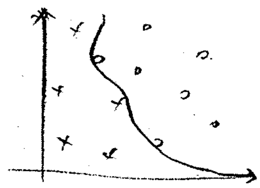
。カーネル法と学習理論

カーネル法を通して学習理論の基本的テクニックを紹介

やりたいこと

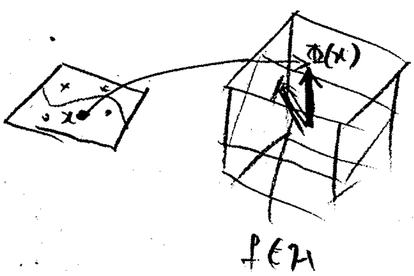
非線形推論



基本的アイディア

あるヒルベルト空間 $\mathcal{H}$ へ x を非線形にマップして、$\mathcal{H}$ 上で線形な操作を行う。

$$\mathcal{X} \ni x \longmapsto \Phi(x) \in \mathcal{H}$$



判別関数

$x^T \beta$

$\Downarrow$

$\langle \Phi(x), f \rangle_{\mathcal{H}}$ （ヒルベルト空間内の内積）

$f \in \mathcal{H}$

特に $\langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}} =: k(x, x')$ をカーネル関数と呼ぶ。

カーネル法の定式化

ある再生核ヒルベルト空間 $\mathcal{H}$ を用いて
（実関数のなすヒルベルト空間）

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2 \quad \text{——} \ ①$$

とする。$\|f\|_{\mathcal{H}}$ は $\mathcal{H}$ における f のノルム。

定義（正定値カーネル）

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ が次をみたすとき、正定値カーネルと呼ぶ。

- 対称性： $k(x, x') = k(x', x) \quad (\forall x, x' \in \mathcal{X})$
- 正値性： $\sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \geq 0$

$(\forall \alpha_i \in \mathbb{R}, \forall x_i \in \mathbb{R} \ (i = 1, \cdots, m))$

定義（再生核ヒルベルト空間）

集合 $\mathcal{X}$ 上の再生核ヒルベルト空間 (Reproducing kernel Hilbert Space, RKHS)
とは $\mathcal{X}$ 上の関数からなるヒルベルト空間 $\mathcal{H}$ で、

$^\forall x \in \mathcal{X}$ に対して、$k_x \in \mathcal{H}$ が存在して（先の $\Phi(x)$ に相当）

（再生性） $\langle f, k_x \rangle_{\mathcal{H}} = f(x) \qquad (^\forall f \in \mathcal{H})$

をみたすものをいう。

特に、$\langle k_x, k_y \rangle_{\mathcal{H}} = k(x,y)$ は正定値カーネルであり、$\mathcal{H}$ に付随した再生核（Reproducing Kernel, カーネル関数）とよぶ。

逆に任意の正定値カーネル k から RKHS が構成できる。

定理

k を正定値カーネルとする。

$\Rightarrow$ k を再生核として持つ RKHS が一意的に存在。

証明

$$\mathcal{H}_0 := \mathrm{span}\{k(x,\cdot) \mid x \in \mathcal{X}\}$$
$$= \left\{ \sum_{i=1}^{n} k(x_i,\cdot)\alpha_i \;\middle|\; \alpha_i \in \mathbb{R},\ x_i \in \mathcal{X}\ (i=1,\cdots,n),\ n=1,2,\cdots \right\}$$

とし、$\mathcal{H}_0$ に"内積"（$\langle \cdot,\cdot \rangle_{\mathcal{H}_0}$）を次のように定める：

$f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$, $g(x) = \sum_{j=1}^{m} \beta_j k(y_j, x)$ に対し、

$$\langle f, g \rangle_{\mathcal{H}_0} := \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, y_j).$$

これが実際に内積であることを示そう。

- 線形性：自明
- 対称性：自明 （$\because$ k は対称）
- 正値性：

まず $f \in \mathcal{H}_0$ に対し、$\langle f, f \rangle_{\mathcal{H}_0} \geq 0$ が k の正値性より示せる。

次いで、$\|f\|_{\mathcal{H}_0} = \sqrt{\langle f, f \rangle_{\mathcal{H}_0}} = 0$ なら $f(x) = 0 \ (^\forall x \in \mathcal{X})$ であることを示す。

↑関数として0

これまで述べた性質より、Cauchy-Shwartz の不等式が成り立ち、

$$|f(x)| = |\langle k(x,\cdot), f \rangle_{\mathcal{H}_0}| \leq \sqrt{k(x,x)}\, \|f\|_{\mathcal{H}_0} = 0.$$

さらに、$\langle \cdot,\cdot \rangle_{\mathcal{H}_0}$ は f の表示のしかたによらず well-defined であることを示す。

$f = \sum_{i=1}^{n} \alpha_i k(x_i,\cdot) = \sum_{i=1}^{n'} \alpha_i' k(x_i',\cdot)$ のとき、すべての $g = \sum_{j=1}^{m} \beta_j k(y_j,\cdot)$ に対して、

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, y_j) = \sum_{j=1}^{m} f(y_j)$$
$$\sum_{i=1}^{n'} \sum_{j=1}^{m} \alpha_i' \beta_j k(x_i', y_j) = \sum_{j=1}^{m} f(y_j)$$

よって一意。

$\langle\cdot,\cdot\rangle_{\mathcal{H}_0}$ が内積であることが示されたので、これで定まるノルムで $\mathcal{H}_0$ を完備化できる。これを $\mathcal{H}$ とする。$f\in\mathcal{H}$ に対して、f に対応するコーシー列 $\{f_n\}_n\in\mathcal{H}_0$ を用いて

$$f(x):=\lim_{n\to\infty}\langle f_n, k(\cdot,x)\rangle_{\mathcal{H}_0}$$

とすれば、$\mathcal{H}$ は実数値関数の集合とみなせる。
これはコーシー列のとり方によらない。実際 $\{f_m\}_m,\{f'_m\}_m\in\mathcal{H}_0$ がともに f に対応するとき、

$$\left|\langle f_m, k(\cdot,x)\rangle_{\mathcal{H}_0}-\langle f'_m, k(\cdot,x)\rangle_{\mathcal{H}_0}\right|$$
$$\leq \|k(\cdot,x)\|_{\mathcal{H}_0}\|f_m-f'_m\|_{\mathcal{H}_0}\longrightarrow 0 \qquad \text{である。}$$

また、$f(x)=0\ ({}^\forall x\in\mathcal{X})$ のとき、$\langle f,k(x,\cdot)\rangle_{\mathcal{H}}=0\ ({}^\forall x\in\mathcal{X})$ で、
${}^\forall h\in\mathcal{H}_0$ に対し、$\langle h,f_n\rangle_{\mathcal{H}_0}\to 0$ である（$\{f_n\}$ は f に対応するコーシー列）。
よって、$\|f\|_{\mathcal{H}}=0$ である。（弱収束するコーシー列は、強収束する。）
このことより $\mathcal{H}\ni f\longmapsto (x\mapsto f(x))\in\mathbb{R}^{\mathcal{X}}$ は単射であり、
$\mathcal{H}$ は実数値関数を要素としてもつヒルベルト空間として well-defined である。

（一意性）
$\mathcal{H}'$ を k に対応するもう一つの RKHS であるとする。
再生核の定義より $\mathrm{span}\{k(x,\cdot)\mid x\in\mathcal{X}\}\subseteq\mathcal{H}'$（関数の集合として）である。

また ${}^\forall f,g\in\mathrm{span}\{k(x,\cdot)\mid x\in\mathcal{X}\}$ に対し

$$\langle f,g\rangle_{\mathcal{H}}=\langle f,g\rangle_{\mathcal{H}'}$$

であることも k の再生性よりわかる。よって

$$\mathcal{H}=\overline{\mathrm{span}\{k(x,\cdot)\mid x\in\mathcal{X}\}}\subseteq\mathcal{H}'\ \text{（関数の集合として）}$$

であり、${}^\forall f,g\in\mathcal{H}$ に対し $\langle f,g\rangle_{\mathcal{H}}=\langle f,g\rangle_{\mathcal{H}'}$ である。
今 $f\perp k(x,\cdot)$ in $\mathcal{H}'$ $({}^\forall x\in\mathcal{X})$ なら

$$f(x)=\langle f,k(x,\cdot)\rangle_{\mathcal{H}'}=0\ ({}^\forall x\in\mathcal{X})\ \text{なので、}\ \|f\|_{\mathcal{H}'}=0.$$

よって $\mathcal{H}=\mathcal{H}'$（関数の集合として）であり、$f\in\mathcal{H}\subseteq\mathcal{H}'$ に対し $\langle f,f\rangle_{\mathcal{H}}=\langle f,f\rangle_{\mathcal{H}'}$ である。 //

表現定理

定理

$$\hat{f} := \underset{f \in \mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^{n} l(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2 \quad \text{—①}$$

は ある $\alpha_i \in \mathbb{R}$ $(i=1,\dots,n)$ に対し、

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$$

と表現される。

証明

$f(x_i) = \langle k(\cdot, x_i), f \rangle_{\mathcal{H}}$ より、ロス関数に関係する部分は

$\mathcal{H}_1 = \mathrm{span}(k(\cdot, x_1), \dots, k(\cdot, x_n))$ への f の射影で決まる。

$f = f_1 + f_2 \in \mathcal{H}_1 + \mathcal{H}_1^{\perp}$ とおくと、 $f(x_i) = f_1(x_i) + f_2(x_i) = f_1(x_i)$ より、

$$\sum_{i=1}^{n} l(y_i, f_1(x_i)) + \lambda \underbrace{\|f_1 + f_2\|_{\mathcal{H}}^2}_{= \|f_1\|_{\mathcal{H}}^2 + \|f_2\|_{\mathcal{H}}^2}$$

よって最小化元は $f_2 = 0$ つまり $f = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$ と書ける。 //

※ よって、

$$① \Leftrightarrow \min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{n} l\left(y_i, \sum_{j=1}^{n} k(x_i, x_j)\alpha_j\right) + \lambda \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j)$$

と有限次元最適化におちる。

カーネル関数の例

- ガウスカーネル

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

- 多項式カーネル

$$k(x, x') = (1 + x^T x')^d$$

- 線形カーネル

$$k(x,x') = x^T x'$$

- ソボレフ空間　$H^m[a,b] = \{ f \in L^2[a,b] \mid f, \dots, f^{(m-1)}$ は絶対連続で $f^{(m)} \in L^2[a,b] \}$

$$k(x,y) = \sum_{\ell=0}^{m-1} \frac{(x-a)^\ell}{\ell!} \frac{(y-a)^\ell}{\ell!} + \int_a^b \frac{(x-t)_+^{m-1}}{(m-1)!} \frac{(y-t)_+^{m-1}}{(m-1)!} dt$$

$$\langle f, g \rangle_H = \sum_{\ell=0}^{m-1} f^{(\ell)}(a) g^{(\ell)}(a) + \int_a^b f^{(m)}(x) g^{(m)}(x) dx .$$

- グラフカーネル
- 時系列カーネル

・多項式カーネルについて、

$x, x' \in \mathbb{R}$ とすると、

$$k(x,x') = 1 + \binom{d}{1} xx' + \binom{d}{2} (xx')^2 + \cdots + (xx')^d$$

より、$\Phi(x) = \begin{bmatrix} 1 \\ \sqrt{d}\, x \\ \sqrt{\binom{d}{2}}\, x^2 \\ \vdots \\ x^d \end{bmatrix}$ とすれば $k(x,x') = \langle \Phi(x), \Phi(x') \rangle$ である。

(他のカーネル法)

さて、カーネル法による推定量 $\hat{f}$ の汎化誤差は $\boxed{n \to \infty で}$ 本当に小さくなるのか？
収束レートは？

$$L(f) := E_{(x)}[\ell(Y, f(X))]$$

$$\hat{L}(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$$

$$Pf := E_{X,Y}[f(X,Y)]$$

$$P_n f := \frac{1}{n} \sum_{i=1}^{n} f(x_i, y_i) \qquad (L(f) = P(\ell \circ f),\ \hat{L}(f) = P_n(\ell \circ f))$$

とおく。

以下、簡単のため、$\mathcal{F} := \{ f \in H \mid \|f\|_H \leq 1 \}$ なる集合を考える。

$$f^* := \operatorname*{argmin}_{f \in \mathcal{F}} L(f)$$

$$\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \hat{L}(f)$$

抑えたいもの：

$$\underline{L(\hat{f}) - L(f^*)} \quad (汎化誤差)$$

※ 汎化誤差をおさえるのが学習理論の基本的問題意識

仮定1

$$\left[\begin{array}{ll} \circ\ k(x,x) \leqq 1 & ({}^\forall x \in \mathcal{X}) \\ \circ\ |\ell(y,f(x)) - \ell(y,f'(x))| \leqq |f(x) - f'(x)| & ({}^\forall x, y,\ {}^\forall f \in \mathcal{F}) \end{array}\right.$$

最も基本的な問題設定で話を進める.
これらの仮定は緩めることもできる.

命題

$$\left[\begin{array}{ll} k(x,x) \leqq 1 \quad ({}^\forall x \in \mathcal{X}) & \text{なら} \\ \|f\|_\infty \leqq \|f\|_{\mathcal{H}} & ({}^\forall f \in \mathcal{H}) \end{array}\right.$$

証明.

$$|f(x)| = |\langle k(x,\cdot), f\rangle_{\mathcal{H}}| \leqq \|k(x,\cdot)\|_{\mathcal{H}} \|f\|_{\mathcal{H}}$$
$$= \sqrt{\langle k(x,\cdot), k(x,\cdot)\rangle_{\mathcal{H}}}\ \|f\|_{\mathcal{H}}$$
$$= \sqrt{k(x,x)}\ \|f\|_{\mathcal{H}} \leqq \|f\|_{\mathcal{H}}$$

//

これより ${}^\forall f \in \mathcal{F}$ は $\|f\|_\infty \leqq 1$ である.

・出発点、

$\hat{f}, f^*$ の定義より

$$L(f^*) \leqq L(\hat{f})$$
$$\hat{L}(\hat{f}) \leqq \hat{L}(f^*)$$

である. これより、

$$0 \leqq L(\hat{f}) - L(f^*) = L(\hat{f}) - \hat{L}(\hat{f}) - L(f^*) + \hat{L}(f^*) + \underbrace{\hat{L}(\hat{f}) - \hat{L}(f^*)}_{\leqq 0}$$

$$\leqq (L(\hat{f}) - \hat{L}(\hat{f})) - (L(f^*) - \hat{L}(f^*))$$

$$= (P - P_n)(\ell \circ \hat{f} - \ell \circ f^*)$$

ここで右辺を次のように抑える.

$$L(\hat{f}) - L(f^*) \leqq \sup_{f \in \mathcal{F}} (P - P_n)(\ell \circ f - \ell \circ f^*)$$

※ ここで、

$$\sup_{f\in\mathcal{F}} P_n(l\circ f) \quad や \quad \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i\, l(y_i, f(x_i))$$

らは以後可測であると仮定する。可測性がない場合は以下の議論は成り立たない。外測度を用いた議論も使えない。

詳しくは Dudley "Uniform Central Limit Theorems" 等.

c.f. Suslin image admissible.

ここで、 $g := l\circ f - l\circ f^*$, ($\overset{\text{つもり}}{z = (x,y)}$, $g(z) = l(y,f(x)) - l(y,f^*(x))$)

$$\mathcal{G} := \{g \mid g = l\circ f - l\circ f^*,\ f\in\mathcal{F}\}$$

とおけば $\sup_{g\in\mathcal{G}}(P-P_n)g$ を抑える問題になる。

Rademacher 複雑さ

定義

$\sigma_i\in\{\pm1\}$ を $P(\sigma_i=1)=P(\sigma_i=-1)=\frac{1}{2}$ なる i.i.d. Rademacher 変数とする。

$$\hat{R}_n(\mathcal{G}) := E_{\{\sigma_i\}_{i=1}^n}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(z_i) \,\middle|\, z_1,\dots,z_n\right]$$

(経験 Rademacher 複雑さ)

$$R_n(\mathcal{G}) := E_{\{z_i\}_{i=1}^n}\left[\hat{R}_n(\mathcal{G})\right]$$

(Rademacher 複雑さ)

定理 (対称化)

$$E_{\{z_i\}}\left[\sup_{g\in\mathcal{G}} (P-P_n)g\right] \leq 2R_n(\mathcal{G})$$

証明

$\{z_i\}$と同じ i.i.d. sequence

$$左辺 = E_{\{z_i\}}\left[\sup_{g\in\mathcal{G}}\left\{E_{\{z_i'\}}\left[\frac{1}{n}\sum_{i=1}^{n} g(z_i')\right] - \frac{1}{n}\sum_{i=1}^{n} g(z_i)\right\}\right]$$

$$\leq E_{\{z_i\},\{z_i'\}}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n}\left(g(z_i') - g(z_i)\right)\right]$$

$$= E_{\sigma_i}\left[E_{z_i,z_i'}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i\left(g(z_i') - g(z_i)\right)\right]\right]$$

(∵ z_i'とz_iは対称)

$\sigma_i \leftarrow -\sigma_i$ としても同じ分布

$$\leq E_{\sigma_i} E_{z_i'}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(z_i')\right] + E_{\sigma_i} E_{z_i}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(z_i)\right]$$

$= 2\, R_n(G)$ （∵ Fubini の定理） //

命題

$$P\left(\sup_{g \in G} (P - P_n) g \geqq 2 R_n(G) + 4\sqrt{\frac{\log(1/\delta)}{2n}} \right) \leqq \delta$$

証明 （$|g| \leqq 2$ と McDiarmid の不等式）

補題

（McDiarmid の不等式）

$X_1, \ldots, X_n$ を $\mathcal{X}$ 上の独立な（独立同一とは限らない）確率変数とする。

$f : \mathcal{X}^n \to \mathbb{R}$ を可測関数とし、

$$|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i$$

が任意の $x_1, \ldots, x_n, x_i' \in \mathcal{X}$ に対して成り立つとき、

$$P\left(f(X_1, \ldots, X_n) - E[f(X_1, \ldots, X_n)] \geq \varepsilon \right) \leq \exp\left(- \frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2} \right)$$

今、$\forall f \in \mathcal{F}$ が $\|f\|_\infty \leq 1$ なので仮定から

$$|g(z)| = |\ell(y, f(x)) - \ell(y, f^*(x))| \leq |f(x) - f^*(x)| \leq 2$$

である。よって

$$h(z_1, \ldots, z_n) = \sup_{g \in G} \left\{ E[g(Z)] - \frac{1}{n} \sum_{i=1}^n g(z_i) \right\}$$

は

$$|h(z_1, \ldots, z_i, \ldots, z_n) - h(z_1, \ldots, z_i', \ldots, z_n)| \leq \frac{4}{n}$$

をみたす。あとは McDiarmid の不等式と先の対称化により、題意を得る。 //

定理（Rademacher 複雑さの性質）

(1) $G \subset G'$ なら $R_n(G) \leqq R_n(G')$

(2) $\phi_i : \mathbb{R} \to \mathbb{R}$ が $|\phi_i(x) - \phi_i(y)| \leqq M |x - y|$ をみたすとき、

$$E_{z_i \sigma_i}\left[\sup_{g \in G} \frac{1}{n} \sum_{i=1}^n \sigma_i \phi_i(g(z_i)) \right] \leqq M\, R_n(G)$$

(3) G の凸包 $\mathrm{co}\, G = \left\{ \sum_{i=1}^m \lambda_i g_i \,\middle|\, \sum_{i=1}^m \lambda_i = 1,\ \lambda_i \geqq 0,\ g_i \in G,\ m = 1, 2, \ldots \right\}$ に対し

$$R_n(G) = R_n(\mathrm{co}\, G)$$

証明

(2)のみ示す。$z_1, \dots, z_n$ 固定で示せれば十分。

$\sigma_2, \dots, \sigma_n$ 固定のもとで、

$$D = \{ (g(z_1), \sum_{i=2}^n \sigma_i \phi_i(g(z_i))) \mid g \in G \} \subseteq \mathbb{R}^2$$

とおく。

すると、

$$E_{\{\sigma_i\}}\left[\sup_{g \in G} \sum_{i=1}^n \sigma_i \phi_i(g(z_i)) \right]$$

$$= E_{\sigma_2, \dots, \sigma_n}\left[E_{\sigma_1}\left[\sup_{g \in G} \sum_{i=1}^n \sigma_i \phi_i(g(z_i)) \mid \sigma_2, \dots, \sigma_n \right] \right]$$

$$= E_{\sigma_2, \dots, \sigma_n}\left[\underbrace{\frac{1}{2} \sup_{(t_1, s_1) \in D} (\phi_1(t_1) + s_1) + \frac{1}{2} \sup_{(t_2, s_2) \in D} (-\phi_1(t_2) + s_2)} \right]$$

$$= \frac{1}{2} (\phi_1(t_1^*) + s_1^* - \phi_1(t_2^*) + s_2^*) \qquad (\exists\, t_1^*, s_1^*, t_2^*, s_2^* \in \mathbb{R})$$

$$\leq \begin{cases} \frac{M}{2}(t_1^* - t_2^*) + \frac{1}{2}(s_1^* + s_2^*) & (t_1^* \geq t_2^*) \\ \frac{M}{2}(t_2^* - t_1^*) + \frac{1}{2}(s_1^* + s_2^*) & (t_1^* < t_2^*) \end{cases}$$

$$\leq \frac{1}{2} \sup_{(t_1, s_1) \in D} (M t_1 + s_1) + \frac{1}{2} \sup_{(t_2, s_2) \in D} (-M t_2 + s_2)$$

$$= E_{\sigma_1}\left[\sup_{g \in G} \sigma_1 M g(z_1) + \sum_{i=2}^n \sigma_i \phi_i(g(z_i)) \right]$$

以上の議論をくり返すことで題意を得る。 //

ここで、$l(y, t)$ のリプシッツ連続性の仮定を用いることで

$$L(\hat{f}) - L(f^*) \leq \sup_{f \in \mathcal{F}} (P - P_n)(l \circ f - l \circ f^*)$$

$$\leq 2 R_n(G) + 4 \sqrt{\frac{g(1/\delta)}{2n}} \quad \text{with prob. } 1-\delta \quad (\delta \in (0,1))$$

$$\leq 2 R_n(\mathcal{F}) + 4 \sqrt{\frac{g(1/\delta)}{2n}}$$

を得る。よって、$R_n(\mathcal{F})$ を上からおさえれば良い。

Rademacher 複雑さの上界

- $\mathcal{F} = \{ f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq 1 \}$, $\sup_x k(x, x) \leq 1$ のとき.

$$R_n(\mathcal{F}) = E_{\{x_i\}} E_{\{\sigma_i\}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \right]$$

$$= E_{\{x_i\}} E_{\{\sigma_i\}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \left\langle \sum_{i=1}^{n} \sigma_i k(x_i, \cdot), f \right\rangle_{\mathcal{H}} \right]$$

$$\leq E_{\{x_i\}} E_{\{\sigma_i\}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \left\| \sum_{i=1}^{n} \sigma_i k(x_i, \cdot) \right\|_{\mathcal{H}} \cdot \underbrace{\|f\|_{\mathcal{H}}}_{\leq 1} \right]$$

$$= E_{\{x_i\}} E_{\{\sigma_i\}} \left[\frac{1}{n} \sqrt{\sum_{i,j} \sigma_i \sigma_j k(x_i, x_j)} \right]$$

$$\leq \frac{1}{n} \sqrt{E_{\{x_i\}} E_{\{\sigma_i\}} \left[\sum_{i,j} \sigma_i \sigma_j k(x_i, x_j) \right]} = \frac{1}{n} \sqrt{n E_x [k(x, x)]}$$

$$\leq \sqrt{\frac{1}{n}}$$

- $\mathcal{F}$ が有限集合のとき

Th. (Massart の定理)

$\mathcal{F} = \{ f_1, \ldots, f_M \}$ なら,

$$R_n(\mathcal{F}) \leq E_{\{z_i\}} \left[\max_m \sqrt{\frac{1}{n} \sum_{i=1}^{n} f_m(z_i)^2} \right] \times \sqrt{\frac{2 \log M}{n}}$$

特に $\|f_m\|_\infty \leq R$ $(\forall m = 1, 2, \ldots, M)$ なら

$$\leq R \sqrt{\frac{2 \log M}{n}}$$

証明

$A \subset \mathbb{R}^n$ を有限集合とする。このとき $R = \max_{a \in A} \sqrt{\sum_{i=1}^{n} a_i^2}$ に対し

$$E_{\{\sigma_i\}} \left[\max_{a \in A} \sum_{i=1}^{n} \sigma_i a_i \right] \leq R \sqrt{2 \log |A|}$$

を示せば十分。

$$\exp\left(s E \left[\max_{a \in A} \sum_{i=1}^{n} \sigma_i a_i \right] \right)$$

$$\leq E \left[\exp\left(s \max \sum_{i=1}^{n} \sigma_i a_i \right) \right] \leq E \left[\sum_{a \in A} \exp\left(s \sum_{i=1}^{n} \sigma_i a_i \right) \right]$$

$$\leq \sum_{a \in A} \prod_{i=1}^{n} E \left[\exp(s \sigma_i a_i) \right]$$

$$\leq \sum_{a \in A} \prod_{i=1}^{n} \exp(s^2 a_i^2 / 2) \quad \text{(Hoeffding の不等式)}$$

$$\leq \exp\left(\log |A| + \frac{s^2}{2} R^2 \right)$$

$s = \frac{\sqrt{2 \log |A|}}{R}$ として 両辺 $\log$ とれば 題意を 示す //

一 $\mathcal{F}$ が VC 次元有限な集合の時.

$\mathcal{F}$ が2値関数族とする. i.e. $\forall f \in \mathcal{F}$ で $f(x) \in \{\pm 1\}$ $(\forall x \in \mathcal{X})$

定義

$\mathcal{F}$ の VC 次元

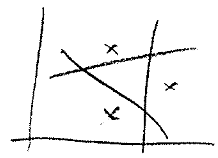
$d_{VC}(\mathcal{F}) := \max\{ n \mid \exists x_1, \dots, x_n \in \mathcal{X}$ で、すべての ± 1 列 $\varepsilon^n \in \{\pm 1\}^n$ に対し、
$\exists f \in \mathcal{F}$ で $\varepsilon_i = f(x_i)$ とできる。$\}$

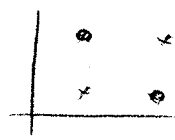
↑ n コのサンプルを完全に識別できる。

例: $\mathcal{F} = \{ f(x) = \mathrm{sign}(x^T \beta + a) \mid \beta \in \mathbb{R}^d, a \in \mathbb{R} \}$

$\Longrightarrow d_{VC}(\mathcal{F}) = d+1$

※ VC 次元は分布によらない。





⇐ 識別できない

定理 ($n \geq d_{VC}$ のとき)

$$R_n(\mathcal{F}) \leq \sqrt{\frac{2 d_{VC}(\mathcal{F}) \log(n) + 2 d_{VC}(\mathcal{F}) \log\left(e / d_{VC}(\mathcal{F})\right)}{n}}$$

$$\left(\leq c \sqrt{\frac{d_{VC}(\mathcal{F}) \log(n)}{n}} \right)$$

定理 (Sauerの定理)

$$\mathcal{F}_{x_1^n} := \{ (f(x_1), \dots, f(x_n)) \in \{\pm 1\}^n \mid f \in \mathcal{F} \}$$

とし、

$$m^{\mathcal{F}}(n) := \max\{ |\mathcal{F}_{x_1^n}| \mid (x_1, \dots, x_n) \in \mathcal{X}^n \} \quad \text{(増大関数)}$$

とする。 $\left(d_{VC}(\mathcal{F}) = \max\{ n \mid |m^{\mathcal{F}}(n)| = 2^n \} \right)$

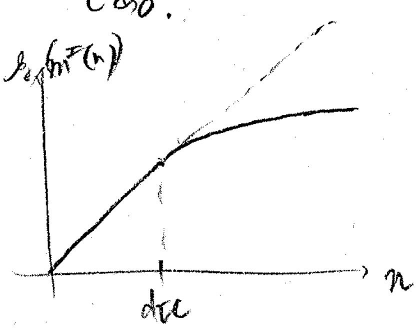
このとき、

$n \leq d_{VC}(\mathcal{F})$ なら $m^{\mathcal{F}}(n) = 2^n$ ← 指数

$n > d_{VC}(\mathcal{F})$ なら $m^{\mathcal{F}}(n) \leq \sum_{i=1}^{d_{VC}} \binom{n}{i} \leq \left(\frac{en}{d_{VC}} \right)^{d_{VC}}$

← 多項式!

である。



先の定理は、Sauerの定理にMassartの定理の証明を当てはめれば得られる

Th.

$d_{VC}(\mathcal{F}) < \infty$ なら $\lim_{n\to\infty} \sup_{f\in\mathcal{F}} |(P_n - P)f| = 0$ P-a.s. ($^\forall P$: prob. meas.)

($\mathcal{F}$は image admissible Suslin)

かつ $\lim_{n\to\infty} \sup_{P} E_P\left[\sup_{f\in\mathcal{F}} |(P_n-P)f|\right] = 0$. ($d_{VC}(\mathcal{F})<\infty$ なら)

↖ これも Donsker-class

Th. (Assouad)

$(\mathcal{X}, \mathcal{B})$: 可測空間

$\sup_{P} E_P^*\left[\sup_{f\in\mathcal{F}} |(P_n - P)f|\right] \longrightarrow 0 \quad (n\to\infty)$

P: $(\mathcal{X},\mathcal{B})$上の万能確率測度　↖外測度による積分

ならば　$\mathcal{F}$はVC-クラス.

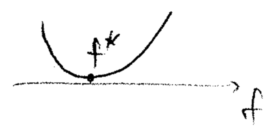
詳しくは、Dudley の "Uniform Central Limit Theorems" を参照のこと.

Local Rademacher 複雑さ

仮定2

- l は強凸：ある $\alpha>0$ が存在して、$\forall f \in \mathcal{F}$ に対し、

$$\frac{\alpha}{2}\|f - f^*\|_{L_2(P_X)}^2 \leq L(f) - L(f^*).$$



- 以前の仮定がみたされている

$L(\hat{f}) - L(f^*) \leq \sup_{g\in G}(P-P_n)g$ はルーズすぎる。

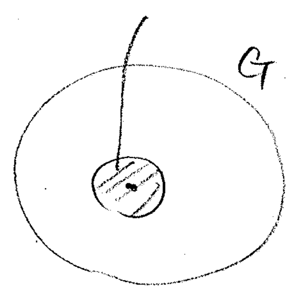
$L(f)$が強凸なら、$\hat{f}$の存在する場所が限定されるはず。

$\sup_{g\in G}$ のかわりに、f^*の近傍で sup をおさえるべきではないか？

$\longrightarrow$ 局所化

Local Rademacher 複雑さ：

$$G_r := \{ g = l\circ f - l\circ f^* \in G \mid Pg \leq r \}$$ として、



$$R_n(G_r)$$

を、local Rademacher 複雑さと言う。

（集合を"局所化"していれば $Pg\leq r$ でなくても $\|g\|_{L_2} \leq r$ としても Local Rademacher 複雑さという）

※ $g = l\circ f - l\circ f^*$ に対し、リプシッツ連続性から

$$\frac{\alpha}{2}\|g\|_{L_2}^2 \leq \frac{\alpha}{2}\|f-f^*\|_{L_2}^2 \leq Pg$$

なので、$Pg \leq r$ なら $\|g\|_{L_2(P_X)} \leq \sqrt{\frac{2r}{\alpha}}$。（$\|f-f^*\|_{L_2(P_X)} \leq \sqrt{\frac{2r}{\alpha}}$ でもある）

定理 (Peeling)

ある $\phi : [0,\infty) \to [0,\infty)$ なる関数が存在し、

$\phi(4r) \leq 2\phi(r)$ かつ

$$R_n(G_r) \leq \phi(r) \quad (r>0)$$

ならば、$\forall r > 0$ に対し、

$$E_{\sigma,Z}\left[\sup_{g\in G}\frac{\frac{1}{n}\sum_{i=1}^n \sigma_i g(z_i)}{Pg + r}\right] \leq \frac{4\phi(r)}{r}$$

である。

証明

$$\sup_{g\in G} \frac{\frac{1}{n}\sum_{i=1}^n \sigma_i g(z_i)}{Pg + r} \leqq \sup_{g\in G_r} \frac{\frac{1}{n}\sum \sigma_i g(z_i)}{r}$$
$$+ \sum_{j=0}^{\infty} \sup_{g\in G_{r4^{j+1}}\setminus G_{r4^j}} \frac{\frac{1}{n}\sum \sigma_i g(z_i)}{r4^j + r}$$

両辺期待値をとると、

$$E[\text{左辺}] \leqq \frac{R_n(G_r)}{r} + \sum_{j=0}^{\infty} \frac{R_n(G_{r4^{j+1}})}{r4^j + r}$$
$$\leqq \frac{\phi(r)}{r} + \frac{1}{r}\sum_{j=0}^{\infty} \frac{\phi(4^{j+1}r)}{4^j+1}$$
$$\leqq \frac{\phi(r)}{r} + \frac{1}{r}\sum_{j=0}^{\infty} \frac{2^{j+1}\phi(r)}{4^j+1}$$
$$\leqq \frac{4\phi(r)}{r}$$

$$\boxed{\begin{array}{l} 2\times\left(\overset{j=0}{\frac{1}{2}} + \sum_{j=1}^{\infty} 2^{-j}\right) \\ = 3 \end{array}}$$

//

定理 (Talagrand の測度集中不等式)

$(Z, \mathcal{A}, P)$: 測度空間

$\tilde{G}$ = 可測関数の集合 on $(Z, \mathcal{A})$ s.t. $\forall g\in\tilde{G}$, $\|g\|_\infty \leqq B$, $E[g]=0$, $E[g^2]\leqq \sigma^2$, $\tilde{G}$ は $\|\cdot\|_\infty$ に関して可分.

すると、$\forall t>0$ に対し、

$$P^n\left(\sup_{g\in\tilde{G}} \frac{1}{n}\sum_{i=1}^n g(z_i) \geqq 2 E_z\left[\sup_{g\in\tilde{G}} \frac{1}{n}\sum_{i=1}^n g(z_i)\right] + \sqrt{\frac{2t\sigma^2}{n}} + \frac{2tB}{n}\right)$$
$$\leqq e^{-t}$$

(定数は最適ではない)

上の $\tilde{G}$ として、$\left\{ \frac{Pg - g(z)}{Pg + r} \,\middle|\, g\in G \right\}$ を代入する。

すると、$\|\tilde{g}\|_\infty = \left\|\frac{Pg - g}{Pg + r}\right\|_\infty \leqq \frac{1}{r}$, $\|\tilde{g}\|_{L_2}^2 = \frac{\|g - Pg\|_{L_2}^2}{(Pg+r)^2} \leqq \frac{P(g^2)}{r\frac{\alpha}{2}P(g^2)} = \frac{2}{\alpha r}$

である。つまり、$B = \frac{1}{r}$, $\sigma^2 = \frac{2}{\alpha r}$ とおける。よって $\forall t>0$ に対し、

$$\sup_{g\in G} \frac{(P-P_n)g}{Pg + r} \leqq 2E\left[\sup_{g\in G} (P-P_n)\tilde{g}\right] + \sqrt{\frac{4t}{\alpha r n}} + \frac{2t}{rn} \quad \text{w.p. } 1-e^{-t}.$$

ここで、前の定理(Peeling)より

$$E\left[\sup_{g\in\mathcal{G}}(P-P_n)g\right] \le 2R_n(\mathcal{G}) \le \frac{8\phi(r)}{r}.$$

である。以上より、

$$L(\hat f)-L(f^*) \le (P-P_n)\overbrace{(\ell\circ\hat f-\ell\circ f^*)}^{\hat g}$$

$$\le (P\hat g + r)\underbrace{\left(\frac{16\phi(r)}{r}+\sqrt{\frac{4t}{\alpha r n}}+\frac{2t}{rn}\right)}_{\psi_n(r)}$$

よって $\psi_n(r)\le\frac{1}{2}$ なる任意の r に対して、($L(\hat f)-L(f^*)=P\hat g$ であることに注意すると)

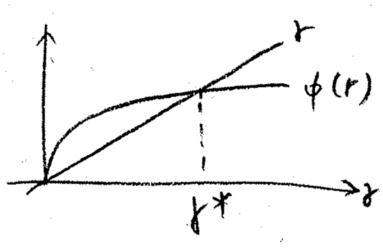
$$L(\hat f)-L(f^*)\le 2r\psi_n(r)\le r$$

が成り立つ。

$$\max\left\{84\phi(r),\ \frac{72t}{\alpha n},\ \frac{12t}{n}\right\}\le r$$

なら良い。

$\phi(r)$ は 凹 関数になることが多い。



$\phi(r)$ と r との交点を求めれば良い

(Fixed Point)

ではφ(r)はどのような形になるであろうか？

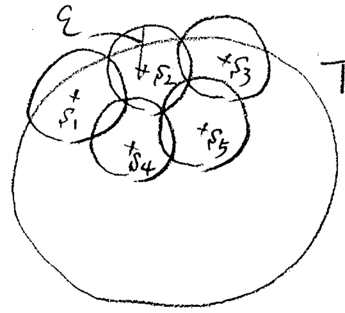
Dudleyのチェイニング

定義(エントロピー数)

(T, d)を距離空間とする。

$$e_i(T,d) := \inf\left\{ \varepsilon > 0 \;\middle|\; \exists s_1, \ldots, s_{2^{i-1}} \in T \text{ s.t. } T \subset \bigcup_{j=1}^{2^{i-1}} B_d(s_j, \varepsilon) \right\}$$

※半径εを指定したときのボールの個数はカバリングナンバーといい、そちらを使うことも多い($N(\varepsilon, T, d)$と書く)。



2^{i-1}個のボールでおおうために必要な最小の半径ε

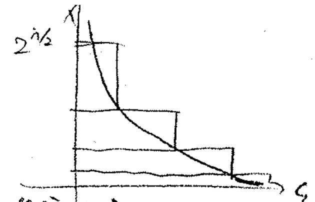
定理(Dudleyのチェイニング)　一般のGで成り立つ

$z_1, \ldots, z_n$を固定する。

$$\hat{R}_n(G) \leq \frac{2}{\sqrt{n}} \left(\sum_{i=1}^{\infty} 2^{i/2} e_{2^i}(G \cup \{0\}, L_2(P_n)) + \sup_{g \in G} \|g\|_{L_2(P_n)} \right)$$

※直感的にはGを有限個の代表点で近似し、その近似精度を粗いのから密なものまで上げてゆく(つなげてゆく)という操作に対応。

※カバリングナンバーを用いれば、$\hat{R}_n(G) \leq \frac{C}{\sqrt{n}} \int_0^{\sup_g \|g\|_{L_2(P_n)}} \sqrt{\log N(\varepsilon, G \cup \{0\}, \|\cdot\|_{L_2(P_n)})}\, d\varepsilon$ である。



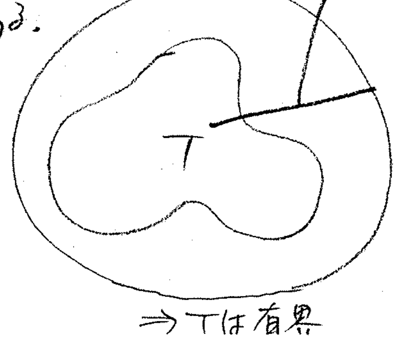
証明

$T = \{(g(z_1), \ldots, g(z_n)) \mid g \in G\} \cup \{(0, \ldots, 0)\} \subset \mathbb{R}^n$,

$d(t, t') = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (t_i - t'_i)^2}$ とすると、(T, d)は可分な距離空間。

以後、$\forall i$で$e_i(T, d)$は有限だとして一般性を失わない。$e_1(T,d) < \infty$

特に、Tは有界で、$i \to \infty$で$e_i(T, d) \to 0$である。



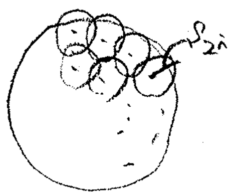
任意の $\varepsilon>0$ に対し、

$$\rho_1 := \sup_{t\in T} d(t,0)$$

$$\rho_{2^i} := (1+\varepsilon)\, e_{2^i}(T,d) \qquad (i=1,2,\cdots)$$

とする。また、$\forall i\geq 1$ に対し、ρ_{2^i}−ネット T_i を $|T_i|\leq 2^{2^i}-1$ なるものとして取って来る。
(エントロピー数の定義より可能)

$\boxed{\bigcup_{t\in T_i} B_d(t,\rho_{2^i}) \supseteq T \text{ なるもの}}$



$T_0=\{0\}$ とする。　すると、$e_j(T,d)\to 0\ (j\to\infty)$ より、$S:=\bigcup_{j=0}^{\infty} T_j$ に対し、

$$\sup_{t\in S}\frac{1}{n}\sum_{i=1}^{n}\sigma_i t_i = \sup_{t\in T}\frac{1}{n}\sum_{i=1}^{n}\sigma_i t_i$$

が成り立つ。よって、T の代わりに S を考えれば良い。

$j\geq 1$, $t\in T$ に対し、$\pi_j(t)\in T_j$ を、$d(t,\pi_j(t))\leq \rho_{2^i}$ なるものとして取って来る。

特に、$t\in T_j$ なら、$\pi_j(t)=t$ としておく。

$j\geq 1$ を1つ固定して、$\gamma_i\ (i\leq j)$ を、

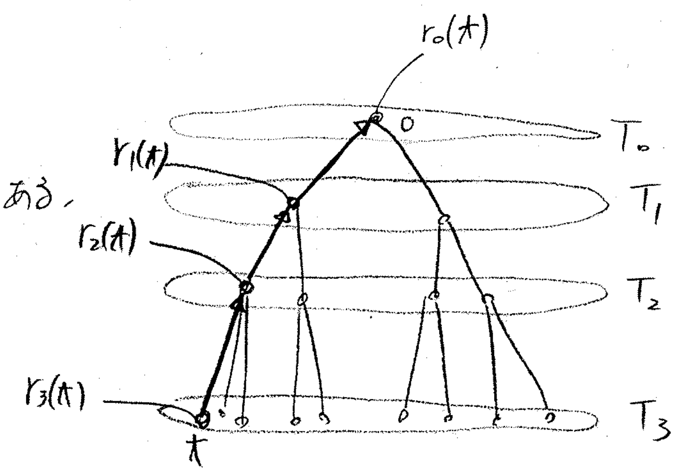
$$\gamma_j=\pi_j$$

$$\gamma_{i-1}=\pi_{i-1}\circ\gamma_i \quad (i=1,\cdots,j)$$

としておく。

当然、$d(\gamma_i(t),\gamma_{i-1}(t))\leq \rho_{2^{i-1}}$ である。

$\boxed{\text{チェイニング}}$



今 $h(t):=\frac{1}{n}\sum_{i=1}^{n}\sigma_i t_i\ (t\in T)$

とおくと、

$t\in T_j$ について、

$$h(t) = h(t)-0 = \sum_{i=1}^{j}\left(h(\gamma_i(t))-h(\gamma_{i-1}(t))\right) \leq \sum_{i=1}^{j}\max_{t\in T_i} h(t)-h(\pi_{i-1}(t))$$

となる。よって、$S_k:=T_0\cup\cdots\cup T_k$ に対しても、

$$\max_{t\in S_k} h(t) \leq \sum_{j=1}^{k}\max_{t\in T_j}\{h(t)-h(\pi_{j-1}(t))\}$$

が成り立つ。すると、有限集合のRademacher複雑さより（Massartの定理）

$$E_{\sigma_i}\left[\max_{t\in S_k} h(t)\right] \leqq \sum_{j=1}^{k} E_{\sigma_i}\left[\max_{t\in T_j} \{h(t) - h(\pi_{j-1}(t))\}\right]$$

$$\leqq \sum_{j=1}^{k} \sqrt{\frac{2\log(2^{2^j-1})}{n}}\ \delta_{2^j-1}$$

$S = \bigcup_{j=0}^{\infty} T_j$ については、$k\to\infty$ とすることで、単調増大性より、

$$E_{\sigma_i}\left[\sup_{t\in S} h(t)\right] \leqq \sum_{j=0}^{\infty} \sqrt{\frac{4\cdot 2^j}{n}}\ \delta_{2^j}$$

あとは、$\varepsilon\to 0$ として、$\delta_1 = \sup_t d(t,0)$ であることに注意すれば題意を示す。//

系

$$\left[\ R_n(G) \leqq \frac{2}{\sqrt{n}}\left(\sum_{j=1}^{\infty} E_{D_n}\left[e_{2^j}(G\cup\{0\}, L_2(P_n))\right] 2^{j/2} + E\left[\sup_{g\in G}\|g\|_{L_2(P_n)}\right]\right.\right.$$

さて、RKHSのエントロピー数は？

定理

カーネル関数が次のように分解されるとする。

$$k(x,x') = \sum_{i=1}^{\infty} \lambda_i\,\phi_i(x)\,\phi_i(x') \qquad (\text{in } (\mathcal{X}\times\mathcal{X},\ P_X\times P_X))$$

ただし、$\lambda_i \geqq 0$, $\{\phi_i\}$ は $L_2(P_X)$ 内での正規直交系。

$$\lambda_i \leqq a\, i^{-\frac{1}{p}} \quad (i\geqq 1)$$

がある $0<p<1$ でなりたつとき、$\mathcal{F} = \{f\in\mathcal{H} \mid \|f\|_{\mathcal{H}} \leqq 1\}$ は、

$$E\left[e_i(\mathcal{F}, L_2(P_n))\right] \leqq C_p\, a\, \min\{i^{\frac{1}{2p}}, n^{\frac{1}{2p}}\}\cdot i^{-\frac{1}{p}} \leqq C_p\, a\, i^{-\frac{1}{2p}}$$

をみたす。

証明は略：Steinwart, Christmann "SVM" の Cor. 7.31.

定理

$\mathcal{F}_r = \{f\in\mathcal{F} \mid \|f\|_{L_2(P_X)} \leqq r\}$ とすると、上の条件のもと、

$$R_n(\mathcal{F}_r) \leqq C(p)\max\left\{\frac{a^p r^{1-p}}{\sqrt{n}},\ n^{-\frac{1}{1+p}} a^{\frac{2p}{1+p}}\right\}$$

（略証）

$k = \min\{j \mid r \geqq C_p a j^{-\frac{1}{2p}}\}$ とおくと、

$$R_n(\mathcal{F}_r) \leqq \sum_{i=1}^{k} \frac{2^{i/2}\cdot r}{\sqrt{n}} + \sum_{i=k+1}^{\infty} \frac{2^{i/2}\cdot C_p a i^{-\frac{1}{2p}}}{\sqrt{n}} \simeq \frac{a^p r^{1-p}}{\sqrt{n}}。$$

$L_2(P_X)$ と $L_2(P_n)$ の間に差がある ので2つめの項がある。//

よって、

$$G_r \subseteq \left\{ \ell\cdot f - \ell\cdot f^* \;\middle|\; f \in \mathcal{F}_{\sqrt{\frac{2r}{\alpha}}} \right\}$$ より、

$$R_n(G_r) \leq R_n\left(\mathcal{F}_{\sqrt{\frac{2r}{\alpha}}}\right) \leq c \max\left(\frac{r^{\frac{1-p}{2}}}{\sqrt{n}},\ n^{-\frac{1}{1+p}} \right)$$

である。　よって、十分大きな η_* に対し、$r = \eta_*\, n^{-\frac{1}{1+p}}$ とおけば、

$$\phi(r) \leq R_n(G_r) \leq r$$

をみたす。よって、　$L(\hat{f}) - L(f^*) \leq c\,\eta_*\, n^{-\frac{1}{1+p}}$　である。

例: ソボレフ空間 (d次元, 滑らかさ m)

$p = \frac{d}{2m}$,　収束レート: $n^{-\frac{2m}{2m+d}}$ ← なめらかなほど速い収束

$(2m>d)$

⟶ これはミニマクス最適.