(x)$. Then there are $V_\ell^T \in \mathbb{R}^{\kappa \times nc_{\ell-1}}$ and $U_\ell \in \mathbb{R}^{nc_\ell \times \kappa}$ being submatrices of the DFT block matrices $F_{\ell-1} \in \mathbb{R}^{nc_{\ell-1} \times nc_{\ell-1}}$ and $F_\ell^T \in \mathbb{R}^{nc_\ell \times nc_\ell}$ respectively, where $\kappa = \mathrm{Rank} Jf_\theta(x_0)$, such that*

$$\sum_{\ell=1}^{L} \|W_\ell - U_\ell S_\ell V_\ell^T\|_F^2 + \|b_\ell\|_F^2 \leq c - 2 \sum_{s_{t,c} \neq 0} \tilde{m}_t^{-2} \log(s_{t,c} \tilde{m}_t)$$

and thus for any $p \in (0,1)$, there are at least $(1-p)L$ layers ℓ with

$$\|W_\ell - U_\ell S_\ell V_\ell^T\|_F^2 + \|b_\ell\|_F^2 \leq \frac{c - 2\sum_{s_{t,c} \neq 0} \tilde{m}_t^{-2} \log(s_{t,c} \tilde{m}_t)}{pL}$$

where $s_{t,c}$ is the (t,c)-th singular value of $Jf_\theta(x_0)$ and $S_\ell \in \mathbb{R}^{\kappa \times \kappa}$ is a diagonal matrix with entries $\in \{\tilde{m}_t^{-1}\}_{t=1}^{n}$.

The assumptions on the norm $\|\theta\|^2$ and the Jacobian $Jf_\theta(x)$ in Theorem 4.1 are there to guarantee that we are in the setting where the upper and lower bounds on $R^{(0)}$ of Theorems 3.1 and 3.2 match up to a constant, and that the network represents f_θ with almost minimal parameter norm. The proof leverages the small gap between the lower and upper bound to prove the bottleneck structure (using the fact that an inequality can only be satisfied with almost equality under certain conditions). We hope that proving the conjecture that $R^{(0)} = \mathrm{Rank}_{\mathrm{CBN}}$ would make it possible to alleviate these assumptions, as there would be a lower bound that matches the upper bound for all functions instead of only some functions.

While at the minimal norm parameters θ, we know the residual term c_1 approaches and is upper bounded by $R^{(1)}(f_\theta;\Omega)$, this result also generalizes to all approximately minimal norm parameters where c_1 is still close to $R^{(1)}(f_\theta;\Omega)$. The fact that it generalizes implies that this bottleneck structure in the weights manifests in an "almost optimal" region around the optimal parameters into which the regularized objective eventually falls.

Bottleneck Structure In Pre-Activations Without Pooling

The fact that almost all weight matrices W_ℓ are supported along only a finite number of frequencies suggests that the corresponding pre-activations $\tilde{\alpha}_\ell(X) = W_\ell \alpha_{\ell-1}(X) + b_\ell$ for any training set X should also be supported along the same frequencies (and possibly also along an additional constant frequency because of the bias term).

This is trivial if the activations remain bounded as the depth L grows, but (Jacot, 2023b) has shown a counterexample: a simple function whose optimal intermediate representations explode in the infinite depth limit. This couterexample can easily be translated to the CNN setup (by applying the same function in parallel to all pixels of a constant input). This implies that to guarantee bounded representations in general, we need another source of bias, in addition to the small parameter norm bias. Following (Jacot, 2023b), we turn to the implicit bias of large learning rates in GD.

We know that GD with a learning rate of η can only converge to a minima $\hat{\theta}$ where the top eigenvalue of the Hessian $\lambda_1(\mathcal{H}\mathcal{L}_\lambda(\hat{\theta}))$ is upper bounded by ${}^2/_\eta$. Other results suggest that SGD is biased towards minima where the trace of the Hessian is small (Damian et al., 2021; Li et al., 2021). The top eigenvalue and trace both are measures of the narrowness of the minimum.

For the MSE loss, the Hessian at a local minimum $\hat{\theta}$ that fits the data (in the sense that $\mathcal{L}_\lambda(\hat{\theta}) = O(\lambda)$) takes the form

$$\mathcal{H}\mathcal{L}_\lambda(\hat{\theta}) = \frac{2}{N}\sum_{i=1}^{N} J_\theta f_\theta(x_i)^T J_\theta f_\theta(x_i) + O(\lambda).$$

The trace of Hessian is then approximately equal to $\frac{2}{N}\sum \|J_\theta f_\theta(x_i)\|_F^2$ and the largest eigenvalue is lower bounded by $\frac{2}{d_{out}N^2 n}\sum \|J_\theta f_\theta(x_i)\|_F^2 - O(\lambda)$ since the first term has rank Nnd_{out}.

The term $\|J_\theta f_\theta(x)\|_F^2$ (which also equals $\mathrm{Tr}[\Theta(x,x)]$ for Θ the NTK (Jacot et al., 2018)) typically scales linearly in depth since it equals the sum over the L terms $\|J_{(W_\ell,b_\ell)} f_\theta(x)\|_F^2$, so a choice of learning rate $\eta = O(L^{-1})$ is natural. This forces convergence to a minimum with $\|J_\theta f_\theta(x)\|_F^2 \leq cL$ which in turns implies that almost all activations are bounded:

Theorem 4.2. *Given a depth L network without pooling, balanced parameters θ with $\|\theta\|^2 \leq Lk + c_1$ for $k = \max_{z\in\Omega_-} \mathrm{Rank}_m(Jf_\theta(z))$, and a point x_0 such that $\mathrm{Rank}\, Jf_\theta(x_0) = \max_{z\in\Omega_-}\mathrm{Rank}_m(Jf_\theta(z))$, then $\|J_\theta f_\theta(x_0)\|_F^2 \leq cL$ implies that,*

$$\sum_{\ell=1}^{L} \|\alpha_{\ell-1}(x_0)\|_2^2 \leq \frac{ce^{\frac{c_1}{k}}}{k|Jf_\theta(x_0)|_+^{2/k}} L.$$

Hence for each $p \in (0,1)$, there are at least $(1-p)L$ layers ℓ with

$$\|\alpha_{\ell-1}(x_0)\|_2^2 \leq \frac{1}{p}\frac{ce^{\frac{c_1}{k}}}{k|Jf_\theta(x_0)|_+^{2/k}}.$$

Theorem 4.2 gives the conditions under which the activations remain bounded, and thereby the pre-activations $\tilde{\alpha}_\ell(X)$ are supported along the same frequencies as W_ℓ (in which Theorem 4.1 proves a bottleneck structure) and possibly also the constant frequency. Note that the results we present in this section do not require the CNNs to be well-trained to (approximately) global minimums.

5. CNNs with up and down-sampling

Given the bottleneck structure in the near-optimal parameters, it is natural to consider implementing down-sampling and up-sampling in CNNs which explicitly enforce a bottleneck structure and save computational cost, as commonly used in practice. In this section, we study CNNs with down-sampling and up-sampling layers.

We only consider CNNs with one down-sampling layer and one up-sampling layer, but the results can be extended to having multiple such layers. To be specific, consider the set of CNNs with the parametrization in Section 2.2 but with one down-sampling layer and one up-sampling layer inserted: Let $\mathcal{N}_{n;m}$ be all possible CNNs with any depth, input dimension n, and pooling m. For each stride $s \in \mathbb{N}$, define the set of stride-CNNs with inner pooling m':

$$\begin{aligned}\mathcal{N}^{(s)}_{n;m,m'} =& \{f_2 \circ \mathrm{Up}_s \circ \hat{f} \circ \mathrm{Down}_s \circ f_1 : \\ & f_1, f_2 \in \mathcal{N}_{n;m}, \hat{f} \in \mathcal{N}_{\lfloor n/s \rfloor;m'}\}\end{aligned} \tag{4}$$

Because of the down-sampling layer, our networks are no longer translationally equivariant; instead, they only represent s-translationally equivariant functions (i.e. invariant under translations by multiples of s). The formal mathematical definitions for down-sampling and up-sampling in (4) are as follows:

Definition 5.1. Define the **down-sampling** operator $\mathrm{Down}_s : \mathbb{R}^{n\times c} \to \mathbb{R}^{\lfloor n/s \rfloor \times c}$ by mapping $(\mathrm{Down}_s(x))_{i,k} = x_{si,k}$, i.e. subsampling every s pixel along each channel.

Define the **up-sampling** operator $\mathrm{Up}_s : \mathbb{R}^{n'\times c} \to \mathbb{R}^{n's\times c}$ by

$$\mathrm{Up}_s(x) = F^*_{n's}[sI_{n'}, 0]^T F_{n'} x$$

i.e. mapping the n' Fourier coefficients of input x to the first n' Fourier coefficients of $\mathrm{Up}_s(x)$ and zeros otherwise. F_N denotes the DFT matrix of dimension $N \times N$.

Remark 5.2. By the Nyquist-Shannon sampling theorem (Shannon, 1949), we have that for $y = \mathrm{Down}_s(x)$, the i-th DFT coefficient is $\tilde{y}_i = \frac{1}{s}\sum_{j=0}^{s-1} \tilde{x}_{i+jn/s \mod n}$. Hence exact reconstruction of x is possible when x is low-frequency i.e. $\tilde{x}_i = 0$ for $i \geq \frac{n}{s}$ (in which case the set of coefficients $\{\tilde{x}_{i+jn/s}\}_{j=0}^{s-1}$ has cardinality ≤ 1 and gives a one-to-one mapping between the coefficients of x and y).

Remark 5.3. The set of all s-stride-CNNs $\mathcal{N}^{(s)}_{n;m,m'}$ is the set of all functions $f = h \circ g$ with bottleneck support only on the first $\frac{n}{s}$ DFT frequencies, cf. Proposition E.1.

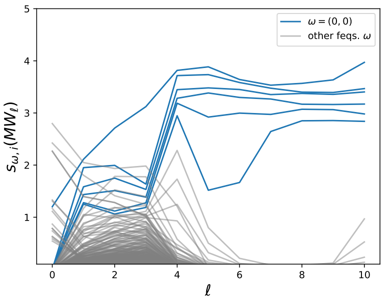
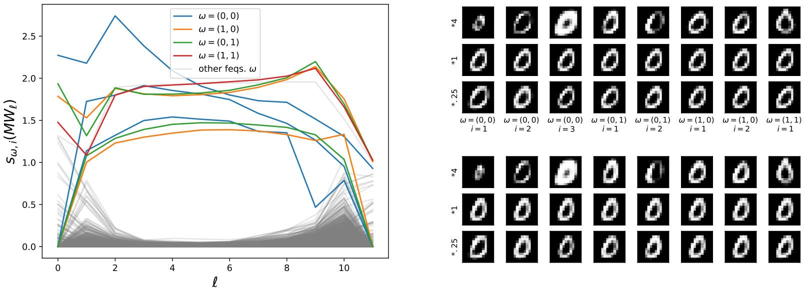


Figure 1. We train a CNN ($L = 11, c_\ell = 60, \lambda = 0.005, \beta = 0.5$) on MNIST. The inputs are 28×28, and scaled down by 2 on the 2nd and 4th layers, with global average pooling and a fully connected layer at the end. We see that for classification, six constant frequencies are kept.



(a) Sing. vals. of MW_ℓ

(b) Latent space interpretation

Figure 2. We train an autoencoder ($L = 12, c_\ell = 50, \lambda = 0.04, \beta = 1.0$) on the 0-digits of MNIST downscaled to the size 13×13. (a) The singular values of MW_ℓ for every layer ℓ, colored by their frequency ω. (b) Along each of the singular values in the 5-th layer, we plot the effect of multiplying the hidden representation along the sing. vector by 2 or 0.5 (for non-constant frequencies we also consider multiplication by complex i and $-i$). We see how each singular value correspond to a (nonlinear) direction of variation of the zeros. For non-constant frequencies the argument encodes the x and y position of the digit.

In other words, if a full-size CNN can be decomposed into two TEPL functions with only low frequencies, then it can be represented by a CNN with down and up-sampling. We thereby have the natural extension of the CBN rank for the stride-CNNs:

$$\mathrm{Rank}^{(s)}_{\mathrm{CBN}}(f;\Omega) \equiv \min_{\substack{f=h^{(s)}\circ g^{(s)} \\ g^{(s)}=g_1^{(s)}\oplus\cdots\oplus g_k^{(s)}}} \sum_{c=1}^{k}\sum_{i\in I_c} \tilde{m}_i'^{-2}$$

where in the decomposition $f = h^{(s)} \circ g^{(s)}$, $g_c^{(s)}$ on each channel c only supports low frequencies $I_c \subseteq [\frac{n}{s}]$. Note that any $f \in \mathcal{N}^{(s)}_{n;m,m'}$ has finite stride-bottleneck rank $\mathrm{Rank}^{(s)}_{\mathrm{CBN}}(f;\Omega) < \infty$. If the inner pooling m' (of size $\frac{n}{s}$) is the same as m (of size n) truncated to the first $\frac{n}{s}$ frequencies, then it is straightforward that $\mathrm{Rank}_{\mathrm{CBN}}(f;\Omega) \leq \mathrm{Rank}^{(s)}_{\mathrm{CBN}}(f;\Omega)$.

Remark 5.4. The reason for having the first $\frac{n}{s}$ frequencies here is due to the choice we made in the up-sampling operator. One can slightly generalize to exact reconstruction of x consisting of another set of $\frac{n}{s}$ frequencies by having a different mapping between the low-dimensional and the high-dimensional Fourier coefficients, as long as the input satisfies $|\{\tilde{x}_{i+jn/s}\}_{j=0}^{s-1}| \leq 1$ for each $0 \leq i < \frac{n}{s}$.

Furthermore, we may recover the upper bound theorem as in Theorem 3.1.

Theorem 5.5. *Let $R_s^{(0)}(f;\Omega)$ denote the rescaled representation cost under the architecture with stride s. Then for any $f \in \mathcal{N}^{(s)}_{n;m,m'}$,*

$$R_s^{(0)}(f;\Omega) \leq \mathrm{Rank}^{(s)}_{CBN}(f;\Omega).$$

Proof. The proof idea follows from that of Theorem 3.1. Suppose $f = h^{(s)} \circ g^{(s)}$ realizes $\mathrm{Rank}^{(s)}_{\mathrm{CBN}}(f;\Omega)$. Observe that $f = h^{(s)} \circ \mathrm{Up}_s \circ \hat{f} \circ \mathrm{Down}_s \circ g^{(s)}$ where $\hat{f}$ consists of $\hat{L}$ identity layers and hence $\mathrm{Up}_s \circ \hat{f} \circ \mathrm{Down}_s = id|_{\mathrm{Im}\, g^{(s)}}$. The bound follows by taking $\hat{L}$ to infinity. □

Remark 5.6. One may also generalize the properties of $R^{(0)}$ to $R_s^{(0)}$ following the same proof ideas.

If the target function possesses a good low-frequency bottleneck structure in the sense that $\mathrm{Rank}_{\mathrm{CBN}} \approx \mathrm{Rank}^{(s)}_{\mathrm{CBN}}$, under the conjecture $R^{(0)} = \mathrm{Rank}_{\mathrm{CBN}}$ and $R_s^{(0)} = \mathrm{Rank}^{(s)}_{\mathrm{CBN}}$, we can see that $R^{(0)} \approx R_s^{(0)}$ (meaning their optimal representation costs are close). Hence one is justified to learn the target with CNNs with enforced down-sampling and up-sampling layers for reduced computation cost and lower-dimensional latent representations in the Euclidean space.

Low Frequency Representation

Although we show exact recovery is possible with appropriate stride s, there remains the question of how to choose the stride for down-sampling in our CNNs a priori. To partially answer this question, under some realistic assumptions on the input domain Ω, we can show that the target function $f : \Omega \to \mathbb{R}^{n\times c_{out}}$ has a low-frequency decomposition $f = h \circ g^{(2)}$ with stride $s = \frac{n}{2}$ (i.e. inner representations only have input size 2 and hence only support 2 frequencies). Yet we remark that having a 2-frequency decomposition does not imply that the optimal stride is of size 2, because low-frequency decomposition may require too many channels for exact recovery, whereas retaining a few more frequencies may be more informative and efficient.

Definition 5.7. The input domain Ω is *translationally* *unique* if $\forall x, y \in \Omega, x = T_p y \implies x = y$ and $p = 0$, where T_p denotes the translation by p along each channel for $p = 0, \dots, n-1$.

In particular, for this kind of domain, $\Omega_- = \emptyset$. Though it is difficult to check or guarantee that all natural images are translationally unique, it seems to hold for the vast majority of images.

Theorem 5.8. *Suppose Ω is translationally unique. Then for any piecewise linear target function $f : \Omega \to \mathbb{R}^{n\times c_{out}}$, $f = h \circ g^{low}$ where h and g^{low} are TEPL functions and $g^{low} : \Omega \to \mathbb{R}^{n\times nc_{in}+1}$ only supports the constant DFT frequency at first nc_{in} channels and the second DFT frequency at the $nc_{in}+1$-th channel.*

Remark 5.9. Theorem 5.8 implies that the identity map on translationally unique domains can be represented using nc_{in} constant frequencies and one 1-periodic frequency. In particular, it gives an upper bound on the bottleneck rank of any TEPL function f on such domain Ω, including id, that

$$\mathrm{Rank}_{\mathrm{CBN}}(f;\Omega) \leq \tilde{m}_2^{-2} + nc_{in}\tilde{m}_1^{-2}.$$

6. Numerical Experiments

For our numerical experiments, we train networks on 4 different tasks, with different depths and ridge parameters. We use filters with full size and cyclic boundaries. The pooling operator is $M_\beta = (1-\beta)I + \beta A_3$, where A_3 is the 3×3 average filter. We use a few different values of β. For the MNIST classification task, we also implement downsampling in the 2nd and 4th layers. The experiments are done for 2D convolution instead of 1D convolution as in the theoretical analysis, but everything translates directly, with the difference that frequencies are indexed by pairs ω.

The emergent bottleneck structure that appears in all the tasks we consider makes these networks highly interpretable. We plot the singular values $s_{\omega,i}(MW_\ell)$ accross the layers $\ell = 1, \dots, L$. We emphasize the singular values that are kept in the bottleneck by coloring them according to their frequency.

MNIST classification: For MNIST classification the CNN features a global pooling layer at the end, followed by a final fully-connected layer. This explains why only constant frequencies are kept in the bottleneck, since any non-constant frequencies in the outputs are killed by the global pooling. Only 6 dimensions are kept, which is sufficient to embed all 10 classes in a linearly separable manner.

Also note that this experiments illustrates a ‘half-bottleneck’, where the representations go from high-dim/high-freq inputs to a low-dim/low-freq bottleneck and remain there until the outputs. This is in contrast to the full bottlenecks that we observe in our other experiments where the representations go back to high-dim/high-freq in the last layers. Note that this half-bottleneck structure (which is common in classification tasks since the outputs of the network are low-dim/low-freq) could explain some aspects of the neural collapse phenomenon (Papyan et al., 2020) as well as other numerical observations (Kornblith et al., 2019).

MNIST digit 0 autoencoder: When training an autoencoder the networks keeps3 constant freq. along with 4 degree 1 freq. and 1 degree two freq. Since the signal inside the bottleneck is almost only supported along low frequencies, the middle layers could have been downsampled before upsampling again (as is usually done with autoencoders), but the L_2-regularization alone recreates the same effect. We believe allowing the network to choose the frequencies it wants to keep and the number of channels is better than forcing it. Of course there are computational advantages to downsampling in the middle of the network.

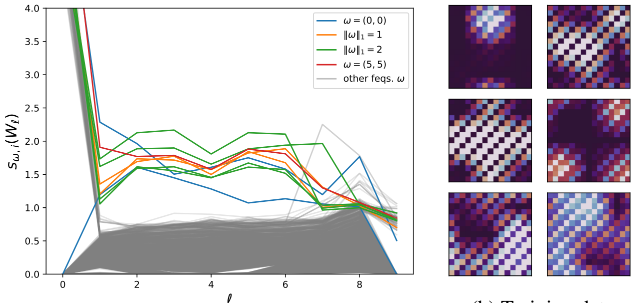
To understand what each of the kept frequencies capture, we plot the effect of multiplying by 4 or 0.25 the signal along each singular value of W_5 and plotting the resulting modified output. The effect along some singular values can be interpreted as capturing e.g. size, boldness, narrowness, angle and more.

Autoencoder on synthetic data: We train an autoencoder on data obtained as the pixelwise multiplication of a low-freq shape with a high-freq repreating pattern (a single freq.-$(5,5)$ Fourier function with random phase). We see that the network disentangles the shape from the pattern in the bottleneck, the shape is encoded in the $\|\omega\|_1 \leq 2$-freqs and the pattern in the single $(5,5)$-freq. This is only possible with non-linear transformations at the beginning and end of the network.

Learning Newtonian Mechanics: We train a network to predict the trajectory of a ball: the inputs to the network are four frames of a ball under gravity (with different frames encoded in different channels) with a random initial position and velocity, from which the network has to predict the next 4 frames. The network keeps two pairs of degree one frequencies (and one constant frequency, which seems to only be there to ensure that the signal remains in R_+ inside the bottleneck; one can check that no information is kept in this constant frequency). The phases of the largest pair of degree one frequencies θ_1 and ϑ_1 encode the (x, y)-position of the ball two frames before the end, and the difference in phases between the largest pair and the smaller pair encodes the (x, y)-velocity at the same frame. Thus the network recognizes that the evolution of the ball is uniquely determined by its position and velocity.

7. Limitations and Discussion

In this paper we focused on describing a bottleneck structure in CNNs with small parameter norm. It still remains to



(a) Singular values of MW_ℓ

(b) Training data

Figure 3. CNN ($L = 10, c_\ell = 60, \lambda = 0.0005, \beta = 0.25$) trained on images that are made up of random low-freq. shapes multiplied with a high frequency ($\omega = (5, 5)$) pattern. In the bottleneck the network keeps track of the shapes in low frequencies ($\|\omega\|_1 \leq 2$) and the pattern in one $\omega = (5, 5)$ frequency. Note that the original images only has signal in high frequencies around $(5, 5)$.

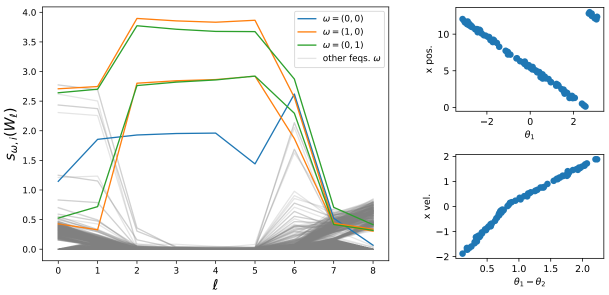
be shown that GD converges under reasonable assumption to such a small parameter solution. The training dynamics of deep nonlinear networks are very difficult to study (outside of the NTK regime (Jacot et al., 2018)), but knowing what kind of structure we expect to appear will probably be helpful.

Our analysis is centered around L_2-regularized networks, but we expect a similar picture to appear in other settings. It has been observed that training with GD on a cross-entropy loss leads to an implicit L_2 regularization (Gunasekar et al., 2018a), thus leading to a BN structure. Similarly a small initialization should bias GD towards solutions with small parameter norm, similar to the dynamics observed in linear nets (Li et al., 2020a; Jacot et al., 2022).

A final limitation is our use of full-size filters and cyclic boundaries. This choices has the obvious advantage of allowing for the use of Fourier analysis, but we expect a similar structure to still appear, though possibly with a different type of sparsity than the sparsity in Fourier basis that we observe. Our analysis only captures the bias induced by the translation invariance of the weights, but the locality of the connections is generally believed to also play an important role.

8. Conclusion

This paper describes a bottleneck structure in CNNs: the network learns functions of the form $f = h \circ g$ where the inner representation is only supported along a few Fourier frequencies, inspired by the appearance of a similar structure in fully-connected networks (Jacot, 2023a;b). Our results provide motivation and justification for the common use of down-sampling in CNNs. This bottleneck structure makes the learned latent features of CNNs highly interpretable, as confirmed by a number of numerical experiments.



(a) Sing. vals. of MW_ℓ

(b) Interpretation

Figure 4. CNN ($L = 9, c_\ell = 60, \lambda = 0.0001, \beta = 0.25$) learns to predict the trajectory of a ball under gravity: the inputs are 4 frames of a ball represented as a dot on a black background, and the outputs are the next four frames. The position appears to be encoded by the phase of the first pair, while the velocity is encoded in the difference between the phases of the two pairs, as confirmed in (b) along the x-axis.

Impact Statement

This work is theoretical in nature and the goal is to advance our understanding in machine learning. It has no direct societal impact.

References

Arora, R., Basu, A., Mianjy, P., and Mukherjee, A. Understanding deep neural networks with rectified linear units. *CoRR*, abs/1611.01491, 2016. URL http://arxiv.org/abs/1611.01491.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R., and Wang, R. On exact computation with an infinitely wide neural net. *Advances in Neural Information Processing Systems*, 32, 2019.

Bietti, A. and Mairal, J. Group invariance, stability to deformations, and complexity of deep convolutional representations. *Journal of Machine Learning Research*, 20(25):1–49, 2019. URL http://jmlr.org/papers/v20/18-190.html.

Cho, Y. and Saul, L. K. Kernel Methods for Deep Learning. In *Advances in Neural Information Processing Systems 22*, pp. 342–350. Curran Associates, Inc., 2009. URL http://papers.nips.cc/paper/3628-kernel-methods-for-deep-learning.pdf.

Cohen, T. S., Geiger, M., and Weiler, M. A general theory of equivariant cnns on homogeneous spaces. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/b9cfe8b6042cf759dc4c0cccb27a6737-Paper.pdf.

Dai, Z., Karzand, M., and Srebro, N. Representation costs of linear neural networks: Analysis and design. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=3oQyjABdbC8.

Damian, A., Ma, T., and Lee, J. D. Label noise sgd provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461, 2021.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. Characterizing implicit bias in terms of optimization geometry. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1832–1841. PMLR, 10–15 Jul 2018a. URL http://proceedings.mlr.press/v80/gunasekar18a.html.

Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. Implicit bias of gradient descent on linear convolutional networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018b. URL https://proceedings.neurips.cc/paper/2018/file/0e98aeeb54acf612b9eb4e48a269814c-Paper.pdf.

Jacot, A. Implicit bias of large depth networks: a notion of rank for nonlinear functions. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=6iDHce-0B-a.

Jacot, A. Bottleneck structure in learned features: Low-dimension vs regularity tradeoff, 2023b.

Jacot, A., Gabriel, F., and Hongler, C. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems 31*, pp. 8580–8589. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8076-neural-tangent-kernel-convergence-and-generalization-in-neural-networks.pdf.

Jacot, A., Ged, F., Şimşek, B., Hongler, C., and Gabriel, F. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity, 2022.

Karantzas, N., Besier, E., Ortega Caro, J., Pitkow, X., Tolias, A. S., Patel, A. B., and Anselmi, F. Understanding robustness and generalization of artificial neural networks through fourier masks. *Frontiers in Artificial Intelligence*, 5, 2022. ISSN 2624-8212. doi: 10.3389/frai.2022.890016. URL https://www.frontiersin.org/articles/10.3389/frai.2022.890016.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529. PMLR, 2019.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278 – 2324, 12 1998. doi: 10.1109/5.726791.

Li, Z., Wang, R., Yu, D., Du, S. S., Hu, W., Salakhutdinov, R., and Arora, S. Enhanced convolutional neural tangent kernels. *arXiv preprint arXiv:1911.00809*, 2019.

Li, Z., Luo, Y., and Lyu, K. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2020a.

Li, Z., Zhang, Y., and Arora, S. Why are convolutional nets more sample-efficient than fully-connected nets? *arXiv preprint arXiv:2010.08515*, 2020b.

Li, Z., Wang, T., and Arora, S. What happens after sgd reaches zero loss?–a mathematical framework. *arXiv preprint arXiv:2110.06914*, 2021.

Mallat, S. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

Mei, S., Misiakiewicz, T., and Montanari, A. Learning with invariances in random features and kernel models. In *Conference on Learning Theory*, pp. 3351–3418. PMLR, 2021.

Misiakiewicz, T. and Mei, S. Learning with convolution and pooling operations in kernel methods. *Advances in Neural Information Processing Systems*, 35:29014–29025, 2022.

Neal, R. M. *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996. ISBN 0387947248.

Papyan, V., Han, X., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Petrini, L., Favero, A., Geiger, M., and Wyart, M. Relative stability toward diffeomorphisms indicates performance in deep nets. *Advances in Neural Information Processing Systems*, 34:8727–8739, 2021.

Shannon, C. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949. doi: 10.1109/JRPROC.1949.232969.

Wang, Z. and Wu, L. Theoretical analysis of the inductive biases in deep convolutional networks. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74289–74338. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/eb1bad7a84ef68a64f1afd6577725d45-Paper-Conference.pdf.

Xiao, L. and Pennington, J. Synergy and symmetry in deep learning: Interactions between the data, model, and inference algorithm. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24347–24369. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/xiao22a.html.

Yarotsky, D. Universal approximations of invariant maps by neural networks. *CoRR*, abs/1804.10306, 2018. URL http://arxiv.org/abs/1804.10306.

A. CNNs as Universal Approximators

Since (fully connected) neural networks are mostly known as universal approximators, one may wonder if the CNNs given by the parameterization in Section 2.2 are universal approximators for translationally equivariant functions. Indeed, when the user-chosen filter m is invertible, universality is guaranteed (Yarotsky, 2018).

Note that the filter being invertible does not prevent it from "shrinking" the high frequencies, since it can have arbitrarily small (but nonzero) singular values at high frequencies. Thereby one may consider it as a smoothened low-pass filter.

B. Properties of $R^{(0)}$ and $R^{(1)}$

We present several interesting properties of the $R^{(0)}$ cost and their proofs here.

Proposition B.1 ($R^{(0)}$ properties)**.** *Write $\bar{M} := \sum_{t=1}^{n} \tilde{m}_t^{-2}$ for simplicity. For any translationally equivariant functions f_1, f_2, we have the following properties:*

1. $R^0(f_2 \circ f_1; \Omega) \leq \min\{R^0(f_2; f_1(\Omega)), R^0(f_1; \Omega)\}$*;*
2. $R^0(f_2 + f_1; \Omega) \leq R^0(f_2; \Omega) + R^0(f_1; \Omega)$*;*
3. $R^0(f_1; \Omega) \leq \min\{c_{in}, c_{out}\}\bar{M}$ *if the filter m is nonnegative and invertible;*
4. $R^0(id; \Omega) = \min\{c_{in}, c_{out}\}\bar{M}$ *if there is an interior point of Ω in Ω_-;*
5. $R^0(f_1; \Omega) = \min\{c_{in}, c_{out}\}\bar{M}$ *if f_1 is a bijection and there is an interior point of Ω in Ω_-;*
6. *If $f : \mathbb{R}^{n\times c_{in}} \to \mathbb{R}^{n\times c_{out}}$ is an affine convolution, namely f is a linear one-layer CNN with weight $w \in \mathbb{R}^{n\times c_{out}\times c_{in}}$ and bias $b \in \mathbb{R}^{c_{out}}$, and there is an interior point of Ω in Ω_-, then*

$$R^0(f; \Omega) = \sum_{c=1}^{\min\{c_{in}, c_{out}\}} \sum_{t\in I_c} \tilde{m}_t^{-2}$$

where I_c is the Fourier frequencies supported by $\{w_{:,c,k}\}_{k=1}^{c_{in}}$, i.e. the indices of nonzero entries of $\{Fw_{:,c,s}\}_{k=1}^{c_{in}}$.

Proof. **1.** Write $f_1(\Omega) \subseteq \mathbb{R}^{n\times c_{mid}}$. Without loss of generality, we may assume $R^0(f_2; f_1(\Omega)) \leq R^0(f_1; \Omega)$. By Lemma G.1, we can fix a CNN with depth $L_1 = \lceil nc_{in} + 1\rceil + 2$ and parameter $\boldsymbol{W}_1$ representing f_1. For any sufficiently large $L > L_1 + \lceil nc_{mid} + 1\rceil + 2$, we have a CNN with depth $L - L_1$ and parameter $\boldsymbol{W}_2$ that represents f_2 with minimal representation cost, i.e. $\|\boldsymbol{W}_2\|^2 = R(f_2; L - L_1, f_1(\Omega))$. Then the norm of the composed CNN is

$$R(f_2 \circ f_1; L, \Omega) \leq \|\boldsymbol{W}_1\|^2 + R(f_2; L - L_1, f_1(\Omega)).$$

Dividing by $L - L_1$ and taking $L \to \infty$ gives the inequality

$$R^0(f_2 \circ f_1; \Omega) \leq \min\{R^0(f_2; f_1(\Omega)), R^0(f_1; \Omega)\}.$$

2. Let f_1 and f_2 be represented by CNNs with some depth L and parameters $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$, respectively, with minimal parameter norms, i.e. $\|\boldsymbol{W}_1\|^2 = R(f_1; L, \Omega)$ and $\|\boldsymbol{W}_2\|^2 = R(f_2; L, \Omega)$. We can construct a network with depth L and parameters $\boldsymbol{W}$ that represents $f_1 + f_2$ by stacking them "in parallel":

At each layer ℓ, we let the number of channels be the sum of the other two networks, i.e. $c_\ell = c_\ell^{(1)} + c_\ell^{(2)}$. For layers $1 \leq \ell < L$, we set the weights and biases as follows:

$$\begin{aligned} (w_\ell)_{:,c,k} &= \mathbb{1}\left[c \leq c_\ell^{(1)} \wedge k \leq c_{\ell-1}^{(1)}\right] (w_\ell^{(1)})_{:,c,k} + \mathbb{1}\left[c > c_\ell^{(1)} \wedge k > c_{\ell-1}^{(1)}\right] (w_\ell^{(2)})_{:,c-c_\ell^{(1)},k-c_{\ell-1}^{(1)}} \\ (b_\ell)_c &= \mathbb{1}\left[c \leq c_\ell^{(1)}\right] (b_\ell^{(1)})_c + \mathbb{1}\left[c > c_\ell^{(1)}\right] (b_\ell^{(2)})_{c-c_\ell^{(1)}} \end{aligned}$$

for $c = 1, \dots, c_\ell$ and $k = 1, \dots, c_{\ell-1}$. We incorporate the sum in the last layer by having

$$\begin{aligned} (w_L)_{:,c,k} &= \mathbb{1}\left[k \le c_{L-1}^{(1)}\right] (w_L^{(1)})_{:,c,k} + \mathbb{1}\left[k > c_{L-1}^{(1)}\right] (w_L^{(2)})_{:,c,k-c_{L-1}^{(1)}} \\ (b_L)_c &= (b_L^{(1)})_c + (b_L^{(2)})_c \end{aligned}$$

for $c = 1, \dots, c_{out}$, $k = 1, \dots, c_{L-1}$.

This CNN represents $f_1 + f_2$ and has parameter norm $\|\boldsymbol{W}\|^2 = \|\boldsymbol{W}_1\|^2 + \|\boldsymbol{W}_2\|^2$. Hence we have

$$R(f_2 + f_1; L, \Omega) \le R(f_2; L, \Omega) + R(f_1; L, \Omega).$$

Dividing by L and taking $L \to \infty$, we obtain the desired inequality

$$R^0(f_2 + f_1; \Omega) \le R^0(f_2; \Omega) + R^0(f_1; \Omega).$$

3. Let ϕ be represented by a depth L_ϕ network with parameter $\boldsymbol{W}_\phi$. For $L \ge L_\phi$, we can let the first $L - L_\phi$ layers of the network be the identity layers: Since Ω is bounded, let $K > 0$ upper bounds $x \in \Omega$ coordinate-wise. For $\ell = 1, \dots, L - L_\phi$, let

$$\begin{aligned} (w_\ell)_{:,c,k} &= m^{-1}\mathbb{1}[c = k] \\ (b_\ell)_c &= K\mathbb{1}[\ell = 1] \\ (b_{L-L_\phi+1})_c &= (b_1^{(\phi)})_c - K(m^{-1} * \mathbf{1}). \end{aligned}$$

Note this CNN represents $\phi \circ id = \phi$. By construction, the parameter norm is $\|\boldsymbol{W}\|^2 = (L - L_\phi)c_{in} \sum_{t=1}^{n} \tilde{m}_t^{-2} + \|\boldsymbol{W}_\phi\|^2 + nK(m^{-1} * \mathbf{1} + 1)$. Dividing by L and taking $L \to \infty$ yields

$$R^0(\phi; \Omega) \le c_{in} \sum_{t=0}^{n-1} \tilde{m}_t^{-2}.$$

Similarly, appending the identity layers after $\boldsymbol{W}_\phi$ yields

$$R^0(\phi; \Omega) \le c_{out} \sum_{t=0}^{n-1} \tilde{m}_t^{-2}.$$

4. Follows from the squeezing bounds Theorem 3.1 and Theorem 3.2.

5. Follows from the observation that

$$\min(c_{in}, c_{out})\bar{M} = R^0(id; \Omega) \le \min\{R^0(\psi; \Omega), R^0(\psi^{-1}; \Omega)\} \le \min(c_{in}, c_{out})\bar{M}.$$

6. Follows from the squeezing bounds Theorem 3.1 and Theorem 3.2 (note for the upper bound decomposition we have $f = id \circ f$). □

Corollary B.2. *Let f be any translationally equivariant function. For any translationally equivariant bijections ϕ and ψ on $\mathbb{R}^{n \times c_{in}}$ and $\mathbb{R}^{n \times c_{out}}$ respectively, we have $R^0(\psi \circ f \circ \phi; \Omega) = R^0(f; \Omega)$.*

With the following proposition, we show that the $R^{(1)}$ correction controls the regularity of the learned function and satisfies subadditivity.

Proposition B.3 ($R^{(1)}$ properties)**.** *For any translationally equivariant functions f and g, we have the following properties:*

1. *For any $x \in \Omega_-$, $R^1(f; \Omega) \ge 2\sum_{s_{t,c} \neq 0} \tilde{m}_t^{-2} \log(s_{t,c}\tilde{m}_t)$ with $s_{t,c}$ being the (t, c)-th singular values of $Jf(x)$ for $t = 1, \dots, n$ and $c = 1, \dots, \min\{c_{in}, c_{out}\}$.*

   In particular, when there is no pooling, i.e. $m = id$, for any $x \in \Omega$, we have $R^1(f; \Omega) \ge 2 \log |Jf(x)|_+$ where $|\cdot|_+$ denotes the pseudo-determinant.

2. *If $R^0(f \circ g; \Omega) = R^0(f; g(\Omega)) = R^0(g; \Omega)$, then $R^1(f \circ g; \Omega) \leq R^1(f; g(\Omega)) + R^1(g; \Omega)$.*
3. *If $R^0(f + g; \Omega) = R^0(f; \Omega) + R^0(g; \Omega)$, then $R^1(f + g; \Omega) \leq R^1(f; \Omega) + R^1(g; \Omega)$.*

Proof. **1.** From the proof of Theorem 3.2 we have $R(f; \Omega, L) \geq L \| M^{1-L} J f_\theta(x) \|_{2/L}^{2/L}$ for any $x \in \Omega_-$. Therefore,

$$\begin{aligned}
R^1(f; \Omega) &= \lim_{L \to \infty} R(f; \Omega, L) - L R^0(f; \Omega) \\
&\geq \lim_{L \to \infty} L \left(\sum_{\substack{t=1 \\ c=1}}^{\substack{n \\ \min\{c_{in}, c_{out}\}}} \tilde{m}_t^{2^{\frac{1-L}{L}}} s_{t,c}^{\frac{2}{L}} - R^0(f; \Omega) \right) \\
&\geq \lim_{L \to \infty} L \sum_{s_{t,c} \neq 0} \tilde{m}_t^{-2} \left(s_{t,c}^{\frac{2}{L}} \tilde{m}_t^{\frac{2}{L}} - 1 \right) \\
&\geq \lim_{L \to \infty} L \sum_{s_{t,c} \neq 0} \tilde{m}_t^{-2} \frac{2}{L} \log(s_{t,c} \tilde{m}_t) \\
&= 2 \sum_{s_{t,c} \neq 0} \tilde{m}_t^{-2} \log(s_{t,c} \tilde{m}_t)
\end{aligned}$$

for all $x \in \Omega_-$ i.e. that are constant along each channel.

Similarly, when there is no pooling, we have for any $x \in \Omega$, $R(f; \Omega, L) \geq L \| J f_\theta(x) \|_{2/L}^{2/L}$ and the result follows from the same reasoning.

2. Since $R(f \circ g; \Omega, L_1 + L_2) \leq R(f; g(\Omega), L_1) + R(g; \Omega, L_2)$, we have:

$$\begin{aligned}
R^1(f \circ g; \Omega) &= \lim_{L_1 + L_2 \to \infty} R(f \circ g; \Omega, L_1 + L_2) - (L_1 + L_2) R^0(f \circ g; \Omega) \\
&\leq \lim_{L_1 \to \infty} R(f; g(\Omega), L_1) - L_1 R^0(f; g(\Omega)) + \lim_{L_2 \to \infty} R(g; \Omega, L_2) - L_2 R^0(g; \Omega) \\
&= R^1(f; g(\Omega)) + R^1(g; \Omega).
\end{aligned}$$

3. Since $R(f + g; \Omega, L) \leq R(f; \Omega, L) + R(g; \Omega, L)$, we have:

$$\begin{aligned}
R^1(f + g; \Omega) &= \lim_{L \to \infty} R(f + g; \Omega, L) - L R^0(f + g; \Omega) \\
&\leq \lim_{L \to \infty} R(f; \Omega, L) - L R^0(f; \Omega) + \lim_{L \to \infty} R(g; \Omega, L) - L R^0(g; \Omega) \\
&= R^1(f; \Omega) + R^1(g; \Omega).
\end{aligned}$$

□

C. Upper and Lower bounds for Rescaled Representation Cost and Correction

In this section, we present the proofs for the CBN upper bound (Theorem 3.1) and the filter-dependent lower bounds (Theorem 3.2) of $R^{(0)}$.

Theorem C.1. *For any translationally equivariant function f with finite $R^{(0)}(f; \Omega)$,*

$$R^{(0)}(f; \Omega) \leq \mathrm{Rank}_{\mathit{CBN}}(f; \Omega).$$

Proof. Let $f = h \circ g$ for any TEPL functions h, g and $g(x) = g_1(x) \oplus \cdots \oplus g_k(x)$, $g_c(x)$ with I_c truncated Fourier coefficient supports for $c = 1, \ldots, k$. Lemma G.1 tells that h and g can be represented by CNNs with parameters $\boldsymbol{W}_h$ and $\boldsymbol{W}_g$ and depths $L_h, L_g \leq \lceil \log(n c_{in} + 1) \rceil + 2$ respectively.

Since Ω is bounded, we can translate $g(\Omega)$ to the first quarter of $\mathbb{R}^{n \times k}$ by adding an extra bias $\bar{b}_g$ in the last layer. Then for any $L > L_h + L_g$, we can efficiently construct a network as follows: first L_g layers are the network representing g with an extra bias $\bar{b}_g$ to translate the output to the first quarter, followed by $L - L_h - L_g$ identity layers as described in the proof of Proposition B.1 and translate the hidden representation α_ℓ back by subtracting $\bar{b}_g$ in the last identity layer, and finally the last L_h layers the network representing h. This construction gives us a bound

$$R(f; L, \Omega) \leq \|\boldsymbol{W}_g\|^2 + (L - L_g - L_h) \sum_{c=1}^{k} \sum_{i \in I_c} \tilde{m}_i^{-2} + 2\|\bar{b}_g\|^2 + \|\boldsymbol{W}_h\|^2$$

for $L > L_g + L_h$. Dividing both side by L and taking $L \to \infty$ gives the inequality

$$R^0(f; \Omega) \leq \sum_{c=1}^{k} \sum_{i \in I_c} \tilde{m}_i^{-2}$$

and the result follows since $h \circ g$ is an arbitrary TEPL decomposition. □

Theorem C.2. *For any translationally equivariant function f, let $Jf(x)$ be the Jacobian of f at x. The following filter-dependent lower bounds hold:*

1. $\frac{1}{\max\{\tilde{m}_{\max}^2, 1\}} \max_{x \in \Omega} \operatorname{Rank}(Jf(x)) \leq R^{(0)}(f; \Omega)$
   In particular, when there is no pooling, $\max_{x \in \Omega} \operatorname{Rank}(Jf(x)) \leq R^{(0)}(f; \Omega)$.
2. $\max_{x \in \Omega_-} \operatorname{Rank}_m(Jf(x)) \leq R^{(0)}(f; \Omega)$
   where the max is now taken over the subset $\Omega_- := \{x \in \Omega \,|\, x_{i,c} = x_{j,c}\ \forall c = 1, \dots, c_{in}, \forall i, j = 1, \dots, n\}$*, i.e. all x that are constant along each channel.*

Proof. **1.** Fix any input $x \in \Omega$, depth L, and the minimal-norm parameter θ with $f_\theta = f$. We can first write

$$Jf_\theta(x) = W_L D_{L-1}(x) M W_{L-1} \cdots D_1(x) M W_1$$

where $D_\ell(x) = \operatorname{diag}(\dot{\sigma}(\alpha_\ell(x))) \in \mathbb{R}^{nc_\ell \times nc_\ell}$ are diagonal matrices with 1 and 0 on the diagonal, $W_\ell \in \mathbb{R}^{nc_\ell \times nc_{\ell-1}}$ are the matrix representation of the convolution filters $w_\ell \in \mathbb{R}^{n \times c_\ell \times c_{\ell-1}}$, and M is that of the channel-wise convolution with $m \in \mathbb{R}^n$ (or simply a convolution filter $\hat{m} \in \mathbb{R}^{n \times c_\ell \times c_\ell}$ with $\hat{m}_{:,c,s} = \mathbb{1}[c = s]m$). From (Jacot, 2023a) and (Dai et al., 2021), we have

$$\begin{aligned}
\|Jf_\theta(x)\|_{2/L}^{2/L} &\leq \frac{1}{L}\left(\|W_L\|_F^2 + \|D_{L-1}(x)MW_{L-1}\|_F^2 + \cdots + \|D_1(x)MW_1\|_F^2\right) \\
&\leq \frac{1}{L}\left(\|W_L\|_F^2 + \|MW_{L-1}\|_F^2 + \cdots + \|MW_1\|_F^2\right) \\
&\leq \frac{1}{L}\max\{\|M\|_2^2, 1\}\left(\|W_L\|_F^2 + \cdots + \|W_1\|_F^2\right) \\
&= \frac{\max\{\tilde{m}_{\max}^2, 1\}}{L}\left(\|W_L\|_F^2 + \cdots + \|W_1\|_F^2\right) \\
&\leq \max\{\tilde{m}_{\max}^2, 1\} R(f; \Omega, L)/L
\end{aligned}$$

Taking $L \to \infty$ on both sides, we have for any input $x \in \Omega$,

$$\frac{1}{\max\{\tilde{m}_{\max}^2, 1\}} \operatorname{Rank}(Jf(x)) \leq R^{(0)}(f; \Omega).$$

2. The key observation in this proof is that if the input x is constant along each channel and W is any translationally equivariant matrix, then $WD_\ell(x)M = MWD_\ell(x)$. This observation follows from the fact that $\alpha_\ell(x)$ is translationally equivariant and hence also channel-wise constant; then $D_\ell(x) = \operatorname{diag}(\dot{\sigma}(\alpha_\ell(x)))$ is either all 0 or all 1 along each channel, and so $D_\ell(x)M = MD_\ell(x)$. The commutativity between M and W always holds and follows from pure algebraic computation. Consequantly, for any $x \in \Omega_-$, depth L, and parameter θ with $f_\theta = f$, we now have

$$\begin{aligned}
Jf_\theta(x) &= W_L D_{L-1}(x) M W_{L-1} \cdots D_1(x) M W_1 \\
&= M^{L-1} W_L D_{L-1}(x) W_{L-1} \cdots D_1(x) W_1
\end{aligned}$$

and

$$\begin{aligned}\left\|M^{1-L}Jf_\theta(x)\right\|_{2/L}^{2/L} &= \|W_L D_{L-1}(x)W_{L-1}\cdots D_1(x)W_1\|_{2/L}^{2/L}\\ &\le \frac{1}{L}\left(\|W_L\|_F^2+\|D_{L-1}(x)W_{L-1}\|_F^2+\cdots+\|D_1(x)W_1\|_F^2\right)\\ &\le \frac{1}{L}\left(\|W_L\|_F^2+\cdots+\|W_1\|_F^2\right)\\ &\le R(f;L,\Omega)/L \end{aligned} \tag{5}$$

Since $Jf_\theta(x)$ is a product of translationally equivariant matrices and so is translationally equivariant, we can index its singular values $s_{t,c}$ by the FDT frequency $t=0,\dots,n-1$ and the channel $c=1,\dots,\min\{c_{in},c_{out}\}$. Then the singular values of $M^{(1-L)}Jf_\theta(x)$ are $\tilde{m}_t^{1-L}s_{t,c}$. Thus we can rewrite

$$\left\|M^{1-L}Jf_\theta(x)\right\|_{2/L}^{2/L}=\sum_{t=1}^{n}\sum_{c=1}^{\min\{c_{in},c_{out}\}}\tilde{m}_t^{2\frac{1-L}{L}}s_{t,c}^{2/L}.$$

Taking $L\to\infty$ on both sides of (5), we have

$$\sum_{t=1}^{n}\sum_{c=1}^{\min\{c_{in},c_{out}\}}\tilde{m}_t^{-2}\mathbb{1}[s_{t,c}\neq 0]\le R^{(0)}(f;\Omega).$$

□

D. Bottleneck Structure in Weights and Activations

Following are the proofs for Theorem 4.1 and Theorem 4.2.

Theorem D.1. *Given $\|\theta\|^2\le L\max_{z\in\Omega_-}\mathrm{Rank}_m(Jf_\theta(z))+c_1$ and $x\in\mathrm{argmax}_{z\in\Omega_-}\mathrm{Rank}_m Jf_\theta(x)$, we have $V_\ell^T\in\mathbb{R}^{\kappa\times nc_{\ell-1}}$ and $U_\ell\in\mathbb{R}^{nc_\ell\times\kappa}$ being submatrices of the DFT block matrices $F_{\ell-1}\in\mathbb{R}^{nc_{\ell-1}\times nc_{\ell-1}}$ and $F_\ell^*\in\mathbb{R}^{nc_\ell\times nc_\ell}$ respectively, where $\kappa=\mathrm{Rank}Jf_\theta(x)$, such that*

$$\sum_{\ell=1}^{L}\|W_\ell-U_\ell S_\ell V_\ell^T\|_F^2+\|b_\ell\|_F^2\le c_1-2\sum_{s_{t,c}\neq 0}\tilde{m}_t^{-2}\log(s_{t,c}\tilde{m}_t)$$

and thus for any $p\in(0,1)$, there are at least $(1-p)L$ layers ℓ with

$$\|W_\ell-U_\ell S_\ell V_\ell^T\|_F^2+\|b_\ell\|_F^2\le\frac{c_1-2\sum_{s_{t,c}\neq 0}\tilde{m}_t^{-2}\log(s_{t,c}\tilde{m}_t)}{pL}$$

where $s_{t,c}$ is the (t,c)-th singular value of $Jf_\theta(x)$ and $S_\ell\in\mathbb{R}^{\kappa\times\kappa}$ is a diagonal matrix with entries $\in\{\tilde{m}_t^{-1}\}_{t=0}^{n-1}$.

Proof. Note for $x\in\Omega_-$ constant input, we have

$$Jf_\theta(x)=W_L D_{L-1}(x)MW_{L-1}\cdots D_1(x)MW_1$$

where all W_ℓ, $D_\ell(x)$, and M are translationally equivariant, and so is $Jf_\theta(x)$ itself. Hence we can decompose $Jf_\theta(x)$ along each Fourier frequency separately: Let P_t be the projection matrix to the signal space consisting of only the t-th frequency. Then we can decompose

$$Jf_\theta(x)=\sum_{t=1}^{n}P_tJf_\theta(x)P_t\,.$$

Now we consider each summand $P_t J f_\theta(x) P_t$. For $x \in \Omega_-$, each P_t represents a single-channel convolution and hence commutes with all translationally equivariant maps. Then we have

$$\begin{aligned} P_t J f_\theta(x) P_t &= P_t W_L D_{L-1}(x) M W_{L-1} \cdots D_1(x) M W_1 P_k \\ &= P_t W_L P_t D_{L-1}(x) P_t M P_t W_{L-1} P_t \cdots D_1(x) P_t M P_t W_1 P_t \\ &= \tilde{m}_t^{L-1} P_t W_L P_t D_{L-1}(x) P_t W_{L-1} P_t \cdots D_1(x) P_t W_1 P_t \\ &= \tilde{m}_t^{L-1} P_t W_L P_t P_{\mathrm{Im}\, J\alpha_{L-1}(x)} P_{\mathrm{Im}\, J(\alpha_{L-1} \to f_\theta)(x)^T} D_{L-1}(x) P_{\mathrm{Im}\, J\tilde{\alpha}_{L-1}(x)} \\ & P_{\mathrm{Im}\, J(\tilde{\alpha}_{L-1} \to f_\theta)(x)^T} P_t W_{L-1} P_t P_{\mathrm{Im}\, J\alpha_{L-2}(x)} \cdots \\ & P_{\mathrm{Im}\, J(\alpha_1 \to f_\theta)(x)^T} D_1(x) P_{\mathrm{Im}\, J\tilde{\alpha}_1(x)} P_{\mathrm{Im}\, J(\tilde{\alpha}_1 \to f_\theta)(x)^T} P_t W_1 P_t \end{aligned}$$

since $P_t M P_t = \tilde{m}_t P_t$.

For general matrices A and B, $|AB|_+ = |A|_+ |B|_+$ when the non-zero pre-image of A matches the image of B, and $|\alpha A|_+ = \alpha^{\mathrm{Rank} A} |A|_+$. Hence we have

$$\begin{aligned} |P_t J f_\theta(x) P_t|_+ = \tilde{m}_t^{(L-1)n_t} & |P_t W_L P_t P_{\mathrm{Im}\, J\alpha_{L-1}(x)}|_+ |P_{\mathrm{Im}\, J(\alpha_{L-1} \to f_\theta)(x)^T} D_{L-1}(x) P_{\mathrm{Im}\, J\tilde{\alpha}_{L-1}(x)}|_+ \\ & |P_{\mathrm{Im}\, J(\tilde{\alpha}_{L-1} \to f_\theta)(x)^T} P_t W_{L-1} P_t P_{\mathrm{Im}\, J\alpha_{L-2}(x)}|_+ \cdots \\ & |P_{\mathrm{Im}\, J(\alpha_1 \to f_\theta)(x)^T} D_1(x) P_{\mathrm{Im}\, J\tilde{\alpha}_1(x)}|_+ |P_{\mathrm{Im}\, J(\tilde{\alpha}_1 \to f_\theta)(x)^T} P_t W_1 P_t|_+ \end{aligned}$$

with $n_t = \sum_{c=1}^{\min\{c_{in}, c_{out}\}} \mathbb{1}[s_{t,c} \neq 0] = \mathrm{Rank}(P_t J f_\theta(x) P_t)$. Then writing $P_{\mathrm{Im}\, J\alpha_0(x)} = I$, we have

$$\begin{aligned} \sum_{\substack{c=1 \\ s_{t,c} \neq 0}}^{\min\{c_{in}, c_{out}\}} \log s_{t,c} = \log |P_t J f_\theta(x) P_t|_+ & \\ &= n_t (L-1) \log \tilde{m}_t + \sum_{\ell=1}^{L-1} |P_{\mathrm{Im}\, J(\alpha_\ell \to f_\theta)(x)^T} D_\ell(x) P_{\mathrm{Im}\, J\tilde{\alpha}_\ell(x)}|_+ \\ &+ \sum_{\ell=1}^{L} |P_{\mathrm{Im}\, J(\tilde{\alpha}_\ell \to f_\theta)(x)^T} P_t W_\ell P_t P_{\mathrm{Im}\, J\alpha_{\ell-1}(x)}|_+ \,. \end{aligned}$$

Observe that

$$-2\tilde{m}_t^{-2} |P_{\mathrm{Im}\, J(\alpha_\ell \to f_\theta)(x)^T} D_\ell(x) P_{\mathrm{Im}\, J\tilde{\alpha}_\ell(x)}|_+ \geq \tilde{m}_t^{-2} \left(\mathrm{Rank} J f_\theta(x) - \|P_{\mathrm{Im}\, J(\alpha_\ell \to f_\theta)(x)^T} D_\ell(x) P_{\mathrm{Im}\, J\tilde{\alpha}_\ell(x)}\|_F^2 \right)$$

which is positive since the eigenvalues of $D_\ell(x)$ is ≤ 1.

Also note that for general matrix A and constants m_i, we have

$$\begin{aligned} & \|A\|_F^2 - \sum_{i=1}^{\mathrm{Rank} A} \left(m_i^{-2} - 2 m_i^{-2} \log m_i - 2 m_i^{-2} \log s_i(A) \right) \\ &= \sum_{i=1}^{\mathrm{Rank} A} s_i(A)^2 - m_i^{-2} (1 + 2 \log m_i + 2 \log s_i(A)) \\ &= \sum_{i=1}^{\mathrm{Rank} A} s_i(A)^2 - m_i^{-2} (1 + 2 \log(m_i s_i(A))) \\ &\geq \sum_{i=1}^{\mathrm{Rank} A} s_i(A)^2 - m_i^{-2} (1 + 2 m_i s_i(A) - 2) \\ &= \sum_{i=1}^{\mathrm{Rank} A} (s_i(A) - m_i^{-1})^2 \end{aligned}$$

Denote $\overline{W}_\ell^{(t)} = P_{\mathrm{Im}\,J(\tilde{\alpha}_\ell \to f_\theta)(x)^T} P_t W_\ell P_t P_{\mathrm{Im}\,J\alpha_{\ell-1}(x)}$ for simplicity. Then we can lower bound the sum

$$\begin{aligned}
&\|\theta\|^2 - \sum_{\ell=1}^{L} \|b_\ell\|_F^2 - L\mathrm{Rank}_m(Jf_\theta(x)) - 2 \sum_{\substack{t,c \\ s_{t,c}\neq 0}} \tilde{m}_t^{-2} \log(s_{t,c}\tilde{m}_t) \\
&= \sum_{\ell=1}^{L} \|W_\ell\|_F^2 - L \sum_{\substack{t,c \\ s_{t,c}\neq 0}} \tilde{m}_t^{-2} - 2 \sum_{\substack{t,c \\ s_{t,c}\neq 0}} \tilde{m}_t^{-2} \log(s_{t,c}\tilde{m}_t) \\
&= \sum_{\ell=1}^{L} \|W_\ell\|_F^2 - L \sum_{\substack{t,c \\ s_{t,c}\neq 0}} \tilde{m}_t^{-2} - 2 \sum_{\substack{t,c \\ s_{t,c}\neq 0}} \tilde{m}_t^{-2} \log s_{t,c} - 2 \sum_{\substack{t,c \\ s_{t,c}\neq 0}} \tilde{m}_t^{-2} \log \tilde{m}_t \\
&\geq \sum_{\ell=1}^{L} \sum_{t=1}^{n} \left(\|P_t W_\ell P_t\|_F^2 - \sum_{\substack{c \\ s_{t,c}\neq 0}} \tilde{m}_t^{-2} - 2\tilde{m}_t^{-2} \sum_{\substack{c \\ s_{t,c}\neq 0}} \log \tilde{m}_t - 2\tilde{m}_t^{-2} \log |\overline{W}_\ell^{(t)}|_+ \right) \\
&= \sum_{\ell=1}^{L} \sum_{t=1}^{n} \left(\|P_t W_\ell P_t - \overline{W}_\ell^{(t)}\|_F^2 + \|\overline{W}_\ell^{(t)}\|_F^2 - \sum_{\substack{c \\ s_{t,c}\neq 0}} \tilde{m}_t^{-2} - 2\tilde{m}_t^{-2} \sum_{\substack{c \\ s_{t,c}\neq 0}} \log \tilde{m}_t - 2\tilde{m}_t^{-2} \log |\overline{W}_\ell^{(t)}|_+ \right) \\
&\geq \sum_{\ell=1}^{L} \sum_{t=1}^{n} \left(\|P_t W_\ell P_t - \overline{W}_\ell^{(t)}\|_F^2 + \sum_{s_{t,c}(\overline{W}_\ell^{(t)})\neq 0} \left(s_{t,c}(\overline{W}_\ell^{(t)}) - \tilde{m}_t^{-1} \right)^2 \right) \\
&\geq \sum_{\ell=1}^{L} \sum_{t=1}^{n} \|P_t W_\ell P_t - U_\ell^{(t)} S_\ell^{(t)} (V_\ell^{(t)})^T\|_F^2 \\
&= \sum_{\ell=1}^{L} \|W_\ell - U_\ell S_\ell V_\ell^T\|_F^2
\end{aligned}$$

since

$$\begin{aligned}
\sum_{t=1}^{n} \|P_t W_\ell P_t - P_{\mathrm{Im}\,J(\tilde{\alpha}_\ell \to f_\theta)(x)^T} P_t W_\ell P_t P_{\mathrm{Im}\,J\alpha_{\ell-1}(x)}\|_F^2 &= \sum_{t=1}^{n} \|P_t W_\ell P_t - P_t P_{\mathrm{Im}\,J(\tilde{\alpha}_\ell \to f_\theta)(x)^T} W_\ell P_{\mathrm{Im}\,J\alpha_{\ell-1}(x)} P_t\|_F^2 \\
&= \|W_\ell - P_{\mathrm{Im}\,J(\tilde{\alpha}_\ell \to f_\theta)(x)^T} W_\ell P_{\mathrm{Im}\,J\alpha_{\ell-1}(x)}\|_F^2.
\end{aligned}$$

Here $U_\ell \Sigma_\ell V_\ell^T$ is the compact SVD decomposition of $P_{\mathrm{Im}\,J(\tilde{\alpha}_\ell \to f_\theta)(x)^T} W_\ell P_{\mathrm{Im}\,J\alpha_{\ell-1}(x)}$. Since we know $P_{\mathrm{Im}\,J(\tilde{\alpha}_\ell \to f_\theta)(x)^T} W_\ell P_{\mathrm{Im}\,J\alpha_{\ell-1}(x)}$ is translationally equivariant, we can let $V_\ell^T \in \mathbb{R}^{\kappa \times nc_{\ell-1}}$ and $U_\ell \in \mathbb{R}^{nc_\ell \times \kappa}$ be submatrices of the DFT block matrices $F_{\ell-1} \in \mathbb{R}^{nc_{\ell-1} \times nc_{\ell-1}}$ and $F_\ell^* \in \mathbb{R}^{nc_\ell \times nc_\ell}$ respectively, and Σ_ℓ correspond to the nonzero singular values of $\kappa = \mathrm{Rank} Jf_\theta(x)$ frequencies. And $S_\ell \in \mathbb{R}^{\kappa \times \kappa}$ consists of the singular values of M^{-1} at corresponding frequencies. This gives

$$\begin{aligned}
\sum_{\ell=1}^{L} \|W_\ell - U_\ell S_\ell V_\ell^T\|_F^2 + \|b_\ell\|_F^2 &\leq \|\theta\|^2 - L\mathrm{Rank}_m(Jf_\theta(x)) - 2 \sum_{s_{t,c}\neq 0} \tilde{m}_t^{-2} \log(s_{t,c}\tilde{m}_t) \\
&\leq c_1 - 2 \sum_{s_{t,c}\neq 0} \tilde{m}_t^{-2} \log(s_{t,c}\tilde{m}_t) \,.
\end{aligned}$$

□

Theorem D.2. *Given a depth L network, balanced parameters θ with $\|\theta\|^2 \leq L \max_{x\in\Omega_-} \mathrm{Rank}_m(Jf_\theta(z)) + c_1$, and a point x_0 with $\mathrm{Rank}\, Jf_\theta(x_0) = k$, then $\|J_\theta f_\theta(x_0)\|_F^2 \leq cL$ implies that,*

$$\sum_{\ell=1}^{L} \|\alpha_{\ell-1}(x_0)\|_2^2 \leq \frac{ce^{\frac{c_1}{k}}}{k|Jf_\theta(x_0)|_+^{2/k}} L$$

Hence for each $p \in (0,1)$*, there are at least* $(1-p)L$ *layers* ℓ *with*

$$\|\alpha_{\ell-1}(x_0)\|_2^2 \leq \frac{1}{p}\frac{ce^{\frac{c_1}{k}}}{k|Jf_\theta(x_0)|_+^{2/k}}\,.$$

Proof. We can write

$$\|J_\theta f_\theta(x_0)\|_F^2 = \mathrm{Tr}\left[\Theta^{(L)}(x_0,x_0)\right] = \sum_{\ell=1}(\|\alpha_{\ell-1}(x_0)\|_2^2+1)\|J(\tilde{\alpha}_\ell \to \alpha_L)(x_0)\|_F^2$$

and our goal is to lower bound $\|J(\tilde{\alpha}_\ell \to \alpha_L)(x_0)\|_F^2$ for each ℓ in the following. We first note that $\mathrm{Rank}(J(\tilde{\alpha}_\ell \to \alpha_L)(x_0)P_\ell) = \mathrm{Rank}\, Jf(x_0) = k$ where P_ℓ is the projection matrix to the image of $J\tilde{\alpha}_\ell(x_0)$.

By AM-GM inequality,

$$\|J(\tilde{\alpha}_\ell \to \alpha_L)(x_0)P_\ell\|_F^2 \geq k|J(\tilde{\alpha}_\ell \to \alpha_L)(x_0)P_\ell|_+^{2/k}\,.$$

Since the parameters are balanced, i.e. $\|w_\ell\|_F^2 + \|b_\ell\|_F^2 = \|w_{\ell+1}\|_F^2$, we have increasing parameter norms $\|W_\ell\|_F^2 \leq \|W_{\ell+1}\|_F^2$ and so

$$\frac{1}{\ell}\sum_{j=1}^{\ell}\|W_j\|_F^2 \leq \frac{1}{L-\ell}\sum_{j=\ell+1}^{L}\|W_j\|_F\,.$$

Thus

$$\begin{aligned}\frac{1}{\ell}\sum_{j=1}^{\ell}\|W_j\|_F^2 &= \frac{1}{L}\sum_{j=1}^{\ell}\|W_j\|_F^2 + \frac{L-\ell}{L}\frac{1}{\ell}\sum_{j=1}^{\ell}\|W_j\|_F^2\\ &\leq \frac{\|\theta\|^2}{L}\end{aligned}$$

and again by AM-GM inequality,

$$\begin{aligned}|P'_\ell J\tilde{\alpha}_\ell(x_0)|_+^{2/kL} &\leq \frac{1}{k}\|P'_\ell J\tilde{\alpha}_\ell(x_0)\|_{2/L}^{2/L}\\ &\leq \frac{1}{k}\frac{\|P'_\ell MW_\ell\|_F^2 + \cdots + \|MW_1\|_F^2}{\ell}\\ &\leq \frac{1}{k}\frac{\|MW_\ell\|_F^2 + \cdots + \|MW_1\|_F^2}{\ell}\\ &\leq \frac{\tilde{m}_{\max}}{k}\frac{\|W_\ell\|_F^2 + \cdots + \|W_1\|_F^2}{\ell}\\ &\leq \frac{\tilde{m}_{\max}\|\theta\|^2}{kL}\\ &\leq \frac{\tilde{m}_{\max}\max_{z\in\Omega_-}\mathrm{Rank}_m(Jf_\theta(z))}{k}\left(1+\frac{c_1}{L\max_{z\in\Omega_-}\mathrm{Rank}_m(Jf_\theta(z))}\right)\end{aligned}$$

where P'_ℓ denotes the projection matrix to the image of $J(\tilde{\alpha}_\ell \to \alpha_L)(x_0)$. For simplicity, we denote $R = \max_{z\in\Omega_-}\mathrm{Rank}_m(Jf_\theta(z))$. Therefore, we have

$$\begin{aligned}\|J(\tilde{\alpha}_\ell \to f_\theta)(x_0)P_\ell\|_F^2 &\geq k|J(\tilde{\alpha}_\ell \to f_\theta)(x_0)P_\ell|_+^{2/k}\\ &= k\frac{|Jf_\theta(x_0)|_+^{2/k}}{|J\tilde{\alpha}_\ell(x_0)|_+^{2/k}}\\ &\geq k\frac{|Jf_\theta(x_0)|_+^{2/k}}{(\tilde{m}_{\max}R/k)^L\left(1+\frac{c_1}{LR}\right)^L}\\ &\geq k|Jf_\theta(x_0)|_+^{2/k}e^{-\left(\frac{c_1}{R}+L(\frac{\tilde{m}_{\max}R}{k}-1)\right)}\\ &= k|Jf_\theta(x_0)|_+^{2/k}e^{-\frac{c_1}{R}}e^{-L(\frac{\tilde{m}_{\max}R}{k}-1)}\end{aligned}$$

and hence

$$\sum_{\ell=1}^{L} \|\alpha_{\ell-1}(x_0)\|_2^2 \leq \frac{ce^{\frac{c_1}{R}} e^{L(\frac{\tilde{m}_{\max} R}{k}-1)}}{k|Jf_\theta(x_0)|_+^{2/k}} L$$

which implies that for each $p \in (0,1)$, there are at most pL layers ℓ with

$$\|\alpha_{\ell-1}(x_0)\|_2^2 \geq \frac{1}{p} \frac{ce^{\frac{c_1}{R}} e^{L(\frac{\tilde{m}_{\max} R}{k}-1)}}{k|Jf_\theta(x_0)|_+^{2/k}} .$$

□

Corollary D.3. *When there is no pooling, x_0 maximizes the rank $\mathrm{Rank} Jf(x_0)$, and the above conditions still hold, we have*

$$\sum_{\ell=1}^{L} \|\alpha_{\ell-1}(x_0)\|_2^2 \leq \frac{ce^{\frac{c_1}{k}}}{k|Jf_\theta(x_0)|_+^{2/k}} L.$$

Hence for each $p \in (0,1)$, there are at least $(1-p)L$ layers ℓ with

$$\|\alpha_{\ell-1}(x_0)\|_2^2 \leq \frac{1}{p} \frac{ce^{\frac{c_1}{k}}}{k|Jf_\theta(x_0)|_+^{2/k}} .$$

E. CNNs with Up-sampling and Down-sampling

Here we present the proofs for characterizing all functions that can be represented by s-stride-CNNs as well as finding a 2-frequency decomposition for translationally unique domains (Theorem 5.8).

Proposition E.1. *Any $f \in \mathcal{N}^{(s)}_{n;m,m'}$ if and only if f has a low-frequency decomposition, i.e. $f = h^{(s)} \circ g^{(s)}$ where $g^{(s)}, h^{(s)}$ are s-translationally equivariant piece-wise linear (s-TEPL) functions, $g^{(s)} = g^{(s)}_1 \oplus \cdots \oplus g^{(s)}_k$, and $g^{(s)}_i$ only supports the first $\frac{n}{s}$ frequencies for $i = 1, \ldots, k$.*

Proof. ($\Longleftarrow$) Note $f = h^{(s)} \circ g^{(s)} = h^{(s)} \circ \mathrm{Up}_s \circ id_{\mathrm{Im}\, g^{(s)}} \circ \mathrm{Down}_s \circ g^{(s)} \in \mathcal{N}^{(s)}_{n;m,m'}$, where $id_{\mathrm{Im}\, g^{(s)}}$ can be the identity layer as we constructed before.

($\Longrightarrow$) To see the other direction, observe that $f \in \mathcal{N}^{(s)}_{n;m,m'}$ gives $f = h^{(s)} \circ (\mathrm{Up}_s \circ \hat{f} \circ \mathrm{Down}_s \circ g^{(s)})$ where the latter is a low-frequency s-TEPL function. □

Theorem E.2. *Suppose Ω is translationally unique. Then for any piecewise linear target function $f : \Omega \to \mathbb{R}^{n \times c_{out}}$, $f = h \circ g^{low}$ where h and g^{low} are TEPL functions and $g^{low} : \Omega \to \mathbb{R}^{n \times nc_{in}+1}$ only supports the constant DFT frequency at first nc_{in} channels and the second DFT frequency at the $nc_{in} + 1$-th channel.*

Proof. Let $\overline{\Omega} = \{T_p x : x \in \Omega, p = 0, \ldots, n-1\}$ be the translational closure of the domain Ω. Since Ω is bounded, so is $\overline{\Omega}$, and hence without loss of generality, we may assume $\overline{\Omega}$ lies in the first quarter and is upper bounded by $Z \geq 1$ coordinate-wise. By Lemma G.1, it suffices to show there exists a TEPL function

$$F : \overline{\Omega} \xrightarrow{G} \left(\Omega \times \{\cos(2\pi p/n)\}_{p=0}^{n-1}\right)^n \xrightarrow{H} \mathbb{R}^{n \times c_{out}}$$

such that $F|_\Omega = f$ and $F = H \circ G$ where G, H are TEPL and G has a low-frequency support at each channel. Define G and H as follows:

$$\begin{aligned} G(T_p x)_{i,0:nc_{in}-1} &= \mathrm{vec}(x) \\ G(T_p x)_{i,nc_{in}} &= \cos(2\pi(p-i)/n) \\ H(G(T_p x)) &= T_p f(x) \end{aligned}$$

for $i \in [n]$. Translational equivariance follows directly from the definition; it remains to verify that G and H are piecewise linear. We first show G^{-1} is piecewise linear by showing it can be represented by a 3-layer no-pooling ConvNet:

(First layer). Denote a threshold $\epsilon = \max_{i\neq p}\cos(2\pi(p-i)/n)$; note $\epsilon < 1$. Let $(w_1)_{i,c,s} = \delta_{i=0}\delta_{c=s}$ for $c \in [nc_{in}]$, where δ is the indicator function (so $w \circledast x = x$). Let $(b_1)_c = -\epsilon\delta_{c=nc_{in}}$ for $c \in [nc_{in}]$. After applying the ReLU, we have activation

$$\begin{aligned}\alpha_1(G(T_px))_{i,0:nc_{in}-1} &= G(T_px)_{i,0:nc_{in}-1}\\ \alpha_1(G(T_px))_{i,nc_{in}} &= \delta_{i=p}(1-\epsilon)\end{aligned}$$

(Second layer). Let $(w_2)_{i,c,s} = \delta_{i=0}\delta_{c=s}$ for $c \in [nc_{in}-1]$ and $\delta_{i=0}Z/(1-\epsilon)$ for $c = nc_{in}$, for $s \in [nc_{in}]$, where δ is the indicator function. Let $(b_2)_c = -Z$ for $c \in [nc_{in}]$. Then we have

$$\begin{aligned}\tilde{\alpha}_2(G(T_px))_{i,0:nc_{in}-1} &= G(T_px)_{i,0:nc_{in}-1} - \delta_{i\neq p}Z\\ \tilde{\alpha}_2(G(T_px))_{i,nc_{in}} &= -\delta_{i\neq p}Z\\ \alpha_2(G(T_px))_{i,0:nc_{in}-1} &= \delta_{i=p}G(T_px)_{i,0:nc_{in}-1}\\ &= \delta_{i=p}\mathrm{vec}(x)\\ \alpha_2(G(T_px))_{i,nc_{in}} &= 0\end{aligned}$$

since x is coordinate-wise upper bounded by $Z > 0$.

(Third layer). Let $(w_3)_{i,c,s} = \delta_{i=c \mod n}\delta_{c=\lfloor s/n\rfloor}$ and $(b_3)_c = 0$ for channel $c \in [c_{in}]$ and $s \in [nc_{in}]$, i.e. translating the s-th channel of the input by $s \mod n$ and then summing every n channels. Then we have the output being

$$\alpha_3(G(T_px)) = T_px.$$

Hence G^{-1} is TEPL and $G = (G^{-1})^{-1}$ is TEPL.

One can see $H = f \circ G^{-1}$ is also TEPL. By letting $g^{low} = G|_\Omega$ and $h = H|_{\mathrm{Im}\, g^{low}}$, we see they are TEPL and each channel of g^{low} only either supports the first or the second DFT coefficient. □

F. Representation Cost in Filter Norm

If one considers the representation cost $\|\tilde{\theta}\|^2$ as the norm of the filters (as opposed to the matrices in Section 2.3) and the biases, by definition one has $\|\tilde{\theta}\|^2 = \frac{1}{n}\|\theta\|^2$ off by a factor of $\frac{1}{n}$. One can then adapt the results and proofs in this paper to the filter norm by substituting $\|\tilde{\theta}\|^2 = \frac{1}{n}\|\theta\|^2$ and get this extra factor in the expressions. For example, let $\widetilde{R}^{(0)}$ and $\widetilde{R}^{(1)}$ denote the costs based on the filter norm; one can get the upper and lower bounds

$$\frac{1}{n}\max_{x\in\Omega_-}\mathrm{Rank}_m(Jf(x)) \leq \widetilde{R}^{(0)}(f;\Omega) \leq \frac{1}{n}\mathrm{Rank}_{\mathrm{CBN}}(f;\Omega)$$

and the regularity control becomes

$$\widetilde{R}^{(1)}(f;\Omega) \geq \frac{2}{n}\sum_{s_{t,c}\neq 0}\tilde{m}_t^{-2}\log(s_{t,c}\tilde{m}_t).$$

G. Auxiliary Lemmas

This section proves a version of Theorem 2.1 in (Arora et al., 2016) for the CNN case and may be of independent interest.

Lemma G.1 (Bounded depth). *For any TEPL function $F : \Omega \subset \mathbb{R}^{n\times c_{in}} \to \mathbb{R}^{n\times c_{out}}$ and $L \geq \lceil\log_2(nc_{in}+1)\rceil + 2$, there is a CNN $f_\theta = F$ with widths $\{c_\ell\}_{\ell=1}^L$ and parameter $\theta = \{w_\ell, b_\ell\}_{\ell=1}^L$.*

Proof. Since the input domain Ω is compact, without loss of generality, we may assume x is positive. Consider $f : \mathbb{R}^{n\times c_{in}} \to \mathbb{R}^{c_{out}}$ with $f(x) = F_1(x) \equiv F(x)_{1,:}$ i.e. the first input at every channel. Note f is piecewise linear and $F_p(x) = f(T_{-p}x)$ where $(T_{-p}x)_{i,:} = x_{i-p,:}$ is the translation of x by $-p$. By Theorem 2.1 in (Arora et al., 2016), there is a ReLU fully-connected network $f_{\boldsymbol{A}}^{FC} = f$ with depth $L-1$, widths $\{n_\ell\}_{\ell=1}^{L-1}$, and parameter $\boldsymbol{A} = \{A_\ell, d_\ell\}_{\ell=1}^L$. Then we can construct an L-layer no-pooling ConvNet to represent F as follows:

(First layer) Construct $w_1 \in \mathbb{R}^{n \times nc_{in} \times c_{in}}$ with $(w_1)_{p,c,s} = \delta_{p=n-\lfloor c/c_{in} \rfloor}$ and $b_1 = 0$ such that

$$\alpha_1(x)_{p,:} = \tilde{\alpha}_1(x)_{p,:} = \text{vec}(T_{-p}x), \quad p = 0, \ldots, n-1.$$

where the first equality follows from assuming x is positive. (i.e. This is done by shifting the identity convolution filter by $-p$ at channel $c \in \{pc_{in}, \ldots, (p+1)c_{in} - 1\}$, then convolve with m^{-1}.) Then we treat each of the $T_{-p}x$ "independently" in the following layers.

(Following layers) For $\ell > 1$, construct $w_\ell \in \mathbb{R}^{n \times n_\ell \times n_{\ell-1}}$ by letting $(w_\ell)_{i,c,s} = (A_\ell)_{c,s}\delta_{i=0}$ and $(b_\ell)_c = (d_\ell)_c$ for $i = 0, \ldots, n-1, c = 1, \ldots, n_\ell$, and $s = 1, \ldots, n_{\ell-1}$. One can verify that for $p = 0, \ldots, n-1$,

$$\begin{aligned}
\tilde{\alpha}_\ell(x)_{p,c} &= (b_\ell)_c + \sum_{s=1}^{n_{\ell-1}} \sum_{i=0}^{n-1} (w_\ell)_{i,c,s}\alpha_{\ell-1}(x)_{p-i,s} \\
&= (d_\ell)_c + \sum_{s=1}^{n_{\ell-1}} (A_\ell)_{c,s}\alpha_{\ell-1}(x)_{p,s} \\
\implies \tilde{\alpha}_\ell(x)_{p,:} &= A_\ell \alpha_{\ell-1}(x)_{p,:} + d_\ell.
\end{aligned}$$

Now at the output layer $\ell = L$, by construction of $\boldsymbol{A}$, we have $\alpha_L(x)_{p,:} = f(\alpha_1(x)_{p,:}) = F_p(x)$. Hence the output of this CNN is $F(x)$. □