# 深層NNによる関数近似

# Hölder, Sobolev, Besov空間



$\Omega = [0,1]^d \subset \mathbb{R}^d$

- **Hölder space** $(\mathcal{C}^\beta(\Omega))$

$$\|f\|_{\mathcal{C}^\beta} = \max_{|\alpha| \le m} \|\partial^\alpha f\|_\infty + \max_{|\alpha|=m} \sup_{x \in \Omega} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{|x-y|^{\beta - m}}$$

- **Sobolev space** $(W_p^k(\Omega))$

$$\|f\|_{W_p^k} = \left( \sum_{|\alpha| \le k} \|D^\alpha f\|_{L^p(\Omega)}^p \right)^{\frac{1}{p}}$$

- **Besov space** $(B_{p,q}^s(\Omega))$ $(0 < p, q \le \infty, 0 < s \le m)$

$$\omega_m(f,t)_p := \sup_{\|h\| \le t} \left\| \sum_{j=0}^m (-1)^{m-j} \binom{m}{j} f(\cdot + jh) \right\|_{L^p(\Omega)},$$

空間的非一様性

$$\|f\|_{B_{p,q}^s(\Omega)} = \|f\|_{L^p(\Omega)} + \left( \int_0^\infty [t^{-s} \omega_m(f,t)_p]^q \frac{\mathrm{d}t}{t} \right)^{1/q}.$$

滑らかさの度合い

# Sobolevクラスの近似理論



$$\Pi_N := \left\{ \sum_{j=1}^{N} \alpha_j \eta(a_j^\top x + b_j) \mid \alpha_j \in \mathbb{R}, a_j \in \mathbb{R}^d, b_j \in \mathbb{R} \right\}$$

中間層の横幅が$N$の
二層ニューラルネットワーク

**定理**

$\eta$がある開区間で無限回微分可能であり，その開区間のある点$b$において

$$\frac{\partial^k \eta}{\partial x^k}(b) \neq 0 \ (\forall k \in \mathbb{Z}, k \geq 0)$$

とする．すると，$\forall f \in W_p^s([0,1]^d)$に対してある$g \in \Pi_N$が存在して，

$$\|f - g\|_p \lesssim N^{-\frac{s}{d}} \|f\|_{W_p^s}$$

(ノード数$N$の中間層を用いた近似誤差)

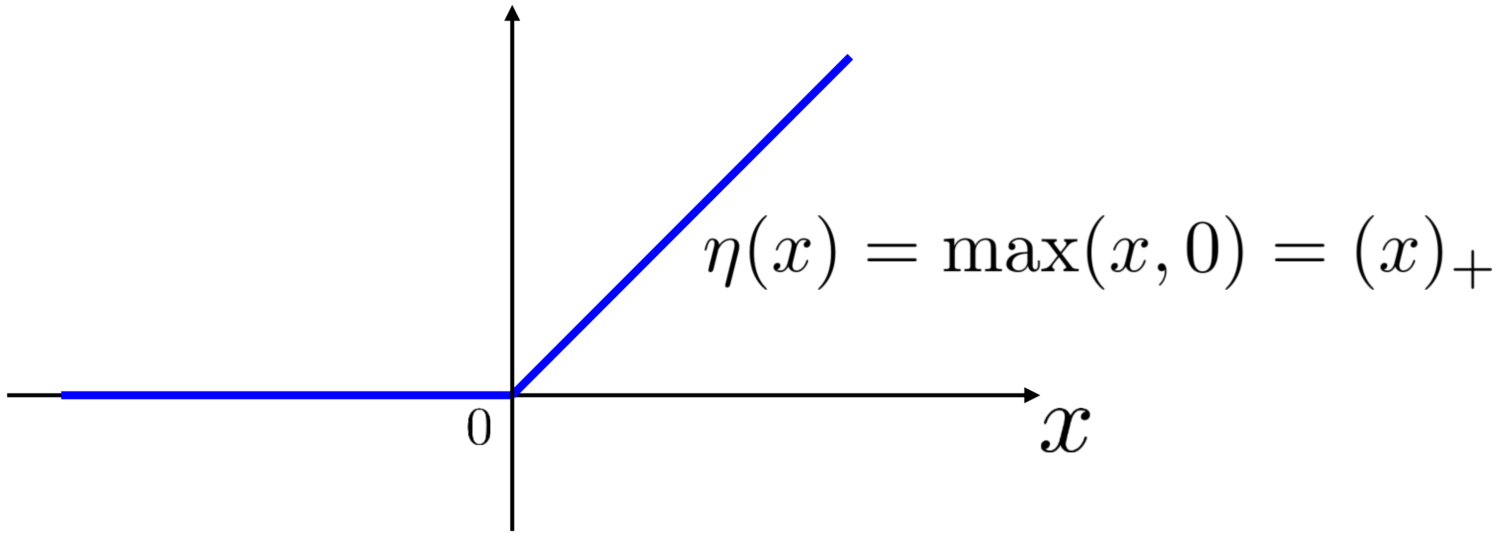
[Mhaskar: Neural networks for optimal approximation of smooth and analytic functions. Neural Computation, 8(1):164–177, 1996]

- この近似誤差は$N$個の基底を用いた近似法の中で最適なオーダーを達成.
- シグモイド関数は条件を満たす．ReLUは満たさない.
- 滑らかな関数はより近似しやすい.

# ReLUの表現力



- ReLU活性化関数




- 現在広く使われている
(LeakyReLUなどの亜種もあるがかなりスタンダード)

- 統計的性質も解明されつつある
  - 万能近似能力あり
  - (区分的) 滑らかな関数の推定
  - 区分線形関数の表現
  - 有理関数の表現
  - 関数のテンソル積，合成関数の表現

**基本：局所多項式近似**

# 滑らかな関数の近似 (Yarotsky, 2016)



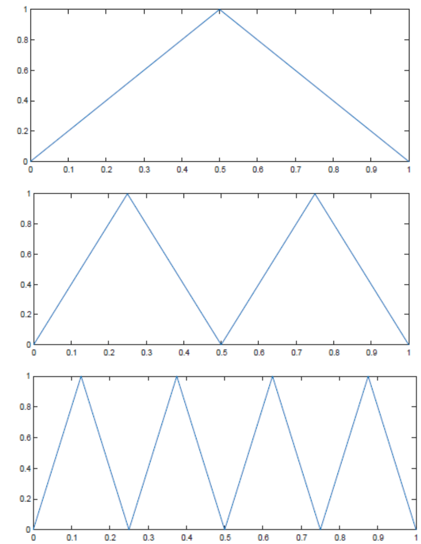
[Yarotsky: Error bounds for approximations with deep ReLU networks. 2016]

- 二次関数の構成

**<u>これが全ての基本</u>**

$$h(x) = \begin{cases} 2x & (0 \le x \le 1/2) \\ 2(1-x) & (1/2 \le x \le 1) \\ 0 & (\text{otherwise}). \end{cases}$$

$$R_k(x) = \underbrace{h \circ h \circ \cdots \circ h}_{k \text{ times}}(x).$$



$h(x)$

$h \circ h(x)$

$h \circ h \circ h(x)$

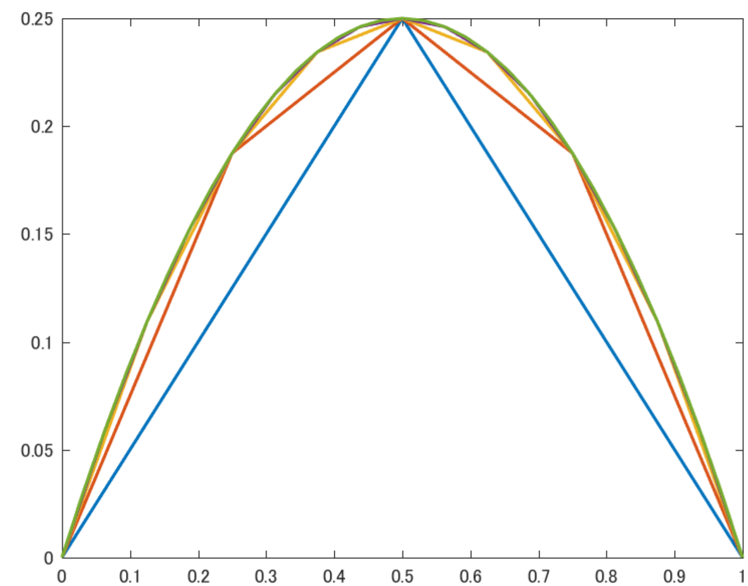
２次関数の近似 (Telgarsky, 2015)

$$\boxed{\left| x(1-x) - \sum_{k=1}^{m} (2^{-2k}) R_k(x) \right| \le 2^{-m}}$$

**層を重ねることで<u>指数的に</u>誤差が減少**

層の数，横幅，ユニット数： $O(\log(1/\epsilon))$

中間層１層の場合： $\Omega(1/\sqrt{\epsilon})$



改善（多項式オーダ→logオーダ）

# 多項式の構成

- 二次関数→掛け算

$$(x+y)^2 - x^2 - y^2 = 2xy$$

（足し算はReLUで実現可能）

- 掛け算→多項式

$$x^m = \underbrace{x \times (x \times (x \times (\cdots)))}_{m\mathrm{times}}$$

（二次関数から構成した
掛け算を繰り返し適用）

$$\sum_{|\alpha| \leq m} c_\alpha (x - x_0)^\alpha$$

（足し算と合わせて多項式を構成）

**→ 滑らかな関数の近似に利用**

# Holder classの近似

$\beta \in (0,\infty),\ m = \lfloor \beta \rfloor$

滑らかな関数のクラス（Holder class）

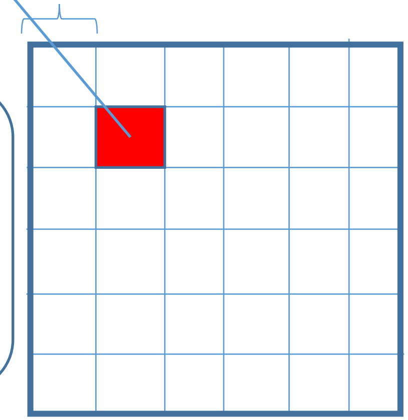
$$\mathcal{F}^{\beta}(K) = \left\{ f \mid \max_{|\alpha|\le m} \|\partial^{\alpha} f\|_{\infty} + \max_{|\alpha|=m} \sup_{x\in[-1,1]^d} \frac{|\partial^{\alpha} f(x) - \partial^{\alpha} f(y)|}{|x-y|^{\beta-m}} \le K \right\}$$

- 滑らかな関数の局所的近似（テイラー展開）

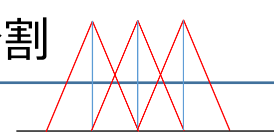
$$\sup_{x:\|x-x_0\|_{\infty}\le\delta} \left| f(x) - \underbrace{\sum_{|\alpha|\le m} \frac{\partial^{\alpha} f(x_0)}{\alpha!} (x-x_0)^{\alpha}}_{m\text{次多項式}\ (R_{x_0} f(x))} \right| \le C\,(d\delta)^{\beta}$$

- 全体の近似

$\delta \simeq \epsilon^{1/\beta} d^{-1}$



$$\sup_{x\in[-1,1]^d} \left| f(x) - \sum_{x_0\in D(\delta)} \underbrace{\prod_{j=1}^{d} (1-\delta^{-1}|x_j - x_{0,j}|)_{+}}_{1\text{の分割}} R_{x_0} f(x) \right| \le C\,(d\delta)^{\beta}$$



(Yarotsky, 2016; Liang&Srikant, 2017)

# 推定理論

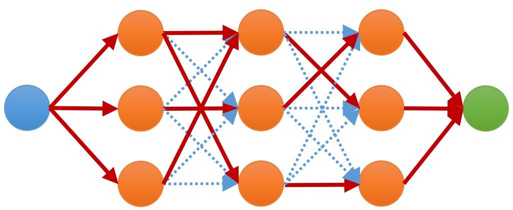


- パラメータ数

$$\underbrace{O\left(\frac{1}{\delta^d}\right)}_{\text{領域分割の数}} \times \underbrace{O(\log(1/\epsilon))}_{\text{分割ごとのパラメータ数}} = O(\epsilon^{-\frac{d}{\beta}} \log(1/\epsilon))$$

横幅 :$O(\epsilon^{-\frac{d}{\beta}} \log(1/\epsilon))$　縦幅：$O(\log(1/\epsilon))$ のネットワークに埋め込める

$\mathcal{F}(L, w, s)$ : 縦幅$L$，横幅$w$，非ゼロパラメータ数$s$ の深層NNモデルの集合



深層学習の汎化誤差（Schmidt-Hieber, 2017）

縦幅$L$ = O(log($n$))，横幅$w$ = O($n^{\frac{d}{2\beta+d}}$log($n$))，非ゼロ要素$s$ = O($n^{\frac{d}{2\beta+d}}$log($n$))

$$\hat{f} = \mathop{\arg\min}_{f \in \mathcal{F}(L,w,s)} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

$$\mathrm{E}[\|\hat{f} - f^*\|_{L_2(P(X))}^2] \leq O(n^{-\frac{2\beta}{2\beta+d}} \log(n)^2)$$

**ミニマックス最適レート**

$$\mathrm{E}[\|\hat{f} - f^*\|_{L_2(P(X))}^2] \leq O\left(\underbrace{\frac{\epsilon^{-d/\beta}\log(\epsilon)^2}{n}}_{\text{バリアンス}} + \underbrace{\epsilon^2}_{\text{バイアス}}\right)$$

$\epsilon = n^{-\frac{\beta}{2\beta+d}}$ でバランス

# Besov空間の近似

# 空間の間の関係



- For $m \in \mathbb{N}$,

$$B^m_{p,1} \hookrightarrow W^m_p \hookrightarrow B^m_{p,\infty},$$
$$B^m_{2,2} = W^m_2.$$

- For $0 < s < \infty$ and $s \notin \mathbb{N}$,

$$\mathcal{C}^s = B^s_{\infty,\infty}.$$

- For $0 < s < m < \infty$, $1 \le p < \infty$, $q \le \infty$,

$$B^s_{p,q} = [L^p, W^m_p]_{s/m,q}.$$

- For $0 < s_1 < s < s_2$, $s = (1-\theta)s_1 + \theta s_2$ and $1 \le q_1, q_2 \le \infty$,

$$B^s_{p,q} = [B^{s_1}_{p,q_1}, B^{s_2}_{p,q_2}]_{\theta,q}.$$

# 補間空間

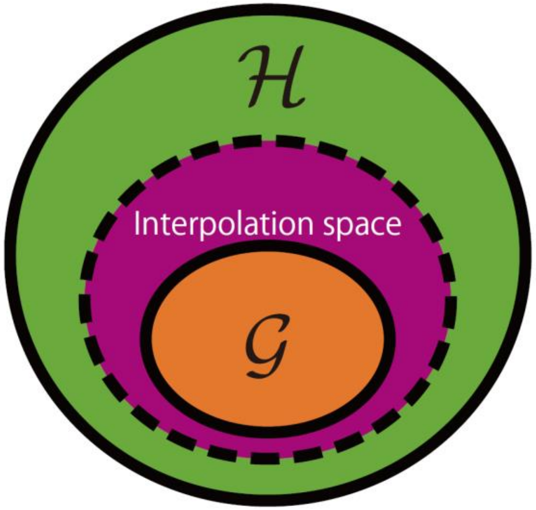


補間空間（線形ノルム空間$\mathcal{H}$と$\mathcal{G}$の間を"補完"，$\mathcal{G} \hookrightarrow \mathcal{H}$とする）：

$$\|f\|_{[\mathcal{H},\mathcal{G}]_{\theta,q}} := \begin{cases} \int_0^\infty [t^{-\theta} \inf_{g\in\mathcal{G}} \{\|f-g\|_{\mathcal{H}} + t\|g\|_{\mathcal{G}}\}]^q \frac{\mathrm{d}t}{t} & (q<\infty), \\ \sup_{t>0} t^{-\theta} \inf_{g\in\mathcal{G}} \{\|f-g\|_{\mathcal{H}} + t\|g\|_{\mathcal{G}}\} & (q=\infty). \end{cases}$$

補間空間は以下を満たす:

$$\mathcal{G} \hookrightarrow [\mathcal{H},\mathcal{G}]_{\theta,q} \hookrightarrow \mathcal{H}.$$

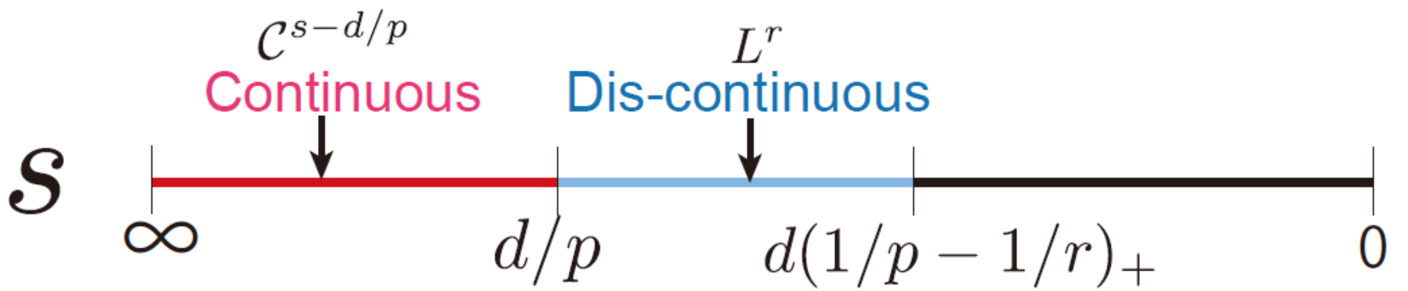


- 連続関数の領域： $s > d/p$

$$B^s_{p,q} \hookrightarrow C^0$$

- $L^r$-可積分な領域：$s > d(1/p - 1/r)_+$

$$B^s_{p,q} \hookrightarrow L^r$$




- 例： $B^1_{1,1}([0,1]) \subset \{\text{bounded total variation}\} \subset B^1_{1,\infty}([0,1])$

# B-spline



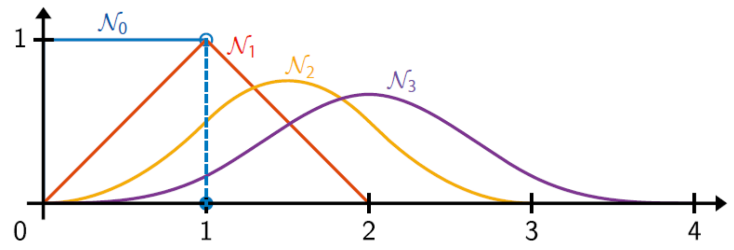
$$\mathcal{N}(x) = \begin{cases} 1 & (x \in [0,1]), \\ 0 & (\text{otherwise}). \end{cases}$$

Cardinal B-spline of order $m$:

$$\mathcal{N}_m(x) = (\underbrace{\mathcal{N} * \mathcal{N} * \cdots * \mathcal{N}}_{m+1 \text{ times}})(x)$$

$$\boxed{f * g(x) = \int f(x-t)g(t)\mathrm{d}t}$$

→ Piece-wise polynomial of order m.




$$\mathcal{N}_{k,j}^{(d)}(x_1, \ldots, x_d) = \prod_{i=1}^{d} \mathcal{N}_m(2^k x_i - j_i)$$

# Cardinal B-spline interpolation (DeVore & Popov, 1988)

- Atomic decomposition:

$f \in B^s_{p,q}$の必要十分条件:

$$f = \sum_{k \in \mathbb{N}+} \sum_{j \in J(k)} \alpha_{k,j} \mathcal{N}^{(d)}_{k,j}$$

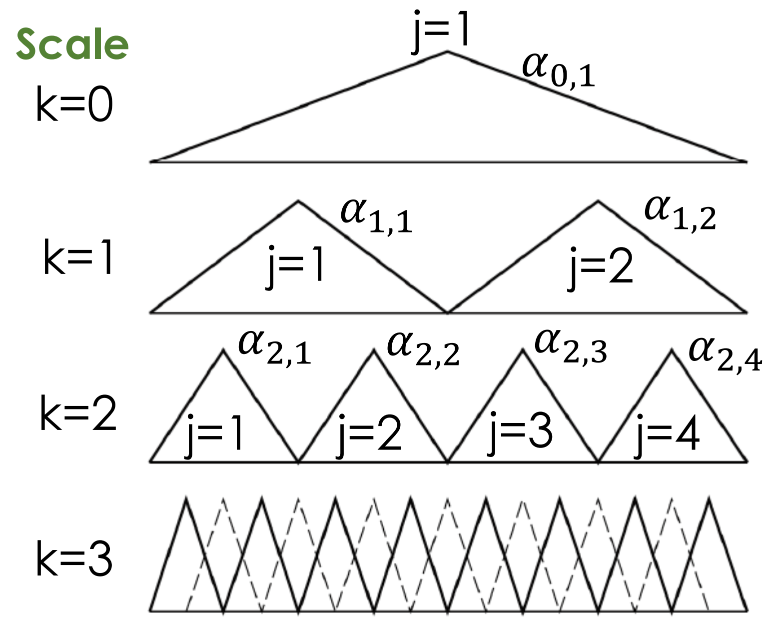
と分解できて
(ただし $J(k) = \{j \in \mathbb{Z}^d \mid -m < j_i < 2^{k_i+1} + m\}$)

$$N(f) = \left[ \sum_{k=0}^{\infty} \{2^{sk} (2^{-kd} \sum_{j \in J(k)} |\alpha_{k,j}|^p)^{1/p}\}^q \right]^{1/q} < \infty$$

$$\|f\|_{B^s_{p,q}} \simeq N(f) \quad \text{(ノルムの同値性)}$$

$$\mathcal{N}^{(d)}_{k,j}(x_1, \dots, x_d) = \prod_{i=1}^{d} \mathcal{N}_m(2^k x_i - j_i)$$

Wavelet/多重解像度展開



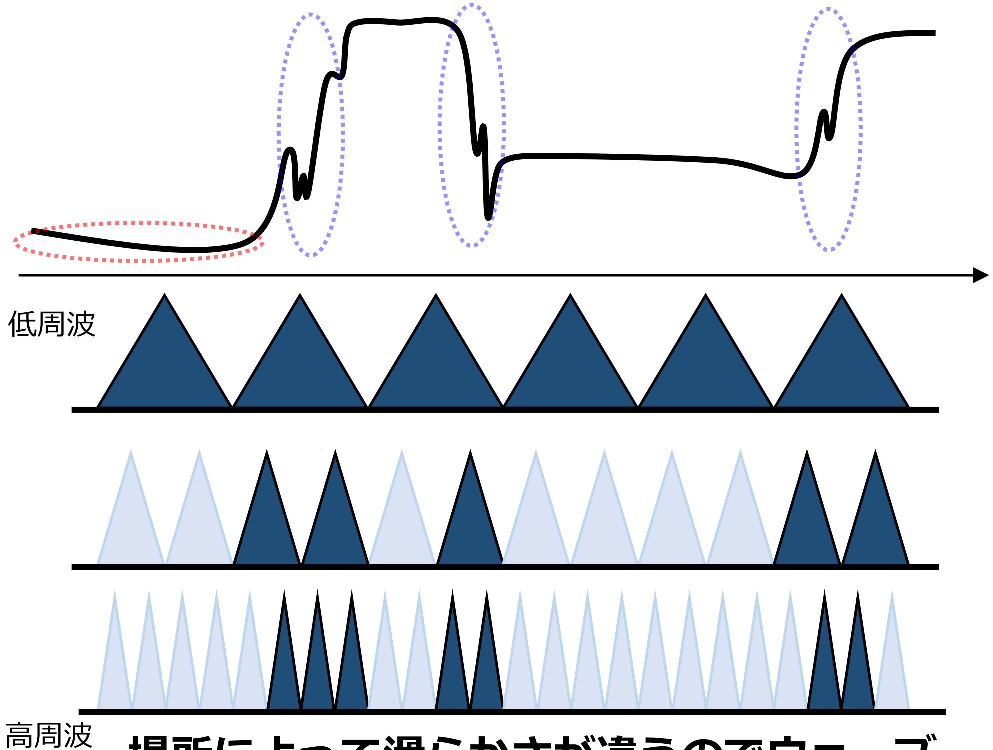

**We show that DNN can approximate each B-spline.**

$$f = \underbrace{\sum_{k,j \in I_N} \alpha_{k,j} \mathcal{N}^{(d)}_{k,j}}_{N \text{ terms}} + O(N^{-s/d})$$

$N$ terms ($f$に応じて適切に選択)

(see also Bolcskei, Grohs, Kutyniok, Petersen: Optimal Approximation with Sparsely Connected Deep Neural Networks. 201

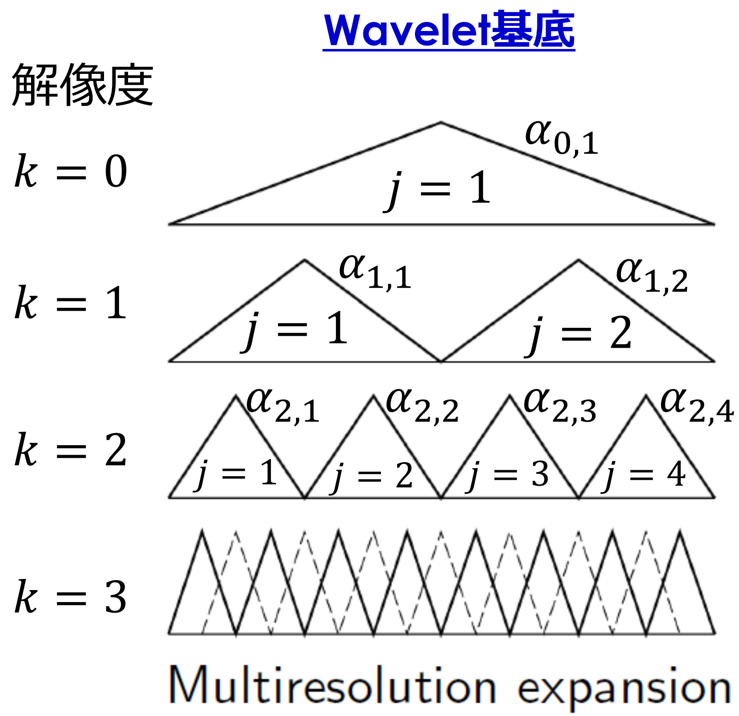
# Besov空間とスパース性との関係




**場所によって滑らかさが違うのでウェーブレット基底のスパースな線形結合が有効**

$$f = \sum_{k \in \mathbb{N}_+} \alpha_k \phi_k$$

Wavelet基底による展開



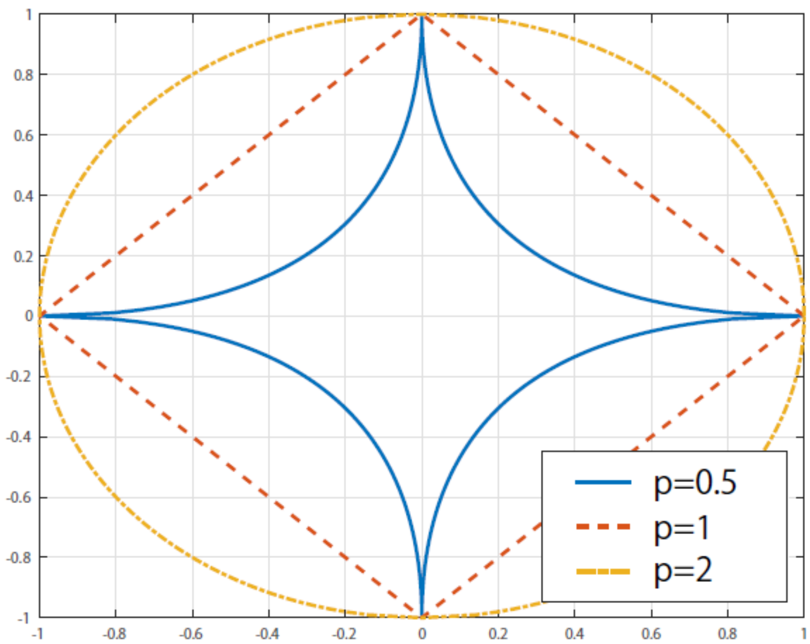

基本的に$\ell^p$-ノルム

$$\|f\|_{B^s_{p,q}} \simeq \left[ \sum_{k=0}^{\infty} \{2^{sk} (2^{-kd} \sum_{j \in J(k)} |\alpha_{k,j}|^p)^{1/p}\}^q \right]^{1/q}$$

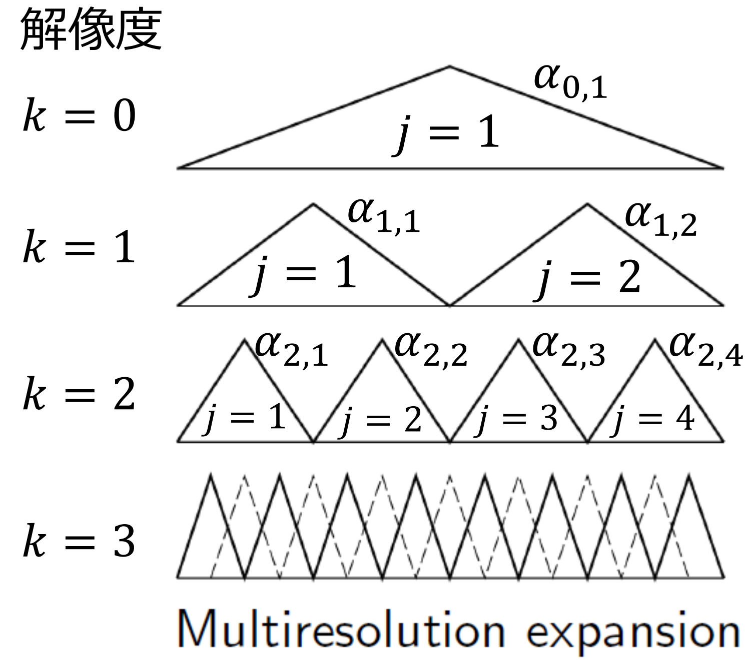
$(0 < p)$

小さな$p$ = スパースな係数 → 空間的な滑らかさの非一様性

# Besov空間とスパース性との関係




Wavelet基底




基本的に$\ell^p$-ノルム

$$f = \sum_{k \in \mathbb{N}_+} \alpha_k \phi_k$$

Wavelet基底による展開

$$\|f\|_{B^s_{p,q}} \simeq \left[ \sum_{k=0}^{\infty} \{2^{sk} (2^{-kd} \sum_{j \in J(k)} |\alpha_{k,j}|^p)^{1/p}\}^q \right]^{1/q}$$

$(0 < p)$

小さな$p$ = スパースな係数 → 空間的な滑らかさの非一様性

- 詳細は手書きノートで説明

# 非適応的な手法との比較

DNN: $s > d(1/p - 1/r)_+$ なる仮定のもとで

$$\inf_{\check{f} \in \mathcal{F}(L,W,S,B)} \sup_{f^\circ \in U(B^s_{p,q}([0,1]^d))} \|f^\circ - \check{f}\|_{L^r([0,1]^d)} \lesssim N^{-s/d}$$

- **<u>適応的</u>**な**<u>非線形</u>**近似が必要 (Dung, 2011)

線形近似 (Linear width)： $\inf_{L_N} \sup_{f \in U(B^s_{p,q})} \|f - L_N(f)\|_r$ ($L_N$はランク$N$の線形作用素)

$$\begin{cases} N^{-s/d + \boxed{(1/p - 1/r)_+}} & \begin{cases} \text{either} & (0 < p \le r \le 2), \\ \text{or} & (2 \le p \le r \le \infty), \\ \text{or} & (0 < r \le p \le \infty), \end{cases} \\ N^{-s/d + \boxed{1/p - 1/2}} & (0 < p < 2 < r < \infty,\ s > d\max(1 - 1/r, 1/p)) \end{cases}$$

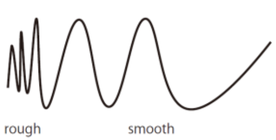
$p \neq r$が重要

非適応的近似（N-term approx., Kolmogorov width）：

$$\inf_{S_N} \sup_{f \in U(B^s_{p,q})} \inf_{\check{f} \in S_N} \|f - \check{f}\|_r$$

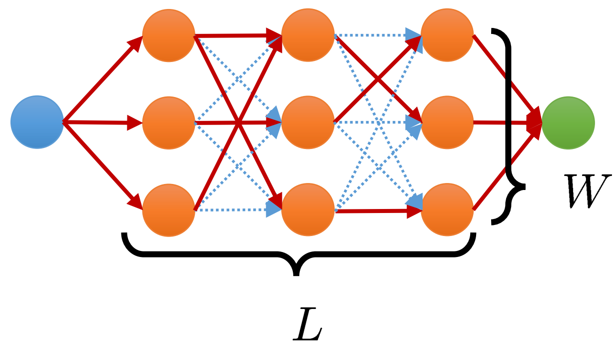
($S_N$は$B^s_{p,q}$内の$N$次元線形部分空間)

$$\begin{cases} N^{-s/d + \boxed{(1/p - 1/r)_+}} & (1 < p < r \le 2,\ s > d(1/p - 1/r)), \\ N^{-s/d + \boxed{1/p - 1/2}} & (1 < p < 2 < r \le \infty,\ s > d/p), \\ N^{-s/d} & (2 \le p < r \le \infty,\ s > d/2), \end{cases}$$




小さな$p$

# 推定誤差の導出：ノーテーション

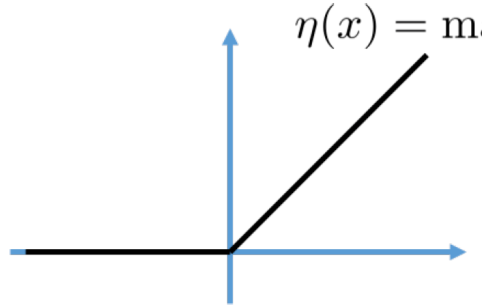



$$f(x) = (W^{(L)}\eta(\cdot) + b^{(L)}) \circ (W^{(L-1)}\eta(\cdot) + b^{(L-1)}) \circ \cdots \circ (W^{(1)}x + b^{(1)})$$

$\mathcal{F}(L, W, S, B)$
- 縦幅：$L$
- 横幅：$W$
- 枝の数：$S$
- 各パラメータの上限：$B$

の深層NNモデルの集合

- 活性化関数はReLUを仮定



# 推定誤差の導出



- 最小二乗解 (訓練誤差最小化)

$$\hat{f} = \mathop{\arg\min}_{\bar{f}: f \in \mathcal{F}(L,W,S,B)} \sum_{i=1}^{n} (y_i - \bar{f}(x_i))^2$$

ただし, $\bar{f} = \min\{\max\{f, -F\}, F\}$ (clipping).

**定理 (推定精度)**

$\|f^{\mathrm{o}}\|_{B^s_{p,q}} \leq 1$ , $\|f^{\mathrm{o}}\|_\infty \leq 1$かつ$0 < p, q \leq \infty$, $s > d(1/p - 1/2)_+$ のとき,

$N \asymp n^{\frac{d}{2s+d}}$ とすることで,

$$\|f^{\mathrm{o}} - \hat{f}\|^2_{L^2(P_X)} \leq n^{-\frac{2s}{2s+d}} \log(n)^3.$$

$p = q = \infty$のとき, Schmidt-Hieber (2017) に帰着.

# 証明: (1) 近似誤差の評価



- $0 < p, q, r \leq \infty$と$0 < s < \infty$が以下を満たすとする:

$$s > d(1/p - 1/r)_+ \qquad (L^r\text{-可積分性})$$

- $m$を$s < \min\{m, m-1+1/p\}$を満たす整数とする.

## 深層ニューラルネットワークの近似誤差

ある自然数$N$と用いて深さ$L$, 横幅$W$, 枝の数$S$, ノルム上界$B$を以下のように定める：

$$L = O(\log(N)), \qquad W = O(N),$$
$$S = O(N\log(N)), \qquad B = O(N^{(d/p-s)_+}),$$

すると，深層NNは以下の誤差でBesov空間の元を近似できる：

大体パラメータ数

$$\sup_{f^\circ \in U(B^s_{p,q}([0,1]^d))} \inf_{\check{f} \in \mathcal{F}(L,W,S,B)} \|f^\circ - \check{f}\|_{L^r([0,1]^d)} \lesssim N^{-s/d}.$$

Pinkus (1999), Mhaskar (1996): $p = r$かつ$1 \leq p$, ReLU活性化関数ではない.
Petrushev (1998): $p = r = 2$, ReLU活性化関数ではない ($s \leq k + 1 + (d-1)/2$).

# 証明: (1) 近似誤差の評価の導出



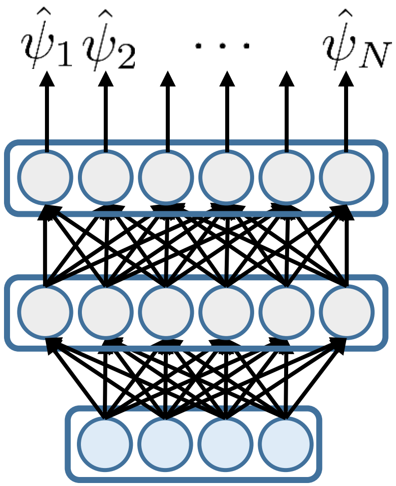
- **Step 1**: Besov空間の基底展開

$$f^\circ \in \mathcal{F} \quad \Rightarrow \quad f^\circ(x) = \sum_{i=1}^{\infty} \alpha_i \psi_i(x)$$

$$f^\circ = \sum_{i=1}^{N} \alpha_i \psi_i + \underbrace{\cancel{\sum_{i=N+1}^{\infty} \alpha_i \psi_i}}_{\|\cdot\|_{L^r} \lesssim N^{-s/d}}$$

∵ B-Splineによる適応的近似
[DeVore & Popov, 1988; Dung, 2011]

- **Step 2**: 各基底をDNNで近似.




$\psi_i \simeq \hat{\psi}_i$ : DNNによる近似.

$$\Rightarrow \quad \check{f} = \sum_{i=1}^{N} \alpha_i \hat{\psi}_i \quad \text{：線形結合}$$

- **Step 3:** 二つの評価を統合

$$\|f^\circ - \check{f}\|_{L^r} \leq \sum_{i=1}^{N} |\alpha_i| \underbrace{\|\psi_i - \hat{\psi}_i\|_{L^r}}_{\lesssim O(e^{-L})} + \underbrace{\|\textstyle\sum_{i=N+1}^{\infty} \alpha_i \psi_i\|_{L^r}}_{\lesssim N^{-s/d}}$$

# 証明: (2) バイアス-バリアンス分解



$$\mathrm{E}[\|f^{\circ} - \hat{f}\|_{L^2(P_X)}^2] \lesssim \underbrace{\frac{S[L\log(BW) + \log(Ln)]}{n}}_{\text{Variance}} + \underbrace{\inf_{f \in \mathcal{F}(L,W,S,B)} \|f - f^{\circ}\|_{L^2(P_X)}^2}_{\text{Bias}}$$

（局所Rademacher complexityを用いて証明）

古典的なノンパラ回帰の方法でOK．DNNに関する評価は[Schmidt-Hieber, 2019; Hayakawa&Suzuki,2020]

深さ　横幅　スパース性 (非零パラメータ数)　各パラメータの絶対値の上界

$$L = O(\log(N)),\, W = O(N),\, S = O(N\log(N)),\, B = O(N^{(d/p-s)_+})$$

なら

$$\text{Bias} = N^{-s/d} \qquad \text{Variance} = \frac{N\log(N)^3}{n}$$

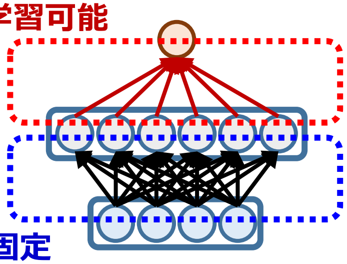
⇒ バイアスとバリアンスのトレードオフをバランスすればよい．

# 線形推定量



**“浅い” 学習法**

**Kernel ridge regression:** 正則化付き最小二乗推定量




$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^\infty} \sum_{i=1}^{n} (y_i - \beta^\top \psi(x_i))^2 + \lambda \beta^\top \beta$$

$\psi : \mathcal{X} \to \mathbb{R}^\infty$ (特徴マップ) **固定**

$K_{X,X} = (\psi(x_i)^\top \psi(x_j))_{i,j=1}^{n,n}$ グラム行列 (カーネル関数)

$$\hat{f}(x) = K_{x,X}(K_{X,X} + \lambda \mathrm{I})^{-1} \underline{Y}$$

(see also [Imaizumi&Fukumizu, 2019])

---

線形推定量: 観測値$Y = (y_i)_{i=1}^n$に対して線形な推定量.

$$X_n = (x_1, \ldots, x_n)$$

$$\hat{f}(x) = \sum_{i=1}^{n} \varphi_i(x; X_n) \underbrace{y_i}_{\text{線形}}$$

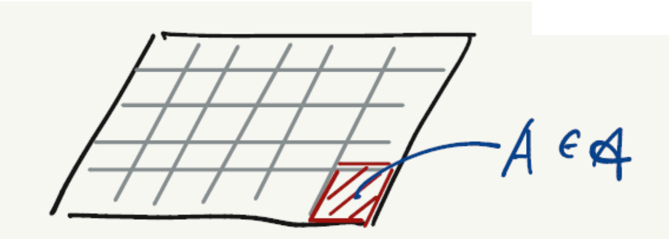
**例**

- Kernel ridge estimator
- Sieve estimator
- Nadaraya-Watson estimator
- k-NN estimator

# Minimax-optimal rate of linear estimators



- $\Omega = [0,1]^d$
- $P_X = \mathrm{Unif}(\Omega)$
- $\mathcal{A}$: decomposition of $\Omega$ s.t. $|\Omega| = 2^K$
- $\mathcal{F}$: function class




**Condition A**

- $\exists r_1, r_2 > 0$ s.t. $n^{-r_1} \leq 2^{-K} \leq 2^{-r_2}$ (polynomial order)
- Event $\mathcal{E}$:
  1. $|\{\, x_i \mid x_i \in A \ (i = 1,2,\ldots,n) \,\}| \leq C' n/2^K$ for all $A \in \mathcal{A}$
  2. $P(\mathcal{E}) \geq 1 - o(1)$

**Condition B**

There exists $\Delta > 0$ that satisfies the following two conditions:

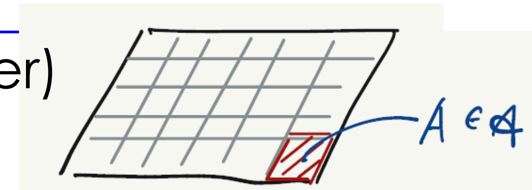
- $\exists F > 0$ s.t. $\forall A \in \mathcal{A}$, $\exists g_A \in \mathcal{F}$, it holds that $g_A(x) \geq \frac{\Delta F}{2} (\forall x \in A)$
- $C'' > 0$ s.t. $\frac{1}{n}\sum_{i=1}^{n} g(x_i)^2 \leq C''\Delta^2 2^{-K}$ $(\forall g \in \mathcal{F})$

# Minimax-optimal rate of linear estimators



**A**

- $\exists r_1, r_2 > 0$ s.t. $n^{-r_1} \le 2^{-K} \le 2^{-r_2}$ (polynomial order)
- Event $\mathcal{E}$:
  1. $|\{x_i \mid x_i \in A \ (i = 1,2,\dots,n)\}| \le C' n / 2^K$ for all $A \in \mathcal{A}$
  2. $P(\mathcal{E}) \ge 1 - o(1)$




**B**

There exists $\Delta > 0$ that satisfies the following two conditions:

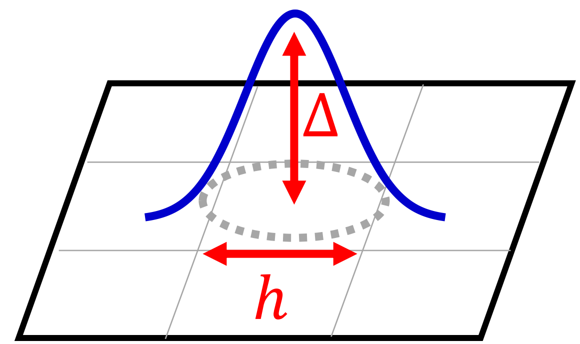
- $\exists F > 0$ s.t. $\forall A \in \mathcal{A}$, $\exists g_A \in \mathcal{F}$, it holds that $g_A(x) \ge \frac{\Delta F}{2}$ $(\forall x \in A)$
- $C'' > 0$ s.t. $\frac{1}{n}\sum_{i=1}^{n} g(x_i)^2 \le C'' \Delta^2 2^{-K}$ $(\forall g \in \mathcal{F})$

$$R^* := \inf_{\hat{f}:\text{Linear}} \sup_{f^\circ \in \mathcal{F}} \mathbb{E}[\|\hat{f} - f^\circ\|_{L^2(P_X)}^2]$$

**Theorem**

Under Conditions A and B, it holds:

$$\min\left\{\frac{F^2}{4CC''}\frac{2^K}{n}, \frac{F^3}{32}\Delta^2 2^{-K}\right\} \le R^*$$



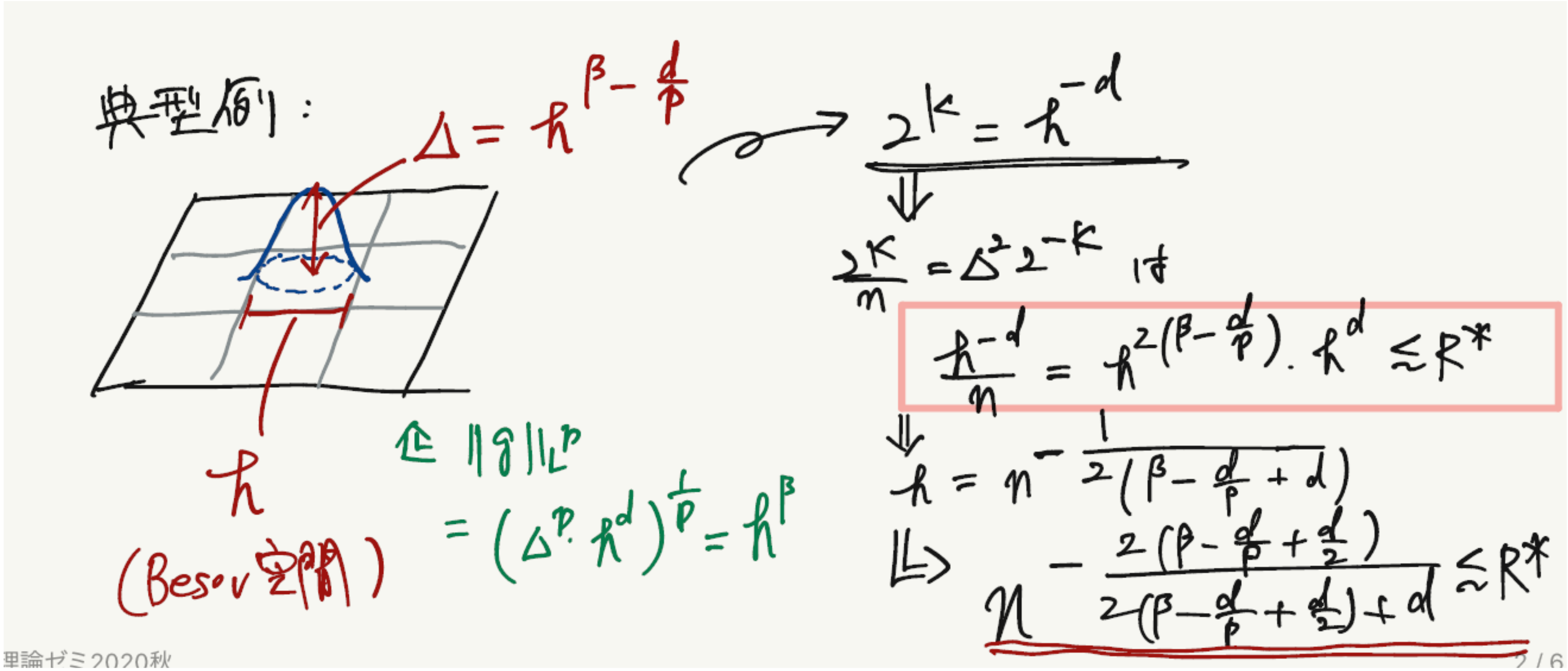

$$2^K = h^{-d}$$

- There is a trade-off between $K$ and $\Delta$.
- Equating first and second terms gives the minimax-rate.

# Typical example



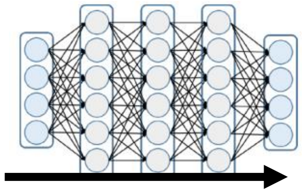
- Besov space: $B^{\beta}_{p,q}([0,1]^d)$



# 「浅い」学習との比較



- 深層学習は場所によって解像度を変える適応力がある
  →学習効率が良い
- 浅い学習は様々な関数を表現できる基底を
  あらかじめ十分用意して“待ち構える”必要がある.
  →学習効率が悪い



**仮定** $f^\circ \in B^s_{p,q}([0,1]^d)$: 真が“Besov空間”に入っている.

[Suzuki, ICLR2019]

### 線形推定量 (非適応的手法)

カーネルリッジ回帰等：

$$n^{-\frac{2s-2d(1/p-1/2)_+}{2s+d-2d(1/p-1/2)_+}}$$

**最適ではない**

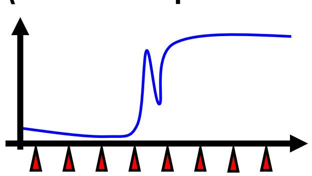
$\gg$

### 深層学習

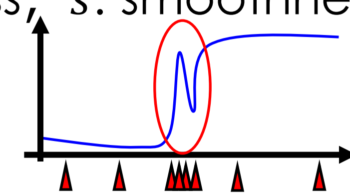
$$n^{-\frac{2s}{2s+d}}$$

**最適**

($n$: sample size, $p$: uniformity of smoothness, $s$: smoothness)



**一様な解像度**



**適応的解像度**

平均二乗誤差 $\mathrm{E}\left[\|\hat{f} - f^*\|^2\right]$がサンプルサイズが増えるにつれ減少するレート

ミニマックス最適性の意味で理論上これ以上改善できない精度を達成できている.

- Wavelet shrinkageより弱い条件
- 基底を用意せず最適化するだけでOK

# 次元の呪いと線形推定量



$$\mathcal{F} = \{f^\circ = g(Wx+b) \mid g \in U(B^s_{p,q}([0,1]^D)),\ W \in \mathbb{R}^{D\times d},\ b \in \mathbb{R}^D\}$$

(s.t. $Wx + b \in [0,1]^D$ for any $x \in [0,1]^d$)

$f^\circ$は$D$-次元部分空間にのみ依存

If $s > \frac{D}{d-D}\left(\frac{d}{2} - \frac{D}{p} + c\right)$

(非適応的)

**深層**

$$n^{-\frac{2s}{2s+D}}$$

($n^{-\frac{2s}{2s+1}}$ when $D=1$)

$\ll$

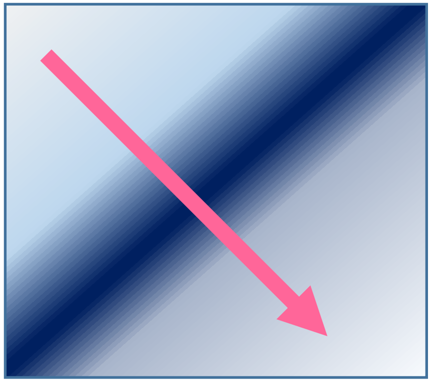
**線形推定量**

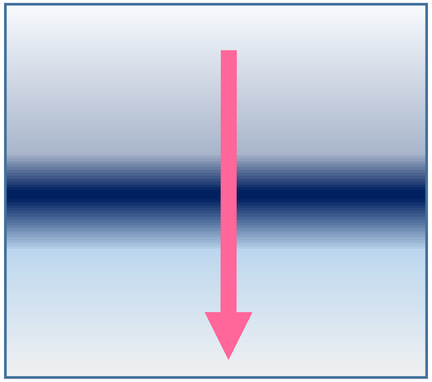
$$n^{-\frac{2(s-D/p+d/2+c)}{2(s-D/p+d/2+c)+d}}$$

$c = 1$ if $D < d/2$, $c = 0$ if $D \geq d/2$.

($n^{-\frac{2s+d}{2s+2d}}$ when $D = 1$ and $p = 1$)

深層にすることで次元の呪いを回避できている.

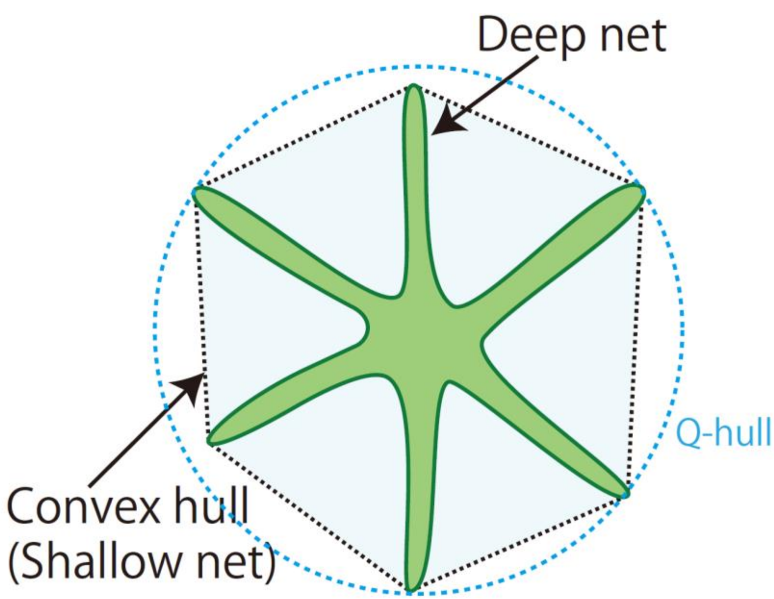




# 導出：凸法の理論

**定理 (凸法の議論) [Hayakawa&Suzuki, 2019] [Donoho & Johnstone, 1994]**

$$\inf_{\hat{f}:\mathrm{Linear}} \sup_{f^\circ \in \mathcal{F}} \mathrm{E}[\|\hat{f} - f^\circ\|_{L_2(P)}^2] = \inf_{\hat{f}:\mathrm{Linear}} \sup_{f^\circ \in \mathrm{conv}(\mathcal{F})} \mathrm{E}[\|\hat{f} - f^\circ\|_{L_2(P)}^2]$$



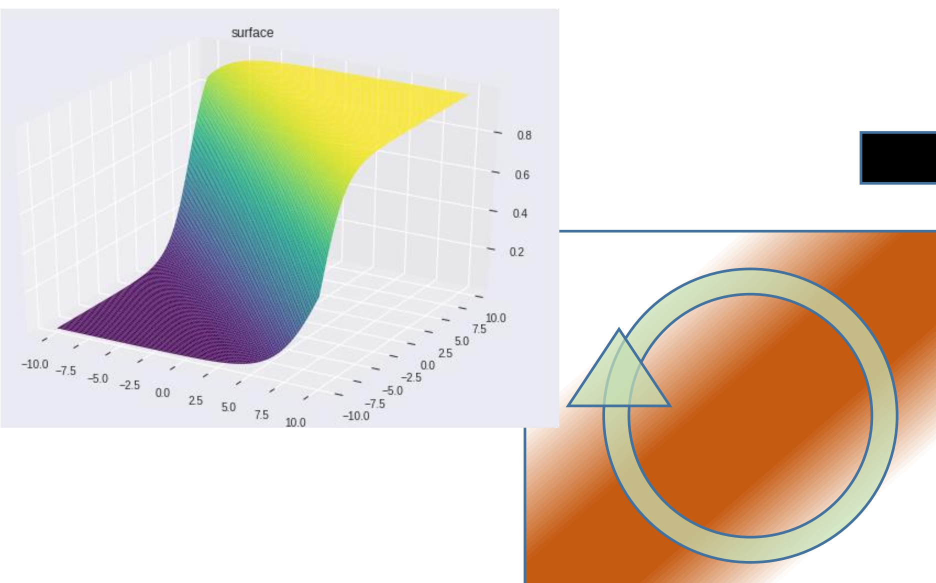

さらに条件を仮定すれば「Q-hull」まで拡張できる.

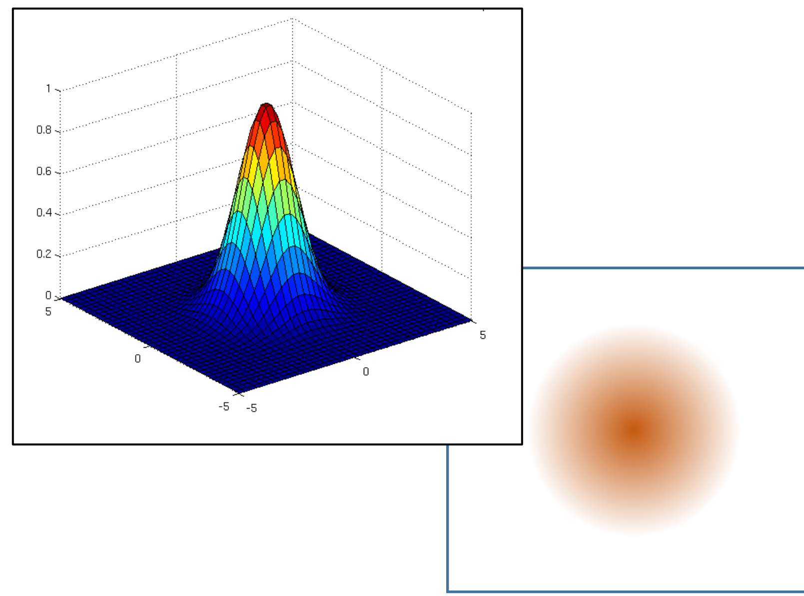
# Approach



- We show that $\mathrm{conv}(\mathcal{F}_\gamma)$ contains a Gaussian kernel with small width.
- The set of neural networks can approximate a ridge shape function.

$$M^D_{k,j}(Ux) \qquad\qquad \exp\left(-\|x\|^2/(2\sigma^2)\right)$$



# Irie-Miyake integral representation



**Theorem (Irie-Miyake integral representation)** [Th.3.1 of Hornik et al. (1990)]

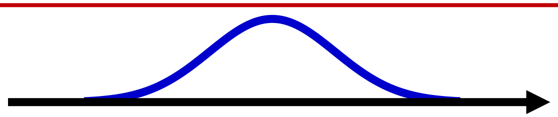
$$f(x) = \int_{\mathbb{R}^d} \exp(\mathrm{i} w^\top x)\hat{f}(w)\mathrm{d}w$$

$$\Rightarrow \quad f(x) = \int_{a \in \mathbb{R}^d} \int_{b \in \mathbb{R}} \psi(a^\top x + b)\mathrm{d}\nu(a,b)$$

for $\mathrm{d}\nu(a,b) = \mathrm{Re}\left(\frac{|w|^d e^{-\mathrm{i}wb}}{2\pi\hat{\psi}(w)}\right)\hat{f}(wa)\mathrm{d}a\mathrm{d}b$

where $w \neq 0$ is any non-zero real value.

$$\psi(x) = \{\sigma(x+1) - \sigma(x-1)\}/2 \quad (x \in \mathbb{R})$$



$$\psi_h(x) = \psi(x/h)$$

$$\psi_h(a^\top(x-c)+b) \in \mathrm{conv}(\mathcal{F}) \Rightarrow \exp\left(-\frac{\|x-c\|^2}{2h^2}\right)/C \in \mathrm{conv}(\mathcal{F})$$

$$\exp\left(-\frac{\|x-c\|^2}{2h^2}\right)$$

$$= \int_{a \in \mathbb{R}^d, b \in \mathbb{R}} \psi_h\left(a^\top(x-c)+b\right) \underbrace{\mathrm{Re}\left(\frac{|w|^d e^{-\mathrm{i}wb}}{2\pi\hat{\psi}_h(w)}\right)\sqrt{\frac{|wh|^d}{(2\pi)^d}}\exp\left(-\frac{(wh)^2\|a\|^2}{2}\right)}_{\text{Integrable}} \mathrm{d}a\mathrm{d}b$$