Wiener and Gaussian Density

Lecture “Introduction to Control and Machine Learning”

Zhengping Ji, Enrique Zuazua

05.03.2026

Friedrich-Alexander-Universität Erlangen-Nürnberg

Wiener's seminal contribution

Gaussians and Finite Data

Other basis functions

Cybenko's contribution

Function Approximation Problem and Density

Density means we can **build** (or approximate) any complex shape using simple "building blocks."

- Think of a complex function $f(x)$ as a big, detailed LEGO sculpture.
- Think of a simple "brick" function $\varphi(x)$ (like a Gaussian "bump") as a single 2×2 LEGO brick.
- **Density** is the idea that with *enough* bricks, placed in the right spots, you can build a shape that is **arbitrarily close** to your complex sculpture.

Question: How to verify if a class of functions as building blocks contain enough "ingredient" to approximate any function?

Fourier Transform as a Tool for Approximation

The Fourier Transform is a "math prism" that breaks a function down into its "ingredients."

- It takes a function $f(x)$ (like a complex sound wave or a beam of white light) and tells you *exactly* how much of each "frequency" ξ (or simple sine/cosine wave, or color) is inside it.
- We write this list of ingredients as $\hat{f}(\xi)$, which is the "spectrum" of the function $f(x)$.

Goal: Showing that as long as a function class contains enough "frequency ingredients", it suffices to approximate any function.

Wiener's seminal contribution

Setup

We work in the Hilbert space $L^2(\mathbb{R})$ with the inner product

$$\langle f, g \rangle = \int_{\mathbb{R}} f(x)\, \overline{g(x)}\, \mathrm{d}x.$$

Fix $\sigma > 0$ and consider the (non-normalized) Gaussian

$$\varphi(x) = e^{-\frac{x^2}{2\sigma^2}}, \qquad x \in \mathbb{R}.$$

For each $a \in \mathbb{R}$ we define the translation

$$\varphi_a(x) := \varphi(x - a) = e^{-\frac{(x-a)^2}{2\sigma^2}}.$$

Let $\mathcal{T} := \{\, \varphi(\cdot - a) \;:\; a \in \mathbb{R} \,\}$ be the set of all the Gaussian translates.

We will prove that: $M := \overline{\operatorname{span} \mathcal{T}}_{L^2}$ is the whole $L^2(\mathbb{R})$.

A More Formal Look at the Fourier Transform

The Fourier Transform $\hat{f}(\xi)$ of a function $f(x)$ is defined by the integral:

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{-ix\xi} \mathrm{d}x$$

What does this integral mean?

- The term $e^{-ix\xi}$ represents a "complex wave" at frequency ξ. (Using Euler's formula: $e^{-ix\xi} = \cos(x\xi) - i\sin(x\xi)$).
- The integral "measures the overlap" (like an inner product) between your function $f(x)$ and that specific frequency wave.
- If $f(x)$ contains high "amplitude" at frequency ξ, the value of $\hat{f}(\xi)$ will be large. If not, it will be small or zero.

A Note on Spaces

This integral is first defined for "well-behaved" functions (e.g., in $L^1(\mathbb{R})$), and then extended to the space of "finite energy" functions ($L^2(\mathbb{R})$) that we use in our density proof.

Fourier transform: basic facts

We use the (formal) Fourier transform

$$\widehat{f}(\xi) = \int_{\mathbb{R}} f(x)\, e^{-ix\xi}\, \mathrm{d}x,$$

extended from $L^1 \cap L^2$ to $L^2(\mathbb{R})$.

Two key facts:

- **Convolution theorem** (for suitable f, g):

$$\widehat{f * g} = \widehat{f}\,\widehat{g}.$$

- **Fourier transform of a Gaussian is a Gaussian:** there exists $C_\sigma > 0$ s.t.

$$\widehat{\varphi}(\xi) = C_\sigma\, e^{-\frac{\sigma^2 \xi^2}{2}}, \qquad \xi \in \mathbb{R}.$$

  In particular, $\widehat{\varphi}(\xi) \neq 0$ for all $\xi \in \mathbb{R}$.

Orthogonality to all Gaussian translates

Using the Fourier transform, we prove the first preliminary result for Gaussian density.

Proposition

Let $f \in L^2(\mathbb{R})$ be such that

$$\int_{\mathbb{R}} f(x)\,\varphi(x-a)\,\mathrm{d}x = 0 \qquad \text{for all } a \in \mathbb{R}.$$

Then $f = 0$ almost everywhere.

Equivalently,

$$\langle f, \varphi(\cdot - a)\rangle = 0 \qquad \text{for all } a \in \mathbb{R}.$$

In other words: the only function in $L^2(\mathbb{R})$ orthogonal to *all* translations of a Gaussian is the zero function.

Proof idea: orthogonality to all translates

By the symmetry of φ, for each $a \in \mathbb{R}$,

$$\int_{\mathbb{R}} f(x)\,\varphi(x-a)\,\mathrm{d}x = \int_{\mathbb{R}} f(x)\,\varphi(a-x)\,\mathrm{d}x = (f * \tilde{\varphi})(a),$$

where $\tilde{\varphi}(x) := \varphi(-x)$. Since φ is even, $\tilde{\varphi} = \varphi$, hence

$$(f * \varphi)(a) = 0 \qquad \text{for all } a \in \mathbb{R},$$

i.e. $f * \varphi \equiv 0$.

Take Fourier transforms and use the convolution theorem:

$$0 = \widehat{f * \varphi}(\xi) = \widehat{f}(\xi)\,\widehat{\varphi}(\xi) \qquad \text{for all } \xi \in \mathbb{R}.$$

Since $\widehat{\varphi}(\xi) \neq 0$ for all ξ, we get $\widehat{f}(\xi) = 0$ for all $\xi \in \mathbb{R}$.

The Fourier transform is injective on $L^2(\mathbb{R})$, so $f = 0$ almost everywhere. This proves the proposition.

Density of Gaussian translates

Theorem

Let $\varphi(x) = e^{-x^2/(2\sigma^2)}$ and

$$\mathcal{T} = \{\varphi(\cdot - a) \ : \ a \in \mathbb{R}\}.$$

Then its closed linear span satisfies

$$\overline{\operatorname{span} \mathcal{T}}_{L^2} = L^2(\mathbb{R}).$$

Equivalently, finite linear combinations

$$\sum_{k=1}^{n} c_k \, \varphi(\cdot - a_k)$$

are dense in $L^2(\mathbb{R})$.

Meaning: any L^2 function on $\mathbb{R}$ can be approximated by linear combination of Gaussian translates.

Proof: density of Gaussian translates

Let $M = \overline{\operatorname{span} \mathcal{T}}_{L^2} \subset L^2(\mathbb{R})$, and denote its orthogonal complement by

$$M^\perp = \{\, f \in L^2(\mathbb{R}) : \langle f, \phi \rangle = 0,\ \forall \phi \in M \,\}$$

If $f \in M^\perp$, then in particular

$$\langle f, \varphi(\cdot - a) \rangle = 0 \qquad \forall a \in \mathbb{R},$$

because each translate $\varphi(\cdot - a)$ belongs to $\mathcal{T} \subset M$.

By the previous proposition, this implies $f = 0$ a.e., hence $M^\perp = \{0\}$.

Therefore $L^2(\mathbb{R}) = M \oplus M^\perp = M \oplus \{0\} = M$, i.e. $\overline{\operatorname{span} \mathcal{T}}_{L^2} = L^2(\mathbb{R})$.

This proves the density of finite linear combinations of Gaussian translates.

Historical remark: Wiener–Tauberian viewpoint

The above argument is a concrete instance of a general principle due to Wiener.

Wiener's Tauberian theorem (informal version)

Let $f \in L^1(\mathbb{R})$. If the Fourier transform $\widehat{f}(\xi)$ never vanishes, then the closed linear span in L^1 of the translates $\{f(\cdot - a) \ : \ a \in \mathbb{R}\}$ is the whole space $L^1(\mathbb{R})$.

Since the Gaussian φ has a nonvanishing Fourier transform, Wiener's theorem implies directly that its translates are dense in $L^1(\mathbb{R})$, and hence also in $L^2(\mathbb{R})$.

Gaussians and Finite Data

Goal: Interpolation Problem

We fix a nontrivial Gaussian on $\mathbb{R}^d$, for example $\varphi(x) = e^{-\gamma|x|^2}, \quad \gamma > 0.$

Interpolation Problem

Given a finite data set

$$\{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R},$$

show that there exists a finite linear combination of translations of φ,

$$f(x) = \sum_{j=1}^{N} \alpha_j \, \varphi(x - c_j),$$

that satisfies the interpolation conditions

$$f(x_i) = y_i, \quad i = 1, \ldots, N.$$

We will see that it is enough to choose $c_j = x_j$ and then solve a finite-dimensional linear system.

Step 1: Choosing the Centres

Let us take the centres at the data locations:

$$c_j = x_j, \quad j = 1, \ldots, N.$$

We consider the ansatz

$$f(x) = \sum_{j=1}^{N} \alpha_j \, \varphi(x - x_j).$$

The interpolation conditions $f(x_i) = y_i$ give

$$\sum_{j=1}^{N} \alpha_j \, \varphi(x_i - x_j) = y_i, \quad i = 1, \ldots, N.$$

In matrix notation, let $\alpha = (\alpha_1, \ldots, \alpha_N)^\top$, $y = (y_1, \ldots, y_N)^\top$ and $K_{ij} = \varphi(x_i - x_j)$. Then the system becomes $K\alpha = y$.

Conclusion: If K is invertible, then there is a unique vector α and hence a finite combination of Gaussians that matches the data.

Step 2: The Gaussian Gram Matrix

Define the *kernel*

$$k(x, z) = \varphi(x - z) = e^{-\gamma |x-z|^2}, \quad \gamma > 0.$$

The matrix K with entries $K_{ij} = k(x_i, x_j)$ is called the *Gram matrix* of the points $\{x_i\}_{i=1}^N$ with respect to this kernel.

Key fact. The Gaussian kernel is *strictly positive definite*, i.e., for any distinct points $x_1, \ldots, x_N$ and any coefficients $\alpha_1, \ldots, \alpha_N$ (not all zero),

$$\sum_{i,j=1}^{N} \alpha_i \alpha_j \, k(x_i, x_j) > 0.$$

This is equivalent to saying that every Gram matrix $K = (k(x_i, x_j))_{i,j}$ is symmetric positive definite and hence invertible.

Positive Definiteness of the Gaussian Kernel (Sketch)

The Gaussian kernel is *translation invariant*: $k(x, z) = \varphi(x - z)$.

Its Fourier transform is again a Gaussian: $\widehat{\varphi}(\xi) = C\, e^{\frac{-|\xi|^2}{4\gamma}}$, $C > 0$, which is *strictly positive for all* $\xi \in \mathbb{R}^d$.

Bochner's theorem (informal)

A continuous, translation-invariant kernel $k(x - z)$ is positive definite if and only if its Fourier transform is a nonnegative (finite) measure. In particular, if $\widehat{\varphi}(\xi) > 0$ for all ξ, then the kernel is strictly positive definite.

Applying Bochner's theorem to the Gaussian:

$$\widehat{\varphi}(\xi) > 0 \quad \Rightarrow \quad k(x, z) = \varphi(x - z)$$

is strictly positive definite. Therefore, each Gram matrix K is symmetric positive definite and hence invertible.

Step 4: Conclusion for Finite Data Sets

Given any finite data set $\{(x_i, y_i)\}_{i=1}^N$ with distinct x_i's:

1. Form the Gaussian Gram matrix

$$K_{ij} = \varphi(x_i - x_j).$$

2. By strict positive definiteness of the Gaussian kernel, K is symmetric positive definite and therefore invertible.
3. Solve the linear system $K\alpha = y$.
4. Define

$$f(x) = \sum_{j=1}^{N} \alpha_j \, \varphi(x - x_j).$$

Then f satisfies the interpolation conditions

$$f(x_i) = y_i, \quad i = 1, \ldots, N.$$

Hence: To represent a finite data set, it is sufficient to consider a *finite* linear combination of translations of a single Gaussian.

Remarks

- The argument is purely *finite-dimensional*: on N points we work in the N-dimensional space spanned by $\{\varphi(\cdot - x_j)\}_{j=1}^{N}$.
- Strict positive definiteness guarantees that these N functions are linearly independent on the set $\{x_1, \ldots, x_N\}$, so we can interpolate arbitrary data values.
- The same reasoning applies to *other strictly positive definite kernels* (e.g., Matérn, inverse multiquadrics).

Other basis functions

Fourier transform convention

Throughout, for $\varphi \in L^1(\mathbb{R}^d)$ we use

$$\widehat{\varphi}(\xi) = \int_{\mathbb{R}^d} \varphi(x)\, e^{-ix\cdot\xi}\, \mathrm{d}x, \qquad \xi \in \mathbb{R}^d.$$

We consider the translation family

$$\mathcal{T}(\varphi) := \{\varphi(\cdot - a) : a \in \mathbb{R}^d\}.$$

Question: what kind of φ can serve as the building blocks for L^p functions?

Wiener–Tauberian theorem (informal form)

Wiener–Tauberian theorem (translation completeness)

Let $\varphi \in L^1(\mathbb{R}^d)$ and assume

$$\widehat{\varphi}(\xi) \neq 0 \quad \text{for all } \xi \in \mathbb{R}^d.$$

Then, for every $1 \leqslant p < \infty$, the closed linear span of its translates satisfies

$$\overline{\operatorname{span}}_{L^p} \mathcal{T}(\varphi) = L^p(\mathbb{R}^d).$$

In particular, if

$$\widehat{\varphi}(\xi) > 0 \quad \forall \xi \in \mathbb{R}^d,$$

then $\widehat{\varphi}$ has no zeros and the translates of φ form a dense set in L^p.

Example 1: Poisson / Cauchy kernel in $\mathbb{R}$

Consider in one dimension

$$\varphi(x) = \frac{1}{\pi(1+x^2)}, \qquad x \in \mathbb{R}.$$

Then $\varphi \in L^1(\mathbb{R})$ and a direct computation gives

$$\widehat{\varphi}(\xi) = \int_{\mathbb{R}} \frac{1}{\pi(1+x^2)} e^{-ix\xi}\,\mathrm{d}x = e^{-|\xi|} > 0 \quad \forall \xi \in \mathbb{R}.$$

- Hence $\widehat{\varphi}$ has no zeros.
- By Wiener–Tauberian, the closed linear span of $\{\varphi(\cdot - a) : a \in \mathbb{R}\}$ is dense in $L^p(\mathbb{R})$ for every $1 \leqslant p < \infty$.

Example 2: Laplace-type kernel $e^{-\alpha|x|}$

In one dimension, let $\alpha > 0$ and

$$\varphi(x) = e^{-\alpha|x|}, \qquad x \in \mathbb{R}.$$

Then $\varphi \in L^1(\mathbb{R})$ and

$$\widehat{\varphi}(\xi) = \int_{\mathbb{R}} e^{-\alpha|x|} e^{-ix\xi} \, \mathrm{d}x = \frac{2\alpha}{\alpha^2 + \xi^2} > 0 \quad \forall \xi \in \mathbb{R}.$$

Consequence:

- $\widehat{\varphi}$ has no zeros $\Rightarrow$ by Wiener–Tauberian, the translates of $e^{-\alpha|\cdot|}$ are dense in $L^p(\mathbb{R})$ for all $1 \leqslant p < \infty$.

Similar kernels in higher dimensions (e.g. $\varphi(x) = e^{-\alpha|x|}$, $x \in \mathbb{R}^d$) also have strictly positive radial Fourier transforms.

Example 3: Radial basis functions with positive FT

A useful class comes from *positive definite radial functions*:

- Let $\psi : [0, \infty) \to \mathbb{R}$ and define $\varphi(x) := \psi(|x|^2)$, $x \in \mathbb{R}^d$.
- If φ is continuous, integrable and positive definite (in Bochner's sense), then there exists a positive finite measure μ such that

$$\varphi(x) = \int_{\mathbb{R}^d} e^{ix\cdot\xi} \mathrm{d}\mu(\xi).$$

If μ has a density $m(\xi)$ with respect to Lebesgue measure, we have

$$\widehat{\varphi}(\xi) = m(\xi) \geq 0.$$

Under mild extra assumptions one often has $m(\xi) > 0$ for all ξ. Then, again, the translates of φ are dense in $L^p(\mathbb{R}^d)$.

Typical RBF examples

Important radial basis functions whose Fourier transforms are strictly positive include:

- **Gaussian:** $\varphi(x) = e^{-\alpha|x|^2}$, $\alpha > 0$.

$$\widehat{\varphi}(\xi) = c_\alpha e^{-|\xi|^2/(4\alpha)} > 0.$$

- **Inverse multiquadric:** $\varphi(x) = (1 + |x|^2)^{-\beta}$ for suitable β. Its Fourier transform is positive and radial.
- **Matérn kernels:** $\varphi(x)$ of the form

$$\varphi(x) = c\,(\nu, d)\,(|x|)^\nu K_\nu(\lambda|x|),$$

  with K_ν the modified Bessel function. Their Fourier transforms are strictly positive rational-type functions in $|\xi|$.

In all these cases, the non-vanishing of $\widehat{\varphi}$ implies that finite linear combinations of translates of φ approximate any L^p function, $1 \leqslant p < \infty$.

Summary

- Wiener–Tauberian theorem: if $\varphi \in L^1(\mathbb{R}^d)$ and $\widehat{\varphi}$ has no zeros, then translates of φ are dense in L^p.
- The condition $\widehat{\varphi}(\xi) > 0$ for all ξ is a convenient sufficient condition.
- Beyond the Gaussian, important examples include:
  - Poisson/Cauchy kernel,
  - Laplace-type kernels $e^{-\alpha|x|}$,
  - Positive definite radial basis functions (Gaussian, inverse multiquadric, Matérn, etc.).
- These kernels are widely used in approximation theory, RBF methods and machine learning, precisely because their translates are so rich.

Cybenko's contribution

Wiener–Tauberian approach revisited

Recall Wiener's theorem (informal form):

Wiener–Tauberian (translation completeness)

Let $\varphi \in L^1(\mathbb{R}^d)$ and assume

$$\widehat{\varphi}(\xi) \neq 0 \quad \forall\, \xi \in \mathbb{R}^d.$$

Then, for $1 \leqslant p < \infty$, the closed linear span of its translates

$$\overline{\operatorname{span}}_{L^p}\{\varphi(\cdot - a) : a \in \mathbb{R}^d\} = L^p(\mathbb{R}^d).$$

Strategy:

- Work in the Fourier domain, using $\widehat{\varphi}$.
- Non-vanishing of $\widehat{\varphi}$ gives invertibility of convolution operators generated by φ.

Application to the approximation using neural networks

Consider a single-hidden-layer network with sigmoid activation, in which the functions are of the form

$$F(x) = \sum_{j=1}^{N} \alpha_j \, \sigma(w_j \cdot x + \theta_j), \quad x \in \mathbb{R}^d.$$

Question: How to show that this family of functions can approximate any given function?

Why this fails for the sigmoid σ (I)

Consider the standard sigmoid

$$\sigma(t) := \frac{1}{1+e^{-t}}, \qquad t \in \mathbb{R}.$$

Issue 1: σ is not in $L^1(\mathbb{R})$.

- As $t \to +\infty$, $\sigma(t) \to 1$; as $t \to -\infty$, $\sigma(t) \to 0$.

Hence

$$\int_{\mathbb{R}} |\sigma(t)|\, dt = \infty$$

and $\sigma \notin L^1(\mathbb{R})$.

Consequence: the standard Fourier transform

$$\widehat{\sigma}(\xi) = \int_{\mathbb{R}} \sigma(t) e^{-it\xi}\, dt$$

does not exist as an L^1-Fourier transform, so Wiener's theorem *cannot be formulated* with σ in the same way as with Gaussians.

Why this fails for the sigmoid σ (II)

One might try to "repair" this by centering or differentiating:

- Centered function $\sigma(t) - \frac{1}{2}$ decays like $e^{-|t|}$, so it *is* in $L^1(\mathbb{R})$.
- The derivative $\sigma'(t) = \sigma(t)\big(1 - \sigma(t)\big)$ also belongs to $L^1(\mathbb{R})$.

However:

- Neural networks use the family

$$\{\sigma(w \cdot x + \theta) : w \in \mathbb{R}^d, \theta \in \mathbb{R}\},$$

  which involves *both* translations and scalings (non-stationary).
- Wiener's theorem is about the closed span of *pure translations* of a *fixed* kernel φ.

Thus, even if we replaced σ by a centered or differentiated version, the usual Wiener analysis (based on convolution and Fourier transform) does not capture the much larger, non-translation-invariant family of functions used in neural networks.

Neural networks with affine transforms

For a single-hidden-layer network with sigmoid activation, we consider functions of the form

$$F(x) = \sum_{j=1}^{N} \alpha_j \, \sigma(w_j \cdot x + \theta_j), \quad x \in \mathbb{R}^d.$$

Observations:

- Each neuron is an *affine* deformation of the base sigmoid (scaling and translation in the input).
- The family

  $$\mathcal{H}_\sigma := \left\{ \sum_{j=1}^{N} \alpha_j \, \sigma(w_j \cdot x + \theta_j) \right\}$$

  is much richer than the span of translations $\{\sigma(\cdot - a)\}_{a \in \mathbb{R}^d}$.
- Density of $\mathcal{H}_\sigma$ in $C([0,1]^d)$ is not a question of convolution invertibility, but of *separating sets* and *functional-analytic duality*.

We therefore need a different approach: this is the Cybenko's theorem.

Cybenko's universal approximation theorem

Definition (sigmoid activation)

A function $\sigma : \mathbb{R} \to \mathbb{R}$ is called a *sigmoid function* if it is bounded, measurable, and

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

and it is not almost everywhere constant.

Cybenko's theorem (1989, informal)

Let σ be a sigmoid function. Consider

$$\mathcal{H}_\sigma = \left\{ x \mapsto \sum_{j=1}^{N} \alpha_j \, \sigma(w_j \cdot x + \theta_j) : N \in \mathbb{N},\ \alpha_j \in \mathbb{R},\ w_j \in \mathbb{R}^d,\ \theta_j \in \mathbb{R} \right\}.$$

Then $\mathcal{H}_\sigma$ is dense in $C([0,1]^d)$ in the uniform norm.

In other words: *a single hidden layer with sigmoid activation* can approximate any continuous function on $[0,1]^d$, arbitrarily well.

Proof strategy: duality and signed measures

The proof follows a standard functional-analytic scheme:

Let $X := C([0,1]^d)$ with the sup norm. Suppose $\mathcal{H}_\sigma$ is *not* dense in X. By Hahn–Banach, there exists a nonzero continuous linear functional $L : X \to \mathbb{R}$ with $L(f) = 0 \quad$ for all $f \in \overline{\mathcal{H}_\sigma}$.

By the Riesz representation theorem, there exists a finite signed measure μ on $[0,1]^d$ such that

$$L(f) = \int_{[0,1]^d} f(x)\mathrm{d}\mu(x), \quad \forall f \in X.$$

In particular, L vanishes on $\mathcal{H}_\sigma$ if and only if

$$\int_{[0,1]^d} \sigma(w \cdot x + \theta)\mathrm{d}\mu(x) = 0 \quad \text{for all } w \in \mathbb{R}^d,\ \theta \in \mathbb{R}.$$

So, to prove density, it is enough to show:

Key lemma

If μ is a finite signed measure on $[0,1]^d$ and $\int \sigma(w \cdot x + \theta)\mathrm{d}\mu(x) = 0 \quad \forall w, \theta$, then $\mu = 0$.

Step 1: Approximating half-space indicators

Fix $w \in \mathbb{R}^d$, $\theta \in \mathbb{R}$, and consider the half-space $H_{w,\theta} := \{x \in \mathbb{R}^d : w \cdot x + \theta \geq 0\}$.

For a large parameter $\lambda > 0$, define the sharply scaled sigmoid $\sigma_\lambda(t) := \sigma(\lambda t)$. Then:

- $\sigma_\lambda(w \cdot x + \theta)$ converges pointwise to $\mathbf{1}_{H_{w,\theta}}(x)$, except on the boundary $\{x : w \cdot x + \theta = 0\}$.
- Since σ is bounded, the family $\sigma_\lambda(w \cdot x + \theta)$ is uniformly bounded.

By dominated convergence theorem,

$$\int \sigma_\lambda(w \cdot x + \theta) \mathrm{d}\mu(x) \xrightarrow[\lambda \to \infty]{} \int \mathbf{1}_{H_{w,\theta}}(x) \mathrm{d}\mu(x) = \mu(H_{w,\theta}).$$

But for each λ,

$$\int \sigma_\lambda(w \cdot x + \theta) \mathrm{d}\mu(x) = 0$$

by assumption (replace w by λw and θ by $\lambda\theta$). Hence $\mu(H_{w,\theta}) = 0$ for all w, θ.

Step 2: From half-spaces to all Borel sets

We have established

$$\mu(H_{w,\theta}) = 0 \quad \text{for all } w \in \mathbb{R}^d,\ \theta \in \mathbb{R}.$$

Now:

- Finite intersections of such half-spaces generate a rich collection of sets, including *rectangles* in $[0,1]^d$.
- By standard measure theoretic arguments (monotone class theorem), the collection of sets on which μ vanishes is a σ-algebra that contains all finite intersections of half-spaces.
- Therefore, this σ-algebra contains all Borel sets of $[0,1]^d$.

Hence

$$\mu(B) = 0 \quad \text{for all Borel sets } B \subset [0,1]^d.$$

This implies $\mu = 0$ as a signed measure, completing the proof of the key lemma.

Conclusion of the proof

Recall:

- If $\mathcal{H}_\sigma$ is not dense in $C([0,1]^d)$, there exists a nonzero continuous linear functional L vanishing on $\mathcal{H}_\sigma$.
- This L is represented by a finite signed measure μ via

$$L(f) = \int f \mathrm{d}\mu.$$

- Vanishing of L on $\mathcal{H}_\sigma$ implies

$$\int \sigma(w \cdot x + \theta) \mathrm{d}\mu(x) = 0 \quad \forall w, \theta.$$

- By the key lemma, this forces $\mu = 0$, hence $L = 0$, contradicting the assumption $L \neq 0$.

Therefore, $\mathcal{H}_\sigma$ is dense in $C([0,1]^d)$, which is exactly Cybenko's universal approximation theorem.

Summary and comparison

- **Wiener approach**: works for a fixed $\varphi \in L^1(\mathbb{R}^d)$ and its translates; relies on non-vanishing of $\widehat{\varphi}$ and convolution invertibility.
- **Sigmoid**: not in L^1, and, more importantly, neural nets use the non-stationary family $\sigma(w \cdot x + \theta)$, not mere translates of a single kernel.
- **Cybenko's theorem**: replaces Fourier analysis by
  - functional-analytic duality (Hahn–Banach, Riesz),
  - approximation of indicator functions of half-spaces by steep sigmoids,
  - measure-theoretic arguments to conclude that the annihilating measure must be zero.
- Conceptually, both Wiener's theorem and Cybenko's theorem say: *if the only continuous linear functional vanishing on your generator family is the zero functional, then you are dense.* The technical tools, however, are quite different.

Reference

- N. Wiener, *Tauberian theorems*, Annals of Mathematics, 33 (1932), 1–100.
- N. Wiener, *The Fourier Integral and Certain of its Applications*, Cambridge University Press, 1933.
- Cybenko, George, *Approximation by superpositions of a sigmoidal function*, Mathematics of Control, Signals and Systems, 2(4), 303–314, 1989.