BROWNIAN HEADS FOR DEEP RELU REPRESENTATIONS: ACTIVATION MASS AND THE COST OF SAME-SAMPLE SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep representation learning often selects hidden features and fits the final predictor on the same sample, so fixed-feature analysis performed after selection can omit selection cost. We study the conditional empirical Rademacher complexity of deep ReLU representations followed by bounded-norm predictors in additive or Lévy–Brownian RKHSs, termed *Brownian heads*. For a fixed representation, we derive an exact dual identity and sharp bounds in terms of activation mass, the average norm of the observed hidden vectors. Under same-sample selection, the representation supremum induces a quadratic Rademacher process. Brownian layer-cake and Gaussian-projection identities reduce it to coordinatewise or signed projected threshold traces, separating realized scale from selection complexity. For samples with pairwise-distinct inputs, explicit scalar ReLU families match the finite-trace and VC rates up to universal constants at the realized trace-and-envelope level. Induced-norm contraction also yields architecture-level bounds for rectangular, rank-deficient ReLU networks. Experiments verify the sharp bounds and rates, exhibit a selection gap at fixed activation mass, and assess the predictive feasibility of Brownian heads.

1 INTRODUCTION AND RELATED WORK

Modern learning architectures are commonly decomposed into a representation map and a comparatively simple prediction head. Given an input space $\mathcal{X}$, such predictors take the form $f = h \circ \phi$, where $\phi\colon \mathcal{X} \to \mathcal{Z}$ extracts task-relevant features and the head h maps the resulting representation to the output space. This viewpoint covers end-to-end neural networks, feature extraction followed by kernel or linear prediction, and the selection of representations from intermediate training checkpoints. Its statistical analysis depends crucially on whether the representation is fixed independently of the sample used to train the head or is selected using that same sample.

Generalization bounds control the discrepancy between empirical and population performance through the Rademacher complexity (RC) of the prediction class (Bartlett & Mendelson, 2002; Mohri et al., 2018). For a representation ϕ fixed independently of the training sample S, the terminal class

$$\mathcal{F}_\phi = \{h \circ \phi : h \in \mathcal{H}\}$$

is fixed, and standard symmetrization applies. If the same sample selects a representation from a family $\Phi(S)$, however, the relevant class is

$$\mathcal{F}_S = \bigcup_{\phi \in \Phi(S)} \{h \circ \phi : h \in \mathcal{H}\}.$$

A fixed-representation bound applied only after selection treats this choice as free and may therefore omit part of the statistical complexity. We call this setting *same-sample representation selection*.

The distinction covers several common regimes. An encoder pretrained without the target sample and fixed before that sample is observed defines a fixed representation; choosing among several such encoders using the target sample already incurs a finite model-selection cost. Fine-tuning, adapter selection, checkpoint selection, and end-to-end training make the representation sample-dependent. A

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

uniform analysis must then cover the representations the procedure could have returned, not only its realized output. This issue occurs within training and is distinct from validation- or test-set leakage.

To isolate this effect, we place a structured terminal class after a deep ReLU representation. Let $z_\theta \colon \mathcal{X} \to \mathbb{R}^m$ denote the output of the last hidden layer of a feedforward ReLU network. For $q \in \{1, 2\}$, we study the complexity of classes containing functions of the form

$$f^{(q)}_{a,\theta}(x) = \langle a, \Phi_q(z_\theta(x)) \rangle_{\mathcal{H}_q}, \qquad \|a\|_{\mathcal{H}_q} \le B, \tag{1}$$

where Φ_q is the canonical feature map of the additive Brownian kernel for $q = 1$ and of the rotation-invariant Lévy–Brownian kernel for $q = 2$. Thus the network produces $z_\theta(x)$, while a bounded-norm predictor in the corresponding Brownian RKHS produces the output. We call the latter a *Brownian head*.

Since $K_q(z, z) = \|z\|_q$, the fixed-representation RC is governed by the empirical *activation mass* $\widehat{M}_{q,S}(z) := n^{-1} \sum_{i=1}^n \|z(x_i)\|_q$. Under same-sample selection, the representation supremum instead produces a quadratic Rademacher process. Brownian layer-cake and Gaussian-projection identities reduce this process to coordinatewise or signed projected threshold traces, separating representation scale from selection complexity. We obtain matching finite-trace and VC rates through explicit scalar ReLU constructions and propagate Brownian feature distance through arbitrary rectangular and rank-deficient ReLU layers using induced matrix norms.

Our bounds are empirical and conditional on the realized representation family. For a sample-independent family $\mathcal{Z}$, standard symmetrization and contraction yield population guarantees for bounded Lipschitz losses; see Corollary B.22. If $\mathcal{Z} = \mathcal{Z}(S)$ is itself sample-dependent, Theorem 3.2 controls the conditional empirical complexity, while a complete population bound additionally requires control of the dependence between the random class and S, for example through a random-class argument (Dupuis et al., 2024). Thus, our results isolate the selection term omitted by a post hoc fixed-representation analysis.

Contributions.

- For a fixed representation z, we derive an exact dual identity and show that the empirical Rademacher complexity of its Brownian-head class scales as $B\sqrt{\widehat{M}_{q,S}(z)/n}$, with optimal lower and upper constants $1/\sqrt{2}$ and 1. Both endpoints are attained by finite Brownian canonical-feature configurations; see Theorem 2.2.
- For a representation family $\mathcal{Z}$ that may be constructed from the same sample, we reduce the representation-indexed quadratic Rademacher process to coordinatewise or signed Gaussian-projected threshold traces. The resulting bounds separate representation scale from threshold-selection complexity; see Lemma 3.1 and Theorem 3.2.
- We prove that the finite-trace and VC dependences are unavoidable, up to universal constants, at the trace-and-envelope level. For pairwise-distinct sample inputs, explicit scalar one-hidden-layer ReLU families attain the corresponding rates; see Theorem 3.3 and Proposition B.14.
- We propagate Brownian feature distance through arbitrary rectangular and rank-deficient exact-ReLU layers using induced matrix norms, without inverse maps or determinant factors, and obtain an explicit deep-network envelope; see Proposition 2.3 and Corollary 3.4.
- Exact finite-sample constructions recover the fixed-representation endpoints and the finite-trace and VC rates. Controlled same-sample adapter families isolate selection at fixed empirical activation mass, while frozen-feature transfer experiments assess the predictive feasibility of Brownian heads; see Sections C and 4.

Related work. **Data-dependent hypothesis classes.** Classical uniform-convergence arguments assume a sample-independent hypothesis class. Foster et al. (2019) study sample-dependent classes through transductive Rademacher complexity and hypothesis-set stability, including stable representation learning; related approaches use algorithm-dependent empirical Rademacher complexity (Sachs et al., 2023) or PAC-Bayesian bounds for random hypothesis sets (Dupuis et al., 2024). These frameworks control the dependence between a random class and its generating sample. Conditional on a realized family $\mathcal{Z}$, we instead exploit Brownian threshold structure to obtain explicit empirical complexity bounds and matching trace-and-envelope lower bounds for exact-ReLU representations.

Brownian and learned kernels. The additive and Lévy–Brownian kernels are classical distance-induced kernels (Sejdinovic et al., 2013). Data-dependent kernel-selection bounds typically control a prespecified family through entropy or kernel pseudodimension (Ying & Campbell, 2010; Cortes et al., 2010; Zhang & Zhang, 2023). Follain & Bach (2025) analyze a shallow Brownian-kernel network that jointly learns a distribution of one-dimensional projections and the predictor, producing a same-sample supremum over Brownian Gram quadratic forms. We place a Brownian RKHS head after a potentially deep ReLU representation and reduce this process to coordinatewise or projected threshold traces, with matching sample-level ReLU constructions.

Composition-operator bounds. Koopman or composition-operator approaches represent layers through $h \mapsto h \circ \varphi$. On Sobolev RKHSs, Hashimoto et al. (2024) obtain determinant-sensitive bounds requiring injective, full-rank layers and smooth bi-Lipschitz activations. Hashimoto et al. (2026) allow bounded domains, changing widths, noninjective weights, and several activations, but their assumptions exclude exact ReLU. Our direct Brownian-metric argument permits rectangular, rank-deficient exact-ReLU layers and controls same-sample selection through threshold traces rather than composition-operator norms.

Recursive RKHS constructions. Brownian Kernel Ladders recursively average Brownian comparisons of preceding-level functions and distinguish a fully adaptive variational family from a fixed ladder or sample-independent finite dictionary (Mohammadigohari et al., 2026). In Section D, we combine the metric and energy recursions of a fixed ladder with our fixed-representation theorem; this is a consequence of the Brownian-head analysis, not part of the same-sample argument. Related layerwise models include Neural Hilbert Ladders and reproducing-kernel chains (Chen, 2023; Heeringa et al., 2025).

Notation. For $p \in \mathbb{N}$, let $[p] := \{1, \dots, p\}$, and for a finite set I, let $2^I := \{T : T \subseteq I\}$ denote its power set. For $v = (v_j)_{j=1}^d \in \mathbb{R}^d$ and $q \in \{1, 2\}$, write $\|v\|_q := \left(\sum_{j=1}^d |v_j|^q\right)^{1/q}$. For $W \in \mathbb{R}^{d' \times d}$, let $\|W\|_{q\to q} := \sup_{v \in \mathbb{R}^d \setminus \{0\}} \|Wv\|_q / \|v\|_q$ denote the induced operator norm. The ReLU function $\mathrm{ReLU}(t) := \max\{t, 0\}$ is applied coordinatewise to vectors, and $\mathrm{Tr}(A)$ denotes the trace of a square matrix A. For nonnegative a and b, we write $a \lesssim b$ if $a \le Cb$ and $a \asymp b$ if $cb \le a \le Cb$, where $0 < c < C$ are universal constants.

2 Brownian Geometry and the Fixed-Representation Baseline

We first identify the empirical capacity of a Brownian head on one representation, then bound it through the network layers.

For $u, v \in \mathbb{R}^m$ and $q \in \{1, 2\}$, define

$$K_q(u, v) = \frac{\|u\|_q + \|v\|_q - \|u - v\|_q}{2}, \qquad \Phi_q(u) = K_q(\cdot, u). \tag{2}$$

These classical positive-definite, distance-induced kernels (Sejdinovic et al., 2013) define RKHSs $\mathcal{H}_q$ with canonical maps $\Phi_q \colon \mathbb{R}^m \to \mathcal{H}_q$. The additive kernel K_1 sums scalar Brownian kernels over coordinates; the rotation-invariant Lévy–Brownian kernel K_2 is a Gaussian mixture of scalar Brownian projection kernels (Proposition B.1). Their basic identities are:

Proposition 2.1 (Distance-induced Brownian geometry)**.** *For every $u, v \in \mathbb{R}^m$ and $q \in \{1, 2\}$,*

$$\|\Phi_q(u)\|_{\mathcal{H}_q}^2 = \|u\|_q, \qquad \|\Phi_q(u) - \Phi_q(v)\|_{\mathcal{H}_q}^2 = \|u - v\|_q. \tag{3}$$

The proof is in Section A.1. These identities connect activation mass to feature size and convert representation distances into squared feature distances.

Fix a map $z \colon \mathcal{X} \to \mathbb{R}^m$, an observed sample $S = (x_i)_{i=1}^n$, and a head radius $B > 0$. The two objects associated with the pair (S, z) are

$$\mathcal{F}_{z,B}^{(q)} = \left\{ x \mapsto \langle a, \Phi_q(z(x)) \rangle_{\mathcal{H}_q} \,\middle|\, a \in \mathcal{H}_q,\ \|a\|_{\mathcal{H}_q} \le B \right\}, \qquad \widehat{M}_{q,S}(z) = \frac{1}{n} \sum_{i=1}^n \|z(x_i)\|_q. \tag{4}$$

Here $\widehat{M}_{q,S}(z)$ is the empirical *activation mass*, the average squared norm of the canonical features by (3).

We measure the size of the fixed terminal class by its *empirical Rademacher* complexity. For a class of real-valued functions $\mathcal{G}$, write

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \frac{1}{n}\mathbb{E}_\varepsilon \sup_{g\in\mathcal{G}} \sum_{i=1}^{n} \varepsilon_i g(x_i), \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\{-1,+1\}. \tag{5}$$

The expectation in (5) is taken over the auxiliary Rademacher variables $\varepsilon_1, \dots, \varepsilon_n$, conditionally on the sample S and the representation z. Hence, the theorem below holds for every realized pair (S, z), including representations constructed from the same observed training data. Whenever a supremum over an uncountable class is not known to be measurable, expectations and probabilities involving that supremum are understood in the outer sense.

Theorem 2.2 (Activation-mass characterization on a fixed representation)**.** *For $q \in \{1, 2\}$,*

$$\widehat{\mathfrak{R}}_S(\mathcal{F}^{(q)}_{z,B}) = \frac{B}{n}\mathbb{E}_\varepsilon \left\| \sum_{i=1}^{n} \varepsilon_i \Phi_q(z(x_i)) \right\|_{\mathcal{H}_q}, \tag{6}$$

with

$$\frac{B}{\sqrt{2}}\sqrt{\frac{\widehat{M}_{q,S}(z)}{n}} \le \widehat{\mathfrak{R}}_S(\mathcal{F}^{(q)}_{z,B}) \le B\sqrt{\frac{\widehat{M}_{q,S}(z)}{n}}. \tag{7}$$

The constants $1/\sqrt{2}$ and 1 are optimal. If $n \ge 2$, both endpoints are attained by nonnegative Brownian canonical feature configurations at every prescribed positive activation mass.

The proof in Section A.3 combines Hilbert-space duality, the Brownian diagonal identity, and the Hilbert-valued Khintchine inequality.

Interpreting the fixed baseline. Let $K_S := \big(K_q(z(x_i), z(x_j))\big)_{i,j=1}^{n}$ be the realized Brownian Gram matrix and let $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$. Then the empirical RC becomes

$$\widehat{\mathfrak{R}}_S(\mathcal{F}^{(q)}_{z,B}) = \frac{B}{n}\mathbb{E}_\varepsilon \sqrt{\varepsilon^\top K_S \varepsilon}, \qquad \mathrm{Tr}(K_S) = n\widehat{M}_{q,S}(z).$$

Thus, the activation mass records the total squared size of the canonical features, whereas the off-diagonal entries of K_S record their pairwise alignment. It therefore determines the sharp universal range in (7), but not the exact complexity. Indeed, one nonzero canonical feature eliminates cancellation and attains the upper endpoint, whereas two identical nonzero features cancel with probability $1/2$ and attain the lower endpoint. The Gram-sensitive bound in Proposition B.2 refines these trace-only bounds when the off-diagonal energy is known.

When hidden states are unavailable, a layer-norm envelope gives an architecture-level bound. Assume that $\mathcal{X} \subseteq \mathbb{R}^{d_0}$ and let $d_L = m$. For each $\ell \in [L]$, let $W_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$, $b_\ell \in \mathbb{R}^{d_\ell}$, and define $z_0(x) = x$, $z_\ell(x) = \mathrm{ReLU}\big(W_\ell z_{\ell-1}(x) + b_\ell\big)$, $\ell \in [L]$. No W_ℓ is assumed square or full rank, and we write $s_{\ell,q} := \|W_\ell\|_{q\to q}$.

Pairwise feature-distance propagation and its one-point, bias-sensitive envelope are stated and proved in Proposition A.2 and Section A.4. The pairwise argument uses coordinatewise 1-Lipschitzness; the one-point bound also uses $\mathrm{ReLU}(0) = 0$. Neither needs determinants, inverse maps, or a positive minimum singular value. Averaging the envelope gives the following consequence.

Proposition 2.3 (Architecture-level fixed-head bound for rectangular ReLU layers)**.** *For every sample $S = (x_i)_{i=1}^n$ and $q \in \{1, 2\}$,*

$$\widehat{\mathfrak{R}}_S(\mathcal{F}^{(q)}_{z_L,B}) \le \frac{B}{\sqrt{n}} \left[\left(\prod_{\ell=1}^{L} s_{\ell,q} \right) \frac{1}{n} \sum_{i=1}^{n} \|x_i\|_q + \sum_{r=1}^{L} \|b_r\|_q \prod_{\ell=r+1}^{L} s_{\ell,q} \right]^{1/2}. \tag{8}$$

No weight matrix is required to be square or full rank.

The bound follows by averaging the one-point estimate (34) and applying (7); see Section A.4. It is necessarily one-sided, since global layer norms need not control the activations realized on the sample; see Proposition B.21.

For a representation fixed independently of the sample, standard symmetrization and loss contraction turn the empirical estimate into the risk bound stated in Corollary B.22. If the same sample is used to select the representation, the conditional empirical identity remains valid, but conditioning on the selected representation cannot justify the same population argument: it removes the selection event from the analysis. The next section retains the corresponding representation supremum explicitly.

3 Same-Sample Representation Selection

Let $\mathcal{Z}$ be a family of representations $z\colon \mathcal{X} \to \mathbb{R}^m$ with a common finite output dimension m, and set

$$\mathcal{F}^{(q)}_{\mathcal{Z},B} := \bigcup_{z\in\mathcal{Z}} \mathcal{F}^{(q)}_{z,B}, \qquad \widehat{M}^{*}_{q,S}(\mathcal{Z}) := \sup_{z\in\mathcal{Z}} \frac{1}{n}\sum_{i=1}^{n} \|z(x_i)\|_q .$$

The family may depend on the observed inputs and training labels, but is held fixed before the auxiliary Rademacher signs are drawn. The upper bounds allow arbitrary, possibly uncountable families for which the displayed quantities are finite. For traces $\mathcal{T} \subseteq 2^{[n]}$, define

$$\rho_n(\mathcal{T}) = \frac{1}{n}\mathbb{E}_\eta \sup_{T\in\mathcal{T}} \left|\sum_{i\in T} \eta_i\right|, \qquad \eta_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\{-1,+1\}. \tag{9}$$

Every realized trace family has at most 2^n elements, even when $\mathcal{Z}$ is uncountable.

Why activation mass alone is insufficient. For pairwise distinct inputs, Proposition B.14 realizes every $T \subseteq [n]$ by a scalar one-hidden-layer ReLU representation $z_T(x_i) = A\mathbf{1}_{\{i\in T\}}$, $A > 0$. The scalar kernels coincide, giving

$$\widehat{\mathfrak{R}}_S\left(\bigcup_{T\in\mathcal{T}} \mathcal{F}^{(q)}_{z_T,B}\right) = B\sqrt{A}\,\rho_n(\mathcal{T}), \qquad q \in \{1,2\}.$$

For even n and the family $\mathcal{T} = \{T : |T| = n/2\}$ fixed before the signs, every candidate has activation mass $A/2$ and fixed-head complexity of order $B\sqrt{A/n}$. Yet $\rho_n(\mathcal{T}) = 1/2$, so the union complexity is $B\sqrt{A}/2$. The fixed-head identity still holds for a selected member; its value alone does not control the selection procedure.

From representation selection to threshold traces. Writing $z_i = z(x_i)$, Hilbert-space duality and the Brownian diagonal identity give

$$\widehat{\mathfrak{R}}_S(\mathcal{F}^{(q)}_{\mathcal{Z},B}) = \frac{B}{n}\mathbb{E}_\varepsilon \sup_{z\in\mathcal{Z}} \left\|\sum_{i=1}^{n} \varepsilon_i \Phi_q(z_i)\right\|_{\mathcal{H}_q},$$

$$\left\|\sum_{i=1}^{n} \varepsilon_i \Phi_q(z_i)\right\|^2_{\mathcal{H}_q} = \sum_{i=1}^{n} \|z_i\|_q + 2\sum_{i<j} \varepsilon_i\varepsilon_j K_q(z_i,z_j). \tag{10}$$

The diagonal is at most $n\widehat{M}^{*}_{q,S}(\mathcal{Z})$. The off-diagonal term has mean zero for each fixed z, but selection inside the expectation can align the kernel values with $\varepsilon_i\varepsilon_j$.

For $q = 1$, assume henceforth that $z(x_i) \in \mathbb{R}^m_+$ for all $z \in \mathcal{Z}$ and $i \in [n]$, as after a final ReLU layer. The coordinatewise layer-cake identity is

$$K_1(z_i,z_j) = \sum_{r=1}^{m} \int_0^\infty \mathbf{1}_{\{z_r(x_i)\ge t\}}\mathbf{1}_{\{z_r(x_j)\ge t\}}\,\mathrm{d}t.$$

For $q = 2$, let $c_0 = \sqrt{\pi/2}$, $G \sim \mathcal{N}(0, I_m)$, and $k_{\mathrm{B}}(\alpha,\beta) = (|\alpha| + |\beta| - |\alpha-\beta|)/2$. The Gaussian-mixture identity of Proposition B.1 yields

$$K_2(u,v) = c_0\mathbb{E}_G k_{\mathrm{B}}(\langle G,u\rangle, \langle G,v\rangle), \qquad k_{\mathrm{B}}(\alpha,\beta) = \sum_{s\in\{-1,+1\}} \int_0^\infty \mathbf{1}_{\{s\alpha\ge t\}}\mathbf{1}_{\{s\beta\ge t\}}\,\mathrm{d}t.$$

Thus coordinates for $q = 1$, or signed projections for $q = 2$, reduce the quadratic kernel process to threshold subsets of the sample:

$$\mathcal{T}_{1,S}(\mathcal{Z}) := \big\{\{i \in [n] : z_j(x_i) \geq t\} : z \in \mathcal{Z},\ j \in [m],\ t > 0\big\}, \tag{11}$$

$$\mathcal{T}_{2,S}(\mathcal{Z}) := \big\{\{i \in [n] : s\langle g, z(x_i)\rangle \geq t\} : z \in \mathcal{Z},\ g \in \mathbb{R}^m,\ s \in \{-1,+1\},\ t > 0\big\}, \tag{12}$$

$$\rho_{q,S}(\mathcal{Z}) := \rho_n\big(\mathcal{T}_{q,S}(\mathcal{Z})\big). \tag{13}$$

Both signs in the projected traces account for the two Brownian half-lines. To separate this selection freedom from the threshold ranges, define

$$\widehat{A}_{1,S}(\mathcal{Z}) := \sup_{z\in\mathcal{Z}} \sum_{j=1}^{m} \max_{i\in[n]} z_j(x_i), \qquad \widehat{W}_{2,S}(\mathcal{Z}) := c_0 \mathbb{E}_G \sup_{\substack{z\in\mathcal{Z}\\ i\in[n]}} |\langle G, z(x_i)\rangle|.$$

These are the total coordinatewise range and expected largest Gaussian projection, respectively; they satisfy $\widehat{M}^*_{1,S}(\mathcal{Z}) \leq \widehat{A}_{1,S}(\mathcal{Z})$ and $\widehat{M}^*_{2,S}(\mathcal{Z}) \leq \widehat{W}_{2,S}(\mathcal{Z})$.

From quadratic chaos to trace complexity.

Lemma 3.1 (Threshold-chaos reduction). *Let $\mathcal{T} \subseteq 2^{[n]}$ and $Q_T(\varepsilon) = n^{-1}\sum_{i<j}\varepsilon_i\varepsilon_j \mathbf{1}_{\{i\in T\}}\mathbf{1}_{\{j\in T\}}$. There is a universal $C_{\mathrm{ch}} > 0$ such that*

$$\mathbb{E}_\varepsilon \sup_{T\in\mathcal{T}} |Q_T(\varepsilon)| \leq C_{\mathrm{ch}}\, n\rho_n(\mathcal{T})^2. \tag{14}$$

Since $Q_T = ((\sum_{i\in T}\varepsilon_i)^2 - |T|)/(2n)$, the lemma controls centered squared trace sums by first-order trace complexity. Applying it inside the threshold integrals gives the following bound; proofs are in Sections A.5 and A.6.

Theorem 3.2 (Uniform threshold control under same-sample selection). *For $q \in \{1, 2\}$, there are universal constants $C_q, C'_q > 0$ such that*

$$\widehat{\mathfrak{R}}_S(\mathcal{F}^{(1)}_{\mathcal{Z},B}) \leq B\left[\frac{\widehat{M}^*_{1,S}(\mathcal{Z})}{n} + C_1\widehat{A}_{1,S}(\mathcal{Z})\rho_{1,S}(\mathcal{Z})^2\right]^{1/2} \leq C'_1 B\sqrt{\widehat{A}_{1,S}(\mathcal{Z})}\,\rho_{1,S}(\mathcal{Z}), \tag{15}$$

$$\widehat{\mathfrak{R}}_S(\mathcal{F}^{(2)}_{\mathcal{Z},B}) \leq B\left[\frac{\widehat{M}^*_{2,S}(\mathcal{Z})}{n} + C_2\widehat{W}_{2,S}(\mathcal{Z})\rho_{2,S}(\mathcal{Z})^2\right]^{1/2} \leq C'_2 B\sqrt{\widehat{W}_{2,S}(\mathcal{Z})}\,\rho_{2,S}(\mathcal{Z}). \tag{16}$$

The first inequalities separate diagonal activation mass from the selection cost, factored into threshold scale and trace complexity. Absorbing the diagonal gives the shorter second bounds, which may be looser when activation mass is much smaller than the envelope. RKHS duality is generic; the diagonal, layer-cake, and Gaussian-projection identities are Brownian-specific. Other kernels would require an analogous controlled threshold-mixture representation.

The bounds are conditional on the sample and family. For sample-independent $\mathcal{Z}$, symmetrization and contraction give population guarantees for bounded Lipschitz losses. A sample-generated $\mathcal{Z}(S)$ needs an additional random-class argument, and the empirical bound must cover all traces available for selection. Selection-safe root-mass and intrinsic-width refinements appear in Propositions B.11 and B.12.

Sharpness of the trace dependence. Write $\mathrm{VCdim}(\mathcal{T})$ for the Vapnik–Chervonenkis dimension (Vapnik & Chervonenkis, 1971). The next result concerns worst-case empirical capacity at the specified trace and envelope level, not statistical minimax excess risk. Its finite-family restriction does not apply to the preceding upper bounds.

Theorem 3.3 (Matching finite-trace and VC dependence). *There are universal constants $0 < c < C$ such that, for every $A, B > 0$, $2 \leq N \leq 2^n$, $1 \leq V \leq n$, and $q \in \{1, 2\}$, the following holds. Each supremum ranges over finite families $\mathcal{Z}$ with common finite output dimension m (which may vary between families), subject to $\widehat{A}_{1,S}(\mathcal{Z}) \leq A$ for $q = 1$ and $\widehat{W}_{2,S}(\mathcal{Z}) \leq A$ for $q = 2$; the $q = 1$*

states are nonnegative. The upper directions hold for every sample S. For pairwise distinct inputs $x_1, \dots, x_n \in \mathbb{R}^d$, the matching lower directions also hold, and hence

$$\sup_{\mathcal{Z}:\,|\mathcal{T}_{q,S}(\mathcal{Z})|\le N} \widehat{\mathfrak{R}}_S(\mathcal{F}^{(q)}_{\mathcal{Z},B}) \asymp B\sqrt{A}\min\left\{1, \sqrt{\frac{\log N}{n}}\right\}, \tag{17}$$

$$\sup_{\mathcal{Z}:\,\mathrm{VCdim}(\mathcal{T}_{q,S}(\mathcal{Z}))\le V} \widehat{\mathfrak{R}}_S(\mathcal{F}^{(q)}_{\mathcal{Z},B}) \asymp B\sqrt{A}\min\left\{1, \sqrt{\frac{V}{n}}\right\}. \tag{18}$$

The saturation follows from $\rho_n(\mathcal{T}) \le 1$. For sharpness, all unions of k disjoint blocks give 2^k traces, VC dimension k, and $\rho_n(\mathcal{T}) \asymp \min\{1, \sqrt{k/n}\}$; take $k \asymp \log N$ or $k \asymp V$. Scalar ReLU realizations apply to both kernels, but may depend on the sample and require width of order n. Thus the result does not claim simultaneous sharpness under all architectural constraints. See Section A.7, corollary B.13, and proposition B.14.

From realized to architecture-level control. Layer norms bound the realized scale via Proposition A.2; ReLU VC bounds control the traces through depth and parameter count. For $q = 2$, the projection contributes m parameters, and the scalar threshold is absorbed into the universal constant.

Corollary 3.4 (Explicit ReLU-family specialization)**.** *Let Θ index a depth-L ReLU representation family with final width m, and set $\mathcal{Z}_\Theta = \{z_L^\theta : \theta \in \Theta\}$. Assume at most P scalar parameters, sample input radius $R_{0,q} = \max_i \|x_i\|_q$, and uniform bounds $\|W_\ell\|_{q\to q} \le S_{\ell,q}$ and $\|b_\ell\|_q \le B_{\ell,q}$. Set*

$$R_{\Theta,q} = R_{0,q}\prod_{\ell=1}^{L} S_{\ell,q} + \sum_{r=1}^{L} B_{r,q}\prod_{\ell=r+1}^{L} S_{\ell,q},$$

with empty products equal to one. Let V_Θ bound the VC dimensions of both realized trace systems; one may take $V_\Theta = C_{\mathrm{net}}(P+m)(L+1)\log(e(P+m))$ for a universal $C_{\mathrm{net}} > 0$. Then

$$\widehat{\mathfrak{R}}_S(\mathcal{F}^{(1)}_{\mathcal{Z}_\Theta,B}) \lesssim B\min\left\{\sqrt{R_{\Theta,1}}, \sqrt{\min\{m,n\}R_{\Theta,1}}\min\{1, \sqrt{V_\Theta/n}\}\right\}, \tag{19}$$

$$\widehat{\mathfrak{R}}_S(\mathcal{F}^{(2)}_{\mathcal{Z}_\Theta,B}) \lesssim B\min\left\{\sqrt{R_{\Theta,2}}, \sqrt{R_{\Theta,2}}m^{1/4}\min\{1, \sqrt{V_\Theta/n}\}\right\}. \tag{20}$$

If all states realized across the family and sample lie in one common r-dimensional subspace, $m^{1/4}$ can be replaced by $r^{1/4}$.

Indeed, $\left\|z_L^\theta(x_i)\right\|_q \le R_{\Theta,q}$ uniformly, so $\widehat{M}^*_{q,S}(\mathcal{Z}_\Theta) \le R_{\Theta,q}$ and

$$\widehat{A}_{1,S}(\mathcal{Z}_\Theta) \le \min\{m,n\}R_{\Theta,1}, \qquad \widehat{W}_{2,S}(\mathcal{Z}_\Theta) \lesssim \sqrt{m}\,R_{\Theta,2}.$$

For $q = 1$, bound coordinate maxima individually or by sample sums. For $q = 2$, Gaussian projection gives $\sqrt{m}$ (or $\sqrt{r}$ in the common subspace). Combine these with $\rho_{q,S}(\mathcal{Z}_\Theta) \lesssim \min\{1, \sqrt{V_\Theta/n}\}$; see Section A.8. The first branch is the selection-safe root-mass bound; the second is useful when $V_\Theta \ll n$. These architectural bounds need no terminal-state inspection; the realized bounds may be sharper. Representer, finite-realization, and recursive-chain results are in Sections B and D.

4 Controlled Experiments

We use three complementary protocols. Exact finite-sample constructions evaluate the theorem endpoints and trace-rate formulas without optimization. Label-dependent one-step rank-one ReLU adapters form nested same-sample families on frozen ResNet-18 features; every candidate is positively rescaled to the same empirical activation mass before complexity is evaluated. Finally, matched linear, stationary-kernel, and Brownian heads are compared on frozen ImageNet features. Complete protocols, all six transfer settings, raw summaries, and secondary diagnostics are in Sections C and C.6 to C.8. The exact formulas attain both endpoints and recover fitted slopes 0.979 for $\sqrt{\log N/n}$ and 0.980 for $\sqrt{V/n}$ ($R^2 = 0.9998$ in both cases).

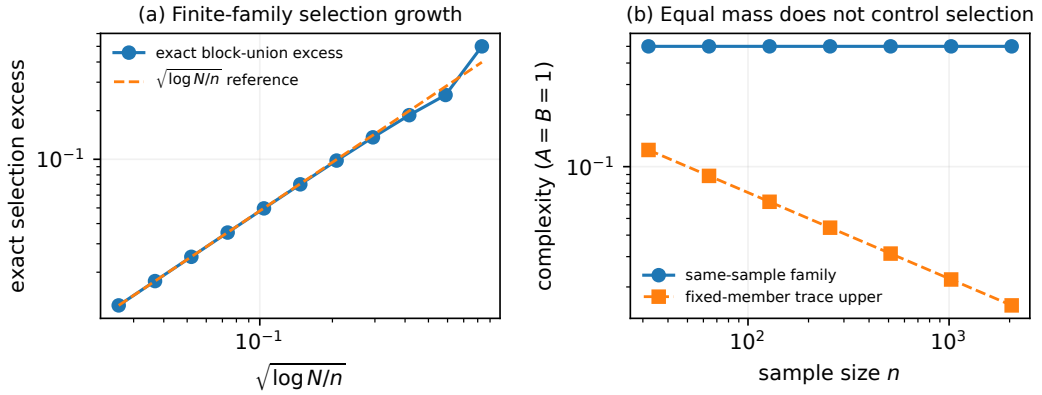

Figure 1: Exact finite-sample constructions supporting the selection theorem. Panel (a) evaluates the closed-form finite-family selection excess. Panel (b) keeps every representation at the same activation mass while the same-sample family complexity remains constant and each fixed-member trace upper decays with n. These are deterministic evaluations, not independent stochastic experiments.

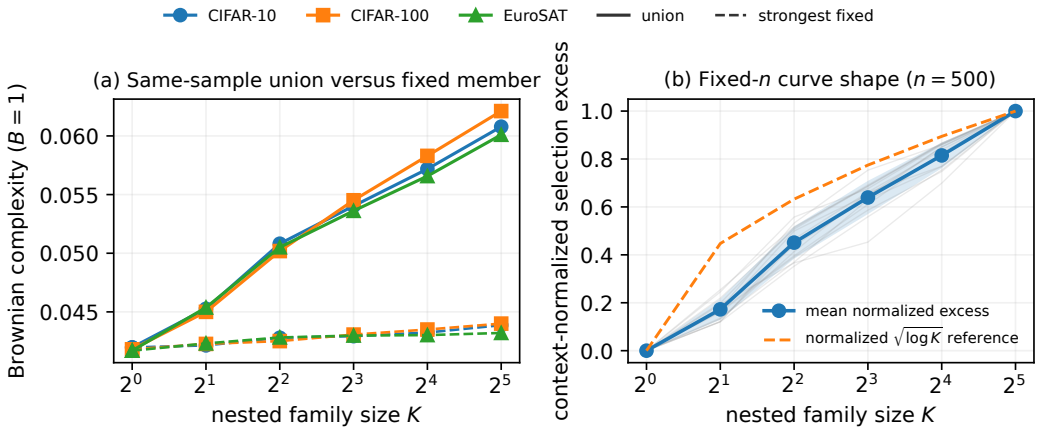

Figure 2: Same-sample learned representation families at the fixed sample size $n = 500$. Panel (a) compares the Brownian union complexity with the strongest fixed member after all candidates are rescaled to identical activation mass. Panel (b) shows the context-normalized union-minus-fixed excess; the dashed curve is a descriptive normalized $\sqrt{\log K}$ reference.

Across three datasets and five paired seeds, the largest predeclared family increases empirical Brownian union complexity by a factor 1.40 over its strongest fixed member (minimum across the 15 contexts: 1.338), despite identical activation mass. At the fixed sample size $n = 500$, the context-normalized selection excess grows consistently with a $\sqrt{\log K}$ curve across the nested families. The dependence on n is established by the separate exact constructions, not by this fixed-n study. This is a controlled selection experiment rather than an end-to-end benchmark.

Across the six dataset–budget combinations, the Lévy–Brownian head ranks first by mean accuracy on four tasks and second on two, with average rank 1.33. Its gains over linear ridge are modest but positive in all six settings, ranging from 0.65 to 1.90 percentage points; it is no more than 0.06 percentage points below the best observed head in any setting. Figure 3 shows the paired differences across all six settings, and Table 1 reports the $n_{\text{lab}} = 1000$ results. Section C.8 gives the protocol and both budgets in Table 5. These frozen-representation experiments do not establish end-to-end superiority or remove the selection term required when the representation is selected using the same sample that trains the head.

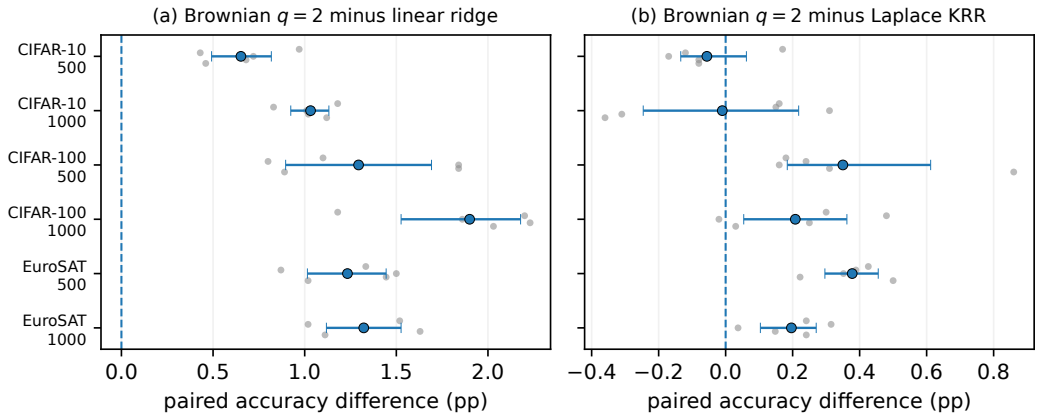

Figure 3: Paired frozen-transfer differences for the Lévy–Brownian head. Small points are the five seed-level paired differences; the large point is their mean with a descriptive paired percentile-bootstrap interval. All methods share the same frozen features and labeled splits. No multiplicity-corrected significance claim is made.

Table 1: Frozen ImageNet-pretrained ResNet-18 transfer with $n_{\mathrm{lab}} = 1000$ labels. Values are mean $\pm$ standard deviation over five paired seeds; bold marks the largest observed mean and is not a significance claim. Ridge regularization is selected by training-set GCV. Test labels are never used for fitting or model selection and are accessed only for final evaluation.

Method	CIFAR-10	CIFAR-100	EuroSAT
Linear ridge	79.63 ± 0.33	45.41 ± 0.75	90.58 ± 0.42
RBF KRR	80.53 ± 0.37	46.55 ± 0.56	91.49 ± 0.33
Laplace KRR	$\mathbf{80.68 \pm 0.45}$	47.10 ± 0.43	91.71 ± 0.30
Brownian $q = 1$	79.67 ± 0.53	45.68 ± 0.54	90.58 ± 0.57
Brownian $q = 2$	80.67 ± 0.23	$\mathbf{47.31 \pm 0.48}$	$\mathbf{91.90 \pm 0.30}$

5 Limitations and Conclusion

Distance-induced Brownian geometry separates two sources of empirical capacity: the activation mass of a fixed representation and the threshold-trace complexity of selecting that representation on the same sample. For a fixed representation, activation mass gives the sharp universal scale of the exact dual complexity. Under same-sample selection, the representation supremum can align threshold traces with the Rademacher signs, and the adaptive bound captures this additional cost. Contraction and induced operator norms extend the analysis to rectangular and rank-deficient ReLU representations. The resulting architecture-level bound is worst-case, while the realized trace and envelope bounds retain the geometry of the candidate family.

The experiments isolate these mechanisms. Exact constructions recover the fixed-representation endpoints and selection rates, while the adapter study exhibits the union gap at fixed activation mass. The frozen-transfer study demonstrates the practical feasibility of Brownian heads, but neither measures the adaptive selection penalty nor establishes predictive superiority.

The scope is limited to Brownian terminal geometries rather than arbitrary heads or a complete theory of end-to-end learning. The matching results concern worst-case empirical capacity, not minimax excess risk, and do not simultaneously saturate all architectural constraints. Moreover, the empirical study uses five seeds, one frozen architecture, and three datasets. Natural directions are to localize the trace cost around a trained representation and to analyze covariance-shaped or recursive Brownian geometries learned from the same data.

AI USE STATEMENT

Generative AI tools were used for literature discovery and comparison; feedback on problem framing, mathematical claims, proof writing and revision, experimental design, research-code implementation and editing, interpretation of experimental results, scientific figures, language editing, and LaTeX formatting. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Complete proofs of all main results and the auxiliary results used by them are provided in Appendices A and B. Appendix C documents the exact finite-sample constructions and experimental protocols, including preprocessing, candidate generation, seed counts, paired-split procedures, hyperparameter grids, complete transfer tables, and secondary diagnostics. The source package contains all LaTeX and figure files required to compile the paper. Experimental code and machine-readable records, including individual random seeds and split/corruption hashes, are separate from this manuscript project.

REFERENCES

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Zhengdao Chen. Multi-layer neural networks as trainable ladders of hilbert spaces. In *Proceedings of the International Conference on Machine Learning*, volume 202, pp. 4294–4329, 2023.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Generalization bounds for learning kernels. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 247–254, 2010.

Benjamin Dupuis, Paul Viallard, George Deligiannidis, and Umut Simsekli. Uniform generalization bounds on data-dependent hypothesis sets via PAC-bayesian theory on random sets. *Journal of Machine Learning Research*, 25(409):1–55, 2024.

Bertille Follain and Francis Bach. Enhanced feature learning via regularisation: Integrating neural networks and kernel methods. *Journal of Machine Learning Research*, 26(172):1–56, 2025.

Dylan J. Foster, Spencer Greenberg, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Hypothesis set stability and generalization. In *Advances in Neural Information Processing Systems*, volume 32, pp. 6729–6739. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/300d1539c3b6aa1793b5678b857732cf-Abstract.html.

Uffe Haagerup. The best constants in the khintchine inequality. *Studia Mathematica*, 70(3):231–283, 1981.

Yuka Hashimoto, Sho Sonoda, Isao Ishikawa, Atsushi Nitanda, and Taiji Suzuki. Koopman-based generalization bound: New aspect for full-rank weights. In *International Conference on Learning Representations*, 2024.

Yuka Hashimoto, Sho Sonoda, Isao Ishikawa, and Masahiro Ikeda. Why high-rank neural networks generalize?: An algebraic framework with RKHSs. arXiv preprint arXiv:2509.21895, 2026.

David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik–chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Tjeerd Jan Heeringa, Len Spek, and Christoph Brune. Deep networks are reproducing kernel chains. arXiv preprint arXiv:2501.03697, 2025.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. doi: 10.1109/JSTARS.2019.2918242.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.

Fanghui Liu, Xiaolin Huang, Chen Gong, Jie Yang, and Li Li. Learning data-adaptive non-parametric kernels. *Journal of Machine Learning Research*, 21(208):1–39, 2020.

Mahdi Mohammadigohari, Giuseppe Di Fatta, Giuseppe Nicosia, and Panos M. Pardalos. Brownian kernel ladders. arXiv preprint arXiv:2606.15812, 2026.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2 edition, 2018.

Adam Paszke et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Fabian Pedregosa et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Sarah Sachs, Tim van Erven, Liam Hodgkinson, Rajiv Khanna, and Umut Şimşekli. Generalization guarantees via algorithm-dependent Rademacher complexity. In *Proceedings of the Thirty-Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 4863–4880. PMLR, 2023. URL https://proceedings.mlr.press/v195/sachs23a.html.

Dino Sejdinovic, Bharath K. Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013. doi: 10.1214/13-AOS1140.

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971. doi: 10.1137/1116025.

Yiming Ying and Colin Campbell. Rademacher chaos complexities for learning the kernel problem. *Neural Computation*, 22(11):2858–2886, 2010. doi: 10.1162/NECO_a_00028.

Yifan Zhang and Min-Ling Zhang. Nearly-tight bounds for deep kernel learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41861–41879, 2023.