- idea: two-layer neural net ≈ combination of nonlinear functions

\[ h_i = \sigma (w_i^T x + b_i) \]

\[ y = c^T h \]

- Q: What functions can be approximated by a 2-layer net?

- Barron’s representor theorem
  - for any function \( f(x) \in \mathcal{L}_1(\mathbb{R}^d) \), there is a function \( g(x) \) representable by a two-layer neural net with \( k \) hidden units such that \( \| f(x) - g(x) \|_2 \leq O \left( \frac{1}{k} \right) \).

- \( \mathcal{L}_1(\mathbb{R}^d) \): set of functions whose Fourier coefficients have \( l_1 \)-norm = 1.
  - in particular: low-frequency, low deg polynomials.

- abstraction of representor theorems

- Given a set \( S \) of functions of norm at most 1.
  - (think: \( S = \) set of functions representable by 1 hidden unit)

- Given a function \( f \in \text{conv} \{ S \} \)

- There exists \( c_1, c_2, \ldots, c_k, f_1, \ldots, f_k \in S \) s.t.
if we set \( g(x) = \frac{1}{K} \sum_{i=1}^{K} C_i f_i(x) \) then
\[
\|f - g\|^2 = \frac{1}{K}
\]

- in general in d-dimensional space, if a point is in convex hull, can be represented by d+1 points, but functions have infinite dimensions.

- Proof: \( f \in \text{conv}(S) \) means there is a distribution \( D \) on \( S \) such that
\[
\mathbb{E}_h h(x) = f(x)
\]

where \( h \sim D \)

sample \( f_i(x) \sim D \),

we know \( \mathbb{E}_i f_i(x) = f(x) \)
\[
\mathbb{E} \|f - f\|^2 \leq 1
\]

so if we let \( g(x) = \frac{1}{K} \sum_{i=1}^{K} f_i(x) \)

we know \( \mathbb{E} g(x) = f(x) \)
\[
\mathbb{E} \|g - f\|^2 = \frac{1}{K}
\]

- problem: representor theorem is not constructive.

Know existence of a representation, but not how to find a representation.
- Greedy approach (Frank-Wolfe algorithm)

Suppose we currently have

\[ g_k = \sum_{i=1}^{K} c_i f_i(x) \]

\[ \|g_k - f\|^2 \leq \frac{1}{K} \]

how to find \( g_{k+1} \) such that \( \|g_{k+1} - f\|^2 \leq \frac{1}{k+1} \)?

**Plan:** let \( g_{k+1} = \lambda g_k + (1-\lambda)f_{k+1}(x) \)

\( \lambda, f_{k+1} \) : unknown

solve the optimization problem

\[
\min_{\lambda, f_{k+1}} \|g_{k+1} - f\|^2 \\
\text{s.t.} \quad \lambda \in [0,1] \\
\quad f_{k+1} \in S.
\]

**Claim:** \( \|g_{k+1} - f\|^2 \leq \frac{1}{k+1} \)

**Proof:** let \( g_{k+1} = \lambda g_k + (1-\lambda)f_{k+1}(x) \)

where \( f_{k+1} \sim D \) (recall \( \mathbb{E}_{D}[h] = f \))

\[
\mathbb{E}[\|g_{k+1} - f\|^2] = \mathbb{E}[\|\lambda(g_k - f) + (1-\lambda)(f_{k+1} - f)\|^2] \\
= \lambda^2 \|g_k - f\|^2 + (1-\lambda)^2 \mathbb{E}[\|f_{k+1} - f\|^2]
\]

no cross term

\[
\leq \frac{\lambda^2 + (1-\lambda)^2}{K} = \frac{(k+1)\lambda^2 - 2\lambda + 1}{k}
\]

let \( \lambda = \frac{1}{k+1} \) we have

\[
\mathbb{E}[\|g_{k+1} - f\|^2] \leq \frac{1}{k+1}.
\]
- Problem: even the simpler problem of optimizing a single unit \( f_{k+1} \) is not easy.

This is still open for neural nets, but in special cases it can be solved/approximated.

Example: if \( \sigma(x) = x^2 \) (nonlinear), then finding optimal \( f_{k+1} \) can be solved by SVD.