Lecture 2 : The curse of dimensionality ; Neural Networks and approximation.

[Bach '21] [Telgarski, '20] (DLT 'Fall20).

Basic Supervised Learning Setup

$\boxed{\text{Goal}}$. Given data $\{(x_i, y_i)\}_{i=1..n}$ with $x_i \in \mathcal{X}$ $\boxed{\text{(high-dimension)}}$, $y_i \in \mathcal{Y}$ labels.

x_i: input, y_i: label.

estimate a mapping $f : \mathcal{X} \to \mathcal{Y}$ that ? generalises to ? unseen data

glorified Interpolation

→ IID Assumption : data is drawn iid from a distribution ν on $\mathcal{X} \times \mathcal{Y}$.

→ Point-wise loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

eg: $\mathcal{Y} \in \{\pm 1\}$: $\ell(y, y') = \log(1 + \exp(-yy'))$

$\ell(y, y') = \frac{1}{2}|y - y'|^2$ ↳ $\mathcal{Y} = \mathbb{R}$

Given any $f : \mathcal{X} \to \mathcal{Y}$, this defines

population risk $R(f) = \mathbb{E}_\nu [\ell(f(x), y)]$

$(\mathbb{E}\, \hat{R}(f) = R(f))$

empirical risk $\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)$ ← unbiased estimator of R.

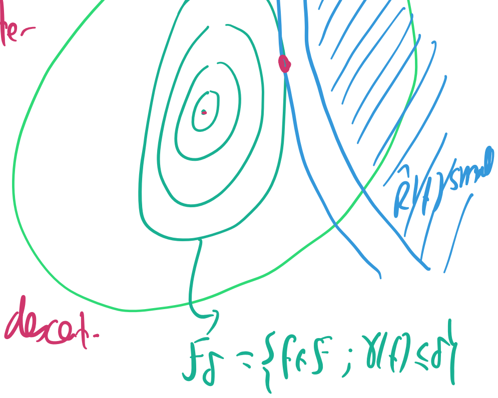
→ Hypothesis Space $\mathcal{F} = \{ f : \mathcal{X} \to \mathcal{Y} \}$

Assume that $\mathcal{F}$ comes with a norm : we can assign to each $f \in \mathcal{F}$ a complexity measure $\gamma(f)$

Example : → norms over the NN parameter weights.

→ number of neurons

→ number of gradient-steps for models f using gradient-descent.



→ Empirical Risk Minimization (ERM).

we search for small empirical risk using small complexity.

(C) $\min\limits_{\gamma(f) \le \delta} \hat{R}(f)$ constrained

(P) $\min\limits_{f \in F} \hat{R}(f) + \lambda \cdot \gamma(f)$ penalized

↳ Lagrange Multiplier.

(I) $\min \gamma(f)$ interpolation

$f \in F$ $\hat{R}(f) = 0$ ← (situations with no noise) $f(x_i) = y_i \quad i = 1 \dots n$.

→ Basic decomposition of error/risk

Let $\hat{f} \in F_\delta$ produced by an arbitrary algorithm → (only sees $\hat{R}$)

ε_{app} = approximation error.

Then

$$R(\hat{f}) - \inf_{f \in F} R(f) = R(\hat{f}) - \inf_{f \in F_\delta} R(f) + \boxed{\inf_{f \in F_\delta} R(f) - \inf_{F} R(f)}$$

$$= \boxed{R(\hat{f}) - \hat{R}(\hat{f})} + \underbrace{\boxed{\hat{R}(\hat{f}) - \inf_{f \in F_\delta} \hat{R}(f)}}_{\varepsilon_{opt}} + \boxed{\inf_{f \in F_\delta} \hat{R}(f) - \inf_{F_\delta} R(f)} + \varepsilon_{app}$$

$\hat{R}(\bar{f}) - R(\bar{f})$ $\bar{f} \in \underset{F_\delta}{\arg\min}\, R(f)$

$$\le \boxed{2 \sup_{f \in F_\delta} |\hat{R}(f) - R(f)|} + \varepsilon_{appr} + \varepsilon_{opt}.$$

truth

↳ $\mathcal{E}$ statistical error.

given function

3 sources of error: → $\mathcal{E}_{app}$; in regression $R(f) = \|f - f^*\|_{L^2}^2$

$\{(x_i, y_i)\}$, with $y_i = f^*(x_i)$

$\inf_{f \in \mathcal{F}_\delta} \|f - f^*\|_{L^2}^2$

↳ As δ increases, $\mathcal{E}_{appr}$ decreases.

→ Statistical error: $\sup_{f \in \mathcal{F}_\delta} |R(f) - \hat{R}(f)|$ ← a random quantity.

measures uniform fluctuations over the ball $\mathcal{F}_\delta$.

Two main quantities drive this error: n: the number of datapoints; δ: size of $\mathcal{F}_\delta$ ball.

What is the expected behavior?

Fix f first. (x_i, y_i) iid.

$$\hat{R}(f) - R(f) = \frac{1}{n} \sum_{i=1}^{n} \left[\ell(f(x_i), y_i) - \mathbb{E}\, \ell(f(x), y) \right]$$

$$\Rightarrow |\hat{R}(f) - R(f)| \sim \frac{\mathrm{std}(\ell(f(x), y))}{\sqrt{n}}$$

(non-asymptotic in n)
(quantified in high-prob using classic tail bounds)

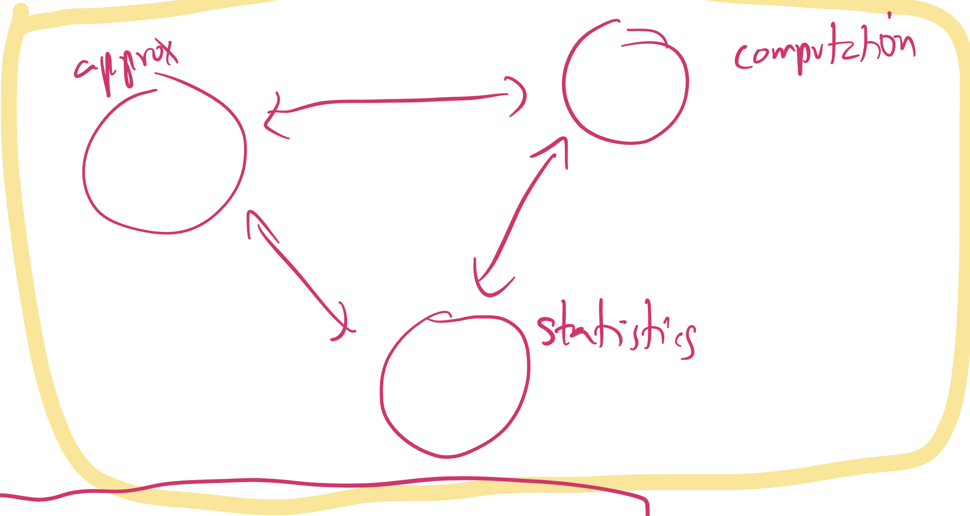
↳ From point-wise control to uniform control, requires some tools from concentration/empirical processes (Rademacher averages, etc.)

$$\mathcal{E}_{stat} \sim \frac{h(\delta, \mathcal{F})}{\sqrt{n}}$$

need to fight each other!

↳ Optim error: $\hat{R}(\hat{f}) - \inf_{f \in \mathcal{F}_\delta} \hat{R}(f)$ measures our ability to solve ERM.

↳ $\mathcal{F}_r$ is non-convex in most practical situations!



The curse of dimensionality : How do approximation and statistical errors behave as input dimension grows?

statistics: we observe $\{(x_i, f^*(x_i))\}_{i=1..n}$ $\quad x_i \sim \mathcal{N}(0, I_d)$, $x_i \in \mathbb{R}^d$.

f^* unknown.

Q: How many samples are needed to estimate f^* up to accuracy ε, ie $\mathbb{E}_x |\hat{f}(x) - f^*(x)|^2 \leq \varepsilon$? (sample complexity).

↳ suppose first f^* is linear: $f^*(x) = \langle x, \theta^* \rangle$ $\quad \theta^* \in \mathbb{R}^d$.

$$\mathcal{F} = \{ f : \mathbb{R}^d \to \mathbb{R} \ ; f(x) = \langle x, \theta \rangle \} \cong \mathbb{R}^d$$

A: $n = d$ are sufficient for exact recovery (and necessary!)
(solve a linear system $\langle x_i, \theta \rangle = \langle x_i, \theta^* \rangle$ $i = 1..d$.

Remark: $f^*(x) = \varphi(\langle x, \theta^* \rangle)$ $\quad \varphi$ is even function, smooth.

$\varphi(t) = |t|$
$= \cos t$.

A: $n = d+1$ sample are sufficient and also necessary

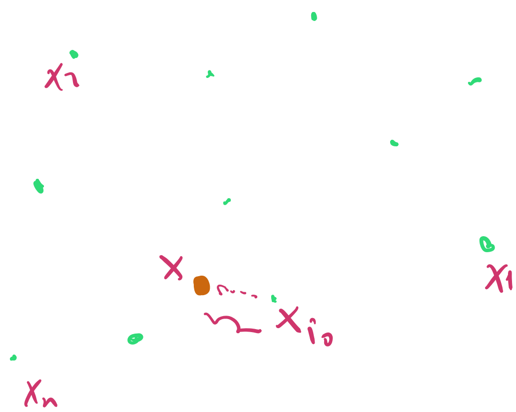
[SZB'21]

↳ Suppose now that f° is only locally linear, ie

β-Lipschitz: $|f(x) - f(x')| \leq \beta \cdot \|x - x'\|_2$ ←

$\mathcal{F} = \{ f : \mathbb{R}^d \to \mathbb{R} \, ; \, f$ bounded Lipschitz $\}$: a Banach space, with norm

$\|f\|_{\mathcal{F}} := \|f\|_\infty + \mathrm{Lip}(f)$.



We consider the smoothest interpolant:

$$\hat{f} = \underset{\mathcal{F}}{\mathrm{argmin}} \{ \mathrm{Lip}(f) \, ; \, f(x_i) = f^\circ(x_i) \; \forall i \}$$

[ERM in interpolant form]

For any x,

$$|\hat{f}(x) - f^\circ(x)| \leq |\hat{f}(x) - \hat{f}(x_{i_0})| + \overbrace{\cancel{|\hat{f}(x_{i_0}) - f^\circ(x_{i_0})|}}^{0}$$
$$+ \underbrace{|f^\circ(x_{i_0}) - f^\circ(x)|}_{\leq \, \beta \|x - x_{i_0}\|}$$

$$\leq 2\beta \|x - x_{i_0}\|$$

$$\Rightarrow \mathbb{E}_{\underbrace{x \sim N(0, I_d)}_{\mu}} |\hat{f}(x) - f^\circ(x)|^2 \leq 4\beta^2 \boxed{\mathbb{E}_{x, \{x_i\}} \|x - x_{i_0}\|^2}$$

training set $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ $\quad \sim W_2^2(\mu, \hat{\mu}_n)$

$\sim n^{-1/d}$ [Dudley '68] [BLG '14].

As a consequence, if $\varepsilon \sim n^{-1/d}$

$\Rightarrow n \sim \varepsilon^{-d}$ we can ensure that we learn with accuracy ε. → $\boxed{\text{cursed by dimension}}$

↳ Is this sample complexity also necessary?

• Consider the box $B = [-1/2, 1/2]^d$

B

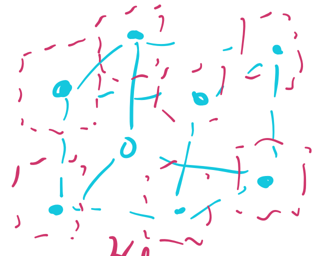
and the function $\Psi : B \to \mathbb{R}$

$$\Psi(x) = \mathrm{dist}(x; \partial B)$$

Exercise: Ψ is 1-Lipschitz.

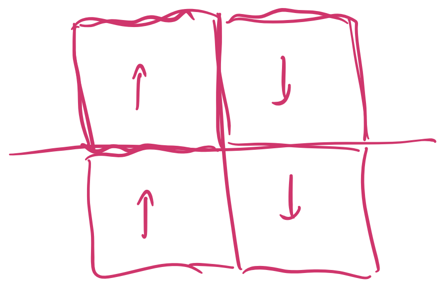
• Consider for each $z \in \{-1/2, 1/2\}^d$

an arbitrary sign $g(z) = \pm 1$



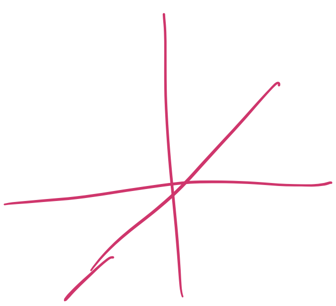
H_d

$$f^*(x) = \sum_{z \in H_d} g(z)\, \Psi(x - z)$$



f^* is 1-Lipschitz and supported in $[-1, 1]^d$.

→ Claim: if $n \ll 2^d$, then any estimator will incur in a relative error



$$\frac{\mathbb{E}_x |\hat{f}(x) - f^*(x)|^2}{\mathbb{E}_x |f^*(x)|^2} = \Theta(1)$$

↳ To summarise:

→ linear functions (or generalised linear functions) $n \sim d$ (easy)

→ Lipschitz functions → $n \sim \varepsilon^{-d}$ (impossible)

In between, Sobolev class:

$$W^{(s)} = \{ f : \mathbb{R}^d \to \mathbb{R} \ ; \ f \text{ has } s\text{-derivatives bounded in } L_p \}$$

sample complexity $\quad n \sim \varepsilon^{-d+2s/s} \quad$ [Tsibakov].

$$\tilde{d} = d/s \qquad \sim \varepsilon^{-\tilde{d}}$$

$\to$ unless $s = \Theta(d)$, no real change.

<u>Conclusion:</u> We will need to search for alternatives.

$\boxed{\text{Shallow}}$ Neural Networks and Approximation

$\to$ Consider $\sigma : \mathbb{R} \to \mathbb{R}$ (Lipschitz) activation function

$$f_K(x;\theta) = \sum_{k=1}^{K} \underbrace{c_k \, \sigma(\langle x, a_k \rangle + b_k)}_{\text{ridge function.}} \qquad x \in \mathbb{R}^d, \ a_k \in \mathbb{R}^d, \ b_k, c_k \in \mathbb{R}.$$

$\theta = \{ a_k, b_k, c_k \}_k$.

$\to$ Space of functions represented with shallow NNs is $\mathcal{H}_\sigma = \{ f_K(\cdot;\theta) \ ; \ \theta \ ; \ K \in \mathbb{N} \}$

<u>Q</u>: How expressive is this set?

<u>Universal Approximation</u>

For a given metric d defined over continuous functions, $\forall f \in C(\mathbb{R}^d)$ and $\forall \varepsilon > 0$, there exist $\hat{f} \in \mathcal{H}_\sigma$

such that $d(f, f) \leq \varepsilon$.

→ To get intuition, let us first consider a 3-layer approximation.

<u>Theorem</u> Let $g: \mathbb{R}^d \to \mathbb{R}$ continuous, $\varepsilon > 0$, $\Omega = [0,1]^d$, and let $\delta > 0$ so that $\|x - x'\|_\infty < \delta$, $x, x' \in \Omega$, then $|g(x) - g(x')| \leq \varepsilon$. Then there exists a 3-layer ReLU network f with $\Theta(\underline{\delta^{-d}})$ and $\|f - g\|_1 \leq 2\varepsilon$.

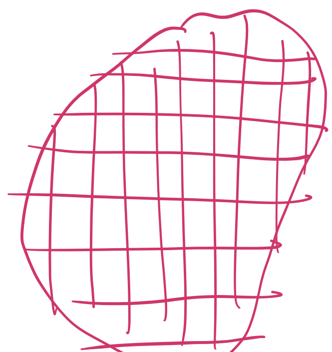
$\|\cdot\|_1 = L^1(\Omega)$

Proof: [Telgarski '20]

<u>Lemma</u>: let $U \subseteq \mathbb{R}^d$, with a partition P of U into rectangles; $P = (R_1 \dots R_N)$, each with lengths $\leq \delta$. Then $\exists$ scalars $\alpha_1 \dots \alpha_N$ with

$\boxed{|P| \sim \delta^{-d}}$ $\sup_{x \in U} |g(x) - h(x)| \leq \varepsilon$, with

$h(x) = \sum_i \alpha_i \mathbb{1}[R_i]$



(in other words, piece-wise constant approximation has small error provided pieces are sufficiently small.

Let P the partition from above of $U = [0, 1+c)^d$ into rectangles $R_i = \prod_j [a_j, b_j)$ with $b_j - a_j \leq \delta$.

$h(x) = \sum_i \alpha_i \mathbb{1}_{R_i}$, so $\|g - h\|_1 \leq \varepsilon$.

. We construct a network of the form

$f(x) = \sum \alpha_i g_i(x)$ where each g_i is a ReLU net that approximates $\mathbb{1}_{R_i}$.

$$\|f-g\|_1 \le \overbrace{\|f-h\|_1}^{\le \varepsilon \text{ from Lemma}} + \|h-g\|_1$$

$$= \Big\| \sum_i \alpha_i (\mathbb{1}_{R_i} - g_i) \Big\|_1 + \varepsilon$$

$$\le \sum_i |\alpha_i| \, \| \mathbb{1}_{R_i} - g_i \|_1 + \varepsilon \quad \le 2\varepsilon$$

If we can build g_i so that

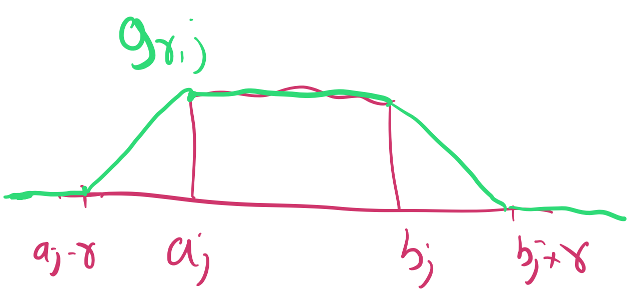
$\to \| \mathbb{1}_{R_i} - g_i \|_1 \le \dfrac{\varepsilon}{\sum_i |\alpha_i|}$, then we are done.

- Fix i, $R_i = \bigotimes [a_j, b_j)$. For $\gamma > 0$, and for each coordinate $j \in \{1..d\}$, let

$$g_{\gamma,j}(z) = \sigma\left(\frac{z-(a_j-\gamma)}{\gamma}\right) - \sigma\left(\frac{z-a_j}{\gamma}\right)$$

$$- \sigma\left(\frac{z-b_j}{\gamma}\right) + \sigma\left(\frac{z-(b_j+\gamma)}{\gamma}\right)$$

↳ $\sigma(t) = \max(0,t)$



$$g_\gamma(x) = \sigma\Big(\boxed{\sum_j g_{\gamma,j}(x)} - (d-1)\Big)$$

↳ We verify that

(i) $g_\gamma(x) = \begin{cases} 1 & \text{if } x \in R_i \\ 0 & \text{if } x \notin \bigotimes_j (a_j-\gamma, b_j+\gamma) \\ [0,1] & \text{otherwise.} \end{cases}$

(ii) $\| g_\gamma - \mathbb{1}_{R_i} \|_1 \leq O(\gamma)$ $\square$

$\boxed{\text{Remark / question.}}$ Are the two hidden layers necessary?

A: Hell no! Recall a classic result in polynomial approximation

Theorem (Stone-Weierstrass) Let $\Omega = [0,1]^d$. Let $\mathcal{F}$ be a function class satisfying:

(i) Each $f \in \mathcal{F}$ is continuous,

(ii) $\forall x \in \Omega$, $\exists f \in \mathcal{F}$ such that $f(x) \neq 0$.

(iii) $\forall x \neq x' \in \Omega$, $\exists f \in \mathcal{F}$ $f(x) \neq f(x')$.

(iv) $\mathcal{F}$ is an algebra: closed under multiplication and vector space operations.

Then $\mathcal{F}$ enjoys universal approx: any continuous $g: \mathbb{R}^d \to \mathbb{R}$ $\forall \varepsilon > 0$, $\exists$ $\underline{f \in \mathcal{F}}$ with $\sup_x |f(x) - g(x)| \leq \varepsilon$.

↳ This theorem can be used to establish UAT for $\mathcal{H}_\sigma$ with general choices of σ:

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

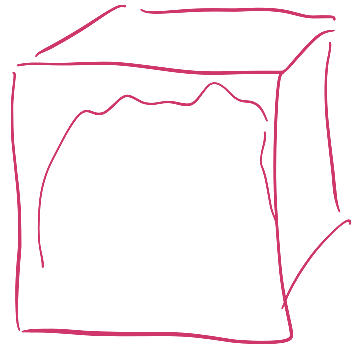
[Hornik et al 89]

↳ $\sigma \neq$ polynomial (Leshno '93)

↳ Is this surprising?
↳ Are the approximation rates cursed using $\mathcal{H}_\sigma$?

The Fourier perspective

→ let $f \in C(\mathbb{R}^d)$ and consider its restriction to a compact Ω (eg $\Omega = [0,1]^d$).



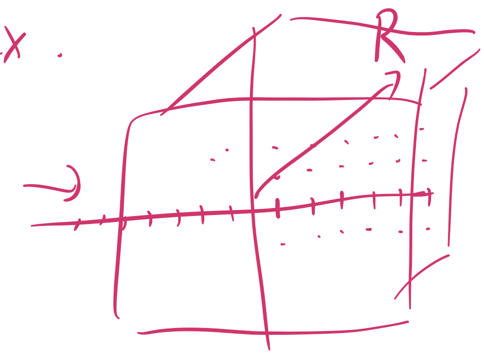
For $\xi \in \mathbb{Z}^d$, consider its Fourier decomposition

$$\hat{f}(\xi) = \left\langle f, e^{i\langle x,\xi\rangle} \right\rangle_{L^2(\Omega)} = \int_\Omega f(x)\, e^{-i\langle x,\xi\rangle}\, dx .$$

↳ Fourier Inversion Lemma:

$$\boxed{f_M(x) = \sum_{\|\xi\| \le R} \hat{f}(\xi)\, e^{i\langle x,\xi\rangle}}$$

$\to f$ in $L^2(\Omega)$ as $M, R \to \infty$



M is the number of frequencies inside $\{\|\xi\| < R\}$.

$$e^{i\langle x,\xi\rangle} = \sigma(\langle x,\xi\rangle) \quad \text{with } \sigma(t) = e^{it} = \cos t + i \sin t.$$

f_M is a shallow NN with M periodic neurons.

We have tight control of how regularity of f

$\updownarrow$

decay of $\hat{f}$ as $\|\xi\|$ grows.

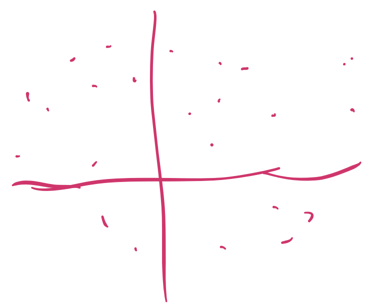
$\boxed{f \in W^s \text{ sobolev}}$ $\Rightarrow$ $\|f - f_M\| = O(M^{-s/d})$ [Tsybakov] [de Vore] etc.

curse of dim in approximation!

↳ Overcome this curse?

Fourier Representation, in L_1

$\hat{f}(\xi)$



$$\hat{f}(\xi) = \int f(x)\, e^{-i\langle x, \xi\rangle}\, dx$$

$$\boxed{f(x) = \int \hat{f}(\xi)\, e^{+i\langle x, \xi\rangle}\, d\xi}$$

If $\|\hat{f}\|_1$ is finite, $\|\hat{f}\|_1 = \int |\hat{f}(\xi)|\, d\xi < +\infty$.

↳

Idea: $f(x) = \int \sigma(\langle x, w\rangle)\, g(w)\, dw$ with g integrable $\|g\|_1 < +\infty$.

$$g(w) = \operatorname{sign}(g(w)) \cdot \|g\|_1 \cdot \underbrace{\frac{|g(w)|}{\|g\|_1}}_{q(w)}$$

$q \geq 0$

$\int q(w)\, dw = 1$

$$f(x) = \int \sigma(\langle x, w\rangle) \cdot \operatorname{sign}(g(w)) \cdot \|g\|_1 \cdot q(w)\, dw$$

$\Rightarrow q$ is a proba density.

$$= \mathbb{E}_q \Big[\underbrace{\sigma(\langle x, w\rangle) \cdot \operatorname{sign}(g(w))}_{\phi(x, w)} \Big] \cdot \|g\|_1$$

$|\phi| \leq 1$

↳ $\hat{f}_M(x) = \frac{1}{M} \sum_{m=1}^{M} \phi(x, w_i)$, $w_i \sim q$ iid.

$\boxed{\text{Monte Carlo estimator}}$

$$\mathbb{E}\, \|f - \hat{f}_M\|^2 \leq \|g\|_1^2 \frac{\mathbb{E}_q \left(\mathbb{E}_x\, \phi(x, w)^2\right)}{M} \leq \frac{\|g\|_1^2 \sup_w \left(\mathbb{E}_x\, \phi^2\right)}{M}$$

Curse of dim is avoided, provided $\|g\|_1 < +\infty$.

Theorem: [Barron '93] Suppose $C = \int \|\widehat{\nabla f}(\xi)\| \, d\xi < +\infty$ with $f, \hat{f} \in L_1$. Then we can approximate f ~~with~~ to accuracy ε with Relu/Sigmoid units with $\sim \frac{C}{\varepsilon^2}$ (No curse!).

↳ Main q: When/Why is C small.

Conclusions / Take-home:

(1) Learning in high-dim efficiently requires new function spaces, adapted to the physical world (images, sounds, etc).