How DNNs break the Curse of Dimensionality: Compositionality and Symmetry Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

We show that deep neural networks (DNNs) can efficiently learn any composition of functions with bounded F_1-norm, which allows DNNs to break the curse of dimensionality in ways that shallow networks cannot. More specifically, we derive a generalization bound that combines a covering number argument for compositionality, and the F_1-norm (or the related Barron norm) for large width adaptivity. We show that the global minimizer of the regularized loss of DNNs can fit for example the composition of two functions $f^* = h \circ g$ from a small number of observations, assuming g is smooth/regular and reduces the dimensionality (e.g. g could be the modulo map of the symmetries of f^*), so that h can be learned in spite of its low regularity. The measures of regularity we consider is the Sobolev norm with different levels of differentiability, which is well adapted to the F_1 norm. We compute scaling laws empirically and observe phase transitions depending on whether g or h is harder to learn, as predicted by our theory.

1 Introduction

One of the fundamental features of DNNs is their ability to generalize even when the number of neurons (and of parameters) is so large that the network could fit almost any function [46]. Actually DNNs have been observed to generalize best when the number of neurons is infinite [8, 21, 20]. The now quite generally accepted explanation to this phenomenon is that DNNs have an implicit bias coming from the training dynamic where properties of the training algorithm lead to networks that generalize well. This implicit bias is quite well understood in shallow networks [11, 36], in linear networks [24, 30], or in the NTK regime [28], but it remains ill-understood in the general deep nonlinear case.

In both shallow networks and linear networks, one observes a bias towards small parameter norm (either implicit [12] or explicit in the presence of weight decay [42]). Thanks to tools such as the F_1-norm [5], or the related Barron norm [44], or more generally the representation cost [14], it is possible to describe the family of functions that can be represented by shallow networks or linear networks with a finite parameter norm. This was then leveraged to prove uniform generalization bounds (based on Rademacher complexity) over these sets [5], which depend only on the parameter norm, but not on the number of neurons or parameters.

Similar bounds have been proposed for DNNs [7, 6, 39, 33, 25, 40], relying on different types of norms on the parameters of the network. But it seems pretty clear that we have not yet identified the ‘right’ complexity measure for deep networks, as there remains many issues: these bounds are typically orders of magnitude too large [29, 23], and they tend to explode as the depth L grows [40].

Two families of bounds are particularly relevant to our analysis: bounds based on covering numbers which rely on the fact that one can obtain a covering of the composition of two function classes from covering of the individual classes [7, 25], and path-norm bounds which extend the techniques behind the F_1-norm bound from shallow networks to the deep case [32, 6, 23].

Another issue is the lack of approximation results to accompany these generalization bounds: many different complexity measures $R(\theta)$ on the parameters θ of DNNs have been proposed along with guarantees that the generalization gap will be small as long as $R(\theta)$ is bounded, but there are often little to no result describing families of functions that can be approximated with a bounded $R(\theta)$ norm. The situation is much clearer in shallow networks, where we know that certain Sobolev spaces can be approximated with bounded F_1-norm [5].

We will focus on approximating composition of Sobolev functions, and obtaining close to optimal rates. This is quite similar to the family of tasks considered [39], though the complexity measure we consider is quite different, and does not require sparsity of the parameters.

1.1 Contribution

We consider Accordion Networks (AccNets), which are the composition of multiple shallow networks $f_{L:1} = f_L \circ \cdots \circ f_1$, we prove a uniform generalization bound $\mathcal{L}(f_{L:1}) - \tilde{\mathcal{L}}_N(f_{L:1}) \lesssim R(f_1, \dots, f_L)\frac{\log N}{\sqrt{N}}$, for a complexity measure

$$R(f_1, \dots, f_L) = \prod_{\ell=1}^{L} Lip(f_\ell) \sum_{\ell=1}^{L} \frac{\|f_\ell\|_{F_1}}{Lip(f_\ell)} \sqrt{d_\ell + d_{\ell-1}}$$

that depends on the F_1-norms $\|f_\ell\|_{F_1}$ and Lipschitz constanst $Lip(f_\ell)$ of the subnetworks, and the intermediate dimensions $d_0, \dots, d_L$. This use of the F_1-norms makes this bound independent of the widths $w_1, \dots, w_L$ of the subnetworks, though it does depend on the depth L (it typically grows linearly in L which is still better than the exponential growth often observed).

Any traditional DNN can be mapped to an AccNet (and vice versa), by spliting the middle weight matrices W_ℓ with SVD USV^T into two matrices $U\sqrt{S}$ and $\sqrt{S}V^T$ to obtain an AccNet with dimensions $d_\ell = \mathrm{Rank} W_\ell$, so that the bound can be applied to traditional DNNs with bounded rank.

We then show an approximation result: any composition of Sobolev functions $f^* = f^*_{L^*} \circ \cdots \circ f^*_1$ can be approximated with a network with either a bounded complexity $R(\theta)$ or a slowly growing one. Thus under certain assumptions one can show that DNNs can learn general compositions of Sobolev functions. This ability can be interpreted as DNNs being able to learn symmetries, allowing them to avoid the curse of dimensionality in settings where kernel methods or even shallow networks suffer heavily from it.

Empirically, we observe a good match between the scaling laws of learning and our theory, as well as qualitative features such as transitions between regimes depending on whether it is harder to learn the symmetries of a task, or to learn the task given its symmetries.

2 Accordion Neural Networks and ResNets

Our analysis is most natural for a slight variation on the traditional fully-connected neural networks (FCNNs), which we call Accordion Networks, which we define here. Nevertheless, all of our results can easily be adapted to FCNNs.

Accordion Networks (AccNets) are simply the composition of L shallow networks, that is $f_{L:1} = f_L \circ \cdots \circ f_1$ where $f_\ell(z) = W_\ell \sigma(V_\ell z + b_\ell)$ for the nonlinearity $\sigma : \mathbb{R} \to \mathbb{R}$, the $d_\ell \times w_\ell$ matrix W_ℓ , $w_\ell \times d_{\ell-1}$ matrix V_ℓ, and w_ℓ-dim. vector b_ℓ, and for the widths $w_1, \dots, w_L$ and dimensions $d_0, \dots, d_L$. We will focus on the ReLU $\sigma(x) = \max\{0, x\}$ for the nonlinearity. The parameters θ are made up of the concatenation of all (W_ℓ, V_ℓ, b_ℓ). More generally, we denote $f_{\ell_2:\ell_1} = f_{\ell_2} \circ \cdots \circ f_{\ell_1}$ for any $1 \leq \ell_1 \leq \ell_2 \leq L$.

We will typically be interested in settings where the widths w_ℓ is large (or even infinitely large), while the dimensions d_ℓ remain finite or much smaller in comparison, hence the name accordion.

If we add residual connections, i.e. $f^{res}_{1:L} = (f_L + id) \circ \cdots \circ (f_1 + id)$ for the same shallow nets $f_1, \dots, f_L$ we recover the typical ResNets.

Remark. The only difference between AccNets and FCNNs is that each weight matrix M_ℓ of the FCNN is replaced by a product of two matrices $M_\ell = V_\ell W_{\ell-1}$ in the middle of the network (such a structure has already been proposed [34]). Given an AccNet one can recover an equivalent FCNN by choosing $M_\ell = V_\ell W_{\ell-1}$, $M_0 = V_0$ and $M_{L+1} = W_L$. In the other direction there could be multiple ways to split M_ℓ into the product of two matrices, but we will focus on taking $V_\ell = U\sqrt{S}$ and $W_{\ell-1} = \sqrt{S}V^T$ for the SVD decomposition $M_\ell = USV^T$, along with the choice $d_\ell = \mathrm{Rank} M_\ell$.

2.1 Learning Setup

We consider a traditional learning setup, where we want to find a function $f : \Omega \subset \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{out}}$ that minimizes the population loss $\mathcal{L}(f) = \mathbb{E}_{x\sim\pi}\left[\ell(x, f(x))\right]$ for an input distribution π and a ρ-Lipschitz and ρ-bounded loss function $\ell(x, y) \in [0, B]$. Given a training set $x_1, \dots, x_N$ of size N we approximate the population loss by the empirical loss $\tilde{\mathcal{L}}_N(f) = \frac{1}{N}\sum_{i=1}^N \ell(x_i, f(x_i))$ that can be minimized.

To ensure that the empirical loss remains representative of the population loss, we will prove high probability bounds on the generalization gap $\tilde{\mathcal{L}}_N(f) - \mathcal{L}(f)$ uniformly over certain functions families $f \in \mathcal{F}$.

For **regression tasks**, we assume the existence of a true function f^* and try to minimize the distance $\ell(x, y) = \|f^*(x) - y\|^p$ for $p \geq 1$. If we assume that $f^*(x)$ and y are uniformly bounded then one can easily show that $\ell(x, y)$ is bounded and Lipschitz. We are particularly interested in the cases $p \in \{1, 2\}$, with $p = 2$ representing the classical MSE, and $p = 1$ representing a L_1 distance. The $p = 2$ case is amenable to 'fast rates' which take advantage of the fact that the loss increases very slowly around the optimal solution f^*, We do not prove such fast rates (even though it might be possible) so we focus on the $p = 1$ case.

For **classification tasks** on k classes, we assume the existence of a 'true class' function $f^* : \Omega \to \{1, \dots, k\}$ and want to learn a function $f : \Omega \to \mathbb{R}^k$ such that the largest entry of $f(x)$ is the $f^*(k)$-th entry. One can consider the hinge cost $\ell(x, y) = \max\{0, 1 - (y_{f^*(k)} - \max_{i\neq f^*(x)} y_i)\}$, which is zero whenever the margin $y_{f^*(k)} - \max_{i\neq f^*(x)} y_i$ is larger than 1 and otherwise equals 1 minus the margin. The hinge loss is Lipschitz and bounded if we assume bounded outputs $y = f(x)$. The cross-entropy loss also fits our setup.

3 Generalization Bound for DNNs

The reason we focus on accordion networks is that there exists generalization bounds for shallow networks [5, 44], that are (to our knowledge) widely considered to be tight, which is in contrast to the deep case, where many bounds exist but no clear optimal bound has been identified. Our strategy is to extend the results for shallow nets to the composition of multiple shallow nets, i.e. AccNets. Roughly speaking, we will show that the complexity of an AccNet f_θ is bounded by the sum of the complexities of the shallow nets $f_1, \dots, f_L$ it is made of.

We will therefore first review (and slightly adapt) the existing generalization bounds for shallow networks in terms of their so-called F_1-norm [5], and then prove a generalization bound for deep AccNets.

3.1 Shallow Networks

The complexity of a shallow net $f(x) = W\sigma(Vx + b)$, with weights $W \in \mathbb{R}^{w\times d_{out}}$ and $V \in \mathbb{R}^{d_{in}\times w}$, can be bounded in terms of the quantity $C = \sum_{i=1}^w \|W_{\cdot i}\| \sqrt{\|V_{i\cdot}\|^2 + b_i^2}$. First note that the rescaled function $\frac{1}{C}f$ can be written as a convex combination $\frac{1}{C}f(x) = \sum_{i=1}^w \frac{\|W_{\cdot i}\|\sqrt{\|V_{i\cdot}\|^2+b_i^2}}{C}\bar{W}_{\cdot i}\sigma(\bar{V}_{i\cdot}x + \bar{b}_i)$ for $\bar{W}_{\cdot i} = \frac{W_{\cdot i}}{\|W_{\cdot i}\|}$, $\bar{V}_{i\cdot} = \frac{V_{i\cdot}}{\sqrt{\|V_{i\cdot}\|^2+b_i^2}}$, and $\bar{b}_i = \frac{b_i}{\sqrt{\|V_{i\cdot}\|^2+b_i^2}}$, since the coefficients $\frac{\|W_{\cdot i}\|\sqrt{\|V_{i\cdot}\|^2+b_i^2}}{C}$ are positive and sum up to 1. Thus f belongs to C times the convex hull

$$B_{F_1} = \mathrm{Conv}\left\{x \mapsto w\sigma(v^T x + b) : \|w\|^2 = \|v\|^2 + b^2 = 1\right\}.$$

We call this the F_1-ball since it can be thought of as the unit ball w.r.t. the F_1-norm $\|f\|_{F_1}$ which we define as the smallest positive scalar s such that[1] $\frac{1}{s}f \in B_{F_1}$. For more details in the single output case, see [5].

The generalization gap over any F_1-ball can be uniformly bounded with high probability:

Theorem 1. *For any input distribution π supported on the L_2 ball $B(0,b)$ with radius b, we have with probability $1-\delta$, over the training samples $x_1,\dots,x_N$, that for all $f \in B_{F_1}(0,R) = R\cdot B_{F_1}$*

$$\mathcal{L}(f) - \tilde{\mathcal{L}}_N(f) \leq \rho b R\sqrt{d_{in}+d_{out}}\frac{\log N}{\sqrt{N}} + c_0\sqrt{\frac{2\log {}^2/_\delta}{N}}$$

This theorem is a slight variation of the one found in [5]: we simply generalize it to multiple outputs, and also prove it using a covering number argument instead of a direct computation of the Rademacher complexity, which will be key to obtaining a generalization bound for the deep case. But due to this change of strategy we pay a $\log N$ cost here and in our later results. We know that the $\log N$ term can be removed in Theorem 1 by switching to a Rademacher argument, but we do not know whether it can be removed in deep nets.

Notice how this bound does not depend on the width w, because the F_1-norm (and the F_1-ball) themselves do not depend on the width. This matches with empirical evidence that shows that increasing the width does not hurt generalization [8, 21, 20].

To use Theorem 1 effectively we need to be able to guarantee that the learned function will have a small enough F_1-norm. The F_1-norm is hard to compute exactly, but it is bounded by the parameter norm: if $f(x) = W\sigma(Vx+b)$, then $\|f\|_{F_1} \leq \frac{1}{2}\left(\|W\|_F^2 + \|V\|_F^2 + \|b\|^2\right)$, and this bound is tight if the width w is large enough and the parameters are chosen optimally. Adding weight decay/L_2-regularization to the cost then leads to bias towards learning with small F_1 norm.

3.2 Deep Networks

Since an AccNet is simply the composition of multiple shallow nets, the functions represented by an AccNet is included in the set of composition of F_1 balls. More precisely, if $\|W_\ell\|^2 + \|V_\ell\|^2 + \|b_\ell\|^2 \leq 2R_\ell$ then $f_{L:1}$ belongs to the set $\{g_L \circ \cdots \circ g_1 : g_\ell \in B_{F_1}(0, R_\ell)\}$ for some R_ℓ, which is width agnostic.

As already noticed in [7], the covering number number is well-behaved under composition, this allows us to bound the complexity of AccNets in terms of the individual shallow nets it is made up of:

Theorem 2. *Consider an accordion net of depth L and widths $d_L,\dots,d_0$, with corresponding set of functions $\mathcal{F} = \{f_{L:1} : \|f_\ell\|_{F_1} \leq R_\ell, Lip(f_\ell) \leq \rho_\ell\}$. With probability $1-\delta$ over the sampling of the training set X from the distribution π supported in $B(0,b)$, we have for all $f \in \mathcal{F}$*

$$\mathcal{L}(f) - \tilde{\mathcal{L}}_N(f) \leq C\rho b\rho_{L:1}\sum_{\ell=1}^{L}\frac{R_\ell}{\rho_\ell}\sqrt{d_\ell + d_{\ell-1}}\frac{\log N}{\sqrt{N}}(1+o(1)) + c_0\sqrt{\frac{2\log {}^2/_\delta}{N}}.$$

Theorem 2 can be extended to ResNets $(f_L + id)\circ\cdots\circ(f_1+id)$ by simply replacing the Lipschitz constant $Lip(f_\ell)$ by $Lip(f_\ell + id)$.

The Lipschitz constants $Lip(f_\ell)$ are difficult to compute exactly, so it is easiest to simply bound it by the product of the operator norms $Lip(f_\ell) \leq \|W_\ell\|_{op}\|V_\ell\|_{op}$, but this bound can often be quite loose. The fact that our bound depends on the Lipschitz constants rather than the operator norms $\|W_\ell\|_{op}, \|V_\ell\|_{op}$ is thus a significant advantage.

This bound can be applied to a FCNNs with weight matrices $M_1,\dots,M_{L+1}$, by replacing the middle M_ℓ with SVD decomposition USV^T in the middle by two matrices $W_{\ell-1} = \sqrt{S}V^T$ and $V_\ell = U\sqrt{S}$, so that the dimensions can be chosen as the rank $d_\ell = \mathrm{Rank}M_{\ell+1}$. The Frobenius norm of the new matrices equal the nuclear norm of the original one $\|W_{\ell-1}\|_F^2 = \|V_\ell\|_F^2 = \|M_\ell\|_*$. Some bounds

[1]This construction can be used for any convex set B that is symmetric around zero ($B=-B$) to define a norm whose unit ball is B. This correspondence between symmetric convex sets and norms is well known.

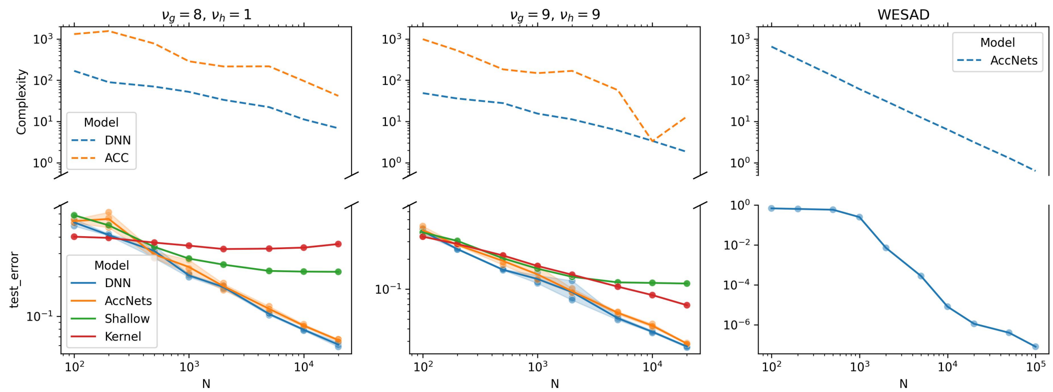

Figure 1: Visualization of scaling laws. We observe that deep networks (either AccNets or DNNs) achieve better scaling laws than kernel methods or shallow networks on certain compositional tasks, in agreement with our theory. We also see that our new generalization bounds approximately recover the right saling laws (even though they are orders of magnitude too large overall). We consider a compositional true function $f^* = h \circ g$ where g maps from dimension 15 to 3 while h maps from 3 to 20, and we denote ν_g, ν_h for the number of times g, h are differentiable. In the first plot $\nu_g = 8, \nu_h = 1$ so that g is easy to learn while h is hard, whereas in the second plot $\nu_g = 9, \nu_h = 9$, so both g and h are relatively easier. The third plot presents the decay in test error and generalization bounds for networks evaluated using the real-world dataset, WESAD [37].

assuming rank sparsity of the weight matrices also appear in [41]. And several recent results have shown that weight-decay leads to a low-rank bias on the weight matrices of the network [27, 26, 19] and replacing the Frobenius norm regularization with a nuclear norm regularization (according to the above mentioned equivalence) will only increase this low-rank bias.

We compute in Figure 1 the upper bound of Theorem 2 for both AccNets and DNNs, and even though we observe a very large gap (roughly of order 10^3), we do observe that it captures rate/scaling of the test error (the log-log slope) well. So this generalization bound could be well adapted to predicting rates, which is what we will do in the next section.

Remark. Note that if one wants to compute this upper bound in practical setting, it is important to train with L_2 regularization until the parameter norm also converges (this often happens after the train and test loss have converged). The intuition is that at initialization, the weights are initialized randomly, and they contribute a lot to the parameter norm, and thus lead to a larger generalization bound. During training with weight decay, these random initial weights slowly vanish, thus leading to a smaller parameter norm and better generalization bound. It might be possible to improve our generalization bounds to take into account the randomness at initialization to obtain better bounds throughout training, but we leave this to future work.

4 Breaking the Curse of Dimensionality with Compositionality

In this section we study a large family of functions spaces, obtained by taking compositions of Sobolev balls. We focus on this family of tasks because they are well adapted to the complexity measure we have identified, and because kernel methods and even shallow networks do suffer from the curse of dimensionality on such tasks, whereas deep networks avoid it (e.g. Figure 1).

More precisely, we will show that these sets of functions can be approximated by a AccNets with bounded (or in some cases slowly growing) complexity measure

$$R(f_1, \ldots, f_L) = \prod_{\ell=1}^{L} Lip(f_\ell) \sum_{\ell=1}^{L} \frac{\|f_\ell\|_{F_1}}{Lip(f_\ell)} \sqrt{d_\ell + d_{\ell-1}}.$$

This will then allow us show that AccNets can (assuming global convergence) avoid the curse of dimensionality, even in settings that should suffer from the curse of dimensionality, when the input dimension is large and the function is not very smooth (only a few times differentiable).

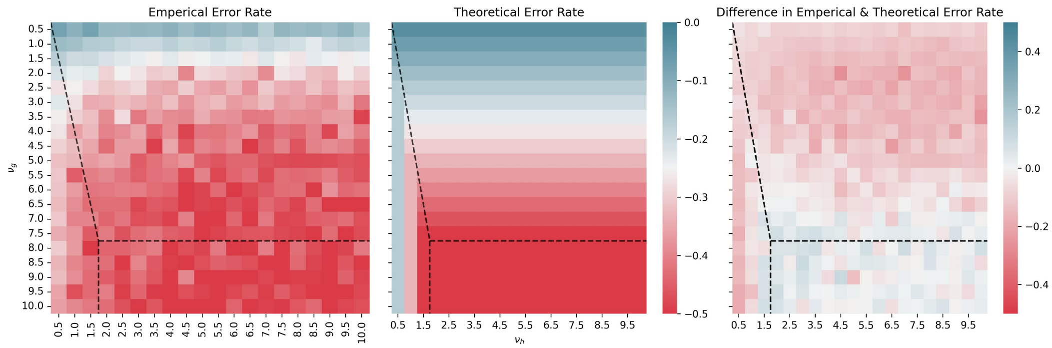

Figure 2: A comparison of empirical and theoretical error rates. The first plot illustrates the log decay rate of the test error with respect to the dataset size N based on our empirical simulations. The second plot depicts the theoretical decay rate of the test error as discussed in Section 4.1, $-\min\{\frac{1}{2}, \frac{\nu_g}{d_{in}}, \frac{\nu_h}{d_{mid}}\}$. The final plot on the right displays the difference between the two. The lower left region represents the area where g is easier to learn than h, the upper right where h is easier to learn than g, and the lower right region where both f and g are easy.

.

4.1 Composition of Sobolev Balls

The family of Sobolev norms capture some notion of regularity of a function, as it measures the size of its derivatives. The Sobolev norm of a function $f : \mathbb{R}^{d_{in}} \to \mathbb{R}$ is defined in terms of its derivatives $\partial_x^\alpha f$ for some d_{in}-multi-index α, namely the $W^{\nu,p}(\pi)$-Sobolev norm with integer ν and $p \geq 1$ is defined as

$$\|f\|_{W^{\nu,p}(\pi)}^p = \sum_{|\alpha| \leq \nu} \|\partial_x^\alpha f\|_{L_p(\pi)}^p .$$

Note that the derivative $\partial_x^\alpha f$ only needs to be defined in the 'weak' sense, which means that even non-differentiable functions such as the ReLU functions can actually have finite Sobolev norm.

The Sobolev balls $B_{W^{\nu,p}(\pi)}(0, R) = \{f : \|f\|_{W^{\nu,p}(\pi)} \leq R\}$ are a family of function spaces with a range of regularity (the larger ν, the more regular). This regularity makes these spaces of functions learnable purely from the fact that they enforce the function f to vary slowly as the input changes. Indeed we can prove the following generalization bound:

Proposition 3. *Given a distribution π with support the L_2 ball with radius b, we have that with probability $1 - \delta$ for all functions $f \in \mathcal{F} = \{f : \|f\|_{W^{\nu,2}} \leq R, \|f\|_\infty \leq R\}$*

$$\mathcal{L}(f) - \tilde{\mathcal{L}}_N(f) \leq 2\rho C_1 R E_{\nu/d}(N) + c_0\sqrt{\frac{2\log {}^2/_\delta}{N}}.$$

where $E_r(N) = N^{-\frac{1}{2}}$ if $r > \frac{1}{2}$, $E_r(N) = N^{-\frac{1}{2}} \log N$ if $r = \frac{1}{2}$, and $E_r(N) = N^{-r}$ if $r < \frac{1}{2}$.

But this result also illustrates the **curse of dimensionality:** the differentiability ν needs to scale with the input dimension d_{in} to obtain a reasonable rate. If instead ν is constant and d_{in} grows, then the number of datapoints N needed to guarantee a generalization gap of at most ϵ scales exponentially in d_{in}, i.e. $N \sim \epsilon^{-\frac{d_{in}}{\nu}}$. One way to interpret this issue is that regularity becomes less and less useful the larger the dimension: knowing that similar inputs have similar outputs is useless in high dimension where the closest training point x_i to a test point x is typically very far away.

4.1.1 Breaking the Curse of Dimensionality with Compositionality

To break the curse of dimensionality, we need to assume some additional structure on the data or task which introduces an 'intrinsic dimension' that can be much lower than the input dimension d_{in}:

Manifold hypothesis: If the input distribution lies on a d_{surf}-dimensional manifold, the error rates typically depends on d_{surf} instead of d_{in} [38, 10].

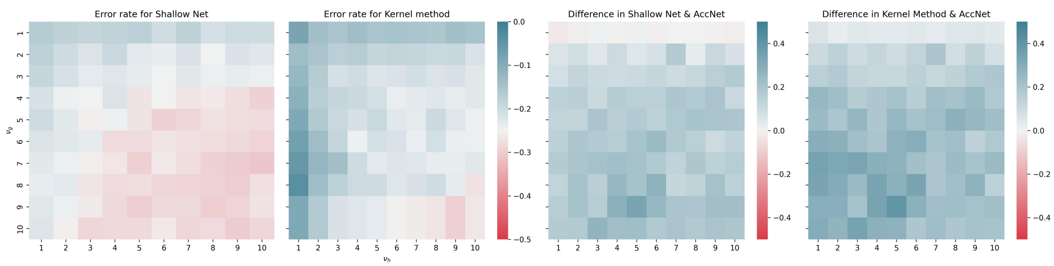

Figure 3: Comparing error rates for shallow and AccNets: shallow nets vs. AccNets, and kernel methods vs. AccNets. The left two graphs shows the empirical decay rate of test error with respect to dataset size (N) for both shallow nets and kernel methods. In contrast to our earlier empirical findings for AccNets, both shallow nets and kernel methods exhibit a slower decay rate in test error. The right two graphs present the differences in log decay rates between shallow nets and AccNets, as well as between kernel methods and AccNets. AccNets almost always obtain better rates, with a particularly large advantage at the bottom and middle-left.

.

Known Symmetries: If $f^*(g \cdot x) = f^*(x)$ for a group action $\cdot$ w.r.t. a group G, then f^* can be written as the composition of a modulo map $g^* : \mathbb{R}^{d_{in}} \to \mathbb{R}^{d_{in}}/G$ which maps pairs of inputs which are equivalent up to symmetries to the same value (pairs x, y s.t. $y = g \cdot x$ for some $g \in G$), and then a second function $h^* : \mathbb{R}^{d_{in}}/G \to \mathbb{R}^{d_{out}}$, then the complexity of the task will depend on the dimension of the modulo space $\mathbb{R}^{d_{in}}/G$ which can be much lower. If the symmetry is known, then one can for example fix g^* and only learn h^* (though other techniques exist, such as designing kernels or features that respect the same symmetries) [31].

Symmetry Learning: However if the symmetry is not known then both g^* and h^* have to be learned, and this is where we require feature learning and/or compositionality. Shallow networks are able to learn translation symmetries, since they can learn so-called low-index functions which satisfy $f^*(x) = f^*(Px)$ for some projection P (with a statistical complexity that depends on the dimension of the space one projects into, not the full dimension [5, 2]). Low-index functions correspond exactly to the set of functions that are invariant under translation along the kernel $\ker P$. To learn general symmetries, one needs to learn both h^* and the modulo map g^* simultaneously, hence the importance of feature learning.

For g^* to be learnable efficiently, it needs to be regular enough to not suffer from the curse of dimensionality, but many traditional symmetries actually have smooth modulo maps, for example the modulo map $g^*(x) = \|x\|^2$ for rotation invariance. This can be understood as a special case of composition of Sobolev functions, whose generalization gap can be bounded:

Theorem 4. *Consider the function set* $\mathcal{F} = \mathcal{F}_L \circ \cdots \circ \mathcal{F}_1$ *where* $\mathcal{F}_\ell = \left\{ f_\ell : \mathbb{R}^{d_{\ell-1}} \to \mathbb{R}^{d_\ell} \text{ s.t. } \|f_\ell\|_{W^{\nu_\ell,2}} \leq R_\ell, \|f_\ell\|_\infty \leq b_\ell, Lip(f_\ell) \leq \rho_\ell \right\}$*, and let* $r_{min} = \min_\ell r_\ell$ *for* $r_\ell = \frac{\nu_\ell}{d_{\ell-1}}$*, then with probability* $1-\delta$ *we have for all* $f \in \mathcal{F}$

$$\mathcal{L}(f) - \tilde{\mathcal{L}}_N(f) \leq \rho C_0 \left(\sum_{\ell=1}^{L} \left(C_\ell \rho_{L:\ell+1} R_\ell \right)^{\frac{1}{r_{min}+1}} \right)^{r_{min}+1} E_{r_{min}}(N) + c_0 \sqrt{\frac{2 \log 2/\delta}{N}},$$

where C_ℓ *depends only on* $d_{\ell-1}, d_\ell, \nu_\ell, b_{\ell-1}$.

We see that only the smallest ratio r_{min} matters when it comes to the rate of learning. And actually the above result could be slightly improved to show that the sum over all layers could be replaced by a sum over only the layers where the ratio r_ℓ leads to the worst rate $E_{r_\ell}(N) = E_{r_{min}}(N)$ (and the other layers contribute an asymptotically subdominant amount).

Coming back to the symmetry learning example, we see that the hardness of learning a function of the type $f^* = h \circ g$ with inner dimension d_{mid} and regularities ν_g and ν_h, the error rate will be (up to log terms) $N^{-\min\{\frac{1}{2}, \frac{\nu_g}{d_{in}}, \frac{\nu_h}{d_{mid}}\}}$. This suggests the existence of three regimes depending on which term attains the minimum: a regime where both g and h are easy to learn and we have $N^{-\frac{1}{2}}$ learning, a regime g is hard, and a regime where h is hard. The last two regimes differentiate between tasks where learning the symmetry is hard and those where learning the function knowing its symmetries is hard.

In contrast, without taking advantage of the compositional structure, we expect f^* to be only $\min\{\nu_g, \nu_h\}$ times differentiable, so trying to learn it as a single Sobolev function would lead to an error rate of $N^{-\min\{\frac{1}{2}, \frac{\min\{\nu_g,\nu_h\}}{d_{in}}\}} = N^{-\min\{\frac{1}{2}, \frac{\nu_g}{d_{in}}, \frac{\nu_h}{d_{in}}\}}$ which is no better than the compositional rate, and is strictly worse whenever $\nu_h < \nu_g$ and $\frac{\nu_h}{d_{in}} < \frac{1}{2}$ (we can always assume $d_{mid} \leq d_{in}$ since one could always choose $d = id$).

Furthermore, since multiple compositions are possible, one can imagine a hierarchy of symmetries that slowly reduce the dimensionality with less and less regular modulo maps. For example one could imagine a composition $f_L \circ \cdots \circ f_1$ with dimensions $d_\ell = d_0 2^{-\ell}$ and regularities $\nu_\ell = d_0 2^{-\ell}$ so that the ratios remain constant $r_\ell = \frac{d_0 2^{-\ell}}{d_0 2^{-\ell+1}} = \frac{1}{2}$, leading to an almost parametric rate of $N^{-\frac{1}{2}} \log N$ even though the function may only be $d_0 2^{-L}$ times differentiable. Without compositionality, the rate would only be $N^{-2^{-L}}$.

Remark. In the case of a single Sobolev function, one can show that the rate $E_{\nu/d}(N)$ is in some sense optimal, by giving an information theoretic lower bound with matching rate. A naive argument suggests that the rate of $E_{\min\{r_1,\dots,r_L\}}(N)$ should similarly be optimal: assume that the minimum r_ℓ is attained at a layer ℓ, then one can consider the subset of functions such that the image $f_{\ell-1:1}(B(0,r))$ contains a ball $B(z,r') \subset \mathbb{R}^{d_{\ell-1}}$ and that the function $f_{L:\ell+1}$ is β-non-contracting $\|f_{L:\ell+1}(x) - f_{L:\ell+1}(y)\| \geq \beta \|x-y\|$, then learning $f_{L:1}$ should be as hard as learning f_ℓ over the ball $B(z,r')$ (more rigorously this could be argued from the fact that any ϵ-covering of $f_{L:1}$ can be mapped to an ϵ/β-covering of f_ℓ), thus forcing a rate of at least $E_{r_\ell}(N) = E_{\min\{r_1,\dots,r_L\}}(N)$.

An analysis of minimax rates in a similar setting has been done in [22].

4.2 Breaking the Curse of Dimensionality with AccNets

Now that we know that composition of Sobolev functions can be easily learnable, even in settings where the curse of dimensionality should make it hard to learn them, we need to find a model that can achieve those rates. Though many models are possible [2], we focus on DNNs, in particular AccNets. Assuming convergence to a global minimum of the loss of sufficiently wide AccNets with two types of regularization, one can guarantee close to optimal rates:

Theorem 5. *Given a true function $f^*_{L^*:1} = f^*_{L^*} \circ \cdots \circ f^*_1$ going through the dimensions $d^*_0, \dots, d^*_{L^*}$, along with a continuous input distribution π_0 supported in $B(0,b_0)$, such that the distributions π_ℓ of $f^*_\ell(x)$ (for $x \sim \pi_0$) are continuous too and supported inside $B(0,b_\ell) \subset \mathbb{R}^{d^*_\ell}$. Further assume that there are differentiabilities ν_ℓ and radii R_ℓ such that $\|f^*_\ell\|_{W^{\nu_\ell,2}(B(0,b_\ell))} \leq R_\ell$, and ρ_ℓ such that $Lip(f^*_\ell) \leq \rho_\ell$. For an infinite width AccNet with $L \geq L^*$ and dimensions $d_\ell \geq d^*_1, \dots, d^*_{L^*-1}$, we have for the ratios $\tilde{r}_\ell = \frac{\nu_\ell}{d^*_\ell + 3}$:*

- *At a global minimizer $\hat{f}_{L:1}$ of the regularized loss $f_1, \dots, f_L \mapsto \tilde{\mathcal{L}}_N(f_{L:1}) + \lambda \prod_{\ell=1}^L Lip(f_\ell) \sum_{\ell=1}^L \frac{\|f_\ell\|_{F_1}}{Lip(f_\ell)} \sqrt{d_{\ell-1} + d_\ell}$, we have $\mathcal{L}(\hat{f}_{L:1}) = \tilde{O}(N^{-\min\{\frac{1}{2}, \tilde{r}_1, \dots, \tilde{r}_{L^*}\}})$.*
- *At a global minimizer $\hat{f}_{L:1}$ of the regularized loss $f_1, \dots, f_L \mapsto \tilde{\mathcal{L}}_N(f_{L:1}) + \lambda \prod_{\ell=1}^L \|f_\ell\|_{F_1}$, we have $\mathcal{L}(\hat{f}_{L:1}) = \tilde{O}(N^{-\frac{1}{2} + \sum_{\ell=1}^{L^*} \max\{0, \tilde{r}_\ell - \frac{1}{2}\}})$.*

There are a number of limitations to this result. First we assume that one is able to recover the global minimizer of the regularized loss, which should be hard in general[3] (we already know from [5] that this is NP-hard for shallow networks and a simple F_1-regularization). Note that it is sufficient to recover a network $f_{L:1}$ whose regularized loss is within a constant of the global minimum, which

[2] One could argue that it would be more natural to consider compositions of kernel method models, for example a composition of random feature models. But this would lead to a very similar model: this would be equivalent to a AccNet where only the W_ℓ weights are learned, while the V_ℓ, b_ℓ weights remain constant. Another family of models that should have similar properties is Deep Gaussian Processes [15].

[3] Note that the unregularized loss can be optimized polynomially, e.g. in the NTK regime [28, 3, 16], but this is an easier task than findinig the global minimum of the regularized loss where one needs to both fit the data, and do it with an minimal regularization term.

might be easier to guarantee, but should still be hard in general. The typical method of training with GD on the regularized loss is a greedy approach, which might fail in general but could recover almost optimal parameters under the right conditions (some results suggest that training relies on first order correlations to guide the network in the right direction [2, 1, 35]).

We propose two regularizations because they offer a tradeoff:

First regularization: The first regularization term leads to almost optimal rates, up to the change from $r_\ell = \frac{\nu_\ell}{d_\ell^*}$ to $r_\ell = \frac{\nu_\ell}{d_\ell^*+3}$ which is negligible for large dimensions d_ℓ and differentiabilities ν_ℓ. The first problem is that it requires an infinite width at the moment, because we were not able to prove that a function with bounded F_1-norm and Lipschitz constant can be approximated by a sufficiently wide shallow networks with the same (or close) F_1-norm and Lipschitz constant (we know from [5] that it is possible without preserving the Lipschitzness). We are quite hopeful that this condition might be removed in future work.

The second and more significant problem is that the Lipschitz constants $Lip(f_\ell)$ are difficult to optimize over. For finite width networks it is in theory possible to take the max over all linear regions, but the complexity might be unreasonable. It might be more reasonable to leverage an implicit bias instead, such as a large learning rate, because a large Lipschitz constant implies that the nework is sensible to small changes in its parameters, so GD with a large learning rate should only converge to minima with a small Lipschitz constant (such a bias is described in [26]). It might also be possible to replace the Lipschitz constant in our generalization bounds, possibly along the lines of [43].

Second regularization: The second regularization term actually does not require an infinite width, only a sufficiently large one. Also its regularization term is equivalent to $\prod(\|W_\ell\|^2 + \|V_\ell\|^2 + \|b_\ell\|^2)$ which is much closer to the traditional L_2-regularization (and actually one could prove the same or very similar rates for L_2-regularization). The issue is that it lead to rates that could be far from optimal depending on the ratios $\tilde{r}_\ell$: it recovers the same rate as the first regularization term if no more than one ratio $\tilde{r}_\ell$ is smaller than $\frac{1}{2}$, but if many of these ratios are above $\frac{1}{2}$, it can be arbitrarily smaller.

In Figure 2, we compare the empirical rates (by doing a linear fit on a log-log plot of test error as a function of N) and the predicted optimal rates $\min\{\frac{1}{2}, \frac{\nu_g}{d_{in}}, \frac{\nu_h}{d_{mid}}\}$ and observe a pretty good match. Though surprisingly, it appears the the empirical rates tend to be slightly better than the theoretical ones.

Remark. As can be seen in the proof of Theorem5, when the depth L is strictly larger than the true depth L^*, one needs to add identity layers, leading to a so-called Bottleneck structure, which was proven to be optimal and observed empirically in [27, 26, 45]. These identity layers add a term that scales linearly in the additional depth $\frac{(L-L^*)d^*_{min}}{\sqrt{N}}$ to the first regularization, and an exponential prefactor $(2d^*_{min})^{L-L^*}$ to the second. It might be possible to remove these factors by leveraging the bottleneck structure, or simply by switching to ResNets.

5 Conclusion

We have given a generalization bound for Accordion Networks and as an extension Fully-Connected networks. It depends on F_1-norms and Lipschitz constants of its shallow subnetworks. This allows us to prove under certain assumptions that AccNets can learn general compositions of Sobolev functions efficiently, making them able to break the curse of dimensionality in certain settings, such as in the presence of unknown symmetries.

References

[1] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022.

[2] Emmanuel Abbe, Enric Boix-Adserà, Matthew Stewart Brennan, Guy Bresler, and Dheeraj Mysore Nagaraj. The staircase property: How hierarchical structure can guide deep learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[3] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. pages 242–252, 2019.

[4] Kendall Atkinson and Weimin Han. *Spherical harmonics and approximations on the unit sphere: an introduction*, volume 2044. Springer Science & Business Media, 2012.

[5] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[6] Andrew R Barron and Jason M Klusowski. Complexity, statistical risk, and metric entropy of deep nets using total path variation. *stat*, 1050:6, 2019.

[7] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

[8] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[9] M. S. Birman and M. Z. Solomjak. Piecewise-polynomial approximations of functions of the classes W_p^α. *Mathematics of The USSR-Sbornik*, 2:295–317, 1967.

[10] Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Nonparametric regression on low-dimensional manifolds using deep relu networks: Function approximation and statistical recovery. *Information and Inference: A Journal of the IMA*, 11(4):1203–1253, 2022.

[11] Lénaïc Chizat and Francis Bach. On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport. In *Advances in Neural Information Processing Systems 31*, pages 3040–3050. Curran Associates, Inc., 2018.

[12] Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1305–1338. PMLR, 09–12 Jul 2020.

[13] Feng Dai. *Approximation theory and harmonic analysis on spheres and balls*. Springer, 2013.

[14] Zhen Dai, Mina Karzand, and Nathan Srebro. Representation costs of linear neural networks: Analysis and design. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[15] Andreas Damianou and Neil D. Lawrence. Deep Gaussian processes. In Carlos M. Carvalho and Pradeep Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 207–215, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR.

[16] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

[17] I. Dumer, M.S. Pinsker, and V.V. Prelov. On coverings of ellipsoids in euclidean spaces. *IEEE Transactions on Information Theory*, 50(10):2348–2356, 2004.

[18] Lawrence C Evans. *Partial differential equations*, volume 19. American Mathematical Society, 2022.

[19] Tomer Galanti, Zachary S Siegel, Aparna Gupte, and Tomaso Poggio. Sgd and weight decay provably induce a low-rank bias in neural networks. *arXiv preprint arXiv:2206.05794*, 2022.

[20] Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401, 2020.

[21] Mario Geiger, Stefano Spigler, Stéphane d'Ascoli, Levent Sagun, Marco Baity-Jesi, Giulio Biroli, and Matthieu Wyart. Jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review E*, 100(1):012115, 2019.

[22] Matteo Giordano, Kolyan Ray, and Johannes Schmidt-Hieber. On the inability of gaussian process regression to optimally learn compositional functions. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[23] Antoine Gonon, Nicolas Brisebarre, Elisa Riccietti, and Rémi Gribonval. A path-norm toolkit for modern networks: consequences, promises and challenges. In *The Twelfth International Conference on Learning Representations*, 2023.

[24] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1832–1841. PMLR, 10–15 Jul 2018.

[25] Daniel Hsu, Ziwei Ji, Matus Telgarsky, and Lan Wang. Generalization bounds via distillation. In *International Conference on Learning Representations*, 2021.

[26] Arthur Jacot. Bottleneck structure in learned features: Low-dimension vs regularity tradeoff. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 23607–23629. Curran Associates, Inc., 2023.

[27] Arthur Jacot. Implicit bias of large depth networks: a notion of rank for nonlinear functions. In *The Eleventh International Conference on Learning Representations*, 2023.

[28] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems 31*, pages 8580–8589. Curran Associates, Inc., 2018.

[29] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

[30] Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2020.

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

[32] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on learning theory*, pages 1376–1401. PMLR, 2015.

[33] Atsushi Nitanda and Taiji Suzuki. Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. In *International Conference on Learning Representations*, 2020.

[34] Greg Ongie and Rebecca Willett. The role of linear layers in nonlinear interpolating networks. *arXiv preprint arXiv:2202.00856*, 2022.

[35] Leonardo Petrini, Francesco Cagnetta, Umberto M Tomasini, Alessandro Favero, and Matthieu Wyart. How deep neural networks learn compositional data: The random hierarchy model. *arXiv preprint arXiv:2307.02129*, 2023.

[36] Grant Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In *Advances in Neural Information Processing Systems 31*, pages 7146–7155. Curran Associates, Inc., 2018.

[37] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, page 400–408, New York, NY, USA, 2018. Association for Computing Machinery.

[38] Johannes Schmidt-Hieber. Deep relu network approximation of functions on a manifold. *arXiv preprint arXiv:1908.00695*, 2019.

[39] Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875 – 1897, 2020.

[40] Mark Sellke. On size-independent sample complexity of relu networks. *Information Processing Letters*, page 106482, 2024.

[41] Taiji Suzuki, Hiroshi Abe, and Tomoaki Nishimura. Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network. In *International Conference on Learning Representations*, 2020.

[42] Zihan Wang and Arthur Jacot. Implicit bias of SGD in l_2-regularized linear DNNs: One-way jumps from high to low rank. In *The Twelfth International Conference on Learning Representations*, 2024.

[43] Colin Wei and Tengyu Ma. Data-dependent sample complexity of deep neural networks via lipschitz augmentation. *Advances in Neural Information Processing Systems*, 32, 2019.

[44] E Weinan, Chao Ma, and Lei Wu. Barron spaces and the compositional function spaces for neural network models. *arXiv preprint arXiv:1906.08039*, 2019.

[45] Yuxiao Wen and Arthur Jacot. Which frequencies do cnns need? emergent bottleneck structure in feature learning. *to appear at ICML*, 2024.

[46] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *ICLR 2017 proceedings*, Feb 2017.

The Appendix is structured as follows:

1. In Section A, we describe the experimental setup and provide a few additional experiments.
2. In Section B, we prove Theorems 1 and 2 from the main.
3. In Section C, we prove Proposition 3 and Theorem 4.
4. In Section D, we prove Theorem 5 and other approximation results concerning Sobolev functions.
5. In Section E, we prove a few technical results on the covering number.

A Experimental Setup[4]

In this section, we review our numerical experiments and their setup both on synthetic and real-world datasets in order to address theoretical results more clearly and intuitively.

A.1 Dataset

A.1.1 Emperical Dataset

The Matérn kernel is considered a generalization of the radial basis function (RBF) kernel. It controls the differentiability, or smoothness, of the kernel through the parameter ν. As $\nu \to \infty$, the Matérn kernel converges to the RBF kernel, and as $\nu \to 0$, it converges to the Laplacian kernel, a 0-differentiable kernel. In this study, we utilized the Matérn kernel to generate Gaussian Process (GP) samples based on the composition of two Matérn kernels, K_g and K_h, with varying differentiability in the range [0.5,10]×[0.5,10]. The input dimension (d_{in}) of the kernel, the bottleneck mid-dimension (d_{mid}), and the output dimension (d_{out}) are 15, 3, and 20, respectively.

This outlines the general procedure of our sampling method for synthetic data:

1. Sample the training dataset $X \in \mathbb{R}^{D \times d_{in}}$
2. From X, compute the $D \times D$ kernel K_g with given ν_g
3. From K_g, sample $Z \in \mathbb{R}^{D \times d_{mid}}$ with columns sampled from the Gaussian $\mathcal{N}(0, K_g)$.
4. From Z, compute K_g with given ν_h
5. From K_h, sample the test dataset $Y \in \mathbb{R}^{D \times d_{out}}$ with columns sampled from the Gaussian $\mathcal{N}(0, K_h)$.

We utilized four AMD Opteron 6136 processors (2.4 GHz, 32 cores) and 128 GB of RAM to generate our synthetic dataset. The maximum possible dataset size for 128 GB of RAM is approximately 52,500; however, we opted for a synthetic dataset size of 22,000 due to the computational expense associated with sampling the Matérn kernel. This decision was made considering the time complexity of $\mathcal{O}(n^3)$and the space complexity of $\mathcal{O}(n^2)$ involved. Out of the 22,000 dataset points, 20,000 were allocated for training data, and 2,000 were used for the test dataset

A.1.2 Real-world dataset: WESAD

In our study, we utilized the Wearable Stress and Affect Detection (WESAD) dataset to train our AccNets for binary classification. The WESAD dataset, which is publicly accessible, provides multimodal physiological and motion data collected from 15 subjects using devices worn on the wrist and chest. For the purpose of our experiment, we specifically employed the Empatica E4 wrist device to distinguish between non-stress (baseline) and stress conditions, simplifying the classification task to these two categories.

After preprocessing, the dataset comprised a total of 136,482 instances. We implemented a train-test split ratio of approximately 75:25, resulting in 100,000 instances for the training set and 36,482 instances for the test set. The overall hyperparameters and architecture of the AccNets model applied to the WESAD dataset were largely consistent with those used for our synthetic data. The primary differences were the use of 100 epochs for each iteration of Ni from Ns, and a learning rate set to 1e-5.

[4]The code used for experiments are publicly available here

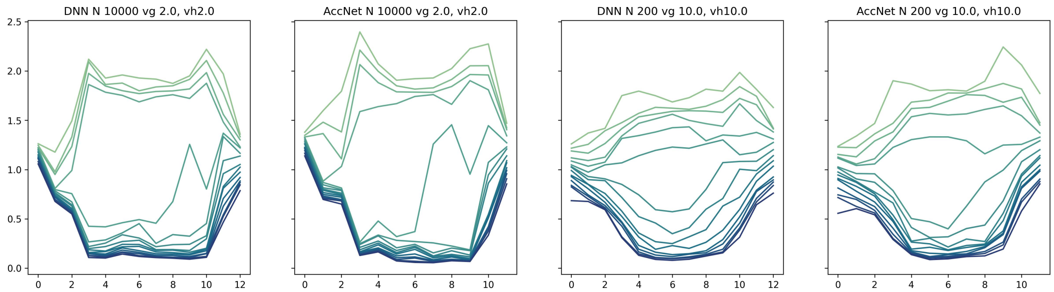

Figure 4: A comparison: singular values of the weight matrices for DNN and AccNets models. The first two plots represent cases where N = 10000 while the right two plots correspond to N = 200.The number of outliers at the top of each plot signifies the rank of each network. The plots with $N = 10000$ datasets demonstrate a clearer capture of the true rank compared to those with $N = 200$ indicating that a higher dataset count provides more accurate rank determination

.

A.2 Model setups

To investigate the scaling law of test error for our synthetic data, we trained models using N_i datapoints from our training data, where $N = [100, 200, 500, 1000, 2000, 5000, 10000, 20000]$. The models employed for this analysis included the kernel method, shallow networks, fully connected deep neural networks (FC DNN), and AccNets. For FC DNN and AccNets, we configured the network depth to 12 layers, with the layer widths set as $[d_{in}, 500, 500, ..., 500, d_{out}]$ for DNNs, and $[d_in, 900, 100, 900, ..., 100, 900, d_{out}]$ for AccNets.

To ensure a comparable number of neurons, the width for the shallow networks was set to 50,000, resulting in dimensions of $[d_{in}, 50000, d_{out}]$.

We utilized ReLU as the activation function and L^1-norm as the cost function, with the Adam optimizer. The total number of batch was set to 5, and the training process was conducted over 3600 epochs, divided into three phases. The detailed optimizer parameters are as follows:

1. For the first 1200 epochs: learning rate $(lr) = 1.5 * 0.001$, weight decay = 0
2. For the second 1200 epochs: $lr = 0.4 * 0.001$, weight decay $= 0.002$
3. For the final 1200 epochs: $lr = 0.1 * 0.001$, weight decay $= 0.005$

We conducted experiments utilizing 12 NVIDIA V100 GPUs (each with 32GB of memory) over a period of 6.3 days to train the synthetic dataset. In contrast, training the WESAD dataset required only one hour on a single V100 GPU.

A.3 Additional experiments

B AccNet Generalization Bounds

The proof of generalization for shallow networks (Theorem 1) is the special case $L = 1$ of the proof of Theorem 2, so we only prove the second:

Theorem 6. *Consider an accordion net of depth L and widths $d_L, \dots, d_0$, with corresponding set of functions $\mathcal{F} = \{f_{L:1} : \|f_\ell\|_{F_1} \le R_\ell, Lip(f_\ell) \le \rho_\ell\}$ with input space $\Omega = B(0, r)$. For any ρ-Lipschitz loss function $\ell(x, f(x))$ with $|\ell(x, y)| \le c_0$, we know that with probability $1 - \delta$ over the sampling of the training set X from the distribution π, we have for all $f \in \mathcal{F}$*

$$\mathcal{L}(f) - \tilde{\mathcal{L}}_N(f) \le C\rho_{L:1}r \sum_{\ell'=1}^{L} \frac{R_{\ell'}}{\rho_{\ell'}} \sqrt{d_{\ell'} + d_{\ell'-1}} \frac{\log N}{\sqrt{N}} (1 + o(1)) + c_0 \sqrt{\frac{2 \log {}^2/_\delta}{N}}.$$

Proof. The strategy is: (1) prove a covering number bound on $\mathcal{F}$ (2) use it to obtain a Rademacher complexity bound, (3) use the Rademacher complexity to bound the generalization error.

(1) We define $f_\ell = V_\ell \circ \sigma \circ W_\ell$ so that $f_\theta = f_{L:1} = f_L \circ \cdots \circ f_1$. First notice that we can write each f_ℓ as convex combination of its neurons:

$$f_\ell(x) = \sum_{i=1}^{w_\ell} v_{\ell,i}\sigma(w_{\ell,i}^T x) = R_\ell \sum_{i=1}^{w_\ell} c_{\ell,i}\bar{v}_{\ell,i}\sigma(\bar{w}_{\ell,i}^T x)$$

for $\bar{w}_{\ell,i} = \frac{w_{\ell,i}}{\|w_{\ell,i}\|}$, $\bar{v}_{\ell,i} = \frac{v_{\ell,i}}{\|v_{\ell,i}\|}$, $R_\ell = \sum_{i=1}^{\ell} \|v_{\ell,i}\| \|w_{\ell,i}\|$ and $c_{\ell,i} = \frac{1}{R_\ell} \|v_{\ell,i}\| \|w_{\ell,i}\|$.

Let us now consider a sequence $\epsilon_k = 2^{-k}$ for $k = 0, \ldots, K$ and define $\tilde{v}_{\ell,i}^{(k)}, \tilde{w}_{\ell,i}^{(k)}$ to be the ϵ_k-covers of $\bar{v}_{\ell,i}, \bar{w}_{\ell,i}$, furthermore we may choose $\tilde{v}_{\ell,i}^{(0)} = \tilde{w}_{\ell,i}^{(0)} = 0$ since every unit vector is within a $\epsilon_0 = 1$ distance of the origin. We will now show that on can approximate f_θ by approximating each of the f_ℓ by functions of the form

$$\tilde{f}_\ell(x) = R_\ell \sum_{k=1}^{K_\ell} \frac{1}{M_{k,\ell}} \sum_{m=1}^{M_{k,\ell}} \tilde{v}_{\ell,i_{\ell,m}^{(k)}}^{(k)} \sigma(\tilde{w}_{\ell,i_{\ell,m}^{(k)}}^{(k)T} x) - \tilde{v}_{\ell,i_{\ell,m}^{(k)}}^{(k-1)} \sigma(\tilde{w}_{\ell,i_{\ell,m}^{(k)}}^{(k-1)T} x)$$

for indices $i_{\ell,m}^{(k)} = 1, \ldots, w_\ell$ choosen adequately. Notice that the number of functions of this type equals the number of $M_{k,\ell}$ quadruples $(\tilde{v}_{\ell,i_{\ell,m}^{(k)}}^{(k)}, \tilde{w}_{\ell,i_{\ell,m}^{(k)}}^{(k)T}, \tilde{v}_{\ell,i_{\ell,m}^{(k)}}^{(k-1)}, \tilde{w}_{\ell,i_{\ell,m}^{(k)}}^{(k-1)T})$ where these vectors belong to the ϵ_k- resp. ϵ_{k-1}-coverings of the d_{in}- resp. d_{out}-dimensional unit sphere. Thus the number of such functions is bounded by

$$\prod_{k=1}^{K_\ell} \left(\mathcal{N}_2(\mathbb{S}^{d_{in}-1}, \epsilon_k)\mathcal{N}_2(\mathbb{S}^{d_{out}-1}, \epsilon_k)\mathcal{N}_2(\mathbb{S}^{d_{in}-1}, \epsilon_{k-1})\mathcal{N}_2(\mathbb{S}^{d_{out}-1}, \epsilon_{k-1})\right)^{M_{k,\ell}},$$

and we have this choice for all $\ell = 1, \ldots, L$. We will show that with sufficiently large $M_{k,\ell}$ this set of functions ϵ-covers $\mathcal{F}$ which then implies that

$$\log \mathcal{N}_2(\mathcal{F}, \epsilon) \leq 2 \sum_{\ell=1}^{L} \sum_{k=1}^{K_\ell} M_{k,\ell} \left(\log \mathcal{N}_2(\mathbb{S}^{d_{in}-1}, \epsilon_{k-1}) + \log \mathcal{N}_2(\mathbb{S}^{d_{in}-1}, \epsilon_{k-1})\right).$$

We will use the probabilistic method to find the right indices $i_{\ell,m}^{(k)}$ to approximate a function $f_\ell = R_\ell \sum_{i=1}^{w_\ell} c_{\ell,i}\bar{v}_{\ell,i}\sigma(\bar{w}_{\ell,i}^T x)$ with a function $\tilde{f}_\ell$. We take all $i_{\ell,m}^{(k)}$ to be i.i.d. equal to the index $i = 1, \cdots, w_\ell$ with probability $c_{\ell,i}$, so that in expectation

$$\begin{aligned}
\mathbb{E}\tilde{f}_\ell(x) &= R_\ell \sum_{k=1}^{K_\ell} \sum_{i=1}^{w_\ell} c_{\ell,i} \left(\tilde{v}_{\ell,i}^{(k)} \sigma(\tilde{w}_{\ell,i}^{(k)T} x) - \tilde{v}_{\ell,i}^{(k-1)} \sigma(\tilde{w}_{\ell,i}^{(k-1)T} x)\right) \\
&= R_\ell \sum_{i=1}^{w_\ell} c_{\ell,i} \tilde{v}_{\ell,i}^{(K)} \sigma(\tilde{w}_{\ell,i}^{(K)T} x).
\end{aligned}$$

We will show that this expectation is $O(\epsilon_{K_\ell})$-close to f_ℓ and that the variance of $\tilde{f}_\ell$ goes to zero as the $M_{\ell,k}$s grow, allowing us to bound the expected error $\mathbb{E}\left\|f_{L:1} - \tilde{f}_{L:1}\right\|_\pi^2 \leq \epsilon^2$ which then implies that there must be at least one choice of indices $i_{\ell,m}^{(k)}$ such that $\left\|f_{L:1} - \tilde{f}_{L:1}\right\|_\pi \leq \epsilon$.

Let us first bound the distance

$$\begin{aligned}
\left\|f_\ell(x) - \mathbb{E}\tilde{f}_\ell(x)\right\| &= R_\ell \left\|\sum_{i=1}^{w_\ell} c_{\ell,i}\left(\bar{v}_{\ell,i}\sigma(\bar{w}_{\ell,i}^T x) - \tilde{v}_{\ell,i}^{(K)}\sigma(\tilde{w}_{\ell,i}^{(K)T}x)\right)\right\| \\
&\leq R_\ell \sum_{i=1}^{w_\ell} c_{\ell,i}\left(\left\|\left(\bar{v}_{\ell,i} - \tilde{v}_{\ell,i}^{(K)}\right)\sigma(\bar{w}_{\ell,i}^T x)\right\| + \left\|\tilde{v}_{\ell,i}^{(K)}\left(\sigma(\bar{w}_{\ell,i}^T x) - \sigma(\tilde{w}_{\ell,i}^{(K)T}x)\right)\right\|\right) \\
&\leq R_\ell \sum_{i=1}^{w_\ell} c_{\ell,i}\left(\left\|\bar{v}_{\ell,i} - \tilde{v}_{\ell,i}^{(K)}\right\|\left\|\bar{w}_{\ell,i}^T x\right\| + \left\|\tilde{v}_{\ell,i}^{(K)}\right\|\left\|\bar{w}_{\ell,i}^T x - \tilde{w}_{\ell,i}^{(K)T}x\right\|\right) \\
&\leq 2R_\ell \sum_{i=1}^{w_\ell} c_{\ell,i}\epsilon_{K_\ell}\left\|x\right\| \\
&= 2R_\ell\epsilon_{K_\ell}\left\|x\right\|.
\end{aligned}$$

Then we bound the trace of the covariance of $\tilde{f}_\ell$ which equals the expected square distance between $\tilde{f}_\ell$ and its expectation:

$$\begin{aligned}
&\mathbb{E}\left\|\tilde{f}_\ell(x) - \mathbb{E}\tilde{f}_\ell(x)\right\|^2 \\
&= \sum_{k=1}^{K_\ell}\frac{R_\ell^2}{M_{k,\ell}^2}\sum_{m=1}^{M_{k,\ell}}\mathbb{E}\left\|\tilde{v}_{\ell,i_{\ell,m}^{(k)}}^{(k)}\sigma(\tilde{w}_{\ell,i_{\ell,m}^{(k)}}^{(k)T}x) - \tilde{v}_{\ell,i_{\ell,m}^{(k)}}^{(k-1)}\sigma(\tilde{w}_{\ell,i_{\ell,m}^{(k)}}^{(k-1)T}x) - \mathbb{E}\left[\tilde{v}_{\ell,i_{\ell,m}^{(k)}}^{(k)}\sigma(\tilde{w}_{\ell,i_{\ell,m}^{(k)}}^{(k)T}x) - \tilde{v}_{\ell,i_{\ell,m}^{(k)}}^{(k-1)}\sigma(\tilde{w}_{\ell,i_{\ell,m}^{(k)}}^{(k-1)T}x)\right]\right\|^2 \\
&\leq \sum_{k=1}^{K_\ell}\frac{R_\ell^2}{M_{k,\ell}^2}\sum_{m=1}^{M_{k,\ell}}\mathbb{E}\left\|\tilde{v}_{\ell,m}^{(k)}\sigma(\tilde{w}_{\ell,m}^{(k)T}x) - \tilde{v}_{\ell,m}^{(k-1)}\sigma(\tilde{w}_{\ell,m}^{(k-1)T}x)\right\|^2 \\
&= \sum_{k=1}^{K_\ell}\frac{R_\ell^2}{M_{k,\ell}}\sum_{i=1}^{w_\ell}c_i\left\|\tilde{v}_{\ell,i}^{(k)}\sigma\left(\tilde{w}_{\ell,i}^{(k)T}x\right) - \tilde{v}_{\ell,i}^{(k-1)}\sigma\left(\tilde{w}_{\ell,i}^{(k-1)T}x\right)\right\|^2 \\
&\leq \sum_{k=1}^{K_\ell}\frac{2R_\ell^2\left\|x\right\|^2}{M_{k,\ell}}\sum_{i=1}^{w_\ell}c_i\left\|\tilde{v}_{\ell,i}^{(k)}\right\|^2\left\|\tilde{w}_{\ell,i}^{(k)} - \tilde{w}_{\ell,i}^{(k-1)}\right\|^2 + c_i\left\|\tilde{v}_{\ell,i}^{(k)} - \tilde{v}_{\ell,i}^{(k-1)}\right\|^2\left\|\tilde{w}_{\ell,i}^{(k-1)}\right\|^2 \\
&\leq \sum_{k=1}^{K_\ell}\frac{4R_\ell^2\left\|x\right\|^2}{M_{k,\ell}}\left(\epsilon_k + \epsilon_{k-1}\right)^2 \\
&\leq \sum_{k=1}^{K_\ell}\frac{36R_\ell^2\left\|x\right\|^2}{M_{k,\ell}}\epsilon_k^2.
\end{aligned}$$

Putting both together, we obtain

$$\begin{aligned}
\mathbb{E}\left\|f_\ell(x) - \tilde{f}_\ell(x)\right\|^2 &\leq 4R_\ell^2\epsilon_{K_\ell}^2\left\|x\right\|^2 + \sum_{k=1}^{K_\ell}\frac{36R_\ell^2\left\|x\right\|^2}{M_{k,\ell}}\epsilon_k^2 \\
&= 4R_\ell^2\left\|x\right\|^2\left(\epsilon_{K_\ell}^2 + 9\sum_{k=1}^{K_\ell}\frac{\epsilon_k^2}{M_{k,\ell}}\right).
\end{aligned}$$

We will now use this bound, together with the Lipschitzness of f_ℓ to bound the error $\mathbb{E}\left\|f_{L:1}(x) - \tilde{f}_{L:1}(x)\right\|^2$. We will do this by induction on the distances $\mathbb{E}\left\|f_{\ell:1}(x) - \tilde{f}_{\ell:1}(x)\right\|^2$. We start by

$$\mathbb{E}\left\|f_1(x) - \tilde{f}_1(x)\right\|^2 \leq 4R_1^2\left\|x\right\|^2\left(\epsilon_{K_\ell}^2 + 9\sum_{k=1}^{K_\ell}\frac{\epsilon_k^2}{M_{k,1}}\right).$$

And for the induction step, we condition on the layers $f_{\ell-1:1}$

$$\begin{aligned}
\mathbb{E}\left\|f_{\ell:1}(x)-\tilde{f}_{\ell:1}(x)\right\|^{2} &= \mathbb{E}\left[\mathbb{E}\left[\left\|f_{\ell:1}(x)-\tilde{f}_{\ell:1}(x)\right\|^{2}|\tilde{f}_{\ell-1:1}\right]\right] \\
&= \mathbb{E}\left\|f_{\ell:1}(x)-\mathbb{E}\left[\tilde{f}_{\ell:1}(x)|\tilde{f}_{\ell-1:1}\right]\right\|^{2}+\mathbb{E}\mathbb{E}\left[\left\|\tilde{f}_{\ell:1}(x)-\mathbb{E}\left[\tilde{f}_{\ell:1}(x)|\tilde{f}_{\ell-1:1}\right]\right\|^{2}|\tilde{f}_{\ell-1:1}\right] \\
&= \mathbb{E}\left\|f_{\ell:1}(x)-f_{\ell}(\tilde{f}_{\ell-1:1}(x))\right\|^{2}+\mathbb{E}\mathbb{E}\left[\left\|\tilde{f}_{\ell:1}(x)-f_{\ell}(\tilde{f}_{\ell-1:1}(x))\right\|^{2}|\tilde{f}_{\ell-1:1}\right] \\
&\leq \rho_{\ell}^{2}\mathbb{E}\left\|f_{\ell-1:1}(x)-\tilde{f}_{\ell-1:1}(x)\right\|^{2}+4R_{\ell}^{2}\mathbb{E}\left\|\tilde{f}_{\ell-1:1}(x)\right\|^{2}\left(\epsilon_{K_{\ell}}^{2}+9\sum_{k=1}^{K_{\ell}}\frac{\epsilon_{k}^{2}}{M_{k,\ell}}\right).
\end{aligned}$$

Now since

$$\begin{aligned}
\mathbb{E}\left\|\tilde{f}_{\ell-1:1}(x)\right\|^{2} &\leq \left\|f_{\ell-1:1}(x)\right\|^{2}+\mathbb{E}\left\|f_{\ell-1:1}(x)-\tilde{f}_{\ell-1:1}(x)\right\|^{2} \\
&\leq \rho_{\ell-1}^{2}\cdots\rho_{1}^{2}\left\|x\right\|^{2}+\mathbb{E}\left\|f_{\ell-1:1}(x)-\tilde{f}_{\ell-1:1}(x)\right\|^{2}
\end{aligned}$$

we obtain that

$$\begin{aligned}
\mathbb{E}\left\|f_{\ell:1}(x)-\tilde{f}_{\ell:1}(x)\right\|^{2} \leq{}& \left(\rho_{\ell}^{2}+4R_{\ell}^{2}\left(\epsilon_{K_{\ell}}^{2}+9\sum_{k=1}^{K_{\ell}}\frac{\epsilon_{k}^{2}}{M_{k,\ell}}\right)\right)\mathbb{E}\left\|f_{\ell-1:1}(x)-\tilde{f}_{\ell-1:1}(x)\right\|^{2} \\
&+4R_{\ell}^{2}\rho_{\ell-1}^{2}\cdots\rho_{1}^{2}\left\|x\right\|^{2}\left(\epsilon_{K_{\ell}}^{2}+9\sum_{k=1}^{K_{\ell}}\frac{\epsilon_{k}^{2}}{M_{k,\ell}}\right).
\end{aligned}$$

We define $\tilde{\rho}_{\ell}^{2}=\rho_{\ell}^{2}\left[1+4\frac{R_{\ell}^{2}}{\rho_{\ell}^{2}}\left(\epsilon_{K_{\ell}}^{2}+9\sum_{k=1}^{K_{\ell}}\frac{\epsilon_{k}^{2}}{M_{k,\ell}}\right)\right]$ and obtain

$$\mathbb{E}\left\|f_{L:1}(x)-\tilde{f}_{L:1}(x)\right\|^{2}\leq 4\sum_{\ell=1}^{L}\tilde{\rho}_{L:\ell+1}^{2}R_{\ell}^{2}\rho_{\ell-1:1}^{2}\left\|x\right\|^{2}\left(\epsilon_{K_{\ell}}^{2}+9\sum_{k=1}^{K_{\ell}}\frac{\epsilon_{k}^{2}}{M_{k,\ell}}\right).$$

Thus for any distribution π over the ball $B(0,r)$, there is a choice of indices $i_{\ell,m}^{(k)}$ such that

$$\left\|f_{L:1}-\tilde{f}_{L:1}\right\|_{\pi}^{2}\leq 4\sum_{\ell=1}^{L}\tilde{\rho}_{L:\ell+1}^{2}R_{\ell}^{2}\rho_{\ell-1:1}^{2}r^{2}\left(\epsilon_{K_{\ell}}^{2}+9\sum_{k=1}^{K_{\ell}}\frac{\epsilon_{k}^{2}}{M_{k,\ell}}\right).$$

We now simply need to choose K_{ℓ} and $M_{k,\ell}$ adequately. To reach an error of 2ϵ, we choose

$$K_{\ell}=\left\lceil-\log\epsilon+\frac{1}{2}\log\left[4\rho_{L:1}^{2}r^{2}\left(\sum_{\ell'=1}^{L}\frac{R_{\ell'}}{\rho_{\ell'}}\sqrt{d_{\ell'}+d_{\ell'-1}}\right)\frac{R_{\ell}}{\rho_{\ell}\sqrt{d_{\ell}+d_{\ell-1}}}\right]\right\rceil$$

where $\rho_{L:1}=\prod_{\ell=1}^{L}\rho_{\ell}$. Notice that that $\epsilon_{K_{\ell}}^{2}\leq\frac{1}{4\rho_{L:1}^{2}r^{2}\left(\sum_{\ell'=1}^{L}\frac{R_{\ell'}}{\rho_{\ell'}}\sqrt{d_{\ell'}+d_{\ell'-1}}\right)}\frac{\rho_{\ell}\sqrt{d_{\ell}+d_{\ell-1}}}{R_{\ell}}\epsilon^{2}$.

Given $s_{0}=\sum_{k=1}^{\infty}\sqrt{k}2^{-k}\approx 1.3473<\infty$, we define

$$M_{k,\ell}=\left\lceil 36\rho_{L:1}^{2}r^{2}s_{0}\left(\sum_{\ell'=1}^{L}\frac{R_{\ell'}}{\rho_{\ell'}}\sqrt{d_{\ell'}+d_{\ell'-1}}\right)\frac{R_{\ell}}{\rho_{\ell}\sqrt{d_{\ell}+d_{\ell-1}}}\frac{2^{-k}}{\sqrt{k}}\frac{1}{\epsilon^{2}}\right\rceil.$$

So that for all ℓ

$$
\begin{aligned}
4\frac{R_\ell^2}{\rho_\ell^2}\left(\epsilon_{K_\ell}^2+9\sum_{k=1}^{K_\ell}\frac{\epsilon_k^2}{M_{k,\ell}}\right) &\leq \frac{\frac{R_\ell}{\rho_\ell}\sqrt{d_\ell+d_{\ell-1}}}{\rho_{L:1}^2 r^2\left(\sum_{\ell'=1}^L \frac{R_\ell}{\rho_\ell}\sqrt{d_\ell+d_{\ell-1}}\right)}\epsilon^2 \\
&\quad+\frac{\frac{R_\ell}{\rho_\ell}\sqrt{d_\ell+d_{\ell-1}}}{\rho_{L:1}^2 r^2\left(\sum_{\ell'=1}^L \frac{R_\ell}{\rho_\ell}\sqrt{d_\ell+d_{\ell-1}}\right)}\epsilon^2\frac{\sum_{k'=1}^{K_\ell}\sqrt{k'}2^{-k'}}{s_0} \\
&\leq 2\frac{\frac{R_\ell}{\rho_\ell}\sqrt{d_\ell+d_{\ell-1}}}{\rho_{L:1}^2 r^2\left(\sum_{\ell'=1}^L \frac{R_\ell}{\rho_\ell}\sqrt{d_\ell+d_{\ell-1}}\right)}\epsilon^2.
\end{aligned}
$$

Now this also implies that

$$
\tilde{\rho}_\ell \leq \rho_\ell \exp\left(2\frac{\frac{R_\ell}{\rho_\ell}\sqrt{d_\ell+d_{\ell-1}}}{\rho_{L:1}^2 r^2\left(\sum_{\ell'=1}^L \frac{R_\ell}{\rho_\ell}\sqrt{d_\ell+d_{\ell-1}}\right)}\epsilon^2\right)
$$

and thus

$$
\tilde{\rho}_{L:\ell+1} \leq \rho_{L:\ell+1}\exp\left(2\frac{\sum_{\ell'=\ell+1}^L\frac{R_\ell}{\rho_\ell}\sqrt{d_\ell+d_{\ell-1}}}{\rho_{L:1}^2 r^2\left(\sum_{\ell'=1}^L \frac{R_\ell}{\rho_\ell}\sqrt{d_\ell+d_{\ell-1}}\right)}\epsilon^2\right) \leq \rho_{L:\ell+1}\exp\left(\frac{2}{\rho_{L:1}^2 r^2}\epsilon^2\right).
$$

Putting it all together, we obtain that

$$
\begin{aligned}
\left\|f_{L:1}-\tilde{f}_{L:1}\right\|_\pi^2 &\leq 4\sum_{\ell=1}^L \tilde{\rho}_{L:\ell+1}^2 R_\ell^2\rho_{\ell-1:1}^2 r^2\left(\epsilon_{K_\ell}^2+9\sum_{k=1}^{K_\ell}\frac{\epsilon_k^2}{M_{k,\ell}}\right) \\
&\leq \exp\left(\frac{2}{\rho_{L:1}^2 r^2}\epsilon^2\right)\rho_{L:1}^2 r^2\sum_{\ell=1}^L 4\frac{R_\ell^2}{\rho_\ell^2}\left(\epsilon_{K_\ell}^2+9\sum_{k=1}^{K_\ell}\frac{\epsilon_k^2}{M_{k,\ell}}\right) \\
&\leq 2\exp\left(\frac{2}{\rho_{L:1}^2 r^2}\epsilon^2\right)\epsilon^2 \\
&= 2\epsilon^2+O(\epsilon^4).
\end{aligned}
$$

Now since $\log\mathcal{N}_2(\mathbb{S}^{d_\ell-1},\epsilon)=d_\ell\log\left(\frac{1}{\epsilon}+1\right)$ and

$$
M_{k,\ell} \leq 36\rho_{L:1}^2 r^2 s_0\left(\sum_{\ell'=1}^L\frac{R_{\ell'}}{\rho_{\ell'}}\sqrt{d_{\ell'}+d_{\ell'-1}}\right)\frac{R_\ell}{\rho_\ell\sqrt{d_\ell+d_{\ell-1}}}\frac{2^{-k}}{\sqrt{k}}\frac{1}{\epsilon^2}+1,
$$

we have

$$
\begin{aligned}
\log\mathcal{N}_2(\mathcal{F},\sqrt{2}\exp\left(\frac{\epsilon^2}{\rho_{L:1}^2 r^2}\right)\epsilon) &\leq 2\sum_{\ell=1}^L\sum_{k=1}^{K_\ell}M_{k,\ell}\left(\log\mathcal{N}_2(\mathbb{S}^{d_\ell-1},\epsilon_{k-1})+\log\mathcal{N}_2(\mathbb{S}^{d_{\ell-1}-1},\epsilon_{k-1})\right) \\
&\leq 2\sum_{\ell=1}^L\sum_{k=1}^{K_\ell}M_{k,\ell}\left(d_\ell+d_{\ell-1}\right)\log(\frac{1}{\epsilon_{k-1}}+1) \\
&\leq 72 s_0\rho_{L:1}^2 r^2\left(\sum_{\ell'=1}^L\frac{R_{\ell'}}{\rho_{\ell'}}\sqrt{d_{\ell'}+d_{\ell'-1}}\right)\sum_{\ell=1}^L\frac{R_\ell}{\rho_\ell}\sqrt{d_\ell+d_{\ell-1}}\sum_{k=1}^{K_\ell}\frac{2^{-k}\log(\frac{1}{\epsilon_{k-1}}+1)}{\sqrt{k}}\frac{1}{\epsilon^2} \\
&\quad+2\sum_{\ell=1}^L\left(d_\ell+d_{\ell-1}\right)\sum_{k=1}^{K_\ell}\log(\frac{1}{\epsilon_{k-1}}+1) \\
&\leq 72 s_0^2\rho_{L:1}^2 r^2\left(\sum_{\ell'=1}^L\frac{R_{\ell'}}{\rho_{\ell'}}\sqrt{d_{\ell'}+d_{\ell'-1}}\right)^2\frac{1}{\epsilon^2}+o(\epsilon^{-2}).
\end{aligned}
$$

The diameter of $\mathcal{F}$ is smaller than $\rho_{L:1}r$, so for all $\delta \geq \rho_{L:1}r$, $\log \mathcal{N}_2(\mathcal{F}, \delta) = 0$. For all $\delta \leq \rho_{L:1}r$ we choose $\epsilon = \frac{\delta}{\sqrt{2e}}$ so that $\sqrt{2}\exp\left(\frac{\epsilon^2}{\rho_{L:1}^2 r^2}\right)\epsilon \leq \delta$ and therefore

$$\log \mathcal{N}_2(\mathcal{F}, \delta) \leq 144 s_0^2 e \rho_{L:1}^2 r^2 \left(\sum_{\ell'=1}^{L} \frac{R_{\ell'}}{\rho_{\ell'}} \sqrt{d_{\ell'} + d_{\ell'-1}}\right)^2 \frac{1}{\delta^2} + o(\delta^{-2}).$$

(2) Our goal now is to use chaining / Dudley's theorem to bound the Rademacher complexity $R(\mathcal{F}(X))$ evaluated on a set X of size N (e.g. Lemma 27.4 in [Understanding Machine Learning]) of our set:

Lemma 7. *Let $c = \max_{f\in\mathcal{F}} \frac{1}{\sqrt{N}} \|f(X)\|$, then for any integer $M > 0$,*

$$R(\mathcal{F}(X)) \leq c2^{-M} + \frac{6c}{\sqrt{N}} \sum_{k=1}^{M} 2^{-k} \sqrt{\log \mathcal{N}(\mathcal{F}, c2^{-k})}.$$

To apply it to our setting, first note that for all $x \in B(0,r)$, $\|f_{L:1}(x)\| \leq \rho_{L:1}r$ so that $c = \max_{f\in\mathcal{F}} \frac{1}{\sqrt{N}} \|f(X)\| \leq \rho_{L:1}r$, we then have

$$\begin{aligned}
R(\mathcal{F}(X)) &\leq c2^{-M} + \frac{6c}{\sqrt{N}} \sum_{k=1}^{M} 2^{-k} 12 s_0 \sqrt{e} \rho_{L:1} r \sum_{\ell'=1}^{L} \frac{R_{\ell'}}{\rho_{\ell'}} \sqrt{d_{\ell'} + d_{\ell'-1}} c^{-1} 2^k (1 + o(1)) \\
&= c2^{-M} + \frac{72}{\sqrt{N}} M s_0 \sqrt{e} \rho_{L:1} r \sum_{\ell'=1}^{L} \frac{R_{\ell'}}{\rho_{\ell'}} \sqrt{d_{\ell'} + d_{\ell'-1}} (1 + o(1)).
\end{aligned}$$

Taking $M = \left\lceil -\log_2 \left(\frac{72}{\sqrt{N}} s_0 \sqrt{e} \sum_{\ell'=1}^{L} \frac{R_{\ell'}}{\rho_{\ell'}} \sqrt{d_{\ell'} + d_{\ell'-1}}\right)\right\rceil$, we obtain

$$\begin{aligned}
R(\mathcal{F}(X)) &\leq \frac{72}{\sqrt{N}} M s_0 \sqrt{e} \rho_{L:1} r \sum_{\ell'=1}^{L} \frac{R_{\ell'}}{\rho_{\ell'}} \sqrt{d_{\ell'} + d_{\ell'-1}} (1 + M(1 + o(1))) \\
&\leq \frac{144}{\sqrt{N}} M s_0 \sqrt{e} \rho_{L:1} r \sum_{\ell'=1}^{L} \frac{R_{\ell'}}{\rho_{\ell'}} \sqrt{d_{\ell'} + d_{\ell'-1}} \left\lceil -\log_2 \left(\frac{72}{\sqrt{N}} s_0 \sqrt{e} \sum_{\ell'=1}^{L} \frac{R_{\ell'}}{\rho_{\ell'}} \sqrt{d_{\ell'} + d_{\ell'-1}}\right)\right\rceil (1 + o(1)) \\
&\leq C \rho_{L:1} r \sum_{\ell'=1}^{L} \frac{R_{\ell'}}{\rho_{\ell'}} \sqrt{d_{\ell'} + d_{\ell'-1}} \frac{\log N}{\sqrt{N}} (1 + o(1)).
\end{aligned}$$

(3) For any ρ-Lipschitz loss function $\ell(x, f(x))$ with $|\ell(x,y)| \leq c_0$, we know that with probability $1 - \delta$ over the sampling of the training set X from the distribution π, we have for all $f \in \mathcal{F}$

$$\begin{aligned}
\mathbb{E}_{x\sim\pi}\left[\ell(x, f(x))\right] - \frac{1}{N}\sum_{i=1}^{N} \ell(x_i, f(x_i)) &\leq 2\mathbb{E}_{X'}\left[R(\ell \circ \mathcal{F}(X'))\right] + c_0 \sqrt{\frac{2\log{}^2/_\delta}{N}} \\
&\leq 2C \rho_{L:1} r \sum_{\ell'=1}^{L} \frac{R_{\ell'}}{\rho_{\ell'}} \sqrt{d_{\ell'} + d_{\ell'-1}} \frac{\log N}{\sqrt{N}} (1 + o(1)) + c_0 \sqrt{\frac{2\log{}^2/_\delta}{N}}.
\end{aligned}$$

□

C Composition of Sobolev Balls

Proposition 8 (Proposition 3 from the main.). *Given a distribution π with support in $B(0,r)$, we have that with probability $1 - \delta$ for all functions $f \in \mathcal{F} = \{f : \|f\|_{W^{\nu,2}} \leq R, \|f\|_\infty \leq R\}$*

$$\mathcal{L}(f) - \tilde{\mathcal{L}}_N(f) \leq 2C_1 R E_{\nu/d}(N) + c_0 \sqrt{\frac{2\log{}^2/_\delta}{N}}.$$

where $E_r(N) = N^{-\frac{1}{2}}$ if $r > \frac{1}{2}$, $E_r(N) = N^{-\frac{1}{2}} \log N$ if $r = \frac{1}{2}$, and $E_r(N) = N^{-r}$ if $r < \frac{1}{2}$.

Proof. (1) We know from Theorem 5.2 of [9] that the Sobolev ball $B_{W^{\nu,2}}(0,R)$ over any d-dimensional hypercube Ω satisfies

$$\log \mathcal{N}_2(B_{W^{\nu,2}}(0,R), \pi, \epsilon) \leq C_0 \left(\frac{R}{\epsilon}\right)^{\frac{d}{\nu}}$$

for a constant c and any measure π supported in the hypercube.

(2) By Dudley's theorem we can bound the Rademacher complexity of our function class $\mathcal{B}(X)$ evaluated on any training set X:

$$\begin{aligned} R(\mathcal{B}(X)) &\leq R2^{-M} + \frac{6R}{\sqrt{N}} \sum_{k=1}^{M} 2^{-k} \sqrt{C_0 \left(\frac{R}{R2^{-k}}\right)^{\frac{d}{\nu}}} \\ &= R2^{-M} + \frac{6R}{\sqrt{N}} \sqrt{C_0} \sum_{k=1}^{M} 2^{k(\frac{d}{2\nu}-1)}. \end{aligned}$$

If $2\nu = d$, we take $M = \frac{1}{2}\log N$ and obtain the bound

$$\frac{R}{\sqrt{N}} + \frac{6R}{\sqrt{N}} \sqrt{C_0} \frac{1}{2} \log N \leq C_1 R \frac{\log N}{\sqrt{N}}.$$

If $2\nu > d$, we take $M = \infty$ and obtain the bound

$$\frac{6R}{\sqrt{N}} \sqrt{C_0} \left(\frac{2^{\frac{d}{2\nu}-1}}{1 - 2^{\frac{d}{2\nu}-1}} \right) \leq C_1 R \frac{1}{\sqrt{N}}.$$

If $2\nu < d$, we take $M = \frac{\nu}{d}\log N$ and obtain the bound

$$R2^{-M} + \frac{6R}{\sqrt{N}} \sqrt{C_0} 2^{\frac{d}{2\nu}-1} \left(\frac{2^{M(\frac{d}{2\nu}-1)} - 1}{2^{\frac{d}{2\nu}-1} - 1} \right) \leq C_1 R N^{-\frac{\nu}{d}}.$$

Putting it all together, we obtain that $R(\mathcal{B}(X)) \leq C_1 E_{\nu/d}(N)$.

(3) For any ρ-Lipschitz loss function $\ell(x, f(x))$ with $|\ell(x,y)| \leq c_0$, we know that with probability $1-\delta$ over the sampling of the training set X from the distribution π, we have for all $f \in \mathcal{F}$

$$\begin{aligned} \mathbb{E}_{x\sim\pi}\left[\ell(x, f(x))\right] - \frac{1}{N}\sum_{i=1}^{N} \ell(x_i, f(x_i)) &\leq 2\mathbb{E}_{X'}\left[R(\ell \circ \mathcal{F}(X'))\right] + c_0 \sqrt{\frac{2\log{}^2\!/\!_\delta}{N}} \\ &\leq 2C_1 E_{\nu/d}(N) + c_0 \sqrt{\frac{2\log{}^2\!/\!_\delta}{N}}. \end{aligned}$$

□

Proposition 9. *Let $\mathcal{F}_1, \ldots, \mathcal{F}_L$ be set of functions mapping through the sets $\Omega_0, \ldots, \Omega_L$, then if all functions in $\mathcal{F}_\ell$ are ρ_ℓ-Lipschitz, we have*

$$\log \mathcal{N}_2(\mathcal{F}_L \circ \cdots \circ \mathcal{F}_1, \sum_{\ell=1}^{L} \rho_{L:\ell+1}\epsilon_\ell) \leq \sum_{\ell=1}^{L} \log \mathcal{N}_2(\mathcal{F}_\ell, \epsilon_\ell).$$

Proof. For any distribution π_0 on Ω there is a ϵ_1-covering $\tilde{\mathcal{F}}_1$ of $\mathcal{F}_1$ with $\left|\tilde{\mathcal{F}}_1\right| \leq \mathcal{N}_2(\mathcal{F}_1, \epsilon_1)$ then for any $\tilde{f}_1 \in \tilde{\mathcal{F}}_1$ we choose a ϵ_2-covering $\tilde{\mathcal{F}}_2$ w.r.t. the measure π_1 which is the measure of $f_1(x)$ if $x \sim \pi_0$ of $\mathcal{F}_2$ with $\left|\tilde{\mathcal{F}}_2\right| \leq \mathcal{N}_2(\mathcal{F}_2, \epsilon)$, and so on until we obtain coverings for all ℓ. Then the set $\tilde{\mathcal{F}} = \left\{ \tilde{f}_L \circ \cdots \circ \tilde{f}_1 : \tilde{f}_1 \in \tilde{\mathcal{F}}_1, \ldots, \tilde{f}_L \in \tilde{\mathcal{F}}_L \right\}$ is a $\sum_{\ell=1}^{L} \rho_{L:\ell+1}\epsilon_\ell$-covering of $\mathcal{F} = \mathcal{F}_L \circ \cdots \circ \mathcal{F}_1$, indeed for any $f = f_{L:1}$ we choose $\tilde{f}_1 \in \tilde{\mathcal{F}}_1, \ldots, \tilde{f}_L \in \tilde{\mathcal{F}}_L$ that cover $f_1, \ldots, f_L$, then $\tilde{f}_{L:1}$ covers $f_{L:1}$:

$$\begin{aligned}\left\| f_{L:1} - \tilde{f}_{L:1} \right\|_\pi &\leq \sum_{\ell=1}^{L} \left\| f_{L:\ell} \circ \tilde{f}_{\ell-1:1} - f_{L:\ell+1} \circ \tilde{f}_{\ell:1} \right\|_\pi \\ &\leq \sum_{\ell=1}^{L} \left\| f_{L:\ell} - f_{L:\ell+1} \circ \tilde{f}_\ell \right\|_{\pi_{\ell-1}} \\ &\leq \sum_{\ell=1}^{L} \rho_{L:\ell+1} \epsilon_\ell,\end{aligned}$$

and log cardinality of the set $\tilde{\mathcal{F}}$ is bounded $\sum_{\ell=1}^{L} \log \mathcal{N}_2(\mathcal{F}_\ell, \epsilon_\ell)$. □

Theorem 10. *Let $\mathcal{F} = \mathcal{F}_L \circ \cdots \circ \mathcal{F}_1$ where $\mathcal{F}_\ell = \left\{ f_\ell : \mathbb{R}^{d_{\ell-1}} \to \mathbb{R}^{d_\ell} \text{ s.t. } \|f_\ell\|_{W^{\nu_\ell,2}} \leq R_\ell, \|f_\ell\|_\infty \leq b_\ell, Lip(f_\ell) \leq \rho_\ell \right\}$, and let $r^* = \min_\ell r_\ell$ for $r_\ell = \frac{\nu_\ell}{d_{\ell-1}}$, then with probability $1-\delta$ we have for all $f \in \mathcal{F}$*

$$\mathcal{L}(f) - \tilde{\mathcal{L}}_N(f) \leq \rho C_0 \left(\sum_{\ell=1}^{L} \left(C_\ell \rho_{L:\ell+1} R_\ell \right)^{\frac{1}{r^*+1}} \right)^{r^*+1} E_{r^*}(N) + c_0 \sqrt{\frac{2 \log {}^2/_\delta}{N}},$$

where C_ℓ depends only on $d_{\ell-1}, d_\ell, \nu_\ell, b_{\ell-1}$.

Proof. (1) We know from Theorem 5.2 of [9] that the Sobolev ball $B_{W^{\nu_\ell,2}}(0, R_\ell)$ over any d_ℓ-dimensional hypercube Ω satisfies

$$\log \mathcal{N}_2(B_{W^{\nu,2}}(0, R_\ell), \pi_{\ell-1}, \epsilon_\ell) \leq \left(C_\ell \frac{R_\ell}{\epsilon_\ell} \right)^{\frac{1}{r_\ell}}$$

for a constant C_ℓ that depends on the size of hypercube and the dimension d_ℓ and the regularity ν_ℓ and any measure $\pi_{\ell-1}$ supported in the hypercube.

Thus Proposition 9 tells us that the composition of the Sobolev balls satisfies

$$\log \mathcal{N}_2(\mathcal{F}_L \circ \cdots \circ \mathcal{F}_1, \sum_{\ell=1}^{L} \rho_{L:\ell+1} \epsilon_\ell) \leq \sum_{\ell=1}^{L} \left(C_\ell \frac{R_\ell}{\epsilon_\ell} \right)^{\frac{1}{r_\ell}}.$$

Given $r^* = \min_\ell r_\ell$, we can bound it by $\sum_{\ell=1}^{L} \left(C_\ell \frac{R_\ell}{\epsilon_\ell} \right)^{\frac{1}{r^*}}$ and by then choosing $\epsilon_\ell = \frac{\rho_{L:\ell+1}^{-1} (\rho_{L:\ell+1} C_\ell R_\ell)^{\frac{1}{r^*+1}}}{\sum_\ell (\rho_{L:\ell+1} C_\ell R_\ell)^{\frac{1}{r^*+1}}} \epsilon$, we obtain that

$$\log \mathcal{N}_2(\mathcal{F}_L \circ \cdots \circ \mathcal{F}_1, \epsilon) \leq \left(\sum_{\ell=1}^{L} \left(\rho_{L:\ell+1} C_\ell R_\ell \right)^{\frac{1}{r^*+1}} \right)^{r^*+1} \epsilon^{-\frac{1}{r^*}}.$$

(2,3) It the follows by a similar argument as in points (2, 3) of the proof of Proposition 8 that there is a constant C_0 such that with probability $1-\delta$ for all $f \in \mathcal{F}$

$$\mathcal{L}(f) - \tilde{\mathcal{L}}_N(f) \leq C_0 \left(\sum_{\ell=1}^{L} \left(\rho_{L:\ell+1} C_\ell R_\ell \right)^{\frac{1}{r^*+1}} \right)^{r^*+1} E_{r^*}(N) + c_0 \sqrt{\frac{2 \log {}^2/_\delta}{N}}$$

□

D Generalization at the Regularized Global Minimum

In this section, we first give the proof of Theorem 5 and then present detailed proofs of lemmas used in the proof. The lemmas are largely inspired by [5] and may be of independent interest.

D.1 Theorem 5 in Section 4.2

Theorem 11 (Theorem 5 in the main)**.** *Given a true function $f^*_{L^*:1} = f^*_{L^*} \circ \cdots \circ f^*_1$ going through the dimensions $d^*_0, \dots, d^*_{L^*}$, along with a continuous input distribution π_0 supported in $B(0, b_0)$, such that the distributions π_ℓ of $f^*_\ell(x)$ (for $x \sim \pi_0$) are continuous too and supported inside $B(0, b_\ell) \subset \mathbb{R}^{d^*_\ell}$. Further assume that there are differentiabilities ν_ℓ and radii R_ℓ such that $\|f^*_\ell\|_{W^{\nu_\ell,2}(B(0,b_\ell))} \leq R_\ell$, and ρ_ℓ such that $Lip(f^*_\ell) \leq \rho_\ell$. For a infinite width AccNet with $L \geq L^*$ and constant width $d \geq d^*_1, \dots, d^*_{L^*-1}$, we have for the ratios $\tilde{r}_\ell = \frac{\nu_\ell}{d^*_\ell + 3}$:*

- *At a global minimizer $\hat{f}_{L:1}$ of the regularized loss $f_1, \dots, f_L \mapsto \tilde{\mathcal{L}}_N(f_{L:1}) + \lambda R(f_1, \dots, f_L)$, we have $\mathcal{L}(\hat{f}_{L:1}) = \tilde{O}(N^{-\min\{\frac{1}{2}, \tilde{r}_1, \dots, \tilde{r}_{L^*}\}})$.*
- *At a global minimizer $\hat{f}_{L:1}$ of the regularized loss $f_1, \dots, f_L \mapsto \tilde{\mathcal{L}}_N(f_{L:1}) + \lambda \prod_{\ell=1}^{L} \|f_\ell\|_{F_1}$, we have $\mathcal{L}(\hat{f}_{L:1}) = \tilde{O}(N^{-\frac{1}{2} + \sum_{\ell=1}^{L^*} \max\{0, \tilde{r}_\ell - \frac{1}{2}\}})$.*

Proof. If $f^* = f^*_{L^*} \circ \cdots \circ f^*_1$ with $L^* \leq L$, intermediate dimensions $d^*_0, \dots, d^*_{L^*}$, along with a continuous input distribution π_0 supported in $B(0, b_0)$, such that the distributions π_ℓ of $f^*_\ell(x)$ (for $x \sim \pi_0$) are continuous too and supported inside $B(0, b_\ell) \subset \mathbb{R}^{d^*_\ell}$. Further assume that there are differentiabilities ν^*_ℓ and radii R_ℓ such that $\|f^*_\ell\|_{W^{\nu^*_\ell,2}(B(0,b_\ell))} \leq R_\ell$.

We first focus on the $L = L^*$ case and then extend to the $L > L^*$ case.

Each f^*_ℓ can be approximated by another function $\tilde{f}_\ell$ with bounded F_1-norm and Lipschitz constant. Actually if $2\nu^*_\ell \geq d^*_{\ell-1} + 3$ one can choose $\tilde{f}_\ell = f^*_\ell$ since $\|f^*_\ell\|_{F_1} \leq C_\ell R_\ell$ by Lemma 14, and by assumption $Lip(\tilde{f}_\ell) \leq \rho_\ell$. If $2\nu^*_\ell < d^*_{\ell-1} + 3$, then by Lemma 13 we know that there is a $\tilde{f}_\ell$ with $\left\|\tilde{f}_\ell\right\|_{F_1} \leq C_\ell R_\ell \epsilon_\ell^{-\frac{1}{2\tilde{r}_\ell}+1}$ and $Lip(\tilde{f}_\ell) \leq C_\ell Lip(f^*_\ell) \leq C_\ell \rho_\ell$ and error

$$\left\|f^*_\ell - \tilde{f}_\ell\right\|_{L_2(\pi_{\ell-1})} \leq c_\ell \left\|f^* - \tilde{f}_\ell\right\|_{L_2(B(0,b_\ell))} \leq c_\ell \epsilon_\ell.$$

Therefore the composition $\tilde{f}_{L:1}$ satisfies

$$\begin{aligned}
\left\|f^*_{L:1} - \hat{f}_{L:1}\right\|_{L_2(\pi_{\ell-1})} &\leq \sum_{\ell=1}^{L} \left\|\tilde{f}_{L:\ell+1} \circ f^*_{\ell:1} - \tilde{f}_{L:\ell} \circ f^*_{\ell-1:1}\right\|_{L_2(\pi)} \\
&\leq \sum_{\ell=1}^{L} Lip(\tilde{f}_{L:\ell+1}) c_\ell \epsilon_\ell \\
&\leq \sum_{\ell=1}^{L} \rho_{L:\ell+1} C_{L:\ell+1} c_\ell \epsilon_\ell.
\end{aligned}$$

For any $L \geq L^*$, dimensions $d_\ell \geq d^*_\ell$ and widths $w_\ell \geq N$, we can build an AccNet that fits eactly $\tilde{f}_{L:1}$, by simply adding zero weights along the additional dimensions and widths, and by adding identity layers if $L > L^*$, since it is possible to represent the identity on $\mathbb{R}^d$ with a shallow network with $2d$ neurons and F_1-norm $2d$ (by having two neurons $e_i \sigma(e_i^T \cdot)$ and $-e_i \sigma(-e_i^T \cdot)$ for each basis e_i). Since the cost in parameter norm of representing the identity scales with the dimension, it is best to add those identity layers at the minimal dimension $\min\{d^*_0, \dots, d^*_{L^*}\}$. We therefore end up with a AccNet with $L - L^*$ identity layers (with F_1 norm $2\min\{d^*_0, \dots, d^*_{L^*}\}$) and L^* layers that approximate each of the f^*_ℓ with a bounded F_1-norm function $\tilde{f}_\ell$.

Since $f^*_{L:1}$ has zero population loss, the population loss of the AccNet $\tilde{f}_{L:1}$ is bounded by $\rho \sum_{\ell=1}^{L} \rho_{L:\ell+1} C_{L:\ell+1} c_\ell \epsilon_\ell$. By McDiarmid's inequality, we know that with probability $1 - \delta$ over the sampling of the training set, the training loss is bounded by $\rho \sum_{\ell=1}^{L} \rho_{L:\ell+1} C_{L:\ell+1} c_\ell \epsilon_\ell + B\sqrt{\frac{2\log 2/\delta}{N}}$.
(1) The global minimizer $\hat{f}_{L:1} = \hat{f}_L \circ \cdots \circ \hat{f}_1$ of the regularized loss (with the first regularization

term) is therefore bounded by

$$\rho\sum_{\ell=1}^{L}\rho_{L:\ell+1}C_{L:\ell+1}c_{\ell}\epsilon_{\ell}+B\sqrt{\frac{2\log{}^{2}/\delta}{N}}$$
$$+\lambda\sqrt{2d}\left[\prod_{\ell=1}^{L^{*}}C_{\ell}\rho_{\ell}\sum_{\ell=1}^{L^{*}}\frac{1}{C_{\ell}\rho_{\ell}}\begin{cases}C_{\ell}R_{\ell} & 2\nu_{\ell}^{*}\geq d_{\ell-1}^{*}+3\\ C_{\ell}R_{\ell}\epsilon_{\ell}^{-\frac{1}{2\tilde{r}_{\ell}}+1} & 2\nu_{\ell}^{*}<d_{\ell-1}^{*}+3\end{cases}+2(L-L^{*})\min\{d_{0}^{*},\dots,d_{L^{*}}^{*}\}\right].$$

Taking $\epsilon_{\ell}=E_{\tilde{r}_{min}}(N)$ and $\lambda=N^{-\frac{1}{2}}\log N$, this is upper bounded by

$$\left[\rho\sum_{\ell=1}^{L}\rho_{L:\ell+1}C_{L:\ell+1}c_{\ell}+C\sqrt{2d}r\prod_{\ell=1}^{L^{*}}C_{\ell}\rho_{\ell}\sum_{\ell=1}^{L^{*}}\frac{R_{\ell}}{\rho_{\ell}}+2(L-L^{*})\min\{d_{0}^{*},\dots,d_{L^{*}}^{*}\}\right]E_{\tilde{r}_{min}}(N)+B\sqrt{\frac{2\log{}^{2}/\delta}{N}}.$$

which implies that at the globla minimizer of the regularized loss, the (unregularized) train loss is of order $E_{\tilde{r}_{min}}(N)$ and the complexity measure $R(\hat{f}_{1},\dots,\hat{f}_{L})$ is of order $\frac{1}{N}E_{\tilde{r}_{min}}(N)$ which implies that the test error will be of order

$$\mathcal{L}(f)\leq\left[2\rho\sum_{\ell=1}^{L}\rho_{L:\ell+1}C_{L:\ell+1}c_{\ell}+2C\sqrt{2d}r\prod_{\ell=1}^{L^{*}}C_{\ell}\rho_{\ell}\sum_{\ell=1}^{L^{*}}\frac{R_{\ell}}{\rho_{\ell}}+2(L-L^{*})\min\{d_{0}^{*},\dots,d_{L^{*}}^{*}\}\right]E_{\tilde{r}_{min}}(N)$$
$$+(2B+c_{0})\sqrt{\frac{2\log{}^{2}/\delta}{N}}.$$

(2) Let us now consider adding the closer to traditional L_2-regularization $\mathcal{L}_{\lambda}(f_{L:1})=\mathcal{L}(f_{L:1})+\lambda\prod_{\ell=1}^{L}\|f_{\ell}\|_{F_{1}}$. ,We see that the global minimizer $\hat{f}_{L:1}$ of the L_2-regularized loss is upper bounded by

$$\rho\sum_{\ell=1}^{L}\rho_{L:\ell+1}C_{L:\ell+1}c_{\ell}\epsilon_{\ell}+B\sqrt{\frac{2\log{}^{2}/\delta}{N}}+\lambda\left[\prod_{\ell=1}^{L^{*}}\begin{cases}C_{\ell}R_{\ell} & 2\nu_{\ell}^{*}\geq d_{\ell-1}^{*}+3\\ C_{\ell}R_{\ell}\epsilon_{\ell}^{-\frac{1}{2\tilde{r}_{\ell}}+1} & 2\nu_{\ell}^{*}<d_{\ell-1}^{*}+3\end{cases}\right](2\min\{d_{0}^{*},\dots,d_{L^{*}}^{*}\})^{(L-L^{*})}.$$

Which for $\epsilon_{\ell}=E_{\tilde{r}_{min}}(N)$ and $\lambda=N^{-\frac{1}{N}}$ is upper bounded by

$$\rho\sum_{\ell=1}^{L}\rho_{L:\ell+1}C_{L:\ell+1}c_{\ell}E_{\tilde{r}_{min}}(N)+B\sqrt{\frac{2\log{}^{2}/\delta}{N}}+N^{-\frac{1}{2}}\left[\prod_{\ell=1}^{L^{*}}C_{\ell}R_{\ell}\sqrt{N}E_{\tilde{r}_{min}}(N)\right](2\min\{d_{0}^{*},\dots,d_{L^{*}}^{*}\})^{(L-L^{*})}.$$

Which implies that both the train error is of order $N^{-\frac{1}{2}}\prod_{\ell=1}^{L^{*}}\sqrt{N}E_{\tilde{r}_{min}}(N)$ and the product of the F_1-norms is of order $\prod_{\ell=1}^{L^{*}}\sqrt{N}E_{\tilde{r}_{min}}(N)$.

Now note that the product of the F_1-norms bounds the complexity measure up to a constant since $Lip(f)\leq\|f\|_{F_{1}}$

$$R(f_{1},\dots,f_{L})=r\prod_{\ell=1}^{L}Lip(f_{\ell})\sum_{\ell=1}^{L}\frac{\|f_{\ell}\|_{F_{1}}}{Lip(f_{\ell})}\sqrt{d_{\ell-1}+d_{\ell}}\leq L\sqrt{2d}\prod_{\ell=1}^{L}\|f\|_{F_{1}}.$$

And since at the global minimum the product of the F_1-norms is of order $\prod_{\ell=1}^{L^{*}}\sqrt{N}E_{\tilde{r}_{min}}(N)$ the test error will of order $\left(\prod_{\ell=1}^{L^{*}}\sqrt{N}E_{\tilde{r}_{\ell}}(N)\right)\frac{\log N}{\sqrt{N}}$.

Note that if there is at a most one ℓ where $\tilde{r}_{\ell}>\frac{1}{2}$ then the rate is up to log term the same as $E_{\tilde{r}_{min}}(N)$. □

D.2 Lemmas on approximating Sobolev functions

Now we present the lemmas used in this proof above that concern the approximation errors and Lipschitz constants of Sobolev functions and compositions of them. We will bound the F_2-norm and note that the F_2-norm is larger than the F_1-norm, cf. [5, Section 3.1].

Lemma 12 (Approximation for Sobolev function with bounded error and Lipschitz constant)**.** *Suppose $g : \mathbb{S}_d \to \mathbb{R}$ is an even function with bounded Sobolev norm $\|g\|^2_{W^{\nu,2}} \le R$ with $2\nu \le d+2$, with inputs on the unit d-dimensional sphere. Then for every $\epsilon > 0$, there is $\hat{g} \in \mathcal{G}_2$ with small approximation error $\|g - \hat{g}\|_{L_2(\mathbb{S}_d)} = C(d,\nu,R)\epsilon$, bounded Lipschitzness $\mathrm{Lip}(\hat{g}) \le C'(d)\mathrm{Lip}(g)$, and bounded norm*

$$\|\hat{g}\|_{F_2} \le C''(d,\nu,R)\epsilon^{-\frac{d+3-2\nu}{2\nu}}.$$

Proof. Given our assumptions on the target function g, we may decompose $g(x) = \sum_{k=0}^{\infty} g_k(x)$ along the basis of spherical harmonics with $g_0(x) = \int_{\mathbb{S}_d} g(y)\mathrm{d}\tau_d(y)$ being the mean of $g(x)$ over the uniform distribution τ_d over $\mathbb{S}_d$. The k-th component can be written as

$$g_k(x) = N(d,k)\int_{\mathbb{S}_d} g(y)P_k(x^T y)\mathrm{d}\tau_d(y)$$

with $N(d,k) = \frac{2k+d-1}{k}\binom{k+d-2}{d-1}$ and a Gegenbauer polynomial of degree k and dimension $d+1$:

$$P_k(t) = (-1/2)^k \frac{\Gamma(d/2)}{\Gamma(k+d/2)}(1-t^2)^{(2-d)/2}\frac{d^k}{dt^k}(1-t^2)^{k+(d-2)/2},$$

known as Rodrigues' formula. Given the assumption that the Sobolev norm $\|g\|^2_{W^{\nu,2}}$ is upper bounded, we have $\|f\|^2_{L_2(\mathbb{S}_d)} \le C_0(d,\nu)R$ for $f = \Delta^{\nu/2}g$ where Δ is the Laplacian on $\mathbb{S}_d$ [18, 5]. Note that g_k are eigenfunctions of the Laplacian with eigenvalues $k(k+d-1)$ [4], thus

$$\|g_k\|^2_{L_2(\mathbb{S}_d)} = \|f_k\|^2_{L_2(\mathbb{S}_d)}(k(k+d-1))^{-\nu} \le \|f_k\|^2_{L_2(\mathbb{S}_d)}k^{-2\nu} \le C_1(d,\nu,R)k^{-2\nu-1} \tag{1}$$

where the last inequality holds because $\|f\|^2_{L_2(\mathbb{S}_d)} = \sum_{k\ge 0}\|f_k\|^2_{L_2(\mathbb{S}_d)}$ converges. Note using the Hecke-Funk formula, we can also write g_k as scaled p_k for the underlying density p of the F_1 and F_2-norms:

$$g_k(x) = \lambda_k p_k(x)$$

where $\lambda_k = \frac{\omega_{d-1}}{\omega_d}\int_{-1}^{1}\sigma(t)P_k(t)(1-t^2)^{(d-2)/2}\mathrm{d}t = \Omega(k^{-(d+3)/2})$ [5, Appendix D.2] and ω_d denotes the surface area of $\mathbb{S}_d$. Then by definition of $\|\cdot\|_{F_2}$, for some probability density p,

$$\|g\|^2_{F_2} = \int_{\mathbb{S}_d}|p|^2\mathrm{d}\tau(v) = \|p\|^2_{L_2(\mathbb{S}_d)} = \sum_{0\le k}\|p_k\|^2_{L_2(\mathbb{S}_d)} = \sum_{0\le k}\lambda_k^{-2}\|g_k\|^2_{L_2(\mathbb{S}_d)}.$$

Now to approximate g, consider function $\hat{g}$ defined by truncating the "high frequencies" of g, i.e. setting $\hat{g}_k = \mathbb{1}[k \le m]g_k$ for some $m > 0$ we specify later. Then we can bound the norm with

$$\begin{aligned}\|\hat{g}\|^2_{F_2} &= \sum_{0\le k:\lambda_k\ne 0}\lambda_k^{-2}\|\hat{g}_k\|^2_{L_2(\mathbb{S}_d)} = \sum_{\substack{0\le k\le m\\ \lambda_k\ne 0}}\lambda_k^{-2}\|g_k\|^2_{L_2(\mathbb{S}_d)}\\ &\overset{\text{(a)}}{\le} C_2(d,\nu,R)\sum_{0\le k\le m}k^{d+2-2\nu}\\ &\overset{\text{(b)}}{\le} C_3(d,\nu,R)m^{d+3-2\nu}\end{aligned}$$

where (a) uses Eq 1 and $\lambda_k = \Omega(k^{-(d+3)/2})$; (b) approximates by integral.

To bound the approximation error,

$$\begin{aligned}\|g-\hat{g}\|^2_{L_2(\mathbb{S}_d)} &= \left\|\sum_{k>m}g_k\right\|^2_{L_2(\mathbb{S}_d)} \le \sum_{k>m}\|g_k\|^2_{L_2(\mathbb{S}_d)}\\ &\le C_4(d,\nu,R)\sum_{k>m}k^{-2\nu-1}\\ &\le C_5(d,\nu,R)m^{-2\nu} \quad \text{by integral approximation.}\end{aligned}$$

Finally, choosing $m = \epsilon^{-\frac{1}{\nu}}$, we obtain $\|g-\hat{g}\|_{L_2(\mathbb{S}_d)} \le C(d,\nu,R)\epsilon$ and

$$\|\hat{g}\|_{F_2} \le C'(d,\nu,R)\epsilon^{-\frac{d+3-2\nu}{2\nu}}.$$

Then it remains to bound $\mathrm{Lip}(\hat{g})$ for our constructed approximation. By construction and by [13, Theorem 2.1.3], we have $\hat{g} = g * h$ with now

$$h(t) = \sum_{k=0}^{m} h_k P_k(t), \quad t \in [-1, 1]$$

by orthogonality of the Gegenbauer polynomial P_k's and the convolution is defined as

$$(g * h)(x) := \frac{1}{\omega_d} \int_{\mathbb{S}_d} g(y) h(\langle x, y \rangle) \mathrm{d}y.$$

The coefficients for $0 \leq k \leq m$ given by [13, Theorem 2.1.3] are

$$h_k \overset{\text{(a)}}{=} \frac{\omega_{d+1}}{\omega_d} \frac{\Gamma(d-1)}{\Gamma(d-1+k)} P_k(1) \frac{k!(k+(d-1)/2)\Gamma((d-1)/2)^2}{\pi 2^{2-d}\Gamma(d-1+k)} \overset{\text{(b)}}{=} O\left(\frac{k}{\Gamma(d-1+k)}\right)$$

where (a) follows from the (inverse of) weighted L_2 norm of P_k; (b) plugs in the unit constant $P_k(1) = \frac{\Gamma(k+d-1)}{\Gamma(d-1)k!}$ and suppresses the dependence on d. Note that the constant factor $\frac{\Gamma(d-1)}{\Gamma(d-1+k)}$ comes from the difference in the definitions of the Gegenbauer polynomials here and in [13]. Then we can bound

$$\begin{aligned}
\|\nabla \hat{g}(x)\|_{op} &\leq \int_{\mathbb{S}_d} \|\nabla g(y)\|_{op} |h(\langle x, y \rangle)| \mathrm{d}y \\
&\leq \mathrm{Lip}(g) \int_{\mathbb{S}_d} |h(\langle x, y \rangle)| \mathrm{d}y \\
&\leq \sqrt{\omega_d} \mathrm{Lip}(g) \left(\int_{\mathbb{S}_d} h(\langle x, y \rangle)^2 \mathrm{d}y \right)^{1/2} \qquad \text{by Cauchy-Schwartz} \\
&= \sqrt{\omega_d} \mathrm{Lip}(g) \left(\sum_{k,j=0}^{m} \int_{\mathbb{S}_d} h_k h_j P_k(\langle x, y \rangle) P_j(\langle x, y \rangle) \mathrm{d}y \right)^{1/2} \\
&= \sqrt{\omega_d} \mathrm{Lip}(g) \left(\sum_{k,j=0}^{m} \int_{-1}^{1} h_k h_j P_k(t) P_j(t) (1-t^2)^{\frac{d-2}{2}} \mathrm{d}t \right)^{1/2} \qquad \text{by [13, Eq A.5.1]} \\
&= \sqrt{\omega_d} \mathrm{Lip}(g) \left(\sum_{k=0}^{m} h_k^2 \int_{-1}^{1} P_k(t)^2 (1-t^2)^{\frac{d-2}{2}} \mathrm{d}t \right)^{1/2} \qquad \text{by orthogonality of } P_k\text{'s w.r.t. this measure} \\
&= \sqrt{\omega_d} \mathrm{Lip}(g) \left(\sum_{k=0}^{m} h_k^2 \frac{\pi 2^{2-d}\Gamma(d-1+k)}{k!(k+(d-1)/2)\Gamma((d-1)/2)^2} \right)^{1/2} \\
&= \sqrt{\omega_d} \mathrm{Lip}(g) \left(O(1) + \sum_{k=1}^{m} O\left(\frac{k}{\Gamma(d-1+k)k!} \right) \right)^{1/2} \\
&= \sqrt{\omega_d} \mathrm{Lip}(g) C(d)
\end{aligned}$$

for some constant $C(d)$ that only depends on d. Hence $\mathrm{Lip}(\hat{g}) = C'(d)\mathrm{Lip}(g)$. □

The next lemma adapts Lemma 12 to inputs on balls instead of spheres following the construction in [5, Proposition 5].

Lemma 13. *Suppose $f : B(0, b) \to \mathbb{R}$ has bounded Sobolev norm $\|f\|^2_{W^{\nu,2}} \leq R$ with $\nu \leq (d+2)/2$ even, where $B(0, b) = \{x \in \mathbb{R}^d : \|x\|_2 \leq b\}$ is the radius-b ball. Then for every $\epsilon > 0$ there exists $f_\epsilon \in \mathcal{F}_2$ such that $\|f - f_\epsilon\|_{L_2(B(0,b))} = C(d, \nu, b, R)\epsilon$, $\mathrm{Lip}(f_\epsilon) \leq C'(b, d)\mathrm{Lip}(f)$, and*

$$\|f_\epsilon\|_{F_2} \leq C''(d, \nu, b, R)\epsilon^{-\frac{d+3-2\nu}{2\nu}}$$

Proof. Define $g(z, a) = f\left(\frac{2bz}{a}\right) a$ on $(z, a) \in \mathbb{S}_d$ with $z \in \mathbb{R}^d$ and $\frac{1}{\sqrt{2}} \leq a \in \mathbb{R}$. One may verify that unit-norm (z, a) with $a \geq \frac{1}{\sqrt{2}}$ is sufficient to cover $B(0, b)$ by setting $x = \frac{bz}{a}$ and solve for (z, a). Then we have bounded $\|g\|^2_{W^{\nu,2}} \leq b^\nu R$ and may apply Lemma 12 to get $\hat{g}$ with $\|g - \hat{g}\|_{L_2(\mathbb{S}_d)} \leq C(d, \nu, b, R)\epsilon$. Letting $f_\epsilon(x) = \hat{g}\left(\frac{ax}{b}, a\right) a^{-1}$ for the corresponding $\left(\frac{ax}{b}, a\right) \in \mathbb{S}_d$ gives the desired upper bounds. □

Lemma 14. *Suppose $f : B(0, b) \to \mathbb{R}$ has bounded Sobolev norm $\|f\|^2_{W^{\nu,2}} \leq R$ with $\nu \geq (d+3)/2$ even. Then $f \in \mathcal{F}_2$ and $\|f\|_{F_2} \leq C(d, \nu) b^\nu R$.*

In particular, $W^{\nu,2} \subseteq \mathcal{F}_2$ for $\nu \geq (d+3)/2$ even.

Proof. This lemma reproduces [5, Proposition 5] to functions with bounded Sobolev L_2 norm instead of L_∞ norm. The proof follows that of Lemma 12 and Lemma 13 and noticing that by Eq 1,

$$\begin{aligned}
\|g\|^2_{F_2} &= \sum_{0 \leq k : \lambda_k \neq 0} \lambda_k^{-2} \|g_k\|^2_{L_2(\mathbb{S}_d)} \\
&\leq \sum_{0 \leq k} k^{d+3-2\nu} \|(\Delta^{\nu/2} g)_k\|^2_{L_2(\mathbb{S}_d)} \\
&\leq \|\Delta^{\nu/2} g\|^2_{L_2(\mathbb{S}_d)} \\
&\leq C_1(d, \nu) \|g\|^2_{W^{\nu,2}} \\
&\leq C_1(d, \nu) R.
\end{aligned}$$

□

Finally, we remark that the above lemmas extend straightforward to functions $f : B(0, b) \to \mathbb{R}^{d'}$ with multi-dimensional outputs, where the constants then depend on the output dimension d' too.

D.3 Lemma on approximating compositions of Sobolev functions

With the lemmas given above and the fact that the F_2-norm upper bounds the F_1-norm, we can find infinite-width DNN approximations for compositions of Sobolev functions, which is also pointed out in the proof of Theorem 5.

Lemma 15. *Assume the target function $f : \Omega \to \mathbb{R}^{d_{out}}$, with $\Omega \subseteq B(0, b) \subseteq \mathbb{R}^{d_{in}}$, satisfies:*

- *$f = g_k \circ \cdots \circ g_1$ a composition of k Sobolev functions $g_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i+1}}$ with bounded norms $\|g_i\|^2_{W^{\nu_i,2}} \leq R$ for $i = 1, \ldots, k$, with $d_1 = d_{in}$;*
- *f is Lipschitz, i.e. $\mathrm{Lip}(g_i) < \infty$ for $i = 1, \ldots, k$.*

If $\nu_i \leq (d_i + 2)/2$ for any i, i.e. less smooth than needed, for depth $L \geq k$ and any $\epsilon > 0$, there is an infinite-width DNN $\tilde{f}$ such that

- *$\mathrm{Lip}(\tilde{f}) \leq C_1 \prod_{i=1}^k \mathrm{Lip}(g_i)$;*
- *$\|\tilde{f} - f\|_{L_2} \leq C_2 \epsilon$;*

the constants C_1 depends on all of the input dimensions d_i (to g_i) and d_{out}, and C_2 depends on $d_i, d_{out}, \nu_i, b, R, k,$ and $\mathrm{Lip}(g_i)$ for all i.

If otherwise $\nu_i \geq (d_i + 3)/2$ for all i, we can have $\tilde{f} = f$ where each layer has a parameter norm bounded by $C_3 R$, with C_3 depending on $d_i, d_{out}, \nu_i,$ and b.

Proof. Note that by Lipschitzness,

$$(g_i \circ \cdots \circ g_1)(\Omega) \subseteq B\left(0, b \prod_{j=1}^{i} \mathrm{Lip}(g_j)\right),$$

i.e. the pre-image of each component lies in a ball. By Lemma 12, for each g_i, if $\nu_i \leq (d_i + 2)/2$, we have an approximation $\hat{g}_i$ on a slightly larger ball $b'_i = b \prod_{j=1}^{i-1} C''(d_j, d_{j+1}) \mathrm{Lip}(g_j)$ such that

- $\|g_i - \hat{g}_i\|_{L_2} \leq C(d_i, d_{i+1}, \nu_i, b'_i, R)\epsilon$;
- $\|\hat{g}_i\|_{F_2} \leq C'(d_i, d_{i+1}, \nu_i, b'_i, R)\epsilon^{\frac{d_i+3-2\nu_i}{2\nu_i}}$;
- $\mathrm{Lip}(\hat{g}_i) \leq C''(d_i, d_{i+1})\mathrm{Lip}(g_i)$;

where d_i is the input dimension of g_i. Write the constants as C_i, C'_i, and C''_i for notation simplicity. Note that the Lipschitzness of the approximations $\hat{g}_i$'s guarantees that, when they are composed, $(\hat{g}_{i-1} \circ \cdots \circ \hat{g}_1)(\Omega)$ lies in a ball of radius $b'_i = b\prod_{j=1}^{i-1} C''_j \mathrm{Lip}(g_j)$, hence the approximation error remains bounded while propagating. While each $\hat{g}_i$ is a (infinite-width) layer, for the other $L-k$ layers, we may have identity layers[5].

Let $\tilde{f}$ be the composed DNN of these layers. Then we have

$$\mathrm{Lip}(\tilde{f}) \leq \prod_{i=1}^{k} C''_i \mathrm{Lip}(g_i) = C''(d_1, \ldots, d_k, d_{out}) \prod_{i=1}^{k} \mathrm{Lip}(g_i)$$

and approximation error

$$\|\tilde{f} - f\|_{L_2} \leq \sum_{i=1}^{k} C_i \epsilon \prod_{j>i} C''_j \mathrm{Lip}(g_j) = O(\epsilon)$$

where the last equality suppresses the dependence on $d_i, d_{out}, \nu_i, b, R, k$, and $\mathrm{Lip}(g_i)$ for $i = 1, \ldots, k$.

In particular, by Lemma 14, if $\nu_i \geq (d_i+3)/2$ for any $i = 1, \ldots, k$, we can take $\hat{g}_i = g_i$. If this holds for all i, then we can have $\tilde{f} = f$ while each layer has a F_2-norm bounded by $O(R)$. $\square$

E Technical results

Here we show a number of technical results regarding the covering number.

First, here is a bound for the covering number of Ellipsoids, which is a simple reformulation of Theorem 2 of [17]:

Theorem 16. *The d-dimensional ellipsoid $E = \{x : x^T K^{-1} x \leq 1\}$ with radii $\sqrt{\lambda_i}$ for λ_i the i-th eigenvalue of K satisfies $\log \mathcal{N}_2(E, \epsilon) = M_\epsilon (1 + o(1))$ for*

$$M_\epsilon = \sum_{i:\sqrt{\lambda_i} \geq \epsilon} \log \frac{\sqrt{\lambda_i}}{\epsilon}$$

if one has $\log \frac{\sqrt{\lambda_1}}{\epsilon} = o\left(\frac{M_\epsilon^2}{k_\epsilon \log d}\right)$ for $k_\epsilon = \left|\left\{i : \sqrt{\lambda_i} \geq \epsilon\right\}\right|$

For our purpose, we will want to cover a unit ball $B = \{w : \|w\| \leq 1\}$ w.r.t. to a non-isotropic norm $\|w\|_K^2 = w^T K w$, but this is equivalent to covering E with an isotropic norm:

Corollary 17. *The covering number of the ball $B = \{w : \|w\| \leq 1\}$ w.r.t. the norm $\|w\|_K^2 = w^T K w$ satisfies $\log \mathcal{N}(B, \|\cdot\|_K, \epsilon) = M_\epsilon (1 + o(1))$ for the same M_ϵ as in Theorem 16 and under the same condition.*

Furthermore, $\log \mathcal{N}(B, \|\cdot\|_K, \epsilon) \leq \frac{\mathrm{Tr} K}{2\epsilon^2}(1 + o(1))$ as long as $\log d = o\left(\frac{\sqrt{\mathrm{Tr} K}}{\epsilon}\left(\log \frac{\sqrt{\mathrm{Tr} K}}{\epsilon}\right)^{-1}\right)$.

Proof. If $\tilde{E}$ is an ϵ-covering of E w.r.t. to the L_2-norm, then $\tilde{B} = K^{-\frac{1}{2}}\tilde{E}$ is an ϵ-covering of B w.r.t. the norm $\|\cdot\|_K$, because if $w \in B$, then $\sqrt{K}w \in E$ and so there is an $\tilde{x} \in \tilde{E}$ such that $\left\|x - \sqrt{K}w\right\| \leq \epsilon$, but then $\tilde{w} = \sqrt{K}^{-1} x$ covers w since $\|\tilde{w} - w\|_K = \left\|x - \sqrt{K}w\right\|_K \leq \epsilon$.

[5] Since the domain is always bounded here, one can let the bias translate the domain to the first quadrant and let the weight be the identity matrix, cf. the construction in [45, Proposition B.1.3].

Since $\lambda_i \leq \frac{\mathrm{Tr}K}{i}$, we have $K \leq \bar{K}$ for $\bar{K}$ the matrix obtained by replacing the i-th eigenvalue λ_i of K by $\frac{\mathrm{Tr}K}{i}$, and therefore $\mathcal{N}\left(B, \|\cdot\|_K, \epsilon\right) \leq \mathcal{N}\left(B, \|\cdot\|_{\bar{K}}, \epsilon\right)$ since $\|\cdot\|_K \leq \|\cdot\|_{\bar{K}}$. We now have the a[proximation $\log \mathcal{N}\left(B, \|\cdot\|_{\bar{K}}, \epsilon\right) = \bar{M}_\epsilon \left(1 + o(1)\right)$ for

$$\bar{M}_\epsilon = \sum_{i=1}^{\bar{k}_\epsilon} \log \frac{\sqrt{\mathrm{Tr}K}}{\sqrt{i}\epsilon}$$
$$\bar{k}_\epsilon = \left\lfloor \frac{\mathrm{Tr}K}{\epsilon^2} \right\rfloor .$$

We now have the simplification

$$\bar{M}_\epsilon = \sum_{i=1}^{k_\epsilon} \log \frac{\sqrt{\mathrm{Tr}K}}{\sqrt{i}\epsilon} = \frac{1}{2} \sum_{i=1}^{\bar{k}_\epsilon} \log \frac{\bar{k}_\epsilon}{i} = \frac{\bar{k}_\epsilon}{2} (\int_0^1 \log \frac{1}{x} dx + o(1)) = \frac{\bar{k}_\epsilon}{2} (1 + o(1))$$

where the $o(1)$ term vanishes as $\epsilon \searrow 0$. Furthermore, this allows us to check that as long as $\log d = o\left(\frac{\sqrt{\mathrm{Tr}K}}{4\epsilon \log \frac{\sqrt{\mathrm{Tr}K}}{\epsilon}}\right)$, the condition is satisfied

$$\log \frac{\sqrt{\mathrm{Tr}K}}{\epsilon} = o\left(\frac{\bar{k}_\epsilon}{4 \log d}\right) = o\left(\frac{\bar{M}_\epsilon^2}{\bar{k}_\epsilon \log d}\right).$$

$\square$

Second we prove how to obtain the covering number of the convex hull of a function set $\mathcal{F}$:

Theorem 18. *Let $\mathcal{F}$ be a set of B-uniformly bounded functions, then for all $\epsilon_K = B2^{-K}$*

$$\sqrt{\log \mathcal{N}_2(\mathrm{Conv}\mathcal{F}, 2\epsilon_K)} \leq \sqrt{18} \sum_{k=1}^{K} 2^{K-k} \sqrt{\log \mathcal{N}_2(\mathcal{F}, B2^{-k})}.$$

Proof. Define $\epsilon_k = B2^{-k}$ and the corresponding ϵ_k-coverings $\tilde{\mathcal{F}}_k$ (w.r.t. some measure π). For any f, we write $\tilde{f}_k[f]$ for the function $\tilde{f}_k[f] \in \tilde{\mathcal{F}}_k$ that covers f. Then for any functions f in $\mathrm{Conv}\mathcal{F}$, we have

$$f = \sum_{i=1}^{m} \beta_i f_i = \sum_{i=1}^{m} \beta_i \left(f_i - \tilde{f}_K[f_i]\right) + \sum_{k=1}^{K} \sum_{i=1}^{m} \beta_i \left(\tilde{f}_k[f_i] - \tilde{f}_{k-1}[f_i]\right) + \tilde{f}_0[f_i].$$

We may assume that $\tilde{f}_0[f_i] = 0$ since the zero function ϵ_0-covers the whole $\mathcal{F}$ since $\epsilon_0 = B$.

We will now use the probabilistic method to show that the sums $\sum_{i=1}^{m} \beta_i \left(\tilde{f}_k[f_i] - \tilde{f}_{k-1}[f_i]\right)$ can be approximated by finite averages. Consider the random functions $\tilde{g}_1^{(k)}, \ldots, \tilde{g}_{m_k}^{(k)}$ sampled iid with $\mathbb{P}\left[\tilde{g}_j^{(k)}\right] = \left(\tilde{f}_k[f_i] - \tilde{f}_{k-1}[f_i]\right)$ with probability β_i. We have $\mathbb{E}[\tilde{g}_j^{(k)}] = \sum_{i=1}^{m} \beta_i \left(\tilde{f}_k[f_i] - \tilde{f}_{k-1}[f_i]\right)$ and

$$\begin{aligned}
\mathbb{E}\left\|\sum_{k=1}^{K} \frac{1}{m_k} \sum_{j=1}^{m_k} \tilde{g}_j^{(k)} - \sum_{k=1}^{K} \sum_{i=1}^{m} \beta_i \left(\tilde{f}_k[f_i] - \tilde{f}_{k-1}[f_i]\right)\right\|_{L_p(\pi)}^p &\leq \sum_{k=1}^{K} \frac{1}{m_k^p} \sum_{j=1}^{m_k} \mathbb{E}\left\|\tilde{g}_j^{(k)}\right\|_{L_p(\pi)}^p \\
&= \sum_{k=1}^{K} \frac{1}{m_k} \sum_{i=1}^{m} \beta_i \left\|\tilde{f}_k[f_i] - \tilde{f}_{k-1}[f_i]\right\|_{L_p(\pi)}^p \\
&\leq \sum_{k=1}^{K} \frac{3^2 \epsilon_k^2}{m_k}.
\end{aligned}$$

Thus if we take $m_k = \frac{1}{a_k}(\frac{3\epsilon_k}{\epsilon_K})^2$ with $\sum a_k = 1$ we know that there must exist a choice of $\tilde{g}_j^{(k)}$s such that

$$\left\| \sum_{k=1}^{K} \frac{1}{m_k} \sum_{j=1}^{m_k} \tilde{g}_j^{(k)} - \sum_{k=1}^{K} \sum_{i=1}^{m} \beta_i \left(\tilde{f}_k[f_i] - \tilde{f}_{k-1}[f_i] \right) \right\|_{L_p(\pi)} \leq \epsilon_K.$$

This implies that finite the set $\tilde{\mathcal{C}} = \left\{ \sum_{k=1}^{K} \frac{1}{m_k} \sum_{j=1}^{m_k} \tilde{g}_j^{(k)} : \tilde{g}_j^{(k)} \in \tilde{\mathcal{F}}_k - \tilde{\mathcal{F}}_{k-1} \right\}$ is an $2\epsilon_K$ covering of $\mathcal{C} = \mathrm{Conv}\mathcal{F}$, since we know that for all $f = \sum_{i=1}^{m} \beta_i f_i$ there are $\tilde{g}_j^{(k)}$ such that

$$\begin{aligned} \left\| \sum_{k=1}^{K} \frac{1}{m_k} \sum_{j=1}^{m_k} \tilde{g}_j^{(k)} - \sum_{i=1}^{m} \beta_i f_i \right\|_{L_p(\pi)} &\leq \left\| \sum_{i=1}^{m} \beta_i \left(f_i - \tilde{f}_K[f_i] \right) \right\|_{L_p(\pi)} \\ &\quad + \sum_{k=1}^{K} \left\| \frac{1}{m_k} \sum_{j=1}^{m_k} \tilde{g}_j^{(k)} - \sum_{i=1}^{m} \beta_i \left(\tilde{f}_k[f_i] - \tilde{f}_{k-1}[f_i] \right) \right\|_{L_p(\pi)} \\ &\leq 2\epsilon_K. \end{aligned}$$

Since $\left|\tilde{\mathcal{C}}\right| = \prod_{k=1}^{K} \left|\tilde{\mathcal{F}}_k\right|^{m_k} \left|\tilde{\mathcal{F}}_{k-1}\right|^{m_k}$, we have

$$\begin{aligned} \log \mathcal{N}_p(\mathcal{C}, 2\epsilon_K) &\leq \sum_{k=1}^{K} \frac{1}{a_k} (\frac{3\epsilon_k}{\epsilon_K})^2 \left(\log \mathcal{N}_p(\mathcal{F}, \epsilon_k) + \log \mathcal{N}_p(\mathcal{F}, \epsilon_{k-1}) \right) \\ &\leq 18 \sum_{k=1}^{K} \frac{1}{a_k} 2^{2(K-k)} \log \mathcal{N}_2(\mathcal{F}, \epsilon_k). \end{aligned}$$

This is minimized for the choice

$$a_k = \frac{2^{(K-k)} \sqrt{\log \mathcal{N}_2(\mathcal{F}, \epsilon_k)}}{\sum 2^{(K-k)} \sqrt{\log \mathcal{N}_2(\mathcal{F}, \epsilon_k)}},$$

which yields the bound

$$\sqrt{\log \mathcal{N}_p(\mathcal{C}, 2\epsilon_K)} \leq \sqrt{18} \sum_{k=1}^{K} 2^{K-k} \sqrt{\log \mathcal{N}_2(\mathcal{F}, \epsilon_k)}$$

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contribution section accurately describes our contributions, and all theorems are proven in the appendix.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.

- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations of our Theorems after we state them.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All assumptions are either stated in the Theorem statements, except for a few recurring assumptions that are stated in the setup section.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The experimental setup is described in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We use openly available data or synthetic data, with a description of how to build this synthetic data.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In the experimental setup section in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The numerical experiments are mostly there as a visualization of the theoretical results, our main goal is therefore clarity, which would be hurt by putting error bars everywhere.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: In the experimental setup section of the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We have read the Code of Ethics and see no issue.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: The paper is theoretical in nature, so it has no direct societal impact that can be meaningfully discussed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Not relevant to our paper.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: In the experimental setup section of the Appendix.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not release any new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Not relevant to this paper.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Not relevant to this paper.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.