CMPUT 654: Mathematical Foundations of Modern AI Systems Fall 2026

Lecture 15: When neural networks approximate efficiently

October 22, 2026

Lecturer: Csaba Szepesvári *Note prepared by: Csaba Szepesvári*

Reading guide. Lecture 14 studied representation questions for maps with discrete inputs and outputs. Lecture 15 changes both the domain and the target: the input belongs to a subset of $\mathbb{R}^d$, the target is real-valued, and the statistical problem is regression. Continuous regression makes three questions distinct. An approximation result asks whether a good member of a function family exists. A constructive approximation result asks whether an algorithm can produce one from specified information about the target. A learning result asks what can be estimated from noisy data.

The material presented in class runs through Section 15.6. The development begins with the progression from approximation to constructive approximation to learning. Shallow ridge and radial-basis-function networks make the questions concrete. The minimax limit for ordinary smooth regression supplies an information-theoretic benchmark. The central Hilbert-space lemma then shows that a function with a bounded integral representation has an accurate finite approximation. The lemma's proof uses random sampling, although the lemma's conclusion is deterministic. The two density corollaries answer the minimum approximation question: radial-basis-function and ridge networks can make the approximation error arbitrarily small in the stated L^2 spaces. Quantitative approximation results for Hölder balls describe how the error decreases with the number of terms, and suitable neural-network constructions give corresponding rates. The integral representation also identifies functions that are easy to approximate by m-term networks. Barron's theorem shows that the same class has the following property: for every fixed m-dimensional linear space, some function in the class is poorly approximated by that space, even though target-dependent ridge atoms attain an $m^{-1/2}$ approximation rate.

The Barron comparison makes target-dependent dictionary selection the central issue. Fixing the dictionary distribution in advance gives a kernel method. Allowing the dictionary elements to depend on the target creates a feature-selection problem: can a constructive or statistical method find the useful elements? Projection pursuit gives a greedy constructive approach conditional on a difficult oracle. The general learning question remains open, so the final two main sections restrict the target to a multi-index model. Under a rank condition, target gradients and a singular-value decomposition identify a predictive subspace. A special two-stage stochastic-gradient procedure then gives a learning result under restrictive assumptions.

The optional-reading section contains the entropy result that was not presented in class, additional classical and neural comparators, and results about depth. The appendices contain the wavelet, Fourier, active-query, RKHS, and staged-SGD details used to support the main development. Lecture 9, Appendix 9.A, reviews $L^2(P_X)$, inner products, Hilbert spaces, orthogonality, and projections.

The lecture's central question is:

Neural networks can choose features for the target. Can a constructive or statistical algorithm find the useful features from the information available to the algorithm?

The lecture gives a general representation theorem and a restricted learning theorem. A general efficient learning theorem remains open.

15.1 From approximation to regression learning

Regression starts from random input–response pairs (X, Y) with $X \in \mathcal{X} \subseteq \mathbb{R}^d$ and $Y \in \mathbb{R}$. Under squared loss, the target is the conditional mean

$$f(x) = \mathbb{E}[Y \mid X = x].$$

The input domain and the target are continuous. Regression leads to a progression from approximation, to constructive approximation, to learning.

Approximation. For a target f and a family $\mathcal{F}_m$, how small can

$$\inf_{g \in \mathcal{F}_m} \|f - g\| \tag{15.1}$$

be? The infimum is an existential quantity and may use complete knowledge of f.

To make the approximation question concrete, consider the function families used throughout the lecture. For an activation $\sigma : \mathbb{R} \to \mathbb{R}$, a width-$m$ one-hidden-layer *ridge network* belongs to

$$\mathcal{N}_m^\sigma = \left\{ x \mapsto c_0 + \sum_{j=1}^m c_j \sigma(a_j^\top x + b_j) : a_j \in \mathbb{R}^d,\ b_j, c_j, c_0 \in \mathbb{R} \right\}. \tag{15.2}$$

For $\psi : \mathbb{R}^d \to \mathbb{R}$, a width-$m$ scale–location network belongs to

$$\mathcal{R}_m^\psi = \left\{ x \mapsto c_0 + \sum_{j=1}^m c_j \psi\big(s_j(x - u_j)\big) : s_j > 0,\ u_j \in \mathbb{R}^d,\ c_j, c_0 \in \mathbb{R} \right\}. \tag{15.3}$$

If $\psi(z) = \rho(\|z\|_2)$, then $\psi(s_j(x - u_j)) = \rho(s_j\|x - u_j\|_2)$, and Eq. (15.3) is a *radial-basis-function network*. The center u_j selects a location, and the scale s_j selects the spatial resolution of the jth atom. These formulas specify representation families. An approximation theorem may use complete knowledge of the target to choose the parameters of each network.

Constructive approximation. Given an information oracle for f, can a finite algorithm call the oracle and produce an accurate g? A point-evaluation oracle returns

$$L_x(f) = f(x) = \int f \, d\delta_x,$$

where δ_x is the Dirac point mass. A general linear-information oracle returns $L(f)$ for specified linear functionals L. Point evaluations are a special case. A numerically stable construction changes little when the oracle values are perturbed. Numerical stability permits a subsequent analysis of observation noise.

Learning. Regression supplies noisy point evaluations. From random noisy observations, can an estimator $\widehat{f}_n$ achieve small prediction risk? For the $L^2(P_X)$ loss, the risk is

$$\mathbb{E}\big[\|\widehat{f}_n - f\|_{L^2(P_X)}^2\big]. \tag{15.4}$$

The expectation averages over the training sample and any randomness used by the estimator.

The norm used in the main results is

$$\langle g, h\rangle_{L^2(P_X)} = \mathbb{E}[g(X)h(X)], \qquad \|h\|^2_{L^2(P_X)} = \mathbb{E}[h(X)^2], \tag{15.5}$$

where $\mathcal{L}(X) = P_X$. Functions that agree P_X-almost everywhere represent the same element of $L^2(P_X)$. The space is a Hilbert space, so independent centered errors add through the inner product exactly as Euclidean vectors do. Lecture 9, Appendix 9.A, gives the required background. Approximation theory also commonly uses the uniform norm $\|h\|_\infty = \sup_x |h(x)|$; the present lecture uses $L^2(P_X)$ because squared-loss regression averages prediction error under the input law.[1]

No choice of representation or algorithm can avoid the information-theoretic limits of regression. To state the standard benchmark, set $\mathcal{X} = [0,1]^d$. Suppose that P_X has a density bounded above and bounded away from zero, and let $(X_i, Y_i)_{i=1}^n$ be independent copies of (X, Y) satisfying

$$Y_i = f(X_i) + \varepsilon_i, \qquad \mathcal{L}(\varepsilon_i) = N(0, \sigma^2), \tag{15.6}$$

where the noise variables are independent of the inputs. A function h is *β-Hölder* for $0 < \beta \le 1$ if

$$|h(x) - h(x')| \le L\|x - x'\|_2^\beta \quad \text{for every } x, x'. \tag{15.7}$$

The case $\beta = 1$ is Lipschitz continuity. When $0 < \beta < 1$, small distances satisfy $r^\beta \ge r$ for $0 \le r \le 1$, so the same constant permits larger local changes than a Lipschitz condition.

Write $\alpha = k + \beta$ with $k \in \mathbb{N}_0$ and $0 < \beta \le 1$. The Hölder ball $\mathcal{H}_d^\alpha(L)$ contains functions whose partial derivatives through order k are bounded by L and whose derivatives of order k are β-Hölder with constant L. The minimax mean integrated squared error over $\mathcal{H}_d^\alpha(L)$ satisfies

$$\inf_{\widehat{f}_n} \sup_{f \in \mathcal{H}_d^\alpha(L)} \mathbb{E}\big[\|\widehat{f}_n - f\|^2_{L^2(P_X)}\big] \asymp n^{-2\alpha/(2\alpha+d)}. \tag{15.8}$$

The infimum ranges over all measurable estimators based on the n observed pairs; the notation does not impose a computational model. Thus Eq. (15.8) is an information-theoretic statement, not a claim about an efficiently implementable algorithm. The exponent deteriorates with the ambient dimension. A faster rate requires a smaller target class, a different observation model, or a different loss. The bound sets the regression benchmark. A complementary entropy result was not developed in class and appears in the optional reading.

15.2 Integral representations and density

The finite network formulas in Eqs. (15.2) and (15.3) both start from a *dictionary*: a parameterized family $\mathcal{D} = (\phi_\theta)_{\theta\in\Theta}$ of functions. For a ridge network, $\theta = (a, b)$ and $\phi_\theta(x) = \sigma(a^\top x + b)$. For a scale–location network, $\theta = (s, u)$ and $\phi_\theta(x) = \psi(s(x - u))$. A finite network selects finitely many dictionary elements. Replacing the finite sum by an integral permits a continuous collection and gives a useful sufficient condition for accurate finite approximation.

Definition 15.2.1 (Finite signed measure and integral representation)**.** Let Θ be a measurable parameter space. A *finite signed measure* will mean a measure of the form

$$d\mu(\theta) = s(\theta)\, d\nu(\theta),$$

where ν is a finite positive measure and $s : \Theta \to \{-1, 1\}$ is measurable. The positive measure $|\mu| := \nu$ is the *total variation measure* of μ, and

$$\|\mu\|_{\mathrm{TV}} := |\mu|(\Theta)$$

is its *total variation norm*. For a dictionary $(\phi_\theta)_{\theta\in\Theta}$ in a Hilbert space $\mathcal{H}$, an *integral representation* of $f \in \mathcal{H}$ is

$$f = \int_\Theta \phi_\theta \, d\mu(\theta). \tag{15.9}$$

The definition through a sign function is equivalent to the usual definition of a finite real-valued signed measure.[2]

A finite sum is a special integral. If $f = \sum_{j=1}^m c_j \phi_{\theta_j}$, take $\mu = \sum_{j=1}^m c_j \delta_{\theta_j}$. Then $|\mu| = \sum_{j=1}^m |c_j| \delta_{\theta_j}$ after repeated atoms have been combined. Integral representations can also describe functions that require a limit of finite sums.

The same formula is used in mean-field descriptions of two-layer networks. If ρ is a probability measure on neuron parameters, then

$$f_\rho(x) = \int a\, \sigma(w^\top x + b)\, d\rho(a, w, b). \tag{15.10}$$

Absorbing the output weight a into a signed measure on (w, b) gives the representation in Eq. (15.9). Mean-field training adds an evolution equation for ρ; the static integral formula alone is a representation statement.

The integral form is useful because every function with a representing measure of bounded total variation has an accurate finite-sum approximation. The following lemma is the central result of the lecture. The lemma's statement requires only Hilbert-space geometry.

Lemma 15.2.2 (Finite approximation of an integral representation)**.** *Let $\mathcal{H}$ be a real Hilbert space, let $\theta \mapsto \phi_\theta \in \mathcal{H}$ be measurable, and suppose that $\|\phi_\theta\|_{\mathcal{H}} \le 1$ for every θ. If $f = \int \phi_\theta \, d\mu(\theta)$ and $B = \|\mu\|_{\mathrm{TV}} < \infty$, then, for every $m \ge 1$, there are $\theta_1, \dots, \theta_m \in \Theta$ and $s_1, \dots, s_m \in \{-1, 1\}$ such that*

$$\left\| f - \frac{B}{m} \sum_{j=1}^m s_j \phi_{\theta_j} \right\|_{\mathcal{H}} \le \frac{B}{\sqrt{m}}. \tag{15.11}$$

Proof. The conclusion is immediate when $B = 0$. Suppose that $B > 0$. Let $\Theta_1, \dots, \Theta_m$ be independent random elements with common law $\pi = |\mu|/B$, let $s = d\mu/d|\mu|$, and define

$$S_j = s(\Theta_j), \qquad Z_j = B S_j \phi_{\Theta_j}.$$

The random elements $Z_1, \dots, Z_m$ are independent, $\mathbb{E}[Z_j] = f$, and $\|Z_j\|_{\mathcal{H}} \le B$. The inner product and independence give

$$\mathbb{E}\left[\left\| \frac{1}{m} \sum_{j=1}^m Z_j - f \right\|_{\mathcal{H}}^2 \right] = \frac{1}{m} \mathbb{E}\big[\|Z_1 - f\|_{\mathcal{H}}^2\big] = \frac{1}{m} \left(\mathbb{E}[\|Z_1\|_{\mathcal{H}}^2] - \|f\|_{\mathcal{H}}^2 \right) \le \frac{B^2}{m}. \tag{15.12}$$

At least one realization has squared error no larger than the expectation. Fixing that realization proves Eq. (15.11). □

The proof is the *probabilistic method*: a random experiment proves that a deterministic choice exists. The proof also exposes the limitation of the construction. The sampling law $|\mu|/B$ depends on a representing measure for f. Knowing that measure can require essentially the same information as knowing the target. The lemma therefore proves economical finite representation. The lemma alone does not supply an implementable oracle algorithm or a regression estimator.

The lemma gives a route from integral representations to universal approximation. If a dense subclass of targets has representing measures of finite total variation, finite networks are dense as well. Applying the integral-representation argument to scale–location atoms and ridge atoms gives the next two corollaries.

For a measurable domain $D \subseteq \mathbb{R}^d$, $L^2(D)$ denotes the space of measurable functions $h : D \to \mathbb{R}$ such that $\int_D |h(x)|^2\,dx < \infty$, where functions that agree Lebesgue-almost everywhere are identified. The $L^2(D)$ norm is $\|h\|_{L^2(D)} = (\int_D |h(x)|^2\,dx)^{1/2}$. For a subset $\mathcal{F} \subseteq L^2(D)$, the notation $\overline{\mathcal{F}}^{L^2(D)}$ denotes the closure of $\mathcal{F}$ in the $L^2(D)$ norm. Thus $\overline{\mathcal{F}}^{L^2(D)} = L^2(D)$ means that for every $h \in L^2(D)$ and every $\epsilon > 0$, there is an $f \in \mathcal{F}$ such that $\|h - f\|_{L^2(D)} < \epsilon$.

Corollary 15.2.3 (Density of Gaussian radial-basis-function networks)**.** *Let* $\psi(x) = \exp(-\|x\|_2^2/2)$. *Then*

$$\overline{\bigcup_{m\geq 1}\left\{\sum_{j=1}^{m} c_j\psi\big(s_j(\cdot - u_j)\big) : c_j \in \mathbb{R},\ s_j > 0,\ u_j \in \mathbb{R}^d\right\}}^{L^2(\mathbb{R}^d)} = L^2(\mathbb{R}^d). \tag{15.13}$$

A wavelet proof chooses a Laplacian-of-Gaussian analysis function and the Gaussian as the synthesis atom. Translation and dilation then give an integral representation, and Lemma 15.2.2 replaces the integral by finitely many scale–location atoms. Appendix 15.A and Appendix 15.B give the coefficient estimate, reconstruction formula, and Fourier verification.

Corollary 15.2.4 (Density of shallow ridge networks on a cube)**.** *Let* $Q = [0,1]^d$, *and let* $\sigma : \mathbb{R} \to [0,1]$ *be measurable with*

$$\lim_{t\to-\infty} \sigma(t) = 0, \qquad \lim_{t\to+\infty} \sigma(t) = 1. \tag{15.14}$$

For

$$\mathcal{N}_\sigma(Q) = \left\{ c_0 + \sum_{j=1}^{m} c_j\sigma(a_j^\top x + b_j)\Big|_Q : m \geq 1,\ c_0, c_j, b_j \in \mathbb{R},\ a_j \in \mathbb{R}^d \right\}, \tag{15.15}$$

Then

$$\overline{\mathcal{N}_\sigma(Q)}^{L^2(Q)} = L^2(Q). \tag{15.16}$$

Fourier inversion writes a smooth target as an integral of oscillations $e^{i\omega^\top x}$. On the cube, each oscillation is a univariate function of $\omega^\top x$, and the fundamental theorem of calculus writes the univariate profile as an integral of thresholds. Steep sigmoids approximate the thresholds. Appendix 15.B makes each step precise.

The two density corollaries answer a basic minimum question: can a finite network make the approximation error arbitrarily small? The answer is yes for the stated RBF and ridge families. Density alone gives no relation between the required network width and the desired error.

For a Hölder ball, approximation theory asks the quantitative version of the same question. Under the input-distribution assumptions used in Eq. (15.8), standard local-polynomial or wavelet spaces V_m with m degrees of freedom can be chosen so that

$$\sup_{f\in\mathcal{H}_d^\alpha(L)} \inf_{g\in V_m} \|f - g\|_{L^2(P_X)} \lesssim m^{-\alpha/d}. \tag{15.17}$$

Suitable neural-network architectures can implement the same local approximations. For example, ReLU networks with $O(m \log m)$ parameters can attain uniform, and hence $L^2(P_X)$, error of order $m^{-\alpha/d}$ by implementing local Taylor approximations and a partition of unity. The parameter count

and logarithmic factors depend on the activation and architecture. For a classical m-dimensional series estimator, the squared approximation error is of order $m^{-2\alpha/d}$ by Eq. (15.17), and the estimation error is of order m/n. Balancing the two terms gives the regression rate $n^{-2\alpha/(2\alpha+d)}$ in Eq. (15.8). The bias– estimation balance connects the quantitative approximation result to the regression benchmark; approximation and regression still answer different questions.

15.3 Target-dependent approximation and fixed features

The density results are useful, but there is a more interesting fact. Some functions have a bounded continuous neural-network representation of the form in Eq. (15.9). Lemma 15.2.2 therefore gives accurate m-term approximations for those functions. Now fix any m basis functions in advance, whether they come from a classical approximation method or from any other proposed basis. Barron's theorem says that some function from the same class is nevertheless poorly approximated by their span. To state the theorem, first define the target class by fixing the Fourier normalization

$$\widehat{f}(\omega) = \int_{\mathbb{R}^d} e^{-i\omega^\top x} f(x)\, dx. \tag{15.18}$$

Definition 15.3.1 (Anchored spectral Barron norm)**.** Let $\mathcal{B}(\mathbb{R}^d)$ be the class of real-valued functions $f : \mathbb{R}^d \to \mathbb{R}$ that admit a representation

$$f(x) = c + \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} (e^{i\omega^\top x} - 1) F(\omega)\, d\omega, \tag{15.19}$$

where $c \in \mathbb{R}$, $F \in L^1(\mathbb{R}^d; \mathbb{C})$, the integral is real for every x, and $\int_{\mathbb{R}^d} \|\omega\|_1 |F(\omega)|\, d\omega < \infty$. Setting $x = 0$ in Eq. (15.19) gives $c = f(0)$. The weighted spectral term assigns value zero to every constant function by taking $F = 0$. Using only the weighted spectral term would therefore give a seminorm. Adding $|c|$ records the constant component. Define

$$\|f\|_{\mathcal{B}(\mathbb{R}^d)} = \inf_{c,F} \left\{ |c| + \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \|\omega\|_1 |F(\omega)|\, d\omega \right\}, \tag{15.20}$$

where the infimum ranges over all representations in Eq. (15.19). The ambient Barron ball and its restriction to $Q = [0,1]^d$ are

$$\mathcal{B}_C(\mathbb{R}^d) = \{f \in \mathcal{B}(\mathbb{R}^d) : \|f\|_{\mathcal{B}(\mathbb{R}^d)} \le C\}, \qquad \mathcal{B}_C(Q) = \{f|_Q : f \in \mathcal{B}_C(\mathbb{R}^d)\} \subseteq L^2(Q). \tag{15.21}$$

Each restriction in $\mathcal{B}_C(Q)$ is identified with its equivalence class in $L^2(Q)$. The quantity in Eq. (15.20) is a norm on the represented functions on $\mathbb{R}^d$.[3]

Theorem 15.3.2 (Barron's adaptive upper bound and fixed-space lower bound)**.** *Let P_X be uniform on $Q = [0,1]^d$, and let σ satisfy Eq. (15.14). There is a universal constant $\kappa > 0$ such that the following statements hold.*

(i) For every $h \in \mathcal{B}_C(Q)$, every $m \ge 1$, and every $\eta > 0$, there is $f_m \in \mathcal{N}_m^\sigma$ satisfying

$$\|h - f_m\|_{L^2(P_X)} \le \frac{2(C+\eta)}{\sqrt{m}}. \tag{15.22}$$

(ii) For every fixed collection $h_1, \dots, h_m \in L^2(P_X)$,

$$\sup_{h \in \mathcal{B}_C(Q)} \inf_{a \in \mathbb{R}^m} \left\| h - \sum_{j=1}^m a_j h_j \right\|_{L^2(P_X)} \ge \kappa \frac{C}{d} m^{-1/d}. \tag{15.23}$$

Upper-bound idea. Choose $f \in \mathcal{B}_C(\mathbb{R}^d)$ satisfying $f|_Q = h$, and choose a representation of f whose cost is at most $C+\eta$. The Fourier-to-threshold calculation used for Corollary 15.2.4 represents $f - c$ by a signed measure over ridge atoms whose total variation is at most twice the spectral moment. Lemma 15.2.2 then selects m atoms. The constant c is represented exactly. The selected ridge directions and biases depend on the target.

Lower-bound idea. Choose $2m$ distinct integer frequencies of size $O(dm^{1/d})$. The associated cosine functions are orthonormal in $L^2(Q)$, and their Barron norms have the same order as their frequencies. For any subspace V of dimension at most m, Bessel's inequality gives

$$\sum_{j=1}^{2m} \|P_V e_j\|^2_{L^2(Q)} \le m.$$

At least one normalized cosine e_j therefore has

$$\inf_{v\in V} \|e_j - v\|^2_{L^2(Q)} = 1 - \|P_V e_j\|^2_{L^2(Q)} \ge \frac{1}{2}.$$

Scaling that cosine into $\mathcal{B}_C(Q)$ gives Eq. (15.23). Appendix 15.B handles the harmless smoothing needed to use an integrable spectral density.

The lower bound is a statement about the *Kolmogorov m-width*

$$d_m(\mathcal{F}; L^2(P_X)) = \inf_{\substack{V\subseteq L^2(P_X)\\ \dim(V)\le m}} \sup_{f\in\mathcal{F}} \inf_{v\in V} \|f - v\|_{L^2(P_X)}. \tag{15.24}$$

Kolmogorov's question is whether one m-dimensional subspace can approximate every target in $\mathcal{F}$. Barron's upper bound allows a different set of m ridge atoms for each target. To reach error ϵ, the upper bound uses $m = O(C^2\epsilon^{-2})$, while a fixed linear space can require $m = \Omega((C/(d\epsilon))^d)$.[4]

The special feature in Barron's upper bound is that the ridge directions and biases are chosen after the target is known. The alternative is to choose the dictionary elements in advance. In the integral representation in Eq. (15.9), choosing the dictionary in advance means fixing a probability measure τ on Θ and permitting only measures of the form $d\mu = p\, d\tau$. Define the resulting function space and its coefficient norm by

$$\begin{aligned} \mathcal{F}_{2,\tau} &= \left\{ f : f(x) = \int_\Theta p(\theta)\phi_\theta(x)\, d\tau(\theta) \text{ for some } p \in L^2(\tau) \right\}, \\ \|f\|_{\mathcal{F}_{2,\tau}} &= \inf \left\{ \|p\|_{L^2(\tau)} : f(x) = \int_\Theta p(\theta)\phi_\theta(x)\, d\tau(\theta) \right\}. \end{aligned} \tag{15.25}$$

The feature locations are distributed according to τ; only the coefficient function p changes with the target. The corresponding kernel is

$$k_\tau(x, x') = \int_\Theta \phi_\theta(x)\phi_\theta(x')\, d\tau(\theta). \tag{15.26}$$

The normed function space in Eq. (15.25) is the reproducing-kernel Hilbert space associated with k_τ. The infimum is needed because different coefficient functions can represent the same function. Conversely, every separable reproducing-kernel Hilbert space admits a feature representation of the form in Eq. (15.25); Appendix 15.D gives a direct construction.

The integral representation in Eq. (15.9) allows any finite signed measure. Define the corresponding variation space and norm by

$$\mathcal{F}_1 = \left\{ f : f(x) = \int_\Theta \phi_\theta(x)\, d\mu(\theta) \text{ for some finite signed measure } \mu \right\},$$
$$\|f\|_{\mathcal{F}_1} = \inf \left\{ \|\mu\|_{\mathrm{TV}} : f(x) = \int_\Theta \phi_\theta(x)\, d\mu(\theta) \right\}. \tag{15.27}$$

The two unit balls are

$$B_{\mathcal{F}_1} = \{f : \|f\|_{\mathcal{F}_1} \le 1\}, \qquad B_{\mathcal{F}_{2,\tau}} = \{f : \|f\|_{\mathcal{F}_{2,\tau}} \le 1\}. \tag{15.28}$$

For either $\mathcal{G} \in \{\mathcal{F}_1, \mathcal{F}_{2,\tau}\}$, every nonzero function $f \in \mathcal{G}$ is a positive scalar multiple of its normalized version in the unit ball: $f = \|f\|_{\mathcal{G}}(f/\|f\|_{\mathcal{G}})$. Thus the unit ball represents every direction in the function space; the norm records the required scale.

The fixed-kernel representation requires $d\mu = p\, d\tau$ with $p \in L^2(\tau)$. Since τ is a probability measure, Cauchy–Schwarz gives

$$\|f\|_{\mathcal{F}_1} \le \int_\Theta |p|\, d\tau \le \|p\|_{L^2(\tau)}. \tag{15.29}$$

Taking the infimum over all coefficient functions representing f yields $\|f\|_{\mathcal{F}_1} \le \|f\|_{\mathcal{F}_{2,\tau}}$, and hence

$$\mathcal{F}_{2,\tau} \subseteq \mathcal{F}_1, \qquad B_{\mathcal{F}_{2,\tau}} \subseteq B_{\mathcal{F}_1}. \tag{15.30}$$

Either inclusion can be strict, depending on the dictionary and feature law. The mathematical distinction is that a variation-space representation may concentrate mass on features chosen for the target, while an RKHS representation must have an $L^2(\tau)$ density relative to the fixed law τ.

The representation results show a possible advantage of neural networks: the dictionary elements can be chosen for the target instead of being fixed in advance. Can a constructive or statistical algorithm find the useful dictionary elements? Projection pursuit is a classical constructive attempt.

15.4 Constructive approximation by projection pursuit

Projection pursuit regression selects target-dependent ridge functions greedily. Projection pursuit approximates a target by

$$f_K(x) = c_0 + \sum_{j=1}^{K} g_j(a_j^\top x), \qquad a_j \in \mathbb{R}^d, \tag{15.31}$$

where both the directions a_j and the univariate profiles g_j are selected for the target. A shallow neural network is a restricted projection-pursuit model with $g_j(t) = c_j\sigma(t + b_j)$.

A stagewise constructive method begins with a residual $r_0 = f$. At step j, the method searches for a direction and profile whose ridge function has large correlation with r_{j-1}, adds that function to the current approximation, and updates the residual. Backfitting may then revise earlier terms. In an abstract Hilbert space, an exact correlation oracle and a bounded variation norm give an $O(K^{-1/2})$ approximation bound. Appendix 15.C states and proves the greedy result.

The oracle is the main computational issue. For a finite dictionary, one can scan every atom. For the continuous ridge dictionary, the oracle must optimize over a direction and a univariate function, and the search is generally nonconvex.

Projection pursuit therefore gives a constructive approximation method only when a strong atom-selection oracle is available. Learning requires an efficient algorithm that finds useful features from noisy data, without access to the target or to such an oracle. An efficient learning algorithm for the general Barron class remains open. The next positive result restricts the target to a multi-index model, where the features form one low-dimensional linear subspace.

15.5 Multi-index models and average derivatives

The multi-index restriction replaces the search for arbitrary ridge sums by the recovery of one predictive subspace. Several common structured models have the form

$$f(x) = \sum_{j=1}^{K} g_j(W_j^\top x), \qquad W_j \in \mathbb{R}^{d \times s_j}. \tag{15.32}$$

An additive model uses coordinate directions. A single-index model has $K = s_1 = 1$. Projection pursuit uses $s_j = 1$ and unknown directions. A one-hidden-layer network is projection pursuit with a specified parametric form for the profiles. A *multi-index model* focuses on one term,

$$f(x) = g(W^\top x), \qquad W \in \mathbb{R}^{d \times s}, \quad W^\top W = I_s, \quad s \ll d. \tag{15.33}$$

Only the column space of W matters: an orthogonal change of basis inside that space can be absorbed into the unknown link g.

Differentiation reveals the subspace. The chain rule gives

$$\nabla f(x) = W \nabla g(W^\top x), \tag{15.34}$$

so every gradient belongs to $\operatorname{span}(W)$. If gradients are available at points $x_1, \dots, x_q$, form

$$G = \begin{bmatrix} \nabla f(x_1) & \cdots & \nabla f(x_q) \end{bmatrix} \in \mathbb{R}^{d \times q}. \tag{15.35}$$

When $\operatorname{rank}(G) = s$, the columns of G span $\operatorname{span}(W)$. For a singular-value decomposition

$$G = \sum_{\ell=1}^{s} \sigma_\ell u_\ell v_\ell^\top, \qquad \sigma_s > 0, \tag{15.36}$$

the vectors $u_1, \dots, u_s$ therefore form an orthonormal basis of the predictive subspace. If target gradients were available, the feature-selection problem would therefore reduce to a singular-value decomposition. Recovering the subspace is sufficient because the link can be redefined in the recovered coordinates.

The subspace gives a dimension reduction. If g has Hölder smoothness α and the subspace were known, the generic regression benchmark would become

$$n^{-2\alpha/(2\alpha+s)} \tag{15.37}$$

instead of the ambient-dimensional rate in Eq. (15.8).

Regression observations do not contain the target gradients. Methods that estimate averages of derivatives from regression data are called *average derivative estimation* methods. A population matrix that can expose every direction is the *average gradient outer product*

$$G_f = \mathbb{E}[\nabla f(X) \nabla f(X)^\top] = W\, \mathbb{E}[\nabla g(W^\top X) \nabla g(W^\top X)^\top] W^\top. \tag{15.38}$$

If the middle matrix is positive definite, the range of G_f equals $\operatorname{span}(W)$. A zero or nearly singular middle matrix means that the available derivatives fail to reveal one or more active directions.

Point evaluations can estimate gradients by finite differences in a noiseless active-query model; Appendix 15.C gives a quantitative construction followed by the SVD argument. Passive noisy regression is harder because the derivative information must first be estimated. Classical two-stage methods estimate the index space by local smoothing and derivative averages and then fit the link in the estimated coordinates. The error of the subspace estimate must not dominate the second-stage link error.

Gradients and a singular-value decomposition would recover the predictive subspace and reduce the regression problem from dimension d to dimension s. Regression data do not provide the gradients directly. The next result asks whether a two-layer network trained from data can exploit the same structure. The positive theorem uses an unusual staged procedure under restrictive model and algorithmic assumptions.

15.6 A theorem for neural feature learning by stochastic gradient descent

The positive result concerns a particular two-stage procedure, not ordinary simultaneous training of both layers. The student predictor is the two-layer network

$$\widehat{y}(x; W, a, b) = \sum_{j=1}^{m} a_j \sigma(w_j^\top x + b_j), \qquad W \in \mathbb{R}^{m \times d}. \tag{15.39}$$

W is the first-layer weight matrix, b contains the hidden-unit biases, and a contains the second-layer or output weights. In the first stage, only W is updated. Every online-SGD update receives a fresh independent pair (X_t, Y_t) and has the form

$$W^{(t+1)} = (1 - \eta_t \lambda) W^{(t)} - \eta_t \nabla_W \ell\big(\widehat{y}(X_t; W^{(t)}, a^{(0)}, b^{(0)}), Y_t\big). \tag{15.40}$$

$a^{(0)}$ and $b^{(0)}$ remain fixed throughout the first stage.[5]

The prediction algorithm then uses a separate second stage. After computing $W^{(T)}$, the algorithm draws independent new biases satisfying $\mathcal{L}(b_j) = \operatorname{Unif}[-\Delta, \Delta]$, freezes both $W^{(T)}$ and b, and updates only the output weights for $t = 0, \dots, T' - 1$:

$$a^{(t+1)} = (1 - \eta'_t \lambda') a^{(t)} - \eta'_t \nabla_a \ell\big(\widehat{y}(X_{I_t}; W^{(T)}, a^{(t)}, b), Y_{I_t}\big), \tag{15.41}$$

where the indices I_t are sampled independently and uniformly from $\{0, \dots, T-1\}$ at the second-stage updates. The second stage therefore reuses the same T observations rather than receiving fresh data; the sample count is T, while the two stages together perform $T + T'$ SGD updates. Conditional on $W^{(T)}$ and b, the second stage is regularized linear prediction in a. The guarantee concerns the staged algorithm and does not cover ordinary simultaneous training. The prediction theorem uses a symmetric first-stage initialization: all rows of $W^{(0)}$ equal the same vector drawn from $N(0, I_d/d)$, all output weights equal the same positive value of order $1/m$, and all initial biases equal the same positive constant. The second-stage resampling of b replaces the initial biases. For the prediction theorem, $\sigma(z) = z_+$, $\tau \geq 1$, and

$$\ell_H(z) = \begin{cases} z^2/2, & |z| \leq 1, \\ |z| - 1/2, & |z| > 1. \end{cases} \qquad R_\tau(h) = \mathbb{E}\big[\ell_H(h(X) - Y) \wedge \tau\big]. \tag{15.42}$$

Theorem 15.6.1 (Single-index prediction by staged SGD, informal form)**.** *Suppose* $\mathcal{L}(X) = N(0, I_d)$*, and suppose the response follows the single-index model*

$$Y = f(u^\top X) + \varepsilon, \qquad \|u\|_2 = 1, \tag{15.43}$$

where the noise ε *is independent of* X*, and where* f *is monotone and twice differentiable,* $|f(0)| < 1$*, the second derivative* f'' *has at most polynomial growth, and* ε *is* 1*-sub-Gaussian. The algorithm uses the two stages in Eqs. (15.40) and (15.41). For a sufficiently large second-stage iteration count* T' *and under the specified initialization and tuning assumptions, its returned predictor* $\widehat{y}_{T,T',m}$ *satisfies, with high probability,*

$$R_\tau(\widehat{y}_{T,T',m}) - \mathbb{E}[\ell_H(\varepsilon)] \le \widetilde{C}\left(\sqrt{\frac{d + \log(1/\delta)}{T}} + \sqrt{\frac{\log(1/\delta)}{m}}\right), \tag{15.44}$$

where R_τ *is truncated population Huber risk,* ℓ_H *is the Huber loss, and* $\widetilde{C}$ *suppresses logarithmic factors and constants determined by the fixed link, truncation level, and tuning parameters.*

For a fixed monotone polynomial link and fixed target accuracy, with m and T' chosen sufficiently large, the required number T of distinct samples is $\widetilde{O}(d)$. In contrast, learning a degree-p target with a rotationally invariant kernel can require $d^{\Omega(p)}$ samples.

The theorem gives a positive learning result for the two-stage procedure. The nearly linear dependence on d requires Gaussian covariates, explicit weight decay, online first-stage samples, bias resampling, and separate training of the two layers. The theorem does not show that ordinary simultaneous training adapts across the structural families in Eq. (15.32).

15.7 Optional reading

The material through Eq. (15.44) was presented in class. The optional sections compare other feature-adaptation mechanisms, state the entropy result that was not presented in class, and summarize what additional depth can change.

15.7.1 Additional feature-adaptation mechanisms

Classical structure-adaptive smoothers alternate local derivative estimation with an update of the smoothing geometry. Under model-specific smoothness, design, and identifiability conditions, such procedures can estimate a multi-index subspace before fitting the low-dimensional link. These classical procedures show that data-dependent directions are not unique to neural networks.

Variation-ball empirical risk minimization treats the representing measure in Eq. (15.27) as the fitted object. Statistical bounds can adapt to projection-pursuit targets with a small hidden dimension, but standard conditional-gradient implementations require a generally nonconvex neuron-selection oracle.

The *recursive feature machine* repeatedly fits a kernel predictor, computes the empirical average gradient outer product of that predictor, and uses the matrix to update the kernel metric. The method makes feature-metric adaptation explicit without training a neural network. The *neural feature ansatz* is the empirical relation

$$W_i^\top W_i \;\propto\; \frac{1}{n}\sum_{r=1}^{n} \nabla f_i(h_i(X_r))\nabla f_i(h_i(X_r))^\top \tag{15.45}$$

between the Gram matrix of a trained layer and the empirical AGOP of the remaining network. Experiments support the relation across several settings; available proofs use more restricted training schemes. The ansatz is an empirically supported hypothesis, not a general theorem about ordinary deep training.

Method	What changes with the data?	Positive statement	Main limitation under the stated assumptions
Fixed linear space or fixed RKHS	Coefficients only	Standard estimation and approximation guarantees	No learned directions or feature metric
Projection pursuit and structure-adaptive smoothing	Ridge directions, term count, or smoothing geometry	Greedy oracle rates; active-query and passive-regression guarantees	Specialized observations and models; continuous atom search is nonconvex
Variation-ball ERM	Representing measure over neurons	Structured statistical rate with hidden dimension s	Neuron-selection oracle is nonconvex
Recursive feature machine	Kernel metric through empirical AGOP	Explicit iterative dimension adaptation	Not ordinary neural training
Two-layer network trained by online SGD	First-layer weights through online SGD	Subspace convergence; single-index learning theorem	Gaussian model, weight decay, staged training

Table 1: Different methods adapt different objects and satisfy different guarantees; "neural networks versus classical methods" is too coarse.

15.7.2 Entropy limits for one regression procedure

The preceding results describe function classes on which particular methods perform well. A universal entropy theorem gives a converse question that applies to every regression procedure: if a method has risk $Cn^{-\gamma}$ at every sample size, how large can the set of targets receiving that guarantee be?

The entropy theorem assumes a known input law P_X. Let $X_1, \dots, X_n$ be independent with common law P_X, and let

$$Y_i = f(X_i) + \varepsilon_i, \qquad \mathcal{L}(\varepsilon_1, \dots, \varepsilon_n) = N(0,1)^{\otimes n}, \tag{15.46}$$

with the noise variables independent of the inputs. A regression procedure $\mathcal{A}$ produces an estimator $\widehat{f}_n$ at every sample size. Define

$$R_{\mathcal{A}}(f, n) = \mathbb{E}\big[\|\widehat{f}_n - f\|_2^2\big] \tag{15.47}$$

and, for fixed $A, C > 0$ and $0 < \gamma \le 1$, define its rate-γ success set

$$\mathcal{S}_{\mathcal{A}}(C, \gamma) = \big\{ f : \|f\|_2 \le A, \quad R_{\mathcal{A}}(f, n) \le Cn^{-\gamma} \text{ for every } n \ge 1 \big\}. \tag{15.48}$$

The requirement at every n matters. A procedure can accidentally perform well on a very large set at one fixed sample size without giving a uniform rate along the sample-size sequence.

For a subset $\mathcal{F} \subseteq L_2(P_X)$, let $M(\epsilon, \mathcal{F})$ be the logarithm of the largest number of elements of $\mathcal{F}$ separated pairwise by more than ϵ in $L_2(P_X)$. Thus M is packing entropy, not packing number.

Theorem 15.7.1 (A universal entropy bound)**.** *For every regression procedure $\mathcal{A}$ and every $0 < \gamma < 1$, there is a constant C' depending on A, C, and γ such that, for all sufficiently small $\epsilon > 0$,*

$$M\big(\epsilon, \mathcal{S}_{\mathcal{A}}(C, \gamma)\big) \le C' \epsilon^{-2(1-\gamma)/\gamma}. \tag{15.49}$$

For $\gamma = 1$, the corresponding upper bound has order $\log(1/\epsilon)$.

The proof turns the sequence of estimates produced by $\mathcal{A}$ into a joint coding distribution. The cumulative risk controls the Kullback–Leibler divergence between the coding distribution and every regression model in the success set. A standard testing lower bound then prevents the success set from containing too many well-separated functions. The argument is information-theoretic and does not depend on whether $\mathcal{A}$ is a neural network, a kernel method, or a classical smoother.

The exponent calibrates the theorem. Substituting the generic smoothness rate

$$\gamma_d = \frac{2\alpha}{2\alpha + d} \tag{15.50}$$

into Eq. (15.49) gives

$$\frac{2(1-\gamma_d)}{\gamma_d} = \frac{d}{\alpha}. \tag{15.51}$$

The packing entropy of a d-dimensional α-smooth ball has the same order $\epsilon^{-d/\alpha}$. A minimax-optimal smoothness-adaptive procedure can therefore attain the same entropy order. Serving that entire ball at its minimax rate does not preclude better performance on smaller subsets. At the slower d-dimensional rate, structured smooth targets may already belong to the full smoothness ball. At a faster rate, only a smaller structured subset receives the stronger guarantee. Even if an additional family lies outside the ball, adding the family need not change the leading entropy order of the success set. Saturating the entropy exponent therefore does not exhaust a procedure's possible successes.

For an s-dimensional multi-index class, the analogous rate $\gamma_s = 2\alpha/(2\alpha + s)$ gives entropy exponent s/α. Allowing all s-dimensional orientations does not change the leading exponent when d is fixed. An ϵ-net over the orientation matrices contributes a finite-dimensional term of order

$$O\big(ds \log(1/\epsilon)\big), \tag{15.52}$$

which is smaller than $\epsilon^{-s/\alpha}$ as $\epsilon \downarrow 0$. The entropy theorem therefore permits one procedure to serve all orientations and all smooth links at the s-dimensional rate. The theorem forbids the same rate on a class with substantially larger packing entropy, such as the full d-dimensional smoothness ball when $s < d$.

The calculation also explains why adaptation to unknown smoothness is possible. One estimator may attain the minimax rate separately on many nested smoothness classes. At any fixed rate, the corresponding success set must satisfy Eq. (15.49); slower rates permit larger sets. The constants need not be uniform across every class. A sharper negative result holds for specially constructed, mutually separated classes: no procedure can achieve each class's minimax risk with one uniform finite-sample constant for infinitely many such classes. The theorem does not rule out adaptation for particular countable collections.

Packing entropy measures the size of a success set but not its membership. Two procedures can both attain the maximal entropy order while one procedure's success set strictly contains the other's. The theorem also does not identify which functions a given procedure favors. The entropy bound therefore places a rate-dependent size limit on the success set of every learning procedure, including a neural network trained by gradient descent.

15.7.3 Maxisets and a concrete open question

A *maxiset* records the targets on which a fixed procedure attains a fixed rate. For a positive sequence $r = (r_n)_{n\geq 1}$, define

$$\mathrm{MS}(\mathcal{A}, r) = \left\{ f : \sup_{n\geq 1} \frac{\mathbb{E}[\|\widehat{f}_n - f\|_2^2]}{r_n} < \infty \right\}. \tag{15.53}$$

An ordinary risk upper bound proves that a named function class is contained in the maxiset. An exact maxiset characterization also needs a converse: targets outside a specified approximation or regularity space cannot attain the rate under that procedure. Such characterizations are known for several wavelet-thresholding and model-selection procedures.

Maxisets at different rates have a natural nesting. If $r_n \leq C_0 s_n$ for every n, then

$$\mathrm{MS}(\mathcal{A}, r) \subseteq \mathrm{MS}(\mathcal{A}, s). \tag{15.54}$$

Faster rates have smaller success sets. A smoothness-adaptive procedure has the relevant smoothness classes inside these nested sets. The same procedure may also contain projection-pursuit or multi-index targets at faster rates when those targets add sufficiently little packing entropy.

For neural networks, the important unresolved problem is algorithm-specific: characterize the rate-indexed maxisets of a fixed architecture, initialization, regularization rule, and gradient-training algorithm. Does ordinary training simultaneously adapt to generic smoothness, ridge structure, and multi-index structure? Do its success sets attain the maximal entropy order? Do they contain the success sets of projection pursuit or a recursive feature machine? Existing approximation theorems and restricted feature-learning theorems do not answer these questions.

15.7.4 Related feature-learning theorems

The first theorem separates one neural feature-learning scheme from a fixed kernel regime.

Theorem 15.7.2 (Polynomial feature learning, informal form)**.** *Let $\mathcal{L}(X) = N(0, I_d)$ and let $f(x) = g(Ux)$ be a degree-p polynomial, where U has a small number of rows. Under a nondegeneracy condition on the expected Hessian, a two-layer training scheme learns the hidden subspace with sample complexity of leading order d^2. The corresponding fixed-kernel regime requires order d^p samples.*

The training scheme removes the linear component of the target, takes one carefully scaled first-layer gradient step, resamples biases, and then trains the output layer. The theorem proves feature learning but does not analyze ordinary joint training.

The second theorem, discussed in the optional material of Lecture 13, takes a different route.

Theorem 15.7.3 (Intrinsic-dimensional rate from pruned gradient training, informal form)**.** *Suppose that the input law is supported in a shrinking neighborhood of a low-dimensional Lipschitz manifold and that the regression function is smooth in the ambient coordinates. A prescribed randomly pruned sigmoid network, output truncation, and full-batch gradient descent with data-dependent step sizes achieve, up to logarithmic factors, the intrinsic-dimensional rate*

$$\mathbb{E}\big[\|\widehat{f}_n - f\|_2^2\big] \leq C(\log n)^c n^{-2p/(2p+d^*)}, \tag{15.55}$$

where p is the assumed smoothness and d^* is the manifold dimension used in tuning the estimator.

The theorem supplies a rate for a concrete gradient-based neural estimator. The assumptions concern input geometry rather than hierarchical composition, and the result does not isolate an advantage caused by depth or by feature movement. The theorem does not compare the neural estimator with non-neural manifold-regression methods, so the rate alone establishes no neural advantage.

15.7.5 What depth changes

Lecture 14 gave an exact representation separation: repeated composition of the tent map has a linear-size deep representation, while a one-hidden-layer ReLU network needs exponentially many units. Approximation theory contains two stronger forms of the separation.

Proposition 15.7.4 (Two depth separations, informal form)**.** *There are oscillatory target families represented by polynomial-size deep ReLU networks for which networks restricted to a specified smaller depth need exponentially many units to attain a fixed approximation error. There are also radial functions in growing dimension that three-layer networks approximate with polynomial size, while every two-layer approximation at the stated accuracy requires exponential width.*

The proposition establishes a representation advantage for depth. The separations do not show that gradient descent finds the deep representation from data.

Hierarchical composition gives a statistical counterpart to the depth separations. Consider a target of the form

$$f = g_q \circ g_{q-1} \circ \cdots \circ g_0, \tag{15.56}$$

where each coordinate of g_i depends on at most t_i coordinates of its input and has Hölder smoothness β_i. Later nonsmooth maps can reduce the effective smoothness of an earlier constituent. Define

$$\beta_i^* = \beta_i \prod_{\ell=i+1}^{q} (\beta_\ell \wedge 1). \tag{15.57}$$

Theorem 15.7.5 (Regression rate for hierarchical compositions, informal form)**.** *Under bounded-domain and noise assumptions, an empirical risk minimizer over a suitably sparse ReLU network class with depth of order* $\log n$ *satisfies for constants* $C, c > 0$*,*

$$\mathbb{E}\big[\|\widehat{f}_n - f\|_2^2\big] \le C(\log n)^c \max_{0\le i\le q} n^{-2\beta_i^*/(2\beta_i^*+t_i)}. \tag{15.58}$$

If $t_i \le \min(d_0, \dots, d_{i-1})$ *for* $i = 1, \dots, q$*, the polynomial part of the rate matches the minimax lower bound for the stated composition class.*

Each part of the network approximates one low-dimensional constituent, and successive layers compose these approximations. Depth therefore converts a specified compositional description into an estimator whose rate is governed by the hardest local constituent rather than by the ambient dimension.

The statistical conclusion has two important boundaries. The estimator is a global empirical risk minimizer over a tuned network class; the theorem does not show that ordinary gradient training discovers an unknown composition. The theorem also does not prove that every shallow neural estimator must have a slower rate. Polynomial-spline estimators can exploit related hierarchical interaction models. The statistical result shows that deep networks can exploit hierarchical composition. The result does not give neural networks exclusive access to the structure.

Depth can also change the regularizer induced by parameter norm without changing the represented function class. The relevant matrix factorization identity makes the effect precise.

Proposition 15.7.6 (Factorization cost)**.** *Let A have rank at most every intermediate width. For $L \geq 2$,*

$$\inf_{A=A_L\cdots A_1} \frac{1}{L}\sum_{\ell=1}^{L} \|A_\ell\|_F^2 = \|A\|_{S_{2/L}}^{2/L}, \tag{15.59}$$

where $\|A\|_{S_p} = (\sum_j \sigma_j(A)^p)^{1/p}$ is the Schatten p-quasi-norm.

For a network with several linear input layers followed by one ReLU layer, collapsing the linear layers leaves a shallow ReLU function. Additional linear layers therefore leave the represented function class unchanged while the sum of squared parameter norms increasingly favors low-rank input matrices as L grows. The formal analysis connects the induced cost to single-index and multi-index functions; experiments also study standard training. No general statistical rate for gradient descent follows from the representation-cost result.

The layerwise neural feature ansatz in Eq. (15.45) supplies a complementary empirical claim: trained layers align with the AGOP of the remaining network. Together, the results suggest two precise roles for depth. Nonlinear composition can reduce representation size, and even additional linear factorization can alter the implicit or explicit regularization. A general theorem converting either mechanism into a statistical and computational advantage under standard deep-network training remains unavailable.

15.8 Take-home messages

1. **Approximation, constructive approximation, and learning answer different questions.** Approximation asks whether a good function exists in the family. Constructive approximation specifies information about the target and an algorithm. Regression learning specifies noisy observations and a prediction risk.

2. **Bounded integral representations have accurate finite approximations.** If $f = \int \phi_\theta \, d\mu(\theta)$ in a Hilbert space and $\|\phi_\theta\| \leq 1$, then m signed atoms attain error at most $\|\mu\|_{\mathrm{TV}}/\sqrt{m}$. Random sampling appears in the proof; the lemma's conclusion is deterministic. The sampling law depends on the unknown representing measure.

3. **Density answers the minimum approximation question.** The RBF and ridge density results show that the error can be made arbitrarily small. For Hölder balls, quantitative approximation theory relates error to the number of terms; suitable neural-network constructions give corresponding smoothness-dependent rates.

4. **Barron's theorem gives a target-dependent representation advantage.** On a spectral Barron ball, a shallow ridge network attains an $m^{-1/2}$ approximation rate, while the Kolmogorov m-width for fixed linear spaces can decay only as $m^{-1/d}$. The comparison depends on the target class, norm, and target-dependent choice of ridge atoms.

5. **A fixed kernel restricts the representing measure.** A fixed feature law τ permits measures of the form $d\mu = p\, d\tau$ with $p \in L^2(\tau)$. A variation representation may use a measure concentrated on features selected for the target. The larger representation class creates a feature-selection problem.

6. **Gradient information can reduce a multi-index problem from dimension d to dimension s.** For $f(x) = g(W^\top x)$, every gradient lies in $\mathrm{span}(W)$. Informative gradients and an SVD

recover that subspace when gradients are available. Passive regression must first estimate the derivative information accurately.

7. **The neural learning theorem covers a staged procedure under restrictive assumptions.** The procedure first updates only the first-layer weights, then freezes those weights, resamples the biases, and trains only the output weights. For a monotone single-index model, the procedure satisfies a prediction bound with nearly linear dependence on the ambient dimension. The result does not show that ordinary simultaneous training realizes the general Barron representation advantage.

Endnotes

1. Uniform error is the appropriate objective when every input must receive a guarantee. Such bounds also arise in dynamic programming when a worst-case value-function error must remain controlled through repeated Bellman updates. The sampling lemma below extends to any Hilbert space, but L^∞ is not a Hilbert space, so the same variance calculation does not directly give an L^∞ bound.

2. The usual definition starts from a countably additive real-valued set function. The Jordan decomposition theorem writes every such measure as $\mu = \mu^+ - \mu^-$ for mutually singular positive measures. Consequently, $|\mu| = \mu^+ + \mu^-$ and $d\mu = s\, d|\mu|$ for a measurable sign function $s \in \{-1, 1\}$, up to a $|\mu|$-null set. Thus the sign-times-positive- measure definition loses no finite signed measures.

3. One may instead use the spectral moment as a norm modulo constants. The network represents c exactly through its output bias, so the approximation rate depends only on the spectral moment. Different Fourier normalizations change constants but not the approximation exponents.

4. The comparison is stated for the Barron ball of radius C. The theorem does not assert that C stays bounded along a high-dimensional sequence of targets.

5. Appendix 15.E gives a more general first-stage result: online SGD contracts the component of W orthogonal to a teacher subspace. The appendix also records the population-gradient identity used in the analysis.

Bibliographical notes

The minimax Hölder benchmark in Eq. (15.8) is due to the nonparametric regression theory developed by Stone [1982]. The infimum there is over statistical decision rules, without a computational restriction.

The Hilbert-space sampling argument is a standard probabilistic-method form of finite approximation in a convex or variation ball. The wavelet route to Corollary 15.2.3 follows Delyon et al. [1995]: a target-dependent continuous wavelet transform supplies the representing measure, and sampling that measure supplies a finite scale–location approximation. The appendices record the reconstruction and Fourier calculations needed for the Gaussian example. Countable dyadic wavelet bases and multiresolution analysis are developed by Mallat [1989]; Daubechies [1988] constructs compactly supported orthonormal wavelets with high regularity. The RBF corollary uses the continuous-wavelet construction rather than the full dyadic theory. Classical quantitative

approximation rates for Hölder and related smoothness classes are discussed by DeVore [1998]; Petersen and Zech [2026] survey corresponding neural-network approximation results.

The shallow-network upper bound and fixed-linear-space lower bound in Theorem 15.3.2 are normalized versions of results in Barron [1993]. Fourier conventions and the precise Barron norm change constants. The anchored norm in Eq. (15.20) includes the constant part so that the displayed quantity is a norm rather than a seminorm. Barron's lower-bound proof uses an orthogonal family of Fourier modes. The resulting fixed-space quantity is a Kolmogorov width. Classical nonlinear m-term approximation and its relation to widths are surveyed by DeVore [1998].

Bach [Bach, 2017] develops the fixed-feature RKHS and target-dependent variation-space comparison used after Theorem 15.3.2. Bach also analyzes structured statistical rates and conditional-gradient selection from an infinite neuron dictionary. The selection step remains a nonconvex optimization problem for the ridge dictionary. The measure-valued formula in Eq. (15.10) is standard in infinite-width two-layer network limits; Mei et al. [2019] analyze corresponding mean-field dynamics.

Projection pursuit regression was introduced by Friedman and Stuetzle [1981]. Jones [1992] proves the Hilbert-space greedy lemma used in Appendix 15.C and applies the method to projection pursuit and neural-network training. Barron et al. [2008] develop several greedy variants and their statistical consequences. The constructive guarantees depend on the quality and cost of the atom-selection step.

The active point-query literature makes the gradient–SVD mechanism quantitative. Cohen et al. [2012] study ridge-function recovery, Fornasier et al. [2012] combine randomized queries with sparse recovery, and Tyagi and Cevher [2012] study dense multi-index and additive-ridge models. Appendix 15.C gives an elementary coordinate finite-difference version and an obstruction showing why an informative-gradient condition is necessary.

Average derivative and structure-adaptive dimension-reduction methods use related gradient information under passive noisy sampling. Hristache et al. [2001], Dalalyan et al. [2008], and Xia et al. [2002] analyze different local-smoothing procedures and different dimension, regularity, and identifiability regimes. The optional discussion of the AGOP, recursive feature machines, and the neural feature ansatz follows Radhakrishnan et al. [2024]; the empirical ansatz and the available restricted theorems must be distinguished.

The population-gradient decomposition, online contraction theorem, and staged single-index prediction theorem are due to Mousavi-Hosseini et al. [2023]. The polynomial feature-learning comparison is from Damian et al. [2022]. The pruned sigmoid-network theorem is from Kohler et al. [2026]. The entropy limit and its adaptive-estimation discussion are from Yang [2007]; the maxiset terminology and comparison with minimax analysis are developed by Kerkyacharian and Picard [2002].

The optional depth separations are due to Telgarsky [2016] and Eldan and Shamir [2016]. The effect of additional linear layers on representation cost is analyzed by Parkinson et al. [2025]. Hierarchical interaction models and corresponding neural or spline estimators are studied by Kohler and Krzyżak [2017], Bauer and Kohler [2019], Schmidt-Hieber [2020], and Kohler and Langer [2021]. Petersen and Zech [2026] provide a broad reference for modern neural-network approximation theory.

Glossary

Dictionary A finite, countable, or continuous collection of candidate atoms used to represent or approximate functions.

Constructive approximation Approximation by an algorithm that receives specified information

about the target, makes finitely many oracle calls and arithmetic operations, and returns an approximating function.

Signed measure A real-valued measure represented in Definition 15.2.1 as a measurable sign function times a finite positive measure.

Total variation measure and norm For $d\mu = s\, d|\mu|$ with $|s| = 1$, the positive measure $|\mu|$ is the total variation measure and $\|\mu\|_{\mathrm{TV}} = |\mu|(\Theta)$ is the total variation norm.

Hölder continuity The condition $|h(x) - h(x')| \le L\|x - x'\|^{\beta}$ for $0 < \beta \le 1$. Lipschitz continuity is the case $\beta = 1$.

Nonlinear m-term approximation Approximation by a sum of at most m dictionary atoms whose identities and coefficients may depend on the target. The selection makes the approximation map nonlinear.

Kolmogorov m-width The smallest worst-case approximation error achieved by an m-dimensional linear space, as defined in Eq. (15.24).

Wavelet A function in a translation–dilation system that satisfies a reconstruction property. Compact support provides localization, while vanishing moments suppress local low-degree polynomial behavior; the resulting coefficients record both location and scale.

Radial-basis-function network A finite sum $\sum_{j=1}^{m} c_j \varphi(s_j(x - t_j))$ with a fixed radial function φ and trainable coefficients, scales, and centers.

Atomic norm The smallest total variation of a signed measure representing a function as an integral of dictionary atoms. For a countable basis with a unique coefficient expansion, the atomic norm is the ℓ_1 norm of the coefficients.

Representation advantage A smaller approximation error or smaller representation size than a named comparison class, without a claim that the representation can be estimated from data.

Learning advantage A risk or sample-complexity improvement for a specified estimator and data model relative to a specified competing method.

Ridge function A function of the form $x \mapsto g(w^{\top} x)$. A multi-index function uses several linear projections.

Projection pursuit A regression method that fits a sum of ridge functions and estimates their directions and link functions.

Atom-selection oracle An operation that returns a dictionary atom maximizing its correlation with a supplied residual. Approximate variants require an additional oracle-error term in the approximation guarantee.

Active point-query model An observation model in which an algorithm chooses each input at which the target function is evaluated. The model differs from passive regression, where the input locations are sampled from a fixed distribution.

Variation space The class of integral feature representations whose signed representing measure has finite total variation.

Spectral Barron class A class of restrictions of ambient Fourier representations with bounded anchored spectral norm, as in Eqs. (15.20) and (15.21).

Variation Barron space A variation space generated by a ridge-feature dictionary. Spectral Barron conditions can imply membership in a variation Barron space, but the two definitions are not identical. The literature uses several feature normalizations and hence several related Barron norms.

Mean-field representation An infinite-width two-layer network written as an integral over a probability measure on neuron parameters. Mean-field training theory also specifies a geometry and an evolution equation for that measure.

Average gradient outer product The matrix $\mathbb{E}[\nabla f(X)\nabla f(X)^\top]$, whose leading eigenspace records directions along which f varies on average.

Average derivative estimation A family of dimension-reduction methods that estimate averages or moments of target derivatives from regression data in order to recover predictive directions.

Recursive feature machine A kernel procedure that alternates predictor fitting with an update of the kernel metric to the fitted predictor's empirical AGOP.

Neural feature ansatz The empirical relation in Eq. (15.45) between a trained layer's Gram matrix and the AGOP of the remaining network.

Feature learning A data-dependent change in the features or feature metric used for prediction. The phrase alone does not specify a statistical guarantee.

Hierarchical composition model A function $f = g_q \circ \cdots \circ g_0$ in which each coordinate of g_i depends on only a bounded number of coordinates from the preceding layer. The local arities can be much smaller than the ambient input dimension.

Packing entropy The logarithm of the largest number of functions separated pairwise by more than a given distance.

Success set The functions on which a fixed procedure attains a stated risk bound at every sample size.

Maxiset The success set of a fixed estimation procedure at a fixed rate, usually characterized by both an upper bound and a converse.

15.9 Exercises

General instructions. Justify every answer, including yes/no answers, even when the question does not explicitly ask for a justification.

Exercise profiles. Filled mountains show difficulty on a four-mountain scale: one is a direct application, two uses a familiar method, three combines ideas, and four is a substantial challenge. Filled cogs show technical load on a three-cog scale: one is light, two is sustained, and three is dense. Filled hourglasses show estimated time: one is under 20 minutes, two is 20–45 minutes, three is 45–90 minutes, and four is more than 90 minutes. Pale pictograms show the unused part of each scale. The work row uses the canonical order proof, calculation, construction, computation, and interpretation; pale icons mark work types not used. ◎ marks the intended

learning outcome. Time estimates cover the non-optional parts. Ratings are not prerequisites or grading weights. A superscript asterisk marks an exercise based on supplementary lecture material; it is not a difficulty rating and does not make the exercise optional when assigned.

Integral representations and density

Exercise 1: Monte Carlo sparsification in a Hilbert space.

Derive the $m^{-1/2}$ approximation rate and adapt the sampling law to nonuniform bounds on the dictionary atoms.

Let $\mathcal{H}$ be a real Hilbert space, let $\|\phi_\theta\|_{\mathcal{H}} \le 1$, and suppose

$$f = \int \phi_\theta \, d\mu(\theta), \qquad B = \|\mu\|_{\mathrm{TV}} \in (0, \infty).$$

Set $\pi = |\mu|/B$ and let $s = d\mu/d|\mu|$. Let $\Theta_1, \dots, \Theta_m$ be independent random elements with common law π, and define

$$Z_j = B s(\Theta_j) \phi_{\Theta_j}, \qquad \widehat{f}_m = \frac{1}{m} \sum_{j=1}^{m} Z_j.$$

(a) Prove that $\mathbb{E}[Z_j] = f$ and $\|Z_j\|_{\mathcal{H}} \le B$ almost surely.

(b) Expand $\mathbb{E}[\|\widehat{f}_m - f\|_{\mathcal{H}}^2]$ using the inner product in $\mathcal{H}$. Show that the cross terms vanish and conclude that

$$\mathbb{E}[\|\widehat{f}_m - f\|_{\mathcal{H}}^2] \le \frac{B^2}{m}.$$

(c) Identify the probabilistic-method step that proves the existence of a deterministic m-term approximation. Then identify two pieces of information that the construction uses but a learner observing noisy input–output pairs does not automatically possess.

(d) Suppose instead that $q : \Theta \to (0, \infty)$ is measurable, $\|\phi_\theta\|_{\mathcal{H}} \le q(\theta)$ for $|\mu|$-almost every θ, and

$$A = \int_\Theta q(\theta) \, d|\mu|(\theta) < \infty.$$

Let $\widetilde{\pi}$ be the probability measure defined by $d\widetilde{\pi} = q \, d|\mu|/A$. For independent $\widetilde{\Theta}_1, \dots, \widetilde{\Theta}_m$ with common law $\widetilde{\pi}$, define

$$\widetilde{Z}_j = A \, s(\widetilde{\Theta}_j) \frac{\phi_{\widetilde{\Theta}_j}}{q(\widetilde{\Theta}_j)}.$$

Prove that the average of the $\widetilde{Z}_j$ is unbiased for f and has expected squared error at most A^2/m. State how the original construction is recovered as a special case.

Exercise 2: Density does not determine an approximation rate.

Distinguish density of an approximation family from the rate for an individual target.

Let $\mathcal{H} = \ell_2(\mathbb{N})$ with standard orthonormal basis $(e_j)_{j \ge 1}$, and let $V_m = \mathrm{span}\{e_1, \dots, e_m\}$.

(a) Prove that $\bigcup_{m \ge 1} V_m$ is dense in $\mathcal{H}$.

(b) Fix $p > 0$ and define $f^{(p)} \in \mathcal{H}$ by $f_j^{(p)} = j^{-p-1/2}$. Prove that

$$\inf_{v \in V_m} \|f^{(p)} - v\|_2 = \left(\sum_{j>m} j^{-2p-1} \right)^{1/2} \asymp m^{-p} \qquad \text{as } m \to \infty.$$

(c) Define $f_1^{(\log)} = 0$ and

$$f_j^{(\log)} = \frac{1}{\sqrt{j} \log j}, \qquad j \geq 2.$$

Prove that $f^{(\log)} \in \mathcal{H}$ and

$$\inf_{v \in V_m} \|f^{(\log)} - v\|_2 \asymp \frac{1}{\sqrt{\log m}} \qquad \text{as } m \to \infty.$$

(d) Explain which statement is common to all targets in parts (b) and (c), and which quantitative conclusion depends on the target. Use the two examples to state precisely why density alone does not imply a polynomial approximation rate.

Hints The best approximation in V_m is the orthogonal projection onto the first m coordinates. Compare each tail sum with the corresponding improper integral.

Exercise 3: From Hölder approximation to regression risk.

Derive the smoothness-dependent regression rate by balancing approximation and estimation errors.

Suppose an estimator based on an m-dimensional approximation space satisfies

$$\sup_{f \in \mathcal{H}_d^\alpha(L)} \mathbb{E}\big[\|\widehat{f}_{n,m} - f\|_{L^2(P_X)}^2\big] \leq A m^{-2\alpha/d} + B\frac{m}{n}, \tag{15.60}$$

where $A, B > 0$ do not depend on m or n.

(a) Balance the two terms in Eq. (15.60). Show that a choice

$$m \asymp n^{d/(2\alpha+d)}$$

gives risk of order $n^{-2\alpha/(2\alpha+d)}$. Explain why rounding m to a positive integer does not change the rate.

(b) Suppose that the target has the form $f(x) = g(W^\top x)$, the matrix $W \in \mathbb{R}^{d \times s}$ is known, and the same approximation and estimation bounds hold in the reduced s-dimensional coordinates. Derive the choices

$$m \asymp n^{s/(2\alpha+s)}, \qquad \mathbb{E}\big[\|\widehat{f}_{n,m} - f\|_{L^2(P_X)}^2\big] \lesssim n^{-2\alpha/(2\alpha+s)}.$$

(c) Explain why the calculation in part (b) proves neither the s-dimensional rate for the full d-dimensional Hölder ball nor the existence of an algorithm that finds an unknown subspace from noisy regression data. Your explanation must identify the target-class restriction and the missing estimation task separately.

Adaptive atoms and fixed features

Exercise 4: Barron approximation versus a fixed linear space.

Prove the orthogonal-family lower-bound mechanism and compare it with target-dependent ridge approximation.

Let $e_1, \dots, e_{2m}$ be orthonormal elements of a Hilbert space $\mathcal{H}$, let $V \subseteq \mathcal{H}$ satisfy $\dim(V) \le m$, and let P_V be the orthogonal projection onto V.

(a) Prove that

$$\sum_{j=1}^{2m} \|P_V e_j\|_{\mathcal{H}}^2 \le m.$$

Conclude that some $j \in \{1, \dots, 2m\}$ satisfies

$$\inf_{v \in V} \|e_j - v\|_{\mathcal{H}} \ge \frac{1}{\sqrt{2}}.$$

(b) Let $\mathcal{F} \subseteq \mathcal{H}$. Suppose that, for every $m \ge 1$, there are $2m$ orthonormal elements $e_1, \dots, e_{2m}$ such that $\rho_m e_j \in \mathcal{F}$ for every j, where $\rho_m > 0$. Prove that the Kolmogorov width satisfies

$$d_m(\mathcal{F}; \mathcal{H}) \ge \frac{\rho_m}{\sqrt{2}}.$$

(c) Assume that every $f \in \mathcal{B}_C(Q)$ has an m-unit ridge approximation satisfying

$$\|f - f_m\|_2 \le 2Cm^{-1/2},$$

and that every fixed m-dimensional linear space V_m satisfies

$$\sup_{f \in \mathcal{B}_C(Q)} \inf_{g \in V_m} \|f - g\|_2 \ge \kappa(C/d) m^{-1/d}.$$

Give a sufficient width for the ridge approximation to attain error at most ϵ. Give a necessary dimension for a fixed linear space to attain worst-case error at most ϵ.

(d) Compare the dependence on ϵ when $d = 2$ and when $d = 20$. Which choices in the ridge approximation are allowed to depend on the target, and which choices in the linear lower bound are fixed?

(e) State why the comparison proves neither a smaller minimax regression risk for a neural-network estimator nor a contradiction with the minimax lower bound for a full d-dimensional smoothness ball. Identify the missing learning statement in the first case and the different target classes in the second case.

HINTS For part (a), expand P_V in an orthonormal basis of V and apply Bessel's inequality to each basis vector.

Exercise 5: A fixed feature law can give a strictly smaller space.

Use a discrete dictionary to distinguish variation representations from fixed-feature RKHS representations.

Let $\mathcal{H} = \ell_2(\mathbb{N})$, let $(e_j)_{j\ge1}$ be its standard orthonormal basis, and set $\tau_j = 2^{-j}$. Use the dictionary $\phi_j = e_j$. Define

$$\mathcal{F}_1 = \left\{ f = \sum_{j\ge1} \mu_j e_j : \sum_{j\ge1} |\mu_j| < \infty \right\},$$
$$\mathcal{F}_{2,\tau} = \left\{ f = \sum_{j\ge1} \tau_j p_j e_j : \sum_{j\ge1} \tau_j p_j^2 < \infty \right\},$$

with the norms defined by the corresponding coefficient costs.

(a) Use uniqueness of coordinates in the orthonormal basis to prove

$$\|f\|_{\mathcal{F}_1} = \sum_{j\ge1} |f_j|, \qquad \|f\|^2_{\mathcal{F}_{2,\tau}} = \sum_{j\ge1} \frac{f_j^2}{\tau_j}.$$

(b) Prove that

$$\|f\|_{\mathcal{F}_1} \le \|f\|_{\mathcal{F}_{2,\tau}} \qquad \text{for every } f \in \mathcal{F}_{2,\tau},$$

and conclude that $B_{\mathcal{F}_{2,\tau}} \subseteq B_{\mathcal{F}_1}$.

(c) Define $f_j = 2^{-j/2}/\sqrt{j}$. Prove that $f \in \mathcal{F}_1$ but $f \notin \mathcal{F}_{2,\tau}$. Set $g = f/\|f\|_{\mathcal{F}_1}$ and prove that $g \in B_{\mathcal{F}_1} \setminus B_{\mathcal{F}_{2,\tau}}$. Conclude that both inclusions in Eq. (15.30) can be strict.

(d) For the function in part (c), identify the signed-measure coefficients μ_j and the density coefficients p_j. Explain why their summability properties exhibit the restriction imposed by the fixed feature law τ.

Recovering low-dimensional structure



Exercise 6: When the AGOP recovers a multi-index subspace.

Prove the population identity behind gradient-based dimension reduction and identify the nondegeneracy condition required for subspace recovery.

Let X be a random vector in $\mathbb{R}^d$, let $W \in \mathbb{R}^{d\times s}$ have full column rank, and suppose $f(x) = g(W^\top x)$ with differentiable g. Assume the displayed expectations below are finite.

(a) Use the chain rule to prove

$$G_f := \mathbb{E}[\nabla f(X)\nabla f(X)^\top] = W G_g W^\top, \qquad G_g := \mathbb{E}[\nabla g(W^\top X)\nabla g(W^\top X)^\top].$$

(b) Prove that $\operatorname{range}(G_f) \subseteq \operatorname{range}(W)$. Show that equality holds when G_g is positive definite.

(c) Give an example in which W has two columns but G_f has rank one. What does the example say about recovering a declared index space from gradients? Explain why an arbitrary interpolant can have a different AGOP from the regression target and why a recursive feature machine therefore needs the AGOP of an accurate predictor.

Appendix-based extensions

Exercise 7: From wavelet coefficient decay to an approximation rate.[*]

Convert scale-wise wavelet coefficient bounds into an m-term approximation rate.

Let $(\psi_{j,k})$ be an orthonormal wavelet basis of $L^2([0,1])$. Suppose that, at every level $j \geq 0$, at most $C_0 2^j$ coefficients are nonzero and

$$|\langle f, \psi_{j,k} \rangle_{L^2}| \leq C_1 2^{-j(r+1/2)}, \qquad r > 0.$$

Let f_J retain the coarse-scale terms and all wavelet coefficients from levels $0, \dots, J-1$.

(a) Use orthonormality to prove

$$\|f - f_J\|_{L^2}^2 \leq \frac{C_0 C_1^2}{1 - 2^{-2r}} \, 2^{-2rJ}.$$

(b) Suppose the number m_J of retained coefficients satisfies $m_J \asymp 2^J$. Prove that

$$\|f - f_J\|_{L^2} = O(m_J^{-r}).$$

(c) State how Proposition 15.A.1 supplies the hypothesis of this exercise when the rth derivative is uniformly bounded on the intervals appearing in that proposition. Identify the step in parts (a)–(b) that converts local smoothness information into a global L^2 approximation bound.

Exercise 8: Greedy approximation with an approximate oracle.[*]

Determine how accurately a correlation oracle must be solved to preserve the $m^{-1/2}$ greedy approximation rate.

Use the setting of Proposition 15.C.1. Thus $\|\phi\|_{\mathcal{H}} \leq 1$ for $\phi \in \mathcal{D}$, $f \in \mathcal{K} = B\,\overline{\operatorname{conv}}(\mathcal{D} \cup -\mathcal{D})$, and $f_0 = 0$. At step $m \geq 1$, write $r_{m-1} = f - f_{m-1}$ and suppose the oracle returns $\phi_m \in \mathcal{D} \cup -\mathcal{D}$ satisfying

$$\langle r_{m-1}, B\phi_m \rangle_{\mathcal{H}} \geq \sup_{\phi \in \mathcal{D} \cup -\mathcal{D}} \langle r_{m-1}, B\phi \rangle_{\mathcal{H}} - \delta_m,$$

where $\delta_m \geq 0$. Set

$$f_m = \left(1 - \frac{1}{m+1}\right) f_{m-1} + \frac{1}{m+1} B\phi_m.$$

(a) Prove by induction that $f_m \in \mathcal{K}$ and hence $\|B\phi_m - f_{m-1}\|_{\mathcal{H}} \leq 2B$.

(b) Prove the recursion

$$\|r_m\|_{\mathcal{H}}^2 \leq \left(1 - \frac{2}{m+1}\right) \|r_{m-1}\|_{\mathcal{H}}^2 + \frac{2\delta_m}{m+1} + \frac{4B^2}{(m+1)^2}.$$

(c) Suppose that $\delta_m \leq B^2/(m+1)$ for every $m \geq 1$. Prove by induction that

$$\|f - f_m\|_{\mathcal{H}}^2 \leq \frac{6B^2}{m+1}.$$

(d) State the required order of the oracle's additive accuracy at iteration m. Explain what the result controls and what additional information would be needed to turn the iteration bound into an arithmetic running-time bound for a continuous ridge dictionary.

HINTS For part (b), compare the approximate oracle value with the value of the linear functional at $f \in \mathcal{K}$.

Exercise 9: Stable finite differences for active subspace recovery.[*]

Quantify the tradeoff between truncation error and observation error in finite-difference gradient estimates.

Use the active-query setting of Theorem 15.C.2. Suppose $\|\nabla^2 f(x)\|_{\mathrm{op}} \le M$ on every line segment used below, where $M > 0$. Instead of $f(x)$, the oracle returns

$$y(x) = f(x) + \xi(x), \qquad |\xi(x)| \le \zeta,$$

where $\zeta > 0$ and the errors may be chosen adversarially. For base points $x_1, \dots, x_q$, define

$$\widetilde{G}_{\ell j} = \frac{y(x_j + he_\ell) - y(x_j)}{h}, \qquad \ell = 1, \dots, d, \quad j = 1, \dots, q.$$

(a) Prove the entrywise bound

$$|\widetilde{G}_{\ell j} - \partial_\ell f(x_j)| \le \frac{Mh}{2} + \frac{2\zeta}{h}.$$

(b) Minimize the right-hand side over $h > 0$. Show that

$$h_\star = 2\sqrt{\frac{\zeta}{M}}, \qquad \frac{Mh_\star}{2} + \frac{2\zeta}{h_\star} = 2\sqrt{M\zeta}.$$

Assume that the query domain permits the step size $h_\star$.

(c) Let G be the exact gradient matrix. Prove

$$\|\widetilde{G} - G\|_{\mathrm{op}} \le 2\sqrt{M\zeta dq}$$

when $h = h_\star$. If $\sigma_k(G) = \lambda_\star$ and $2\sqrt{M\zeta dq} < \lambda_\star$, use Theorem 15.C.2 to prove

$$\|\widehat{P} - P\|_{\mathrm{op}} \le \frac{2\sqrt{M\zeta dq}}{\lambda_\star - 2\sqrt{M\zeta dq}}.$$

(d) Compare the noiseless limit $\zeta = 0$ with the case $\zeta > 0$. Explain why decreasing h improves the truncation error but can worsen the effect of inaccurate point evaluations.

Exercise 10: Contraction toward a teacher subspace.[*]

Separate contraction of irrelevant weight components from recovery and prediction guarantees.

Use the population-gradient decomposition in Eq. (15.140). Let $U \in \mathbb{R}^{s \times d}$ satisfy $UU^\top = I_s$, set $\Pi_\perp = I_d - U^\top U$, and suppose population gradient descent satisfies

$$W^{(t+1)} = W^{(t)} - \eta\big(A_t W^{(t)} + D_t U\big),$$

where $W^{(t)} \in \mathbb{R}^{m \times d}$, $D_t \in \mathbb{R}^{m \times s}$, and each $A_t \in \mathbb{R}^{m \times m}$ is symmetric with

$$\mu I_m \preceq A_t \preceq L I_m, \qquad 0 < \mu \le L, \qquad 0 < \eta \le \frac{1}{L}.$$

Define $W_\perp^{(t)} = W^{(t)}\Pi_\perp$.

(a) Prove

$$W_\perp^{(t+1)} = (I_m - \eta A_t) W_\perp^{(t)}.$$

Identify the algebraic property that removes the term $D_t U$.

(b) Prove

$$\|W_\perp^{(t+1)}\|_{\mathrm{F}} \le (1-\eta\mu)\|W_\perp^{(t)}\|_{\mathrm{F}}$$

and hence

$$\|W_\perp^{(T)}\|_{\mathrm{F}} \le e^{-\eta\mu T}\|W_\perp^{(0)}\|_{\mathrm{F}}.$$

(c) Give one matrix whose rows lie in the teacher subspace but do not span that subspace. Use the example to explain why $\|W_\perp^{(T)}\|_{\mathrm{F}} \to 0$ does not prove recovery of every teacher direction. Then state which additional network quantities and error criterion would have to be controlled to obtain a prediction guarantee.

Further results from the literature

Exercise 11: Balancing variation-space approximation and estimation.*

Recover the hidden-dimension exponent by balancing a variation-norm budget against approximation error.

Let

$$\beta = \alpha + \frac{s-1}{2}, \qquad B_n(D) = A\left(\frac{D}{\sqrt{n}} + D^{-1/\beta}\log D\right), \qquad D \ge e,$$

where A is independent of D and n.

(a) Ignore the logarithmic factor and balance the two remaining terms. Derive the resulting choice of D and rate in n.

(b) Restore the logarithm at the level of order notation and show that the rate is

$$\widetilde{O}\left(n^{-1/(2\alpha+s+1)}\right).$$

For $\alpha = 1$, compare the exponents for a one-dimensional projection and for a generic d-dimensional target.

(c) Explain why replacing d by s is an adaptation statement rather than a guarantee over the full d-dimensional smoothness class. Identify the continuous atom-selection optimization required by a conditional-gradient implementation and explain why the statistical analysis does not supply a polynomial-time algorithm for that optimization.

Exercise 12: What the universal entropy bound permits.*

Use the entropy exponent to distinguish permissible structural adaptation from an impossible uniform rate on a larger class.

Suppose a regression procedure has squared-risk success set $\mathcal{S}_\gamma$ at rate $n^{-\gamma}$. The entropy theorem gives

$$M(\epsilon, \mathcal{S}_\gamma) \lesssim \epsilon^{-2(1-\gamma)/\gamma}, \qquad 0 < \gamma < 1.$$

Assume that the packing entropy of an r-dimensional α-smooth ball has order $\epsilon^{-r/\alpha}$.

(a) Substitute $\gamma_r = 2\alpha/(2\alpha+r)$ into the entropy exponent and show that the upper bound matches the smoothness-ball entropy order.

(b) Consider all functions $x \mapsto g(W^\top x)$, where g ranges over an s-dimensional α-smooth ball and W ranges over orthonormal $d \times s$ matrices. Suppose d is fixed and that covering the possible orientations contributes $O(ds\log(1/\epsilon))$ to the entropy. Explain why the entropy theorem permits the rate $n^{-2\alpha/(2\alpha+s)}$ uniformly over all orientations.

(c) Show why the same s-dimensional rate cannot hold uniformly over a full d-dimensional smoothness ball when $s < d$.

(d) Explain how one procedure can adapt to several unknown smoothness levels without contradicting the entropy theorem. Your explanation should use rate-indexed success sets and constants in addition to asymptotic notation.

15.A A basic wavelet coefficient estimate

The following one-dimensional estimate makes the role of vanishing moments explicit. Tensor-product and multivariate wavelets satisfy analogous bounds with multi-index derivatives.

Proposition 15.A.1 (Decay in a smooth region)**.** *Let $\psi \in L_1(\mathbb{R})$ be compactly supported and satisfy*

$$\int_{\mathbb{R}} u^q \psi(u)\, du = 0, \qquad q = 0, \dots, r-1. \tag{15.61}$$

Define $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$. If f is r times continuously differentiable on $I_{j,k} = \operatorname{conv}(\{2^{-j}k\} \cup \operatorname{supp}(\psi_{j,k}))$, then

$$\left|\langle f, \psi_{j,k}\rangle_{L_2}\right| \le \frac{2^{-j(r+1/2)}}{r!} \left(\int_{\mathbb{R}} |u|^r |\psi(u)|\, du\right) \sup_{x \in I_{j,k}} |f^{(r)}(x)|. \tag{15.62}$$

Proof. Let $x_0 = 2^{-j}k$, and let P_{r-1} be the Taylor polynomial of f at x_0 through degree $r-1$. The moment conditions and the change of variable $u = 2^j x - k$ give $\langle P_{r-1}, \psi_{j,k}\rangle_{L_2} = 0$. Taylor's theorem therefore implies

$$\begin{aligned}
\left|\langle f, \psi_{j,k}\rangle_{L_2}\right| &= \left|\langle f - P_{r-1}, \psi_{j,k}\rangle_{L_2}\right| \\
&\le \frac{\sup_{x \in I_{j,k}} |f^{(r)}(x)|}{r!} \int_{\mathbb{R}} |x - x_0|^r |\psi_{j,k}(x)|\, dx \\
&= \frac{2^{-j(r+1/2)}}{r!} \left(\int_{\mathbb{R}} |u|^r |\psi(u)|\, du\right) \sup_{x \in I_{j,k}} |f^{(r)}(x)|.
\end{aligned}$$

□

The estimate is local: the derivative supremum is taken over the interval joining the wavelet support to its translation point. Smooth regions therefore produce rapidly decaying fine-scale coefficients, while a singularity influences nearby wavelets. When the dyadic wavelet family forms an orthonormal basis, Parseval's identity converts coefficient-tail bounds into L_2 approximation bounds.

15.B Fourier facts used by the approximation arguments

The wavelet and shallow-network arguments use the same Fourier transform for different purposes. The wavelet calculation checks a reconstruction identity through a Fourier multiplier. The shallow-network calculation first writes a target as an integral of Fourier modes and then writes each mode as an integral of ridge functions. This appendix states the required Fourier facts with the normalization fixed in Eq. (15.18).

15.B.1 Inversion, Plancherel, and elementary identities

A function $f : \mathbb{R}^d \to \mathbb{C}$ belongs to the *Schwartz class* $\mathcal{S}(\mathbb{R}^d)$ if f is infinitely differentiable and

$$\sup_{x \in \mathbb{R}^d} |x^\alpha \partial^\beta f(x)| < \infty \quad \text{for every pair of multi-indices } \alpha, \beta. \tag{15.63}$$

Every smooth compactly supported function is a Schwartz function. The changes of variables, differentiations, and integrations by parts below are therefore justified for Schwartz functions and have no boundary terms.

Proposition 15.B.1 (Elementary Fourier identities)**.** *Let $f, g \in \mathcal{S}(\mathbb{R}^d)$, let $t \in \mathbb{R}^d$, and let $s > 0$. With*

$$(f * g)(x) = \int_{\mathbb{R}^d} f(y) g(x - y)\, dy, \tag{15.64}$$

the Fourier transform satisfies

$$\widehat{f(\cdot - t)}(\omega) = e^{-i\omega^\top t} \widehat{f}(\omega), \qquad \widehat{f(s\,\cdot)}(\omega) = s^{-d} \widehat{f}(\omega/s), \tag{15.65}$$

$$\widehat{f * g}(\omega) = \widehat{f}(\omega)\widehat{g}(\omega), \qquad \widehat{\partial^\alpha f}(\omega) = (i\omega)^\alpha \widehat{f}(\omega). \tag{15.66}$$

In particular,

$$\widehat{(-\Delta)^r f}(\omega) = \|\omega\|_2^{2r} \widehat{f}(\omega), \qquad r = 0, 1, \ldots. \tag{15.67}$$

If f is real-valued, then

$$\widehat{f}(-\omega) = \overline{\widehat{f}(\omega)}. \tag{15.68}$$

Proof. The first two identities follow by the substitutions $u = x - t$ and $u = sx$. Fubini's theorem followed by the substitution $u = x - y$ proves the convolution identity. Integration by parts gives $\widehat{\partial_j f}(\omega) = i\omega_j \widehat{f}(\omega)$; iteration proves the multi-index formula and Eq. (15.67). Taking the complex conjugate of Eq. (15.18) proves Eq. (15.68). □

Theorem 15.B.2 (Fourier inversion and Plancherel)**.** *For every $f \in \mathcal{S}(\mathbb{R}^d)$,*

$$f(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{i\omega^\top x} \widehat{f}(\omega)\, d\omega. \tag{15.69}$$

For every $f, g \in \mathcal{S}(\mathbb{R}^d)$,

$$\int_{\mathbb{R}^d} f(x)\overline{g(x)}\, dx = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \widehat{f}(\omega)\overline{\widehat{g}(\omega)}\, d\omega. \tag{15.70}$$

Consequently, the Fourier transform extends uniquely from $\mathcal{S}(\mathbb{R}^d)$ to $L_2(\mathbb{R}^d)$ and satisfies

$$\|f\|_{L_2}^2 = \frac{1}{(2\pi)^d} \|\widehat{f}\|_{L_2}^2. \tag{15.71}$$

Proof sketch. Let $g_\epsilon(x) = (2\pi\epsilon)^{-d/2} \exp(-\|x\|_2^2/(2\epsilon))$. Completing the square in the Gaussian integral gives $\widehat{g}_\epsilon(\omega) = \exp(-\epsilon\|\omega\|_2^2/2)$. Substituting the definition of $\widehat{f}$ into the right-hand side below, applying Fubini's theorem, and evaluating the inverse Gaussian integral give

$$(f * g_\epsilon)(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{i\omega^\top x} e^{-\epsilon\|\omega\|_2^2/2} \widehat{f}(\omega)\, d\omega. \tag{15.72}$$

The functions g_ϵ have integral one and concentrate their mass in every neighborhood of zero as $\epsilon \downarrow 0$. Uniform continuity of a Schwartz function therefore gives $(f * g_\epsilon)(x) \to f(x)$. The

right-hand side converges to the right-hand side of Eq. (15.69) by dominated convergence. Indeed, integration by parts gives $(1+\|\omega\|_2^2)^k \widehat{f}(\omega) = \widehat{(I-\Delta)^k f}(\omega)$, so $\widehat{f}$ is integrable for sufficiently large k. Substituting the inversion formula for g into $\int f\overline{g}$ and applying Fubini proves Eq. (15.70). Smooth compactly supported functions are dense in $L_2(\mathbb{R}^d)$ and belong to $\mathcal{S}(\mathbb{R}^d)$, so $\mathcal{S}(\mathbb{R}^d)$ is dense in $L_2(\mathbb{R}^d)$. Continuity in Eq. (15.70) gives the unique extension and Eq. (15.71). □

The density argument for ridge networks uses one quantitative consequence of smoothness: a compactly supported smooth target has every polynomial spectral moment finite.

Proposition 15.B.3 (Rapid Fourier decay of smooth compactly supported functions)**.** *If $f \in C_c^\infty(\mathbb{R}^d)$ and $N \geq 0$, then there is a finite constant $A_{f,N}$ such that*

$$|\widehat{f}(\omega)| \leq A_{f,N}(1+\|\omega\|_2)^{-N}, \qquad \omega \in \mathbb{R}^d. \tag{15.73}$$

In particular,

$$\int_{\mathbb{R}^d} \|\omega\|_1 |\widehat{f}(\omega)|\, d\omega < \infty. \tag{15.74}$$

Proof. For every integer $k \geq 0$, Eq. (15.67) gives

$$(1+\|\omega\|_2^2)^k |\widehat{f}(\omega)| = |\widehat{(I-\Delta)^k f}(\omega)| \leq \|(I-\Delta)^k f\|_{L_1}. \tag{15.75}$$

Choose k with $2k \geq N$ to obtain Eq. (15.73). Choosing instead $2k > d+1$ makes $\|\omega\|_1(1+\|\omega\|_2^2)^{-k}$ integrable and proves Eq. (15.74). □

15.B.2 Fourier verification of the wavelet reconstruction

Let $\psi, \varphi \in \mathcal{S}(\mathbb{R}^d)$. For a target f, define the continuous analysis transform

$$u_f(s,t) = \int_{\mathbb{R}^d} f(x)\psi(s(x-t))\, dx, \qquad s > 0, \quad t \in \mathbb{R}^d. \tag{15.76}$$

The desired scale–location reconstruction is

$$f(y) = \int_0^\infty \int_{\mathbb{R}^d} u_f(s,t)\varphi(s(y-t)) s^{2d-1}\, dt\, ds \quad \text{in } L^2(\mathbb{R}^d). \tag{15.77}$$

The next proposition converts Eq. (15.77) into a scalar condition at each nonzero frequency. Truncating the scale integral makes the Fourier calculation ordinary; Plancherel then controls the limit.

Proposition 15.B.4 (Fourier criterion for a dual scale–location pair)**.** *Let $\psi, \varphi \in \mathcal{S}(\mathbb{R}^d)$. For $f \in \mathcal{S}(\mathbb{R}^d)$ and $0 < \epsilon < R < \infty$, define*

$$T_{\epsilon,R} f(y) = \int_\epsilon^R \int_{\mathbb{R}^d} u_f(s,t)\varphi(s(y-t)) s^{2d-1}\, dt\, ds, \tag{15.78}$$

where u_f is defined in Eq. (15.76). Then $T_{\epsilon,R}$ is a Fourier multiplier with

$$\widehat{T_{\epsilon,R} f}(\omega) = M_{\epsilon,R}(\omega)\widehat{f}(\omega), \tag{15.79}$$

$$M_{\epsilon,R}(\omega) = \int_\epsilon^R \widehat{\psi}(-\omega/s)\widehat{\varphi}(\omega/s)\, \frac{ds}{s} = \int_{1/R}^{1/\epsilon} \widehat{\psi}(-a\omega)\widehat{\varphi}(a\omega)\, \frac{da}{a}. \tag{15.80}$$

Suppose that

$$\sup_{0<\epsilon<R<\infty} \|M_{\epsilon,R}\|_{L_\infty(\mathbb{R}^d)} < \infty \tag{15.81}$$

and that

$$\lim_{\epsilon\downarrow 0,\ R\uparrow\infty} M_{\epsilon,R}(\omega) = 1 \quad \textit{for almost every } \omega \in \mathbb{R}^d. \tag{15.82}$$

Then the multiplier formula extends $T_{\epsilon,R}$ to $L_2(\mathbb{R}^d)$ and $T_{\epsilon,R}f \to f$ in $L_2(\mathbb{R}^d)$ for every $f \in L_2(\mathbb{R}^d)$. The limit is the reconstruction identity in Eq. (15.77).

Proof. For fixed s, direct integration in t gives

$$\widehat{u_f(s,\cdot)}(\omega) = s^{-d}\widehat{f}(\omega)\widehat{\psi}(-\omega/s). \tag{15.83}$$

The synthesis function $z \mapsto \varphi(sz)$ has Fourier transform $s^{-d}\widehat{\varphi}(\omega/s)$. The convolution identity in Eq. (15.66) now gives Eq. (15.79). Multiplying the s^{-d} factor from the analysis transform by the s^{-d} factor from the synthesis kernel cancels the s^{2d} factor in $s^{2d-1}ds$, leaving ds/s. The substitution $a = 1/s$ gives the second expression in Eq. (15.80). By Plancherel, Eq. (15.81) makes $T_{\epsilon,R}$ uniformly bounded on $\mathcal{S}(\mathbb{R}^d)$ and hence gives a unique bounded extension to $L_2(\mathbb{R}^d)$. For any $f \in L_2(\mathbb{R}^d)$, Eqs. (15.71) and (15.82) and dominated convergence then give $\|T_{\epsilon,R}f - f\|_{L_2} \to 0$. □

Proposition 15.B.5 (Gaussian dual wavelet pair)**.** *Let*

$$\varphi(x) = e^{-\|x\|_2^2/2}, \qquad \psi = c_r(-\Delta)^r\varphi, \qquad c_r = \frac{2}{(2\pi)^d\Gamma(r)}, \tag{15.84}$$

where $r \geq 1$ is an integer and $\Gamma(r) = (r-1)!$. Then the hypotheses of Proposition 15.B.4 hold.

Proof. The Gaussian Fourier transform and Eq. (15.67) give

$$\widehat{\varphi}(\omega) = (2\pi)^{d/2}e^{-\|\omega\|_2^2/2}, \qquad \widehat{\psi}(\omega) = c_r\|\omega\|_2^{2r}\widehat{\varphi}(\omega). \tag{15.85}$$

Both transforms are real and even. For $\omega \neq 0$, the full multiplier is

$$\begin{aligned}\int_0^\infty \widehat{\psi}(-a\omega)\widehat{\varphi}(a\omega)\,\frac{da}{a} &= c_r(2\pi)^d\int_0^\infty (a\|\omega\|_2)^{2r}e^{-a^2\|\omega\|_2^2}\,\frac{da}{a}\\ &= \frac{c_r(2\pi)^d}{2}\Gamma(r) = 1.\end{aligned} \tag{15.86}$$

The truncated multipliers lie between 0 and 1, so they satisfy Eq. (15.81). The L_2 reconstruction identity follows from Proposition 15.B.4. □

Proposition 15.B.6 (Wavelet proof of Gaussian RBF density)**.** *The finite Gaussian scale–location sums in Eq. (15.13) are dense in $L^2(\mathbb{R}^d)$.*

Proof. Choose an integer $r > d/4$ and the pair (ψ,φ) from Proposition 15.B.5. Because $C_c^\infty(\mathbb{R}^d)$ is dense in $L^2(\mathbb{R}^d)$, proving the claim for $f \in C_c^\infty(\mathbb{R}^d)$ suffices. For $0 < s \leq 1$, Fubini's theorem gives

$$\int_{\mathbb{R}^d} |u_f(s,t)|\,dt \leq s^{-d}\|f\|_{L^1}\|\psi\|_{L^1}. \tag{15.87}$$

For $s \geq 1$, the identity $\psi = c_r(-\Delta)^r\varphi$ and integration by parts give

$$\int_{\mathbb{R}^d} |u_f(s,t)|\,dt \leq c_r s^{-d-2r}\|(-\Delta)^r f\|_{L^1}\|\varphi\|_{L^1}. \tag{15.88}$$

Because $2r > d/2$, the two estimates imply

$$B_f := \|\varphi\|_{L^2} \int_0^\infty \int_{\mathbb{R}^d} |u_f(s,t)| s^{3d/2-1}\, dt\, ds < \infty. \tag{15.89}$$

Normalize the synthesis atoms as

$$\phi_{s,t}(y) = \frac{s^{d/2}}{\|\varphi\|_{L^2}} \varphi(s(y-t)), \qquad \|\phi_{s,t}\|_{L^2} = 1.$$

The reconstruction formula in Eq. (15.77) is then an integral representation by the atoms $\phi_{s,t}$ with total variation B_f. Lemma 15.2.2 supplies finite sums converging to f in L^2. Absorbing the normalizing factors into the coefficients gives finite Gaussian radial-basis-function networks. □

15.B.3 From Fourier modes to ridge functions

Fourier inversion represents a target through the functions $x \mapsto e^{i\omega^\top x}$. Each such function is a complex-valued ridge function, but its univariate profile is oscillatory rather than sigmoidal. The next identity converts the oscillatory profile into thresholds, which are limits of steep sigmoids.

Proposition 15.B.7 (Fourier modes as limits of sigmoidal ridge functions)**.** *Let $Q = [0,1]^d$, let P_X be uniform on Q, and let σ satisfy Eq. (15.14). For $\omega \in \mathbb{R}^d$, define*

$$\alpha_\omega = \sum_{j:\,\omega_j<0} \omega_j, \qquad \beta_\omega = \sum_{j:\,\omega_j>0} \omega_j. \tag{15.90}$$

Then $\omega^\top x \in [\alpha_\omega, \beta_\omega]$ for $x \in Q$, $\beta_\omega - \alpha_\omega = \|\omega\|_1$, and

$$e^{it} - 1 = \int_{\alpha_\omega}^{\beta_\omega} ie^{ib} \big(\mathbb{I}\{t \ge b\} - \mathbb{I}\{0 \ge b\}\big)\, db \quad \text{for every } t \in [\alpha_\omega, \beta_\omega]. \tag{15.91}$$

Moreover, for every nonzero ω and every $b \in \mathbb{R}$,

$$\sigma\big(\lambda(\omega^\top x - b)\big) \longrightarrow \mathbb{I}\{\omega^\top x \ge b\} \quad \text{in } L_2(P_X) \text{ as } \lambda \to \infty. \tag{15.92}$$

Consequently, for every nonzero ω there are finite complex-valued networks

$$g_n(x) = c_{0,n} + \sum_{j=1}^{m_n} c_{j,n} \sigma(a_{j,n}^\top x + b_{j,n}) \tag{15.93}$$

such that

$$\|g_n - (e^{i\omega^\top(\cdot)} - 1)\|_{L_2(P_X)} \longrightarrow 0, \qquad |c_{0,n}| + \sum_{j=1}^{m_n} |c_{j,n}| \le 2\|\omega\|_1 + o(1). \tag{15.94}$$

Proof. The formulas for α_ω and β_ω follow by minimizing and maximizing each coordinate contribution $\omega_j x_j$ separately. The fundamental theorem of calculus gives

$$e^{it} - 1 = \int_0^t ie^{ib}\, db, \tag{15.95}$$

and writing the oriented interval from 0 to t with indicators gives Eq. (15.91). The absolute coefficient mass of the two indicator terms is at most $2(\beta_\omega - \alpha_\omega) = 2\|\omega\|_1$.

For nonzero ω, the hyperplane $\{x \in Q : \omega^\top x = b\}$ has Lebesgue measure zero. The two limits in Eq. (15.14) therefore give pointwise convergence almost everywhere in Eq. (15.92); boundedness of σ gives $L_2(P_X)$ convergence by dominated convergence. Moreover,

$$\left\|\mathbb{I}\{\omega^\top(\,\cdot\,) \geq b\} - \mathbb{I}\{\omega^\top(\,\cdot\,) \geq b'\}\right\|_{L_2(P_X)}^2 = P_X\Big(\omega^\top X \in [\min\{b,b'\}, \max\{b,b'\})\Big). \tag{15.96}$$

The probability on the right tends to zero as $b' \to b$ because $\omega^\top X$ has no atoms. Thus the integrand in Eq. (15.91) is a continuous $L_2(P_X)$-valued function of b, and its Riemann sums converge in $L_2(P_X)$. Replace the finitely many thresholds in each Riemann sum by steep sigmoids using Eq. (15.92). Keep the constant indicator $\mathbb{I}\{0 \geq b\}$ in $c_{0,n}$. The Riemann-sum and sigmoid approximations preserve the coefficient bound up to an $o(1)$ term, proving Eq. (15.94). □

For a representation satisfying Eq. (15.19), set $d\nu(\omega) = (2\pi)^{-d}F(\omega)\,d\omega$. Integrating the threshold representations in Eq. (15.91) against ν produces a complex signed measure on the threshold ridge parameters. Taking real parts produces a real signed ridge-parameter measure because the represented target is real. The total variation of the real ridge-parameter measure is at most

$$2\int_{\mathbb{R}^d} \|\omega\|_1\, d|\nu|(\omega). \tag{15.97}$$

The right-hand side is twice the spectral-moment term in Eq. (15.20). Applying Lemma 15.2.2 to the resulting real signed measure proves the upper-bound representation used in Theorem 15.3.2. Rapid Fourier decay from Proposition 15.B.3, followed by approximation of an arbitrary $L^2(Q)$ target by smooth compactly supported functions, also proves Corollary 15.2.4.

15.B.4 Measure-valued spectral representations

The density formulation in Definition 15.3.1 is sufficient for the main theorem. A measure-valued formulation enlarges the class and includes pure Fourier modes without smoothing. For a real-valued $f : \mathbb{R}^d \to \mathbb{R}$, let $\mathfrak{R}(f)$ contain every finite complex Borel measure ν such that

$$f(x) = f(0) + \int_{\mathbb{R}^d} (e^{i\omega^\top x} - 1)\, d\nu(\omega) \quad \text{for every } x \in \mathbb{R}^d, \tag{15.98}$$

where the integral is real. Define the anchored measure cost

$$\|f\|_{\mathcal{B},\mathrm{meas}} = |f(0)| + \inf_{\nu \in \mathfrak{R}(f)} \int_{\mathbb{R}^d} \|\omega\|_1\, d|\nu|(\omega). \tag{15.99}$$

The radius-(C) class on (Q) consists of restrictions of ambient functions whose anchored measure cost is at most (C), in parallel with Eq. (15.21). Every density F in Eq. (15.19) defines $d\nu = (2\pi)^{-d}F\,d\omega$, so the density-defined Barron ball is contained in the measure-defined ball with the same radius. The Fourier-to-threshold construction in Proposition 15.B.7 applies directly to ν and gives the same neural upper bound. The main text uses the density class because smooth spectral approximations to cosine modes also give the fixed-space lower bound inside that smaller class.

15.B.5 Fourier modes on the cube

The fixed-space lower bound uses the integer Fourier modes as an orthogonal collection under the uniform law on the cube.

Proposition 15.B.8 (Orthogonality of integer Fourier modes)**.** *For $k, \ell \in \mathbb{Z}^d$,*

$$\int_{[0,1]^d} e^{2\pi i (k-\ell)^\top x}\, dx = \mathbb{I}\{k = \ell\}. \tag{15.100}$$

Consequently, if $k_1, \dots, k_N$ are nonzero and no two are equal up to sign, then the real functions

$$x \longmapsto \sqrt{2}\cos(2\pi k_j^\top x), \qquad j = 1, \dots, N, \tag{15.101}$$

are orthonormal in $L_2([0,1]^d)$.

Proof. Fubini's theorem factors the integral in Eq. (15.100) into d one-dimensional integrals. For an integer q, $\int_0^1 e^{2\pi i q x}\, dx$ equals 1 when $q = 0$ and 0 otherwise. The cosine claim follows from $2\cos u \cos v = \cos(u - v) + \cos(u + v)$. □

Proposition 15.B.9 (Smooth spectral approximations to cosine modes)**.** *For every nonzero $\omega \in \mathbb{R}^d$ and every $0 < \delta \le 1$, there is a real-valued function $u_{\omega,\delta}$ satisfying Eq. (15.19) such that*

$$\left\| u_{\omega,\delta} - \sqrt{2}\cos(\omega^\top \cdot) \right\|_{L_\infty(Q)} \le \sqrt{2}\,\delta, \tag{15.102}$$

$$\|u_{\omega,\delta}\|_{\mathcal{B}(\mathbb{R}^d)} \le \sqrt{2}\big(1 + \|\omega\|_1 + \delta\big). \tag{15.103}$$

Proof. Choose a nonnegative, even function $\rho \in C_c^\infty(\mathbb{R}^d)$ supported on $\{z : \|z\|_1 \le 1\}$ and satisfying $\int_{\mathbb{R}^d} \rho(z)\, dz = 1$. Set $\rho_\delta(z) = \delta^{-d}\rho(z/\delta)$ and define the Fourier density

$$\widehat{u}_{\omega,\delta}(\xi) = \frac{(2\pi)^d}{\sqrt{2}} \left(\rho_\delta(\xi - \omega) + \rho_\delta(\xi + \omega)\right). \tag{15.104}$$

Fourier inversion gives

$$u_{\omega,\delta}(x) = \sqrt{2}\cos(\omega^\top x) \int_{\mathbb{R}^d} e^{iz^\top x} \rho_\delta(z)\, dz. \tag{15.105}$$

The last integral is real because ρ_δ is even. For $x \in Q$ and z in the support of ρ_δ, $|z^\top x| \le \|z\|_1 \|x\|_\infty \le \delta$. Therefore

$$\left| \int_{\mathbb{R}^d} (e^{iz^\top x} - 1)\rho_\delta(z)\, dz \right| \le \delta, \tag{15.106}$$

which proves Eq. (15.102). The triangle inequality also gives the spectral-moment bound

$$\begin{aligned} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \|\xi\|_1 |\widehat{u}_{\omega,\delta}(\xi)|\, d\xi &= \frac{1}{\sqrt{2}} \int_{\mathbb{R}^d} \left(\|\omega + z\|_1 + \| - \omega + z\|_1\right) \rho_\delta(z)\, dz \\ &\le \sqrt{2}\big(\|\omega\|_1 + \delta\big). \end{aligned} \tag{15.107}$$

Because $u_{\omega,\delta}(0) = \sqrt{2}$, adding the anchored constant cost proves Eq. (15.103). □

15.C Greedy and point-query recovery of ridge structure

The results in this appendix separate two conditional constructions. The first result turns the $m^{-1/2}$ convex-hull approximation bound into a sequential greedy method, provided that a correlation oracle is available; the result is not a near-best m-term oracle inequality. The second result estimates an active linear subspace from noiseless point evaluations, conditional on supplied base points whose gradient matrix is full rank and well conditioned. A final example shows why smoothness and a ridge representation alone do not guarantee such informative base points.

15.C.1 Greedy approximation with a correlation oracle

The following proposition gives the guarantee summarized in Section 15.4. The proof also identifies the operation that can be difficult for a continuous ridge dictionary.

Proposition 15.C.1 (Greedy approximation with a correlation oracle)**.** *Let $\mathcal{D}$ be a collection of elements of a Hilbert space $\mathcal{H}$ satisfying $\|\phi\|_{\mathcal{H}} \le 1$ for every $\phi \in \mathcal{D}$. Fix $B > 0$, and suppose that*

$$f \in \mathcal{K} := B\,\overline{\mathrm{conv}}(\mathcal{D} \cup -\mathcal{D}). \tag{15.108}$$

Set $f_0 = 0$. At step $m \ge 1$, assume that the supremum displayed below is attained and let a correlation oracle return $\phi_m \in \mathcal{D} \cup -\mathcal{D}$ satisfying

$$\langle f - f_{m-1}, B\phi_m\rangle_{\mathcal{H}} = \sup_{\phi \in \mathcal{D}\cup -\mathcal{D}} \langle f - f_{m-1}, B\phi\rangle_{\mathcal{H}}, \tag{15.109}$$

and update

$$f_m = \left(1 - \frac{1}{m+1}\right) f_{m-1} + \frac{1}{m+1} B\phi_m. \tag{15.110}$$

Then f_m is a linear combination of at most m atoms and

$$\|f - f_m\|_{\mathcal{H}} \le \frac{2B}{\sqrt{m+1}}. \tag{15.111}$$

Proof. The set $\mathcal{K}$ is symmetric and convex, contains 0, and lies in the closed ball of radius B. Hence $f_m \in \mathcal{K}$ for every m. Write $r_{m-1} = f - f_{m-1}$ and $\gamma_m = 1/(m+1)$. Since $f \in \mathcal{K}$, linearity and Eq. (15.109) give

$$\langle r_{m-1}, B\phi_m\rangle_{\mathcal{H}} \ge \langle r_{m-1}, f\rangle_{\mathcal{H}}. \tag{15.112}$$

Consequently,

$$\begin{aligned} \|f - f_m\|_{\mathcal{H}}^2 &= \|r_{m-1} - \gamma_m(B\phi_m - f_{m-1})\|_{\mathcal{H}}^2 \\ &\le (1 - 2\gamma_m)\|r_{m-1}\|_{\mathcal{H}}^2 + 4\gamma_m^2 B^2. \end{aligned}$$

The last term uses $\|B\phi_m - f_{m-1}\|_{\mathcal{H}} \le 2B$. Induction, starting from $\|f - f_0\|_{\mathcal{H}} \le B$, now gives $\|f - f_m\|_{\mathcal{H}}^2 \le 4B^2/(m+1)$. The update adds at most one new atom at every step, so f_m uses at most m atoms. □

The proposition concerns the number of selected atoms. The computational content of the proposition depends on Eq. (15.109). The oracle is elementary for an explicitly listed finite dictionary. For the ridge dictionary in Eq. (15.31), the oracle must choose both a direction and a univariate profile. Discretizing all directions can require a set whose size is exponential in d at a fixed angular resolution. The rate in Eq. (15.111) therefore does not by itself control the arithmetic cost of an iteration.

15.C.2 Recovering a multi-index subspace from point values

Noiseless active queries permit a different construction when the target is known to have a multi-index form. Write

$$\mathcal{B}_d(R) = \{x \in \mathbb{R}^d : \|x\|_2 \le R\}. \tag{15.113}$$

The next theorem uses coordinate finite differences so that the conditional subspace-estimation algorithm and its cost are explicit. The theorem assumes that informative base points have already been supplied.

Theorem 15.C.2 (Active-subspace recovery from noiseless point queries)**.** *Let $f : \mathcal{B}_d(1+\bar{h}) \to \mathbb{R}$ satisfy*

$$f(x) = g(Ax), \qquad A \in \mathbb{R}^{k\times d}, \quad AA^\top = I_k, \tag{15.114}$$

where g is twice continuously differentiable. Define

$$P = A^\top A, \quad L = \sup_{z\in\mathcal{B}_d(1+\bar{h})} \|\nabla f(z)\|_2, \quad M = \sup_{z\in\mathcal{B}_d(1+\bar{h})} \|\nabla^2 f(z)\|_{\mathrm{op}}. \tag{15.115}$$

Choose $x_1, \dots, x_q \in \mathcal{B}_d(1)$ and form

$$G = \begin{bmatrix} \nabla f(x_1) & \cdots & \nabla f(x_q) \end{bmatrix}. \tag{15.116}$$

Assume that $\operatorname{rank}(G) = k$ *and write $\lambda_\star = \sigma_k(G) > 0$. For $0 < h \le \bar{h}$, use the $q(d+1)$ point values at x_j and $x_j + he_\ell$ to define*

$$\widetilde{G}_{\ell j} = \frac{f(x_j + he_\ell) - f(x_j)}{h}, \qquad \ell = 1, \dots, d, \quad j = 1, \dots, q. \tag{15.117}$$

Let $\widehat{P}$ be the orthogonal projector onto the leading k left singular vectors of $\widetilde{G}$, and set

$$\eta = \frac{Mh\sqrt{dq}}{2}. \tag{15.118}$$

If $\eta < \lambda_\star$, then

$$\|\widehat{P} - P\|_{\mathrm{op}} \le \frac{\eta}{\lambda_\star - \eta}. \tag{15.119}$$

The reduced function $\widehat{f}(x) = f(\widehat{P}x)$ consequently satisfies

$$\sup_{x\in\mathcal{B}_d(1)} |f(x) - \widehat{f}(x)| \le L\frac{\eta}{\lambda_\star - \eta}. \tag{15.120}$$

Proof. The chain rule gives

$$\nabla f(x) = A^\top \nabla g(Ax). \tag{15.121}$$

Each column of G is therefore a linear combination of the rows of A. The columns of G span a k-dimensional space by the rank assumption, and the k rows of A are orthonormal. Consequently, the columns of G and the rows of A span the same subspace of $\mathbb{R}^d$. The matrix $P = A^\top A$ is the orthogonal projector onto that subspace.

Taylor's theorem along the coordinate direction e_ℓ gives

$$\left|\widetilde{G}_{\ell j} - \partial_\ell f(x_j)\right| \le \frac{Mh}{2}. \tag{15.122}$$

For $E = \widetilde{G} - G$, summing the squared entrywise bounds yields

$$\|E\|_{\mathrm{op}} \le \|E\|_{\mathrm{F}} \le \frac{Mh\sqrt{dq}}{2} = \eta. \tag{15.123}$$

Let $\widehat{U}\widehat{\Sigma}\widehat{V}^\top$ contain the leading k singular triplets of $\widetilde{G}$. The singular-value perturbation inequality gives $\sigma_k(\widetilde{G}) \ge \lambda_\star - \eta > 0$. Since $(I - P)G = 0$,

$$(I - P)\widehat{U} = (I - P)E\widehat{V}\widehat{\Sigma}^{-1}. \tag{15.124}$$

Therefore

$$\|(I - P)\widehat{U}\|_{\mathrm{op}} \le \frac{\eta}{\lambda_\star - \eta}. \tag{15.125}$$

The two subspaces have the same dimension, so the left-hand side of Eq. (15.125) equals $\|\widehat{P} - P\|_{\mathrm{op}}$. The one-sided subspace bound therefore proves Eq. (15.119).

Finally, $AP = A$ implies $f(x) = f(Px)$. Both Px and $\widehat{P}x$ belong to $\mathcal{B}_d(1)$ when $x \in \mathcal{B}_d(1)$. The gradient bound in Eq. (15.115) and the mean-value theorem give

$$|f(x) - f(\widehat{P}x)| = |f(Px) - f(\widehat{P}x)| \le L\|(P - \widehat{P})x\|_2,$$

which proves Eq. (15.120). □

If $\widehat{U}$ has orthonormal columns spanning the estimated subspace, then

$$\widehat{f}(x) = \widehat{g}(\widehat{U}^\top x), \qquad \widehat{g}(y) = f(\widehat{U}y), \qquad y \in \mathcal{B}_k(1). \tag{15.126}$$

Each query of $\widehat{g}$ requires one additional query of f. Suppose that a grid, spline, or another k-dimensional approximation method uses such queries to construct $\widetilde{g}$ satisfying $\sup_{y\in\mathcal{B}_k(1)} |\widehat{g}(y) - \widetilde{g}(y)| \le \delta$. Then

$$\widetilde{f}(x) = \widetilde{g}(\widehat{U}^\top x) \quad \Longrightarrow \quad \sup_{x\in\mathcal{B}_d(1)} |f(x) - \widetilde{f}(x)| \le L\frac{\eta}{\lambda_\star - \eta} + \delta. \tag{15.127}$$

The additional query count is the cost of obtaining the k-dimensional interpolation error δ. For a generic smooth link, the interpolation cost can depend exponentially on the intrinsic dimension k, but not on the ambient dimension d.

The rank assumption in Theorem 15.C.2 is the finite-sample form of an average-gradient condition. If a random element X has a declared law supported on $\mathcal{B}_d(1)$, define

$$H_f = \mathbb{E}[\nabla f(X)\nabla f(X)^\top] = A^\top \mathbb{E}[\nabla g(AX)\nabla g(AX)^\top]A. \tag{15.128}$$

A positive kth eigenvalue means that the gradients reveal every active direction on average. Strict positivity gives identifiability; a useful finite-sample guarantee requires a quantitative lower bound on that eigenvalue. Randomized recovery results combine such a lower bound with matrix concentration to choose enough base points. For sparse or compressible gradients, the directional differences

$$\frac{f(x_j + h\varphi_i) - f(x_j)}{h} \approx \varphi_i^\top \nabla f(x_j) \tag{15.129}$$

form a compressed-sensing problem. Sparse recovery followed by the same SVD can then use fewer than d directions per base point.

15.C.3 Why an informative-link condition is necessary

The next example shows that active queries do not make every smooth ridge class tractable. The obstruction is informational: most queries return zero, so the algorithm receives no evidence about the hidden direction.

Proposition 15.C.3 (A point-query obstruction for smooth ridge functions)**.** *There are numerical constants $c_0, c_1 > 0$ such that the following statement holds for every sufficiently large d. Define*

$$g(t) = 8(t - 1/2)_+^3, \qquad f_a(x) = g(a^\top x), \quad a \in \mathbb{S}^{d-1}, \quad x \in \mathcal{B}_d(1). \tag{15.130}$$

The link g is twice continuously differentiable and $\|f_a\|_\infty = 1$. Any deterministic adaptive algorithm using at most $c_0 e^{c_1 d}$ noiseless point queries in $\mathcal{B}_d(1)$ has the same query transcript on the zero function and on some f_a. Consequently, the algorithm cannot guarantee uniform error strictly smaller than $1/2$ on

$$\{0\} \cup \{f_a : a \in \mathbb{S}^{d-1}\}. \tag{15.131}$$

Proof. Let U be uniform on $\mathbb{S}^{d-1}$. Rotational invariance and the density of the first coordinate of U imply that, for every fixed $x \in \mathcal{B}_d(1)$,

$$\mathbb{P}(U^\top x > 1/2) \le C\sqrt{d}\,(3/4)^{(d-3)/2} \le C_0 e^{-c_1 d} \tag{15.132}$$

for numerical constants $C, C_0, c_1 > 0$. If $x = 0$, the probability is zero. Otherwise, rotate $x/\|x\|_2$ to the first coordinate. Since $\|x\|_2 \le 1$, the event in Eq. (15.132) implies that the first coordinate of U exceeds $1/2$. The density of the first coordinate is proportional to $(1-t^2)^{(d-3)/2}$ on $[-1,1]$. Integrating the first-coordinate density gives the first inequality.

Run the algorithm along the transcript in which every query returns zero. The resulting query points $x_1, \dots, x_q$ are fixed. If $qC_0 e^{-c_1 d} < 1$, the union bound applied to Eq. (15.132) gives a direction a satisfying $a^\top x_j \le 1/2$ for every j. Hence $f_a(x_j) = 0$ for every query, and the algorithm produces the same output, say h, for the targets 0 and f_a. Since $f_a(a) = 1$,

$$1 = \|f_a - 0\|_\infty \le \|f_a - h\|_\infty + \|h - 0\|_\infty.$$

At least one of the two errors is at least $1/2$. Reducing c_0 if necessary ensures $qC_0 e^{-c_1 d} < 1$ for every $q \le c_0 e^{c_1 d}$. □

The lower bound does not conflict with Theorem 15.C.2. Fix $a \in \mathbb{S}^{d-1}$ and let the random element X be uniform on $\mathbb{S}^{d-1}$. For the target f_a in Eq. (15.130),

$$0 < \lambda_1(H_{f_a}) = \mathbb{E}\Big[g'(a^\top X)^2\Big] \le 36\mathbb{P}(a^\top X > 1/2) \le 36 C_0 e^{-c_1 d}. \tag{15.133}$$

Thus the informative eigenvalue is positive but exponentially small. The recovery guarantee in Theorem 15.C.2 is useful only when the target supplies gradient information at base points that the query strategy can obtain with the desired cost.

15.D Feature representations of reproducing-kernel Hilbert spaces

The next construction characterizes kernels that admit the feature representation in Eq. (15.26) and identifies changes of feature law that preserve the RKHS norm.

Proposition 15.D.1 (Every separable RKHS has a scalar feature representation)**.** *Let $\mathcal{H}_k$ be a separable real reproducing-kernel Hilbert space with kernel k, and let $(e_j)_{j\in J}$ be an orthonormal basis of $\mathcal{H}_k$, where J is finite or countably infinite. Choose numbers $\tau_j > 0$ satisfying $\sum_{j\in J} \tau_j = 1$, and define*

$$\phi_j(x) = \frac{e_j(x)}{\sqrt{\tau_j}}. \tag{15.134}$$

Then

$$k(x,x') = \sum_{j\in J} \tau_j \phi_j(x)\phi_j(x'), \tag{15.135}$$

and every $f \in \mathcal{H}_k$ has a representation

$$f(x) = \sum_{j\in J} \tau_j p_j \phi_j(x), \qquad \|f\|_{\mathcal{H}_k}^2 = \sum_{j\in J} \tau_j p_j^2. \tag{15.136}$$

Proof. For $x \in \mathcal{X}$, the reproducing property gives $e_j(x) = \langle e_j, k(x,\cdot)\rangle_{\mathcal{H}_k}$. Expanding both kernel sections in the orthonormal basis gives

$$k(x,x') = \langle k(x,\cdot), k(x',\cdot)\rangle_{\mathcal{H}_k} = \sum_{j\in J} e_j(x)e_j(x'),$$

which becomes Eq. (15.135) after substituting Eq. (15.134). If $f = \sum_j c_j e_j$, set $p_j = c_j/\sqrt{\tau_j}$. Then

$$\sum_j \tau_j p_j \phi_j(x) = \sum_j c_j e_j(x) = f(x), \qquad \sum_j \tau_j p_j^2 = \sum_j c_j^2 = \|f\|_{\mathcal{H}_k}^2.$$

□

Different bases and weights give different feature representations of the same kernel. More generally, suppose $d\widetilde{\tau} = q\,d\tau$, where $q > 0$ τ-almost everywhere and $\int q\,d\tau = 1$. The simultaneous changes

$$\widetilde{\phi}_v = \frac{\phi_v}{\sqrt{q}(v)}, \qquad \widetilde{p}(v) = \frac{p(v)}{\sqrt{q}(v)} \tag{15.137}$$

preserve the kernel, the represented function, and the coefficient norm:

$$\begin{aligned}
\int \widetilde{\phi}_v(x)\widetilde{\phi}_v(x')\,d\widetilde{\tau}(v) &= \int \phi_v(x)\phi_v(x')\,d\tau(v),\\
\int \widetilde{p}(v)\widetilde{\phi}_v(x)\,d\widetilde{\tau}(v) &= \int p(v)\phi_v(x)\,d\tau(v),\\
\int |\widetilde{p}(v)|^2\,d\widetilde{\tau}(v) &= \int |p(v)|^2\,d\tau(v).
\end{aligned}$$

Changing τ while holding the features fixed generally changes the kernel. For a redundant fixed feature representation, several coefficient functions p may represent the same f; the RKHS norm is the infimum of $\|p\|_{L_2(\tau)}$ over those coefficient functions.

15.E First-stage contraction for the staged-SGD result

The prediction theorem in Section 15.6 uses a specialized single-index model. The first-stage analysis applies more generally to a multiple-index response. Let $\mathcal{L}(X) = N(0, I_d)$, let the noise ε be independent of X, and suppose

$$Y = g(UX;\varepsilon), \qquad U \in \mathbb{R}^{s\times d}, \quad UU^\top = I_s, \tag{15.138}$$

where $z \mapsto g(z;\varepsilon)$ is weakly differentiable for every fixed noise value. The rows of U span the teacher subspace. The student is the two-layer network in Eq. (15.39).

The contraction theorem has two assumption regimes. The smooth-activation regime uses squared loss, an almost-sure bound on Y, and uniform bounds on σ, σ', and σ''. The ReLU regime instead assumes

$$0 \le \partial_{11}^2 \ell \le 1, \qquad |\partial_1 \ell| \le 1, \qquad |\partial_{12}^2 \ell| \le 1. \tag{15.139}$$

Appropriately scaled Huber and logistic losses satisfy the ReLU conditions. Monotonicity of the link is not required for the contraction theorem.

The key identity comes from Gaussian integration by parts.

Proposition 15.E.1 (Population-gradient decomposition, informal form)**.** *Under the smooth-activation assumptions, or under the corresponding conditions for a ReLU network and a convex Lipschitz loss, the gradient of the regularized population risk has the form*

$$\nabla_W R_\lambda(W) = \big(H(W) + \lambda I_m\big)W + D(W)U, \tag{15.140}$$

where $H(W) \in \mathbb{R}^{m\times m}$ and $D(W) \in \mathbb{R}^{m\times s}$ depend on the current network and the data distribution. Let $\Pi_\perp$ be the orthogonal projection onto the complement of the row space of U, and write $W_\perp = W\Pi_\perp$. Right-multiplying Eq. (15.140) by $\Pi_\perp$ removes the target term:

$$\nabla_W R_\lambda(W)\Pi_\perp = \big(H(W) + \lambda I_m\big)W_\perp. \tag{15.141}$$

If the symmetric part of $H(W) + \lambda I_m$ remains uniformly positive along the trajectory and the step size is sufficiently small, Eq. (15.141) contracts $W_\perp$. Only the component inside the teacher subspace receives the forcing term $D(W)U$.

Theorem 15.E.2 (Contraction toward the teacher subspace by online SGD, informal form)**.** *The estimator updates W according to Eq. (15.40), using a fresh independent pair (X_t, Y_t) at every update. At initialization, the entries of $W^{(0)}$ are independent $N(0, 1/d)$ variables, the coordinates of $a^{(0)}$ are independent and uniform on $[-1/m, 1/m]$, and the coordinates of $b^{(0)}$ are independent uniform signs. The output weights $a^{(0)}$ and biases $b^{(0)}$ remain fixed. For the specified loss, weight decay, and step-size schedules, the decaying-step version satisfies, for $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$\frac{\|W_\perp^{(T)}\|_F}{\sqrt{m}} \lesssim \sqrt{\frac{d + \log(1/\delta)}{T}}, \tag{15.142}$$

up to constants determined by the regularization schedule. The guarantee requires $m \gtrsim \log(1/\delta)$.

The theorem controls only the component orthogonal to the teacher subspace. The conclusion allows $W^{(T)}$ to approach the zero matrix and does not show that the rows of $W^{(T)}$ span every teacher direction. The theorem also gives neither convergence of the component inside the subspace nor accurate prediction for every link g.

References

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017. https://jmlr.org/papers/v18/14-546.html.

Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993. https://doi.org/10.1109/18.256500.

Andrew R. Barron, Albert Cohen, Wolfgang Dahmen, and Ronald A. DeVore. Approximation and learning by greedy algorithms. *The Annals of Statistics*, 36(1):64–94, 2008. https://doi.org/10.1214/009053607000000631.

Benedikt Bauer and Michael Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261–2285, 2019. https://doi.org/10.1214/18-AOS1747.

Albert Cohen, Ingrid Daubechies, Ronald DeVore, Gérard Kerkyacharian, and Dominique Picard. Capturing ridge functions in high dimensions from point queries. *Constructive Approximation*, 35(2):225–243, 2012. https://doi.org/10.1007/s00365-011-9147-6.

Arnak S. Dalalyan, Anatoly Juditsky, and Vladimir Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. *Journal of Machine Learning Research*, 9:1647–1678, 2008. https://jmlr.org/papers/v9/dalalyan08a.html.

Alex Damian, Jason D. Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 5413–5452, 2022. https://proceedings.mlr.press/v178/damian22a.html.

Ingrid Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 41(7):909–996, 1988. https://doi.org/10.1002/cpa.3160410705.

Bernard Delyon, Anatoli Juditsky, and Albert Benveniste. Accuracy analysis for wavelet approximations. *IEEE Transactions on Neural Networks*, 6(2):332–348, 1995. https://doi.org/10.1109/72.363469.

Ronald A. DeVore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998. https://doi.org/10.1017/S0962492900002816.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Proceedings of the 29th Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 907–940, 2016. https://proceedings.mlr.press/v49/eldan16.html.

Massimo Fornasier, Karin Schnass, and Jan Vybíral. Learning functions of few arbitrary linear parameters in high dimensions. *Foundations of Computational Mathematics*, 12(2):229–262, 2012. https://doi.org/10.1007/s10208-012-9115-y.

Jerome H. Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823, 1981. https://doi.org/10.1080/01621459.1981.10477729.

Marian Hristache, Anatoli Juditsky, Jörg Polzehl, and Vladimir Spokoiny. Structure adaptive approach for dimension reduction. *The Annals of Statistics*, 29(6):1537–1566, 2001. https://doi.org/10.1214/aos/1015345954.

Lee 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. https://doi.org/10.1214/aos/1176348546.

Gérard Kerkyacharian and Dominique Picard. Minimax or maxisets? *Bernoulli*, 8(2):219–253, 2002. https://projecteuclid.org/journals/bernoulli/volume-8/issue-2/Minimax-or-maxisets/10.3150/bj/1068128976.full.

Michael Kohler and Adam Krzyżak. Nonparametric regression based on hierarchical interaction models. *IEEE Transactions on Information Theory*, 63(3):1620–1630, 2017. https://doi.org/10.1109/TIT.2016.2634401.

Michael Kohler and Sophie Langer. On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231–2249, 2021. https://doi.org/10.1214/20-AOS2034.

Michael Kohler, Adam Krzyżak, and Vincent Molinero Römer. Learning of deep neural network regression estimates using gradient descent with pruning. 2026. https://arxiv.org/abs/2608.25743.

Stéphane G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989. https://doi.org/10.1109/34.192463.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: Dimension-free bounds and kernel limit. In *Proceedings of the 32nd Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2388–2464, 2019. https://proceedings.mlr.press/v99/mei19a.html.

Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A. Erdogdu. Neural networks efficiently learn low-dimensional representations with SGD. 2023. https://arxiv.org/abs/2209.14863.

Suzanna Parkinson, Greg Ongie, and Rebecca Willett. ReLU neural networks with linear layers are biased towards single- and multi-index models. *SIAM Journal on Mathematics of Data Science*, 7(3):1021–1052, 2025. https://doi.org/10.1137/24M1672158.

Philipp Petersen and Jakob Zech. Mathematical theory of deep learning. 2026. https://arxiv.org/abs/2407.18384.

Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. Mechanism for feature learning in neural networks and backpropagation-free machine learning models. *Science*, 383(6690):1461–1467, 2024. https://doi.org/10.1126/science.adi5639.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020. https://doi.org/10.1214/19-AOS1875.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982. https://doi.org/10.1214/aos/1176345969.

Matus Telgarsky. Benefits of depth in neural networks. In *Proceedings of the 29th Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1517–1539, 2016. https://proceedings.mlr.press/v49/telgarsky16.html.

Hemant Tyagi and Volkan Cevher. Active learning of multi-index function models. In *Advances in Neural Information Processing Systems 25*, pages 1475–1483, 2012. https://papers.nips.cc/paper/4820-active-learning-of-multi-index-function-models.

Yingcun Xia, Howell Tong, W. K. Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B*, 64(3):363–410, 2002. https://www.stat.purdue.edu/~yuzhu/stat598m3/Papers/AdaptiveEstimation.pdf.

Yuhong Yang. How powerful can any regression learning procedure be? In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 636–643, 2007. https://proceedings.mlr.press/v2/yang07a.html.