HARMONIC ANALYSIS ON THE SPHERE AND DEPTH SEPARATION FOR NEURAL NETWORKS

MIKAEL DE LA SALLE

The aim of this note (not for publication) is to give a self-contained operator-theoretic proof of a theorem in *Depth Separation for Neural Networks* by Amit Daniely, which allows to slightly improve the results.[1]

1. Daniely's theorem

If $\sigma\colon \mathbf{R}\to\mathbf{R}$ is a function, Daniely says that a function $G\colon \Omega\subset\mathbf{R}^n\to\mathbf{R}$ is a depth 2 and width r neural network with activation σ if there are affine maps $T_1\colon\mathbf{R}^n\to\mathbf{R}^r$ and $T_2\colon\mathbf{R}^r\to R$ such that $G(x)=T_2\circ\sigma_r\circ T_1(x)$ on Ω, where $\sigma_r\colon\mathbf{R}^r\to\mathbf{R}^r$ denotes the coordinate-wise map $x=(x_i)_{1\leq i\leq r}\mapsto(\sigma(x_i))_{1\leq i\leq r}$. This terminology is not completely standard, as some authors like to also compose at the end by the activation function, while other authors call these these depth 1 neural networks.

Let $\mathbf{S}^{d-1}$ denote the unit sphere in the euclidean space $\mathbf{R}^d$, equipped with the uniform probability measure that we simply denote dx. Denote also by $N_{d,n}$ the integer $\frac{(2n+d-2)(n+d-3)!}{n!(d-2)!}$. Let us define by $\mu_d=c_d(1-t^2)^{\frac{d-3}{2}}dt$ the image of dx under the map $x=(x_1,\dots,x_d)\in\mathbf{S}^{d-1}\mapsto x_1\in[-1,1]$. For $f\in L_2([-1,1],\mu_d)$ define

$$A_{d,n}(f)=\max\{\|f-p\|_{L_2([-1,1],\mu_d)}\mid p\text{ polynomial of degree }\leq n\}.$$

Theorem 1.1. *(Daniely) Let $f\in L_2([0,1],\mu_d)$ and $F\colon\mathbf{S}^{d-1}\times\mathbf{S}^{d-1}\to\mathbf{R}$ defined by $F(x,y)=f(\langle x,y\rangle)$. Then for depth 2 neural network $G\colon\mathbf{S}^{d-1}\times\mathbf{S}^{d-1}\to\mathbf{R}$ of width r,*

$$\|F-G\|_{L_2(\mathbf{S}^{d-1}\times\mathbf{S}^{d-1})}\geq\sqrt{\sum_{N_{d,n}\geq r}(A_{d,n}(f)^2-A_{d,n+1}(f)^2)\left(1-\frac{r}{N_{d,n}}\right)}$$
$$\geq\sup_{N_{d,n}\geq r}A_{d,n}(f)\sqrt{1-\frac{r}{N_{d,n}}}.$$

Let us make a few observations. Let $F\colon\mathbf{S}^{d-1}\times\mathbf{S}^{d-1}\to\mathbf{R}$ be a measurable square-integrable function.

- F is of the form $F(x,y)=f(\langle x,y\rangle)$ for some $f\colon[-1,1]\to\mathbf{R}$ if and only if F satisfies

$$F(Ux,Uy)=F(x,y)\forall U\in\mathrm{SO}(d),\forall x,y\in\mathbf{S}^{d-1}.\tag{1}$$

  In that case,

$$\|F\|_{L_2(\mathbf{S}^{d-1}\times\mathbf{S}^{d-1})}=\|f\|_{L_2([-1,1],\mu_d)}.$$

- If T_F denotes the linear map $L_2(\mathbf{S}^{d-1})\to L_2(\mathbf{S}^{d-1})$

$$(T_F\varphi)(x)=\int_{\mathbf{S}^{d-1}}F(x,y)\varphi(y)dy,\tag{2}$$

  then F satisfies (1) if and only if

$$T_F\circ\lambda_U=\lambda_U\circ T_F\forall U\in\mathrm{SO}(d),\tag{3}$$

  where λ_U is the unitary operator in $L_2(\mathbf{S}^{d-1})$ $(\lambda_U\varphi)(x)=\varphi(U^{-1}x)$.
- Moreover, if one denotes by $\|T\|_{HS}$ the Hilbert-Schmidt norm of a linear operator T on a Hilbert space $\mathcal{H}$ (see §1.1 for reminders), then we have

$$\|T_F\|_{HS}=\|F\|_{L_2(\mathbf{S}^{d-1}\times\mathbf{S}^{d-1})}.$$

Date: 13 octobre 2019.

[1] We have better bounds on r, and no hypothesis on the weights.

So Daniely's theorem is a result about the difficulty of approximating, in Hilbert-Schmidt norm, an operator commuting with all λ_U's by some special kind of operators. It is therefore usefull to understand the spectral decompositions of such operators. This is the content of the next result, which will be proved in the next Section using some very classical facts on the representation theory of $\mathrm{SO}(d)$.

Proposition 1.2. *There is an orthogonal decomposition $L_2(\mathbf{S}^{d-1}) = \oplus_{n\geq 0}\mathcal{H}_{d,n}$ such that, if P_n denotes the orthogonal projection on $\mathcal{H}_{d,n}$, then*

- *if F is as in Daniely's Theorem, the operrator T_F is of the form $T_F = \sum_n \alpha_n P_n$ where*

$$A_{d,n}(f)^2 - A_{d,n+1}(f)^2 = N_{d,n}|\alpha_n|^2. \tag{4}$$

- *If G is a depth 2 width r neural network, then P_nT_G is a rank r operator.*

Before we prove this Proposition, let us explain how it implies Daniely's Theorem. For that, we need to recall some basic facts on Hilbert-Schmidt operators (see for example [2] for a standard reference).

1.1. Background on linear algebra.

Let A be a linear map on a Hilbert space $\mathcal{H}$. The quantity

$$\sum_n \|Ae_n\|^2$$

does not depend on the orthonormal basis $(e_n)_n$. Indeed, if (f_n) is another orthonormal basis of $\mathcal{H}$, we can write by the Plancherel formula and the definition of the adjoint, $\|Ae_n\|^2 = \sum_m |\langle Ae_n, f_m\rangle|_n^2 = \sum_m |\langle e_n, A^* f_n\rangle|^2$, so we have

$$\sum_n \|Ae_n\|^2 = \sum_m \|A^* f_m\|^2.$$

The right-hand side does not depend on $(e_n)_n$, so neither does the left-hand side. When this quantity is finite, we say that A is a *Hilbert-Schmidt operator* and denote $\|A\|_{HS}$ the square root of $\sum_n \|A_n e_n\|^2$.

Remark 1.3. Finite rank operators are Hilbert-Schmidt.

If $\mathcal{H} = L_2(X,\mu)$, an operator A is Hilbert-Schmidt if and only if there is a function $f \in L_2(X\times X, \mu\otimes\mu)$ such that

$$A\varphi(x) = \int f(x,y)\varphi(y)d\mu(y)$$

for every $\varphi \in L_2(X,\mu)$. In that case, $\|A\|_{HS} = \|f\|_{L_2}$. Indeed, if (e_n) is an orthonormal basis of $L_2(X,\mu)$, then

$$\|A\|_{HS}^2 = \sum_n \|Ae_n\|^2 = \sum_n\sum_m |\langle Ae_n, e_m\rangle|^2 = \sum_n\sum_m |\langle f, e_n\otimes e_m\rangle|^2 = \|f\|_{L_2}^2$$

where on the last line we used that $(e_n\otimes e_m)_{n,m}$ is an orthonormal basis of $L_2(X\times X, \mu\otimes\mu)$.

Lemma 1.4. *Let $\mathcal{H}$ be a Hilbert space and P be a rank N orthogonal projection. Then for every operator B of rank r, we have*

$$\|P - B\|_{HS} \geq \sqrt{\max(N-r, 0)}.$$

Proof. We assume $N > r$. We are free to choose an orthonormal basis to compute the Hilbert-Schmidt norm of $P - B$. In particular we can choose one which contains $N-r$ unit vectors $e_1, \ldots, e_{N-r}$ in the image of P and in the kernel of B, because the intersection of a space of dimension N – the image of P – with a space of codimension r – the image of B – has dimension $\geq N-r$. In that case we have

$$\|P-B\|_{HS}^2 \geq \sum_{j=1}^{N-r} \|Pe_j - Be_j\|^2 = \sum_{j=1}^{N-r} \|e_j\|^2 = N - r.$$

□

1.2. **Proof of Daniely's theorem.** It follows from Proposition 1.2 that the operator T_F defined in (2) can be written as $T_F = \sum_n \alpha_n P_n$ with complex numbers α_n satisfying (4), and that $P_n T_G$ has rank $\leq r$. So we have

$$\begin{aligned}\|F-G\|_2^2 &= \|T_F - T_G\|_{HS}^2\\ &= \sum_n \|P_n T_F - \sum_{j=1}^r P_n B_j\|_{HS}^2\\ &= \sum_n |\alpha_n|^2 \|P_n - \sum_{j=1}^r \frac{1}{\alpha_n} P_n B_j\|_{HS}^2\\ &\geq \sum_n |\alpha_n|^2 \max((N_{d,n}-r),0),\end{aligned}$$

where on the last line we used Lemma 1.4. The Theorem follows using (4).

2. Background on spherical analysis

There are plenty of ways of proving Proposition 1.2. We make the choice to give a presentation based on the representation theory of $\mathrm{SO}(d)$.

2.1. **Generalities.** A *unitary representation* $(\pi, \mathcal{H})$ of a topological group G is a map $\pi\colon G \to U(\mathcal{H})$ from G to the unitary operators on a Hilbert space $\mathcal{H}$ satisfying

- $\pi(gg') = \pi(g) \circ \pi(g')$ for every $g, g' \in G$, and
- $\lim_{g\to 1_G} \|\pi(g)\xi - \xi\| = 0$ for every $\xi \in \mathcal{H}$.

For example, the map $U \mapsto \lambda_U$ of the preceding section is a unitary representation of $\mathrm{SO}(d)$ on $L_2(\mathbf{S}^{d-1})$.

Recall that a closed subspace $\mathcal{K} \subset \mathcal{H}$ is called *a subrepresentation* if $\pi(g)\mathcal{H} = \mathcal{H}$ for every $g \in G$. A representation $(\pi, \mathcal{H})$ is called irreducible if $\dim(\mathcal{H}) > 0$ and if the only suprepresentations $\mathcal{K} \subset \mathcal{H}$ are $\mathcal{K} = \mathcal{H}$ and $\mathcal{K} = 0$.

Peter-Weyl's theorem is a general abstract result stating that any unitary representation of a compact group decomposes as a direct sum of irreducible subrepresentations. The next Proposition describes this decomposition explicitely for the representation λ on $L_2(\mathbf{S}^{d-1})$.

Proposition 2.1. *There is an orthogonal decomposition $L_2(\mathbf{S}^{d-1}) = \oplus_{n\geq 0}\mathcal{H}_{d,n}$ such that*

- *Each $\mathcal{H}_{d,n}$ is an irreducible subrepresentation.*
- *For each n, the space of functions $\varphi \in \mathcal{H}_{d,n}$ satisfying $\lambda_U\varphi = \varphi$ for every $U \in \begin{pmatrix}1 & 0\\ 0 & \mathrm{SO}(d-1)\end{pmatrix}$ is one dimensional (and consists of polynomials of degree n in x_1).*
- $\dim(\mathcal{H}_{d,n}) = \frac{(d+n-3)!(d+2n-2)}{(d-2)!n!} =: N_{d,n}$.

What is crucial for the application to Machine Learning is that the dimension of $\mathcal{H}_{d,n}$ grows with n, faster and faster as d is large. Note also that the dimension of $\mathcal{H}_1$ is d, so grows also with d. I do not know how meaningful this is here, but the general phenomenon of having groups whose second smallest irreducible representation is large is a central phenomenon in modern mathematics (called *quasirandomness* by Gowers), see for example the intense work around the work of Sarnak and Xue, Bourgain and Gamburd, Gowers etc on expansion.

Schur's Lemma is a general fact on representation theory, which states that the space of linear equivariant maps between irreducible representations are 0 or 1-dimensional, being 1-dimensional if and only if the two representations are isomorphic. In our specific setting, this becomes.

Lemma 2.2 (Schur's Lemma)**.** *If an operator T on $L_2(\mathbf{S}^{d-1})$ commutes with $\lambda(k)$ for every k, then there is a bounded sequence $\alpha_n \in \mathbf{C}$ such that $T = \sum_n \alpha_n P_n$ where P_n is the orthogonal projection on $\mathcal{H}_{d,n}$.*

Observe that Proposition 1.2 is a rather direct consequence of the Peter-Weyl Theorem and of Schur's Lemma. Indeed, if F is as in Daniely's Theorem, we have already explained that the operator T_F commutes with λ_U, so the existence of a decomposition $T_F = \sum_n \alpha_n P_n$ is Schur's Lemma. The formula (4) holds because, by the second bullet point in Proposition 2.1, $\sum_{n\leq n_0} \alpha_n P_n$ corresponds to $T_{F_{n_0}}$, where $F_{n_0}(x,y) = f_{n_0}(\langle x,y\rangle)$ for f_{n_0} the orthogonal projection (in $L_2([-1,1],\mu_d)$) of f on the space of polynomials of degree $\leq$

n_0. So $A_{d,n}(f)$ is the Hilbert-Schmidt norm of $\sum_{n>n_0} \alpha_n P_n$, that is $(\sum_{n\geq n_0} |\alpha_n|^2 N_{d,n})^{\frac{1}{2}}$. This is exactly (4). For the statement about T_G, observe that the assumption on G implies that there exist $x_1,\ldots,x_r,y_1,\ldots,y_r \in \mathbf{S}^{d-1}$ and functions $g_1,\ldots,g_r\colon \mathbf{R}^2 \to \mathbf{C}$ such that $G(x,y) = \sum_{j=1}^r g_j(\langle x,x_j\rangle,\langle x,y_j\rangle)$ (an actually $g_j(r,s)$ is of the form $\lambda_j\sigma(a_jr+b_js+c_j)$ for scalars λ_j,a_j,b_j,c_j, but we will not need that). So we can write $T_G = \sum_{j=1}^r B_j$ for B_j the operator

$$B_j\varphi(x) = \int g_j(\langle x,x_j\rangle,\langle x,y_j\rangle)\varphi(y)dy.$$

Oberve that $B_j\varphi(x)$ depends only on $\langle x,x_j\rangle$, so it is enough to prove that for such an operator B_j, P_nB_j has rank ≤ 1. By replacing B_j by $\lambda_U B_j$ for $U \in \mathrm{SO}(d)$ satisfying $Uv = e_1$, we can assume $v = e_1$ (because $P_nB_j = \lambda_{U^{-1}}P_n\lambda_U B_j$). Then for every $U \in \begin{pmatrix}1 & 0\\ 0 & \mathrm{SO}(d-1)\end{pmatrix}$, $\lambda_U P_nB_j = P_n\lambda_U B_j = P_nB_j$ because a function on $\mathbf{S}^{d-1}$ depending only on x_1 is U-invariant. So the image of P_nB_j is contained in the space of $\begin{pmatrix}1 & 0\\ 0 & \mathrm{SO}(d-1)\end{pmatrix}$-invariant elements of $\mathcal{H}_{d,n}$, which is one-dimensional by the Proposition. This proves that P_nB_j has rank one, which concludes the proof of Proposition 1.2.

2.2. **Proof of Proposition 2.1.** Proposition 2.1 is extremely classical, and is covered is numerous textbooks, for example [1]. We give a proof for the audience's convenience.

Define $\mathcal{P}_{d,n}$ the space of complex polynomials in d variables which are homogeneous of degree n. A basis of $\mathcal{P}_{d,n}$ consists of the monomials $x^\alpha = \prod_{i=1}^d x_i^{\alpha_i}$ for $\alpha \in \mathbf{N}^n$, $\sum_{i=1}^d \alpha_i = n$, so $\mathcal{P}_{d,n}$ is a complex vector space of dimension $\binom{n+d-1}{n}$.

We define $\mathcal{H}_{d,n} \subset \mathcal{P}_{d,n}$ the subspace of harmonic polynomials ($\Delta P = 0$), that we see as a subspace of $L_2(\mathbf{S}^{d-1})$. It is easy to see that Δ maps $\mathcal{P}_{d,n}$ onto $\mathcal{P}_{d,n-2}$, so

$$\dim(\mathcal{H}_{d,n}) = \binom{n+d-1}{n} - \binom{n+d-3}{n-2} = N_{d,n}.$$

It is clear that each $\mathcal{H}_{d,n}$ is a subrepresentation of $(\lambda, L_2(\mathbf{S}^{d-1}))$, because Δ commutes with λ_U for every U.

Lemma 2.3. *For each n, $\mathcal{H}_{d,n}$ (and every nonzero subrepresentation of $\mathcal{H}_{d,n}$ contains a nonzero $\begin{pmatrix}1 & 0\\ 0 & \mathrm{SO}(d-1)\end{pmatrix}$-invariant function.*

Proof. Let $\mathcal{K} \subset \mathcal{H}_{d,n}$ be a nonzero subrepresentation. Let $\varphi_0 \in \mathcal{K}$ be nonzero. There is $x_0 \in \mathbf{S}^{d-1}$ such that $\varphi_0(x_0) \neq 0$. By replacing φ_0 by $\lambda_U\varphi_0$ for some U satisfying $U^{-1}e_1 = x_0$, we can assume that $x = e_1$. Then $\varphi(x) := \int_{\mathrm{SO}(d-1)} \varphi_0(\begin{pmatrix}1 & 0\\ 0 & U\end{pmatrix} x)dU = \int_{\begin{pmatrix}1 & 0\\ 0 & \mathrm{SO}(d-1)\end{pmatrix}} \lambda_U\varphi(x)dU$ (integral with respect to the Haar measure on $\mathrm{SO}(d-1)$) belongs to $\mathcal{K}$ and is $\begin{pmatrix}1 & 0\\ 0 & \mathrm{SO}(d-1)\end{pmatrix}$-invariant. φ is nonzero because its value at e_1 is $\varphi_0(e_1) \neq 0$. □

To prove the uniqueness, we will need the following.

Lemma 2.4. *Every element of $\mathcal{P}_{d,n}$ can be written uniquely as*

$$P = \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} (x_1^2 + \cdots + x_d^2)^{2i} P_i$$

for $P_i \in \mathcal{H}_{d,n-2i}$.

Proof. We claim that P can be written uniquely as

$$P = P_0 + (x_1^2 + \cdots + x_d^2)Q$$

for $P_0 \in \mathcal{H}_{d,n}$ and $Q \in \mathcal{P}_{d,n-2}$. The Lemma then follows by an induction argument. We have to show that the map $(P_0,Q) \mapsto P_0 + (x_1^2 + \cdots + x_d^2)Q$ is an isomorphism. By dimension counting, it is enough to show that this map is injective. Let (P_0, Q) belong to its kernel. Then if $P_0 = \sum_{|\alpha|=n} a_\alpha x^\alpha$, we have

$$\sum_{|\alpha|=n} \alpha! |a_\alpha|^2 = \overline{P_0}(\frac{\partial}{\partial x}) P_0 = -\overline{Q}(\frac{\partial}{\partial x}) \Delta(P_0) = 0,$$

so $P_0 = 0$ and $Q = 0$. □

Proposition 2.1 easily follows from this lemma.

Step 1. For each n the space of $\begin{pmatrix} 1 & 0 \\ 0 & \mathrm{SO}(d-1) \end{pmatrix}$-invariant functions in $\mathcal{H}_{d,n}$ is one dimensional. We have already proved that it is at least one dimensional. If its dimension was strictly larger, Lemma 2.4 would imply that the space of degree n homogeneous polynomials depending only on $\langle x, v\rangle$ would be of dimension $> 2 + \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} 1 = \lfloor \frac{n}{2} \rfloor + 1$, which is absurd (as it coincides with the space of degree n polynomials with the same parity as x^n, which is of dimension $\lfloor \frac{n}{2} \rfloor$).

Step 2. $\mathcal{H}_{d,n}$ is an irreducible representation. Indeed, if φ is a nonzero $\begin{pmatrix} 1 & 0 \\ 0 & \mathrm{SO}(d-1) \end{pmatrix}$-invariant element of $\mathcal{H}_{d,n}$, then Step 1 and Lemma 2.3 imply every nonzero subrepresentation of $\mathcal{H}_{d,n}$ contains φ, and therefore contains $\mathcal{K}$, the smallest subrepresentation containing φ. The orthogonal of $\mathcal{K}$ in $\mathcal{H}_{d,n}$ is a subrepresentation not containing φ, so is the zero representation, *i.e.* $\mathcal{K} = \mathcal{H}_{d,n}$. This proves that $\mathcal{H}_{d,n}$ is irreducible.

Step 3. There is an orthogonal decomposition $L_2(\mathbf{S}^{d-1}) = \oplus_{n \geq 0} \mathcal{H}_{d,n}$. Indeed, by irreducibility of $\mathcal{H}_{d,n}$, we have that $\mathcal{H}_{d,n} \cap \mathcal{H}_{d,n'} = \{0\}$ if $n \neq n'$. The fact that $\mathcal{H}_{d,n}$ and $\mathcal{H}_{d,n'}$ are orthogonal follows from Schur's Lemma : if P denotes the first coordinate projection $\mathcal{H}_{d,n} \oplus \mathcal{H}_{d,n'} \to \mathcal{H}_{d,n}$, then the fact that P commutes with λ_U implies that PP^* is a linear map on $\mathcal{H}_{d,n}$ commuting with λ_U, so by Schur's Lemma P^*P is a multiple of the identity. Necessarily, P^*P is the identity on $\mathcal{H}_{d,n}$, which implies that the decomposition is orthogonal. By Lemma 2.4, $\oplus_{n \leq N} \mathcal{H}_{d,n}$ coincides with the restrictions to $\mathbf{S}^{d-1}$ of the polynomials of degree $\leq N$. So the density of $\oplus_n \mathcal{H}_{d,n}$ in $L_2(\mathbf{S}^{d-1})$ follows from the density of the polynomials in the space of continuous functions on $\mathbf{S}^{d-1}$.

References

[1] Jacques Faraut. *Analysis on Lie groups*, volume 110 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2008. An introduction.

[2] Barry Simon. *Trace ideals and their applications*, volume 120 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, second edition, 2005.