hexagon:2611.00249v1

PROPAGATION AND OPTIMIZATION OF LOW-FREQUENCY SPECTRAL ALIGNMENT IN RELU NETWORKS ON A FIXED SAMPLE GRAPH

CAMERON RYAN

Abstract. We study how low-frequency structure, measured with respect to the Laplacian of a fixed graph built on the training samples, propagates through a ReLU network to the least-squares readout g, and how individual neurons $h = S\Phi\theta$ can be optimized to capture low-frequency eigenvectors. The readout error splits into a low-frequency part controlled by *coverage* (how well the feature span contains the low eigenvectors) and a high-frequency part controlled by the *smoothness* (Dirichlet energy) of the target. We prove an error bound that makes this precise, with coverage measured by the coverage gap, the largest principal-angle sine between the low eigenvectors and the feature span, and show that coverage can always be preserved across ReLU layers. Finally, we define a count N_K of captured low-frequency features, relate it to our error bounds, and prove that error can be reduced to a smoothness component when N_K captures every low frequency feature. Lastly, we show that a ReLU layer can strictly increase N_K whenever the previous layer keeps the samples distinct and coverage is not yet full, that $2K + 2$ neurons suffice to preserve coverage, that greedily chosen neurons drive the coverage gap down at rate $O(1/t)$ governed by a ReLU variation norm of the low eigenvectors, and that t neurons can reduce the coverage gap no further than the $(t + 1)$-st principal-angle sine.

1. Setting and notation

Let $X \in \mathbb{R}^{d\times n}$ be the sample and $f \in \mathbb{R}^n$ the target values. Let $\mathcal{G}(X)$ be a weighted undirected graph on the n samples with symmetric nonnegative weights w_{ij}, degrees $d_i = \sum_j w_{ij} > 0$, $D = \mathrm{diag}(d_1, \dots, d_n)$ and $W = (w_{ij})$. We consider the use of Laplacians such as the combinatorial or symmetric normalized:

$$L = D - W, \qquad \mathcal{L}_{\mathrm{sym}} = D^{-1/2} L D^{-1/2}$$

see [2, 3]. Their quadratic forms are

$$x^\top L x = \sum_{i<j} w_{ij}(x_i - x_j)^2, \qquad x^\top \mathcal{L}_{\mathrm{sym}} x = \sum_{i<j} w_{ij}\Big(\frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}}\Big)^2. \tag{1}$$

Throughout, $\mathcal{L}$ denotes either L or $\mathcal{L}_{\mathrm{sym}}$ unless stated otherwise. Write $\mathcal{L} = U\Lambda U^\top$ with $U = [u_1, \dots, u_n]$ orthogonal and $0 = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$. Let $U_K = [u_1, \dots, u_K]$, $P_K = U_K U_K^\top$ and $P_K^\perp = I - P_K$.

Definition 1.1 (Energy and Rayleigh quotient)**.** The *Dirichlet energy* of $x \in \mathbb{R}^n$ is $E(x)^2 = x^\top \mathcal{L} x$, and $E(x) = \sqrt{x^\top \mathcal{L} x} = \|\mathcal{L}^{1/2} x\|$ is the *Dirichlet energy seminorm*. For $x \neq 0$ the *Rayleigh quotient* is

$$R(x) = \frac{E(x)^2}{\|x\|^2} = \frac{x^\top \mathcal{L} x}{x^\top x} = \frac{\sum_k \lambda_k \hat{x}_k^2}{\sum_k \hat{x}_k^2}, \qquad \hat{x} = U^\top x,$$

the average eigenvalue weighted by the spectral mass of x, so that $\lambda_1 \le R(x) \le \lambda_n$, with $R(u_k) = \lambda_k$.

Date: November 9, 2026.

Network. Let $\sigma(t) = \max(t, 0)$ act entrywise. Set $H_0 = X^\top \in \mathbb{R}^{n\times d}$ and, for $\ell = 1, \dots, p$,

$$H_\ell = \sigma\big(H_{\ell-1}W_\ell + \mathbf{1}b_\ell^\top\big), \qquad W_\ell \in \mathbb{R}^{m_{\ell-1}\times m_\ell},\ b_\ell \in \mathbb{R}^{m_\ell},\ m_0 = d.$$

The readout features are $Z = [H_p, \mathbf{1}]$, $\beta = Z^+ f$, $g = Z\beta = \Pi_Z f$ where Π_Z is the orthogonal projector onto $\mathrm{col}(Z)$, and the residual is $r = g - f$.

Single neuron. For a neuron in layer ℓ let $\Phi = [H_{\ell-1}, \mathbf{1}] \in \mathbb{R}^{n\times m}$, $\theta \in \mathbb{R}^m$, $z = \Phi\theta$ the pre-activation and

$$S = S(\theta) = \mathrm{diag}\left(\mathbb{1}[(\Phi\theta)_i > 0]\right), \qquad h = \sigma(z) = S\Phi\theta. \tag{2}$$

For vectors and matrices, $|\cdot|$ is the entrywise absolute value and inequalities between vectors are entrywise.

Coverage. For a matrix Z define the *coverage gap*

$$s_K(Z) = \|(I - \Pi_Z)U_K\|_2 \in [0, 1].$$

When $\mathrm{rank}\, Z \geq K$, $s_K(Z)$ is the sine of the largest principal angle between $\mathrm{span}(U_K)$ and its closest K-dimensional subspace of $\mathrm{col}(Z)$ [6, 7, 5].

2. Preliminaries

Lemma 2.1 (High-frequency mass)**.** *If $\lambda_{K+1} > 0$ then for every $x \in \mathbb{R}^n$, $\|P_K^\perp x\|^2 \leq x^\top \mathcal{L}x/\lambda_{K+1}$.*

Proof. With $\hat{x} = U^\top x$, Parseval gives $\|P_K^\perp x\|^2 = \sum_{k>K} \hat{x}_k^2 \leq \lambda_{K+1}^{-1} \sum_{k>K} \lambda_k \hat{x}_k^2 \leq \lambda_{K+1}^{-1} \sum_k \lambda_k \hat{x}_k^2 = x^\top \mathcal{L}x/\lambda_{K+1}$. □

3. Coverage of the low frequencies

The least-squares error splits into a low-frequency part, controlled by how well $\mathrm{col}(Z)$ contains $\mathrm{span}(U_K)$ as measured by the coverage gap $s_K(Z)$, and a high-frequency part, controlled by the smoothness of f.

Corollary 3.1 (Target-only coverage bound)**.** *Let $s = s_K(Z)$. Then*

$$\|g - f\| \leq s\|P_K f\| + \|P_K^\perp f\| \leq s\|f\| + \frac{E(f)}{\sqrt{\lambda_{K+1}}},$$

the last inequality requiring $\lambda_{K+1} > 0$.

Proof. Since $r = -(I - \Pi_Z)f$ and $f = P_K f + P_K^\perp f$, the triangle inequality gives $\|r\| \leq \|(I - \Pi_Z)U_K U_K^\top f\| + \|(I - \Pi_Z)P_K^\perp f\| \leq s\|U_K^\top f\| + \|P_K^\perp f\|$, and $\|U_K^\top f\| = \|P_K f\|$. The second inequality uses $\|P_K f\| \leq \|f\|$ and Lemma 2.1. □

Lemma 3.2 (Monotonicity of the coverage gap)**.** *If $\mathrm{col}(A) \subseteq \mathrm{col}(B)$ then $s_K(B) \leq s_K(A)$.*

Proof. $\Pi_B\Pi_A = \Pi_A$, so $(I - \Pi_B) = (I - \Pi_B)(I - \Pi_A)$ and $\|(I - \Pi_B)U_K\|_2 \leq \|I - \Pi_B\|_2\|(I - \Pi_A)U_K\|_2 \leq s_K(A)$. □

Proposition 3.3 (Coverage can always be preserved across ReLU layers)**.**

- *(a) For every $z \in \mathbb{R}^n$, $z = \sigma(z) - \sigma(-z)$; hence $z \in \mathrm{span}\{\sigma(z), \sigma(-z)\}$.*
- *(b) For $\ell \geq 2$, the entries of $H_{\ell-1}$ are nonnegative, so a layer containing the $m_{\ell-1}$ neurons with weights e_k and zero bias reproduces $H_{\ell-1}$ exactly, and $\mathrm{col}([H_\ell, \mathbf{1}]) \supseteq \mathrm{col}([H_{\ell-1}, \mathbf{1}])$.*
- *(c) For $\ell = 1$, a layer containing the $2d$ neurons with weights $\pm e_k$ and zero bias satisfies $\mathrm{col}([H_1, \mathbf{1}]) \supseteq \mathrm{col}([X^\top, \mathbf{1}])$.*

Consequently, if $m_1 \geq 2d$ and $m_\ell \geq m_{\ell-1}$ for $\ell \geq 2$, there exist weights for which $s_K([H_\ell, \mathbf{1}])$ is nonincreasing in ℓ, with the remaining neurons free to add coverage (Section 5 shows when and how fast they do).

Proof. (a) holds entrywise: $\max(t,0)-\max(-t,0)=t$. (b) $\sigma(H_{\ell-1}e_k)=H_{\ell-1}e_k$ since $H_{\ell-1}\geq 0$. (c) By (a), $X^\top e_k=\sigma(X^\top e_k)-\sigma(-X^\top e_k)$ lies in $\mathrm{col}(H_1)$. The consequence follows from Lemma 3.2. □

Definition 3.4 (Captured low-frequency features)**.** For $H\in\mathbb{R}^{n\times N}$ with orthonormal basis Q_H of $\mathrm{col}(H)$, define

$$N_K(H)=\mathrm{tr}(\Pi_H P_K)=\|U_K^\top Q_H\|_F^2.$$

Lemma 3.5 (Properties of N_K)**.**

(i) $N_K(H)=\sum_i \cos^2\theta_i$*, where* θ_i *are the principal angles between* $\mathrm{col}(H)$ *and* $\mathrm{span}(U_K)$*.*
(ii) $0\leq N_K(H)\leq \min(K,\mathrm{rank}\,H)$*.*
(iii) For a single nonzero vector, $N_K(h)=\|P_K h\|^2/\|h\|^2\leq 1$*.*
(iv) $\|(I-\Pi_H)U_K\|_F^2=K-N_K(H)$*, so* $s_K(H)^2\leq K-N_K(H)$ *and* $s_K(H)^2\geq (K-N_K(H))/K$*.*

Proof. (i) The singular values of $U_K^\top Q_H$ are the cosines of the principal angles [6]. (ii) There are $\min(K,\mathrm{rank}\,H)$ principal angles and each $\cos^2\theta_i\in[0,1]$. (iii) Take $Q_H=h/\|h\|$. (iv) $\|\Pi_H U_K\|_F^2=\mathrm{tr}(U_K^\top\Pi_H U_K)=N_K(H)$ and $\|U_K\|_F^2=K$, so Pythagoras gives the identity; then $\|\cdot\|_F^2/K\leq\|\cdot\|_2^2\leq\|\cdot\|_F^2$. □

4. Optimizing neurons to capture low frequencies

By Lemma 3.5(iii) a single neuron captures at most one low-frequency dimension; capturing many is a property of a layer. We first bound what a single gated neuron can achieve, then give a fixed-gating formulation for a single neuron and for the marginal gain of adding a neuron to a layer.

Proposition 4.1 (Limits imposed by the active set)**.** *Let* $h\neq 0$ *be supported on* $A\subseteq\{1,\dots,n\}$ *(for a ReLU neuron, its active set) and let* B *be the complement. Let* $\ell_i=\|U_K^\top e_i\|^2$ *be the leverage scores of* U_K [9]*. Then*

$$N_K(h)\leq\|U_K[A,:]\|_2^2\leq\min\Big(1,\sum_{i\in A}\ell_i\Big),\qquad R(h)\geq\lambda_{\min}(\mathcal{L}_{AA}),$$

where $\mathcal{L}_{AA}$ *is the principal submatrix on* A*. Moreover* $\lambda_{\min}(\mathcal{L}_{AA})\leq\lambda_{|B|+1}$*.*

Proof. Write h_A for the restriction of h to A. Then $\|P_K h\|^2=\|U_K[A,:]^\top h_A\|^2\leq\|U_K[A,:]\|_2^2\|h\|^2$. The spectral norm is at most 1 because U_K has orthonormal columns, and at most the Frobenius norm, whose square is $\sum_{i\in A}\ell_i$. Next, $h^\top\mathcal{L}h=h_A^\top\mathcal{L}_{AA}h_A\geq\lambda_{\min}(\mathcal{L}_{AA})\|h\|^2$. Finally, Cauchy interlacing [4] for an $|A|\times|A|$ principal submatrix gives $\lambda_1(\mathcal{L}_{AA})\leq\lambda_{1+n-|A|}(\mathcal{L})=\lambda_{|B|+1}$. □

Remark 4.2. Since $\sum_i\ell_i=K$, for delocalized eigenvectors $\ell_i\approx K/n$ and a neuron active on $|A|$ vertices captures at most about $K|A|/n$. Small active sets cap low-frequency capture, formalizing why dead or very sparse ReLUs hurt coverage. The bound $R(h)\geq\lambda_{\min}(\mathcal{L}_{AA})$ is a Dirichlet eigenvalue: the inactive set B acts as a boundary on which $h(B)=0$, in the spirit of the harmonic-extension theorem [1, Thm. 11.6.1]. Conversely, by Proposition 3.3(a), the pair $\sigma(u_k),\sigma(-u_k)$, supported on the positive and negative nodal domains of u_k, recovers u_k; discrete nodal domain theorems [13] bound how many such domains u_k has.

Proposition 4.3 (Fixed-gating formulation for one neuron)**.** *Fix a* $0/1$ *diagonal matrix* S *and let* $C_S=\{\theta:(2S-I)\Phi\theta\geq 0\}$*, a polyhedral cone. For* $\theta\in C_S$ *we have* $\sigma(\Phi\theta)=S\Phi\theta$*, and if* $S\Phi\theta\neq 0$*,*

$$N_K\big(\sigma(\Phi\theta)\big)=\frac{\theta^\top A_S\theta}{\theta^\top B_S\theta},\qquad A_S=\Phi^\top S P_K S\Phi,\quad B_S=\Phi^\top S\Phi.$$

Moreover

$$\sup_{\theta\in C_S,\ S\Phi\theta\neq 0}N_K\big(\sigma(\Phi\theta)\big)\ \leq\ \|U_K^\top Q_S\|_2^2=\cos^2\theta_{\min}\big(\mathrm{col}(S\Phi),\mathrm{span}(U_K)\big),$$

where Q_S is an orthonormal basis of $\operatorname{col}(S\Phi)$*, with equality if a maximizer of the unconstrained ratio lies in C_S. The global problem* $\max_\theta N_K(\sigma(\Phi\theta))$ *equals the maximum of the cone-constrained problems over the finitely many gating patterns realized by* Φ*.*

Proof. If $S_{ii} = 1$ and $(\Phi\theta)_i \geq 0$ then $\sigma(\Phi\theta)_i = (\Phi\theta)_i = (S\Phi\theta)_i$; if $S_{ii} = 0$ and $(\Phi\theta)_i \leq 0$ then $\sigma(\Phi\theta)_i = 0 = (S\Phi\theta)_i$. So on C_S the neuron is the linear map $\theta \mapsto S\Phi\theta$, and Lemma 3.5(iii) with $S^2 = S = S^\top$ gives the ratio. Dropping the cone constraint, the ratio ranges over $\|P_K v\|^2/\|v\|^2$ for nonzero $v \in \operatorname{col}(S\Phi)$, whose maximum is $\|U_K^\top Q_S\|_2^2$, the squared cosine of the smallest principal angle [6]. Every θ lies in $C_{S(\theta)}$, which proves the last claim; the number of realized patterns is finite and, for fixed $\operatorname{rank}\Phi$, polynomial in n [10, 11]. □

Remark 4.4. A maximizer $v^\star = Q_S y^\star$ of the unconstrained ratio ($y^\star$ a top right singular vector of $U_K^\top Q_S$) is realizable in C_S iff, for one choice of sign, $\pm v_A^\star \geq 0$ and the linear program $\{\theta : \Phi_A\theta = \pm v_A^\star,\ \Phi_B\theta \leq 0\}$ is feasible, where A, B are the active and inactive index sets of S. Maximizing a Rayleigh quotient over a cone is a cone-constrained (Pareto) eigenvalue problem [12]. The cone C_S and the enumeration over patterns are the same objects as in the convex reformulation of two-layer ReLU networks [11].

Proposition 4.5 (Marginal gain of a new neuron)**.** *Let H have orthonormal basis Q, let $h \notin \operatorname{col}(H)$, and set $\tilde{h} = (I - \Pi_H)h$ and $M = (I - \Pi_H)U_K$. Then*

$$N_K([H, h]) - N_K(H) = \frac{\|U_K^\top \tilde{h}\|^2}{\|\tilde{h}\|^2} = \frac{\|M^\top h\|^2}{\|(I - \Pi_H)h\|^2}.$$

For fixed gating S and $\theta \in C_S$ this is the ratio $\theta^\top\Phi^\top SMM^\top S\Phi\theta \,/\, \theta^\top\Phi^\top S(I - \Pi_H)S\Phi\theta$, so Proposition 4.3 applies verbatim with these matrices.

Proof. Gram–Schmidt gives the orthonormal basis $[Q, q]$ of $\operatorname{col}([H, h])$ with $q = \tilde{h}/\|\tilde{h}\|$, so

$$N_K([H, h]) = \|U_K^\top Q\|_F^2 + \|U_K^\top q\|^2.$$

Finally $U_K^\top \tilde{h} = U_K^\top (I - \Pi_H)h = M^\top h$. □

5. Increasing coverage across layers

Proposition 3.3 shows that coverage need not decrease from one layer to the next. This section asks when, and how fast, the remaining neurons of a layer can *increase* it. Fix a layer ℓ and $\Phi = [H_{\ell-1}, \mathbf{1}]$. For a feature matrix H built in this layer write

$$M_H = (I - \Pi_H)U_K, \qquad \delta(H) = K - N_K(H) = \|M_H\|_F^2, \qquad G_H(h) = \frac{\|M_H^\top h\|^2}{\|(I - \Pi_H)h\|^2},$$

so that $\delta(H)$ is the coverage deficit (Lemma 3.5(iv)) and, by Proposition 4.5, $G_H(h)$ is the increase of N_K obtained by appending $h \notin \operatorname{col}(H)$. Since $G_H(h) = \|M_H^\top \hat{y}\|^2$ for the unit vector $\hat{y} = (I - \Pi_H)h/\|(I - \Pi_H)h\|$, we always have $G_H(h) \leq \|M_H\|_2^2 = s_K(H)^2$. We also use that N_K is monotone: if $\operatorname{col}(A) \subseteq \operatorname{col}(B)$ then $\Pi_B - \Pi_A \succeq 0$, so $N_K(B) - N_K(A) = \operatorname{tr}\big((\Pi_B - \Pi_A)P_K\big) \geq 0$.

5.1. When can a layer increase coverage?

What a layer can add is determined by which samples it can still tell apart.

Theorem 5.1 (Span of ReLU features)**.** *Let $\mathcal{P}$ be the partition of $\{1, \ldots, n\}$ into classes of samples with identical rows in $H_{\ell-1}$, let q be the number of classes, and let $V_\mathcal{P} \subseteq \mathbb{R}^n$ be the q-dimensional subspace of vectors that are constant on each class, with orthogonal projector $\Pi_\mathcal{P}$. Then*

(a) $\sigma(\Phi\theta) \in V_\mathcal{P}$ for every θ;

(b) $\operatorname{span}\{\sigma(\Phi\theta) : \theta \in \mathbb{R}^m\} = V_\mathcal{P}$. In fact q neurons sharing one weight direction suffice.

In particular, if the rows of $H_{\ell-1}$ are pairwise distinct, ReLU features of layer ℓ span $\mathbb{R}^n$.

Proof. (a) Samples with identical rows receive identical pre-activations. (b) Choose one representative per class. Their rows are distinct, so a generic $w \in \mathbb{R}^{m_{\ell-1}}$ (one avoiding the finitely many hyperplanes on which two representatives have equal projection) gives distinct projections, which we order as $t_1 < \cdots < t_q$. Choose $b_1 < t_1$ and $b_k \in (t_{k-1}, t_k)$ for $k \geq 2$, and let $v_k = \sigma(H_{\ell-1}w - b_k\mathbf{1})$, the neuron with $\theta = (w, -b_k)$. On class c, v_k equals $(t_c - b_k)_+$, which is 0 for $c < k$ and positive for $c \geq k$. In the basis of class indicators, $v_1, \ldots, v_q$ therefore form a triangular matrix with positive diagonal, so they are linearly independent and span the q-dimensional space $V_{\mathcal{P}}$. This is the interpolation construction of [14]. □

Corollary 5.2 (Coverage ceiling)**.** *Samples with identical rows in $H_{\ell-1}$ have identical rows in every later layer. Hence the readout features of every network that agrees with the given one up to layer $\ell - 1$ satisfy* $\operatorname{col}(Z) \subseteq V_{\mathcal{P}}$*, and*

$$N_K(Z) \leq N_K(V_{\mathcal{P}}) = K - \|(I - \Pi_{\mathcal{P}})U_K\|_F^2.$$

If the rows of $H_{\ell-1}$ are distinct then $V_{\mathcal{P}} = \mathbb{R}^n$ and the ceiling is K. Otherwise, any variation of the low eigenvectors within a collapsed class is lost for all subsequent layers.

Proof. Identical inputs to a layer give identical outputs, so by induction every column of every later $H_{\ell'}$, and the bias column, lies in $V_{\mathcal{P}}$. Monotonicity of N_K gives the inequality, and Lemma 3.5(iv) applied to $V_{\mathcal{P}}$ gives the identity. □

Theorem 5.3 (Strict increase of coverage)**.** *Let H consist of outputs of layer-ℓ neurons and possibly the bias column (for instance the copy neurons of Proposition 3.3), so that* $\operatorname{col}(H) \subseteq V_{\mathcal{P}}$*. Then*

$$N_K(V_{\mathcal{P}}) - N_K(H) = \|(\Pi_{\mathcal{P}} - \Pi_H)U_K\|_F^2,$$

and

- *(a) if $N_K(H) < N_K(V_{\mathcal{P}})$, there exists θ with $N_K\big([H, \sigma(\Phi\theta)]\big) > N_K(H)$;*
- *(b) if $N_K(H) = N_K(V_{\mathcal{P}})$, no neuron of layer ℓ, and no choice of later layers, can increase N_K beyond $N_K(H)$.*

Proof. Since $\operatorname{col}(H) \subseteq V_{\mathcal{P}}$, the projectors Π_H and $\Pi_{\mathcal{P}}$ commute and $\Pi_{\mathcal{P}} - \Pi_H = \Pi_{\mathcal{P}}(I - \Pi_H)$ is an orthogonal projector, so $N_K(V_{\mathcal{P}}) - N_K(H) = \operatorname{tr}\big((\Pi_{\mathcal{P}} - \Pi_H)P_K\big) = \|(\Pi_{\mathcal{P}} - \Pi_H)U_K\|_F^2$. (a) A neuron h has positive gain iff $M_H^\top h \neq 0$; note $h \notin \operatorname{col}(H)$ is then automatic, since $M_H^\top h = U_K^\top(I - \Pi_H)h$. By Theorem 5.1(b), such a neuron exists iff $M_H^\top$ does not vanish on $V_{\mathcal{P}}$, i.e. iff $\Pi_{\mathcal{P}} M_H = (\Pi_{\mathcal{P}} - \Pi_H)U_K \neq 0$, which is exactly the condition that the gap is positive. (b) This follows from Corollary 5.2. □

Corollary 5.4 (A layer can increase coverage)**.** *Suppose the rows of $H_{\ell-1}$ are pairwise distinct and $N_K(\Phi) < K$. Then there is a layer of width $m_{\ell-1} + 1$ (or $2d + 1$ when $\ell = 1$) with $N_K([H_\ell, \mathbf{1}]) > N_K(\Phi)$ and $s_K([H_\ell, \mathbf{1}]) \leq s_K(\Phi)$.*

Proof. Take the copy neurons of Proposition 3.3(b) or (c). Together with $\mathbf{1}$ they give H with $\operatorname{col}(H) \supseteq \operatorname{col}(\Phi)$ (with equality when $\ell \geq 2$), so $N_K(H) \geq N_K(\Phi)$ and, by Lemma 3.2, $s_K(H) \leq s_K(\Phi)$. Here $V_{\mathcal{P}} = \mathbb{R}^n$ and $N_K(V_{\mathcal{P}}) = K$. If $N_K(H) = K$, the copies already give a strict increase over $N_K(\Phi) < K$. Otherwise Theorem 5.3(a) supplies one more neuron with positive gain, so $N_K([H_\ell, \mathbf{1}]) > N_K(H) \geq N_K(\Phi)$. Appending it does not increase s_K, by Lemma 3.2. □

The copy neurons of Proposition 3.3 cost $m_{\ell-1}$ (or $2d$) units. For coverage, far fewer suffice.

Proposition 5.5 (Preserving coverage with $2K + 2$ neurons)**.** *Let $r = \min(K, \operatorname{rank}\Phi)$, let Q_Φ be an orthonormal basis of* $\operatorname{col}(\Phi)$*, let $y_1, \ldots, y_r$ be the top right singular vectors of $U_K^\top Q_\Phi$, and let $p_i = Q_\Phi y_i = \Phi\theta_i$ be the corresponding principal vectors in* $\operatorname{col}(\Phi)$*. Suppose the rows of $H_{\ell-1}$ are distinct, and let w be such that Φw has distinct entries. Then any layer containing the $2r + 2$ neurons $\sigma(\pm\Phi\theta_i)$, $i \leq r$, and $\sigma(\pm\Phi w)$ satisfies*

$$N_K([H_\ell, \mathbf{1}]) \geq N_K(\Phi), \qquad s_K([H_\ell, \mathbf{1}]) \leq s_K(\Phi),$$

and H_ℓ has pairwise distinct rows.

Proof. By Proposition 3.3(a), $p_i = \sigma(p_i) - \sigma(-p_i) \in \mathrm{col}(H_\ell)$, so $W = \mathrm{span}\{p_1, \dots, p_r\} \subseteq \mathrm{col}(H_\ell)$. The matrix $U_K^\top Q_\Phi$ has at most r nonzero singular values $\cos\theta_i$, and $U_K^\top Q_W = U_K^\top Q_\Phi[y_1, \dots, y_r]$ retains all of them, so $N_K(W) = \sum_{i \le r} \cos^2\theta_i = N_K(\Phi)$ by Lemma 3.5(i). If $r = K$, then $s_K(W)^2 = 1 - \sigma_{\min}(U_K^\top Q_W)^2 = 1 - \sigma_{\min}(U_K^\top Q_\Phi)^2 = s_K(\Phi)^2$; if $r < K$, then $\mathrm{rank}\,\Phi < K$, so some unit vector in $\mathrm{span}(U_K)$ is orthogonal to $\mathrm{col}(\Phi)$ and $s_K(\Phi) = 1$. Monotonicity of N_K and Lemma 3.2 give the two inequalities. A direction w exists because the rows of Φ are distinct and a generic w avoids finitely many hyperplanes. If rows $i \neq j$ of H_ℓ were equal, then $(\Phi w)_i = \sigma(\Phi w)_i - \sigma(-\Phi w)_i$ would equal $(\Phi w)_j$, a contradiction. □

Remark 5.6. A layer of width $2K + 2 + t$ therefore preserves coverage, keeps Theorem 5.1 applicable at the next layer, and leaves t neurons free to add coverage. Under the hypotheses of Corollary 5.4, width $2r + 1$ already gives a strict increase: either the $2r$ principal-vector neurons alone give $N_K = K > N_K(\Phi)$, or Theorem 5.3(a) applies to them. The inequality $N_K([H_\ell, \mathbf{1}]) \ge N_K(\Phi)$ is typically strict, because the pieces $\sigma(\pm p_i)$ span more than p_i.

Remark 5.7 (Skip connections). Propositions 3.3 and 5.5 spend $m_{\ell-1}$, $2d$ or $2K + 2$ neurons of every layer only on preserving coverage. Skip connections, as in residual networks [18], do this for free. We consider the concatenated form of [19],

$$H_\ell = [H_{\ell-1}, R_\ell], \qquad R_\ell = \sigma(\Phi\Theta_\ell), \quad \Phi = [H_{\ell-1}, \mathbf{1}],$$

with $\Theta_\ell \in \mathbb{R}^{(m_{\ell-1}+1)\times k_\ell}$, so that the columns of R_ℓ are k_ℓ neurons of the form (2). The skip path preserves coverage, while the ReLU branch tries to increase it. Indeed, $\mathrm{col}([H_\ell, \mathbf{1}]) \supseteq \mathrm{col}(\Phi)$ for every Θ_ℓ, so s_K is nonincreasing (Lemma 3.2) and N_K nondecreasing in ℓ. Rows that are distinct in $H_{\ell-1}$ stay distinct, so the ceiling of Corollary 5.2 never drops and Theorem 5.1 applies at every layer. The branch contributes exactly the marginal gains G_Φ of Proposition 4.5: all k_ℓ branch neurons are free to add coverage, Theorem 5.3(a) with $H = \Phi$ guarantees that one can unless $N_K(\Phi) = N_K(V_\mathcal{P})$, and Theorem 5.10 applies with $H^{(0)} = \Phi$. The additive skip $H_\ell = H_{\ell-1} + R_\ell\Omega_\ell$ of residual networks mixes the branch into the existing features, so $\mathrm{col}([H_\ell, \mathbf{1}])$ need not contain $\mathrm{col}(\Phi)$; we leave its analysis to future work.

5.2. **How fast can a layer increase coverage?** Theorem 5.3 guarantees a positive gain but not its size. Width q trivially attains the ceiling (Theorem 5.1(b)), and for distinct rows $q = n$, which makes $\mathrm{col}(Z) = \mathbb{R}^n$ and $g = f$. The relevant question is therefore how much coverage $t \ll n$ neurons provide. This depends on how economically ReLU features represent the low eigenvectors.

Definition 5.8 (ReLU variation norm)**.** Let $\mathcal{A}_\Phi = \{\sigma(\Phi\theta)/\|\sigma(\Phi\theta)\| : \sigma(\Phi\theta) \neq 0\}$ be the normalized neurons of layer ℓ, and for $v \in \mathbb{R}^n$ let

$$\|v\|_\mathcal{A} = \inf\Big\{ \sum_j |a_j| : \; v = \sum_j a_j d_j, \; d_j \in \mathcal{A}_\Phi \Big\}$$

(finite sums; $\inf\emptyset = \infty$). By Theorem 5.1, $\|v\|_\mathcal{A} < \infty$ iff $v \in V_\mathcal{P}$. We say U_K is (κ, ε)-*representable* by layer ℓ if every unit vector $u \in \mathrm{span}(U_K)$ can be written $u = v + e$ with $\|v\|_\mathcal{A} \le \kappa$ and $\|e\| \le \varepsilon$.

If the rows of $H_{\ell-1}$ are distinct, then $V_\mathcal{P} = \mathbb{R}^n$, so $\|u_k\|_\mathcal{A} < \infty$ for every k. Since $\|\cdot\|_\mathcal{A}$ is a seminorm, every unit $u = \sum_k c_k u_k$ satisfies $\|u\|_\mathcal{A} \le \sum_k |c_k| \|u_k\|_\mathcal{A} \le \sqrt{K}\max_k \|u_k\|_\mathcal{A}$. Hence U_K is $(\kappa, 0)$-representable with $\kappa = \sqrt{K}\max_k \|u_k\|_\mathcal{A} < \infty$, although this κ may be very large.

Lemma 5.9 (Best single neuron)**.** *Let U_K be (κ, ε)-representable by layer ℓ, let H be any feature matrix, and let $s = s_K(H)$. Then*

$$G^* := \sup\big\{ G_H(\sigma(\Phi\theta)) : \; \sigma(\Phi\theta) \notin \mathrm{col}(H) \big\} \;\ge\; \frac{s^2(s-\varepsilon)_+^2}{\kappa^2}.$$

Proof. For every $d \in \mathcal{A}_\Phi$ we have $\|M_H^\top d\| \le \sqrt{G^*}\,\|(I-\Pi_H)d\| \le \sqrt{G^*}$; this also holds if $d \in \mathrm{col}(H)$, since then $M_H^\top d = U_K^\top(I-\Pi_H)d = 0$. Let x be a unit vector with $M_H^\top M_H x = s^2 x$, and write $U_K x = v + e$ with $v = \sum_j a_j d_j$, $\sum_j |a_j| \le \kappa + \eta$ for arbitrary $\eta > 0$, and $\|e\| \le \varepsilon$. Since $M_H^\top M_H = U_K^\top(I-\Pi_H)U_K = M_H^\top U_K$,

$$s^2 = \|M_H^\top U_K x\| \le \sum_j |a_j|\,\|M_H^\top d_j\| + \|M_H\|_2\|e\| \le (\kappa+\eta)\sqrt{G^*} + s\varepsilon.$$

Let $\eta \to 0$ and rearrange. □

Theorem 5.10 (Greedy coverage rate)**.** *Let U_K be (κ,ε)-representable by layer ℓ and fix $\alpha \in (0,1]$. Let $H^{(0)}$ be any initial features of layer ℓ (for example from Proposition 3.3 or Proposition 5.5), and for $t \ge 0$ let $H^{(t+1)} = [H^{(t)}, \sigma(\Phi\theta_{t+1})]$, where the appended neuron achieves at least αG^* for $H = H^{(t)}$ in Lemma 5.9. Write $\delta_t = \delta(H^{(t)})$ and $s_t = s_K(H^{(t)})$. Then, as long as $s_t \ge 2\varepsilon$ (with the convention $1/0 = \infty$),*

$$\delta_t \le \Big(\frac{1}{\delta_0} + \frac{\alpha t}{4K^2\kappa^2}\Big)^{-1},$$

and for all $t \ge 1$,

$$s_t^2 \le \max\Big\{4\varepsilon^2,\ \frac{4K^2\kappa^2}{\alpha t}\Big\}.$$

Consequently, if the readout features Z contain $H^{(t)}$, Corollary 3.1 gives $\|g-f\| \le s_t\|P_K f\| + \|P_K^\perp f\|$ with s_t^2 bounded as above.

Proof. By Lemma 3.2, s_t is nonincreasing, so the steps with $s_t \ge 2\varepsilon$ form an initial segment. On it, $s_t - \varepsilon \ge s_t/2$, and Lemma 3.5(iv) with $\operatorname{rank} M_{H^{(t)}} \le K$ gives $s_t^2 \ge \delta_t/K$. Lemma 5.9 then gives a gain of at least $\alpha s_t^4/(4\kappa^2) \ge c\,\delta_t^2$ with $c = \alpha/(4K^2\kappa^2)$, so $\delta_{t+1} \le \delta_t - c\delta_t^2$. If $\delta_{t+1} > 0$ then $c\delta_t < 1$ and

$$\frac{1}{\delta_{t+1}} \ge \frac{1}{\delta_t(1-c\delta_t)} \ge \frac{1+c\delta_t}{\delta_t} = \frac{1}{\delta_t} + c,$$

and induction gives the bound on δ_t. If $\delta_{t+1} = 0$, all later deficits and gaps vanish. Hence on the initial segment $s_t^2 \le \delta_t \le 1/(ct) = 4K^2\kappa^2/(\alpha t)$, and after it $s_t < 2\varepsilon$. The final claim is Corollary 3.1 combined with Lemma 3.2, which gives $s_K(Z) \le s_t$. □

Remark 5.11. Theorem 5.10 is the $O(1/t)$ rate of greedy approximation in Hilbert space [15, 16, 17], applied to subspace coverage rather than to a single target. The candidate neurons can be produced by the fixed-gating formulation of Propositions 4.3 and 4.5, but the theorem requires an α-approximate maximizer of the gain over all gating patterns, which a local method does not certify. The bound is informative only when κ is small enough that $4K^2\kappa^2/\alpha \ll n$.

Example 5.12 (Clusters cut off by hyperplanes)**.** Suppose the samples are partitioned into clusters $C_1, \dots, C_K$, and each C_c is cut off from the rest by a hyperplane with a margin in layer-$(\ell-1)$ features: for some w_c, with $z = H_{\ell-1}w_c$, no entry of z lies in an interval (a_c, a_c'), and $C_c = \{i : z_i \ge a_c'\}$. Then, exactly on the sample,

$$\mathbb{1}_{C_c} = \frac{\sigma(z - a_c\mathbf{1}) - \sigma(z - a_c'\mathbf{1})}{a_c' - a_c}.$$

Both ReLU terms vanish off C_c and have entries at most $z_{\max} - a_c$ on C_c. Hence the normalized indicator $\chi_c = \mathbb{1}_{C_c}/\sqrt{|C_c|}$ satisfies $\|\chi_c\|_{\mathcal{A}} \le 2\omega_c$, where $\omega_c = (z_{\max} - a_c)/(a_c' - a_c)$ is a spread-to-margin ratio. The χ_c are orthonormal, so every unit $u \in \mathrm{span}(U_K)$ splits as $u = \Pi_\chi u + (I - \Pi_\chi)u$ with $\|\Pi_\chi u\|_{\mathcal{A}} \le 2\sqrt{K}\max_c \omega_c$. Thus U_K is (κ,ε)-representable with

$$\kappa = 2\sqrt{K}\max_c \omega_c, \qquad \varepsilon = \|(I-\Pi_\chi)U_K\|_2.$$

Here ε measures how far the low eigenvectors are from piecewise constant on the clusters, and Davis–Kahan bounds [8] control it through the between-cluster edge weight relative to λ_{K+1}. By Theorem 5.10, about $16K^3 \max_c \omega_c^2/(\alpha s^2)$ greedy neurons reach coverage gap $s \geq 2\varepsilon$, independently of n.

5.3. **What no layer can do.**

Theorem 5.13 (Limits per neuron)**.** *Let $\eta_1 \geq \cdots \geq \eta_K$ be the singular values of $M_\Phi = (I - \Pi_\Phi)U_K$ (when $\operatorname{rank}\Phi \geq K$, the sines of the principal angles between $\operatorname{span}(U_K)$ and $\operatorname{col}(\Phi)$), and set $\eta_i = 0$ for $i > K$. If $\operatorname{col}(Z) \subseteq \operatorname{col}(\Phi) + T$ for a subspace T with $\dim T \leq t$, for example when $\ell \geq 2$ and Z consists of the copy neurons of Proposition 3.3(b), t further neurons and $\mathbf{1}$, then*

$$N_K(Z) - N_K(\Phi) \leq \sum_{i \leq t} \eta_i^2, \qquad s_K(Z) \geq \eta_{t+1}.$$

Both bounds are attained, without the ReLU constraint, by taking T spanned by the top t left singular vectors of M_Φ.

Proof. By monotonicity of N_K and Lemma 3.2, it suffices to treat $\operatorname{col}(Z) = \operatorname{col}(\Phi) + T$. Let Q_T be an orthonormal basis of $(I - \Pi_\Phi)T$, so Q_T has at most t columns and $\Pi_Z = \Pi_\Phi + Q_T Q_T^\top$. Then $I - \Pi_Z = (I - Q_T Q_T^\top)(I - \Pi_\Phi)$, so $M_Z = (I - Q_T Q_T^\top)M_\Phi$ and

$$M_Z^\top M_Z = M_\Phi^\top M_\Phi - M_\Phi^\top Q_T Q_T^\top M_\Phi,$$

a positive semidefinite downdate of rank at most t. Weyl's inequality [4] gives $\lambda_i(M_Z^\top M_Z) \geq \lambda_{i+t}(M_\Phi^\top M_\Phi) = \eta_{i+t}^2$ for every i. Taking $i = 1$ gives the bound on s_K. Summing over $i \leq K$ gives $\delta(Z) \geq \sum_{i>t} \eta_i^2$, and since $\delta(\Phi) = \sum_i \eta_i^2$, this is the bound on N_K. For attainment, the left singular vectors of M_Φ lie in $\operatorname{col}(\Phi)^\perp$, and removing the top t of them leaves exactly the singular values $\eta_{t+1}, \eta_{t+2}, \ldots$. □

Remark 5.14. Increasing N_K and decreasing s_K are different goals. A single neuron can strictly decrease s_K only if the largest singular value of M_Φ is simple, and pushing s_K below a level c requires at least as many new neurons as there are $\eta_i > c$. Theorem 5.13 is the unconstrained (Ky Fan) benchmark; Theorem 5.10 quantifies how closely ReLU neurons approach it. Together with Theorem 5.3, this is the precise sense in which the neurons left free by Proposition 3.3 add coverage.

6. Concluding remarks

Remark 6.1 (Reduction to the smoothness of the target, and future work). By Lemma 3.5(iv), $s_K(Z)^2 \leq K - N_K(Z)$, so Corollary 3.1 gives, for $\lambda_{K+1} > 0$,

$$\|g - f\| \leq \sqrt{K - N_K(Z)}\,\|f\| + \frac{E(f)}{\sqrt{\lambda_{K+1}}}.$$

Hence $s_K(Z) \to 0$ as $N_K(Z) \to K$, and the error reduces to the smoothness component: once $N_K(Z) = K$, $\|g - f\| \leq E(f)/\sqrt{\lambda_{K+1}}$. By Corollary 5.2 this is attainable whenever the rows of H_{p-1} are distinct, and Theorem 5.10 bounds how many greedily chosen neurons it takes. Two directions remain for future work. First, the remaining term $E(f)/\sqrt{\lambda_{K+1}}$ depends only on f and the graph, so we will investigate how to choose the graph, and hence the Laplacian $\mathcal{L}$, to minimize this smoothness component; since $\mathcal{L}$ also determines U_K, this has to be balanced against the coverage the network can achieve. Second, Theorem 5.10 assumes an α-approximate maximizer of the gain over all gating patterns (Remark 5.11); we will develop methods that find maximal-gain neurons using the fixed-gating formulation of Proposition 4.3, applied to the marginal gain through Proposition 4.5, by searching over gating patterns and solving the resulting cone-constrained generalized eigenvalue problems.

References

[1] D. A. Spielman. *Spectral and Algebraic Graph Theory*. Book draft, Yale University, 2019.

[2] F. R. K. Chung. *Spectral Graph Theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, 1997.

[3] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[4] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2013.

[5] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4th edition, 2013.

[6] Å. Björck and G. H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973.

[7] G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

[8] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[9] P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13:3475–3506, 2012.

[10] T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334, 1965.

[11] M. Pilanci and T. Ergen. Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[12] A. Seeger. Eigenvalue analysis of equilibrium processes defined by linear complementarity conditions. *Linear Algebra and its Applications*, 292(1–3):1–14, 1999.

[13] E. B. Davies, G. M. L. Gladwell, J. Leydold, and P. F. Stadler. Discrete nodal domain theorems. *Linear Algebra and its Applications*, 336:51–60, 2001.

[14] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

[15] L. K. Jones. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *The Annals of Statistics*, 20(1):608–613, 1992.

[16] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

[17] R. A. DeVore and V. N. Temlyakov. Some remarks on greedy algorithms. *Advances in Computational Mathematics*, 5:173–187, 1996.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[19] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.