◦ RKHSのRademacher複雑度

$$\mathcal{F} = \{ f \in H \mid \|f\|_H \leq R \}$$

$k(x,x) \leq 1 \quad ({}^\forall x \in \mathcal{X})$ のとき、

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

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

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

$$\leq E_\sigma\left[\sup_{f \in \mathcal{F}} \|f\|_H \sqrt{\left\langle \frac{1}{n} \sum_{i=1}^{n} \sigma_i k(x_i, \cdot), \frac{1}{n} \sum_{i=1}^{n} \sigma_i k(x_i, \cdot) \right\rangle_H} \right]$$

$$\leq R \, E_\sigma\left[\sqrt{\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_i \sigma_j k(x_i, x_j)} \right]$$

$$\leq R \sqrt{E_\sigma\left[\frac{1}{n^2} \sum_i \sum_j \sigma_i \sigma_j k(x_i, x_j) \right]}$$

$$= R \sqrt{\frac{1}{n^2} \sum_{i=1}^{n} k(x_i, x_i)} \leq \frac{R}{\sqrt{n}} \quad //$$

◦ カーネル法の学習効率

モデル： $y_i = f^\circ(x_i) + \varepsilon_i$ $\quad (i=1,\dots,n)$

$x_i \sim P_X,\ \varepsilon_i \sim N(0,\sigma^2)\quad$ (i.i.d.)

n ← サンプルサイズ

カーネルリッジ回帰： $\hat{f} = \underset{f\in\mathcal{H}}{\operatorname{argmin}}\ \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda\|f\|_{\mathcal{H}}^2$

推定誤差： $\|\hat{f}-f^\circ\|_{L^2(P_X)}^2 := \mathbb{E}[(\hat{f}(X) - f^\circ(X))^2]$

← テストデータ: $X\sim P_X$ で期待値取る

Q：サンプルサイズ $n\to\infty$ でどれだけ速く $\|\hat{f}-f^\circ\|_{L^2}^2 \to 0$ か？

2乗損失を用いると解が陽に書けるので
作用素を用いた解析を紹介

$\Sigma : \mathcal{H}\to\mathcal{H}$, $\hat{\Sigma} : \mathcal{H}\to\mathcal{H}$ をそれぞれ次のように定義：

← $k_x^* f = \langle k_x, f\rangle_{\mathcal{H}}$ とする

$$\Sigma = E_x[k_x k_x^*], \quad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} k_{x_i} k_{x_i}^*$$

つまり $f\in\mathcal{H}$ に対し
$$\begin{cases} \Sigma f = \int k_x k_x^* f\, dP_X(x) = \int k_x f(x)\, dP_X(x) \\ \hat{\Sigma} f = \frac{1}{n}\sum_{i=1}^{n} k_{x_i} f(x_i) \end{cases}$$
である。

$g^\circ\in\mathcal{H}$, $\hat{g}\in\mathcal{H}$ を次のように定義： $\quad$ ($f^\circ\in L^2(P_X)$ は仮定しておく)

$$g^\circ := E_x[k_x f^\circ(x)] = \int k_x f^\circ(x)\, dP_X(x)$$
$$\hat{g} = \frac{1}{n}\sum_{i=1}^{n} k_{x_i} y_i$$

$$\Longrightarrow \begin{cases} f_\lambda := (\Sigma + \lambda I)^{-1} g^\circ \\ \hat{f} = (\hat{\Sigma} + \lambda I)^{-1}\hat{g} \end{cases}$$
が定義される。

← $\hat{f}$ の定義より

以後，簡単のため $f^\circ \in \mathcal{H}$ とし，$\mathcal{H} \hookrightarrow L^2(P_X)$ とする．
また，$f \in \mathcal{H}$ も自然な埋め込みを用いて $f \in L^2(P_X)$ とみなす．

$$\|\hat{f} - f^\circ\|_{L^2}^2 \leq 2\Big(\underbrace{\|f^\circ - f_\lambda\|_{L^2}^2}_{\text{バイアス} \,=\, \boxed{A(\lambda)}} + \underbrace{\|f_\lambda - \hat{f}\|_{L^2}^2}_{\text{バリアンス} \,=\, V(\lambda)}\Big)$$

と「バイアス-バリアンス分解」をしておく．

バリアンスの評価をする．　$f_\lambda = (\Sigma + \lambda I)^{-1}\Sigma f^\circ$ に注意すると

$$\begin{aligned}
f_\lambda - \hat{f} &= f_\lambda - (\hat{\Sigma} + \lambda I)^{-1}\hat{g} \\
&= f_\lambda - (\hat{\Sigma} + \lambda I)^{-1}(\hat{g} - \hat{\Sigma} f^\circ) - (\hat{\Sigma} + \lambda I)^{-1}\hat{\Sigma} f^\circ \\
&= f_\lambda - (\hat{\Sigma} + \lambda I)^{-1}\hat{\Sigma} f^\circ - (\hat{\Sigma} + \lambda I)^{-1}(\hat{g} - \hat{\Sigma} f^\circ) \\
&= (\hat{\Sigma} + \lambda I)^{-1}\big[(\hat{\Sigma} + \lambda I)(f_\lambda - f^\circ) + \lambda f^\circ\big] \\
&\qquad - (\hat{\Sigma} + \lambda I)^{-1}(\hat{g} - \hat{\Sigma} f^\circ) \\
&= (\hat{\Sigma} + \lambda I)^{-1}(\hat{\Sigma} - \Sigma)(f_\lambda - f^\circ) \quad \leftarrow x\text{の不確定性} \\
&\qquad - (\hat{\Sigma} + \lambda I)^{-1}(\hat{g} - \hat{\Sigma} f^\circ) \quad \uparrow y\text{の不確定性}
\end{aligned}$$

$$\lambda f^\circ = (\Sigma + \lambda) f^\circ - \Sigma f^\circ = (\Sigma + \lambda) f^\circ - (\Sigma + \lambda) f_\lambda = (\Sigma + \lambda)(f^\circ - f_\lambda)$$

となり，

$$\begin{aligned}
\|f_\lambda - \hat{f}\|_{L^2}^2 &\leq 2\|(\hat{\Sigma} + \lambda I)^{-1}(\hat{\Sigma} - \Sigma)(f_\lambda - f^\circ)\|_{L^2}^2 \quad \leftarrow V_1(\lambda) \\
&\quad + 2\|(\hat{\Sigma} + \lambda I)^{-1}(\hat{g} - \hat{\Sigma} f^\circ)\|_{L^2}^2 \quad \leftarrow V_2(\lambda)
\end{aligned}$$

とおく

を得る．

つまり

$$\|\hat{f} - f^\circ\|_{L^2} \leq 2A(\lambda) + 4V_1(\lambda) + 4V_2(\lambda).$$

定義

- $B(\lambda) := \|f_\lambda - f^\circ\|_{\mathcal{H}}$　（再構成ゴサ）
- $N(\lambda) := \mathrm{Tr}\big[(\Sigma + \lambda I)^{-1}\Sigma\big]$　（自由度）degrees of freedom

$$\boxed{\sum_{j=0}^{\infty} \frac{\mu_j}{\mu_j + \lambda}}$$

$\mu_j = 1 \quad (j = 1, \dots, d)$
$\mu_j = 0 \quad (j > d)$
$\Rightarrow N(\lambda) = \frac{d}{1+\lambda}$

（以後，$k(x,x) \leq 1$ とする）

$\boxed{T_1}$ 十分大きなサンプルサイズのもと.

$$T_1(\lambda) \leqq c\left[\frac{\lambda^{-1}A(\lambda)}{n} + \frac{\lambda^{-1}B(\lambda)}{n^2}\right]\log\left(\frac{1}{\delta}\right)$$

が確率 $1-\delta$ で成立つ ($\delta>0$). なお, cは定数.

$\boxed{T_2}$ ← こちらの方が主要項

$$T_2(\lambda) \leqq c\left[\frac{\sigma^2 N(\lambda)}{n} + \frac{\lambda^{-1}}{n^2}\right]\log^2\left(\frac{1}{\delta}\right)$$

が確率 $1-\delta$ で成り立つ.

T_2 を示す方だけ解説する.

Σの定義より,

$$T_2(\lambda) = \left\|\sqrt{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}(\hat{g}-\hat{\Sigma}f^\circ)\right\|_{\mathcal{H}}^2$$

$$= \left\|\sqrt{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}(\Sigma+\lambda I)^{\frac{1}{2}}(\Sigma+\lambda I)^{-\frac{1}{2}}(\hat{g}-\hat{\Sigma}f^\circ)\right\|_{\mathcal{H}}^2$$

$$\leqq \left\|\sqrt{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}(\Sigma+\lambda I)^{\frac{1}{2}}\right\|_{\mathcal{L}(\mathcal{H})} \leftarrow \text{作用素ノルム}$$

$$\times \left\|(\Sigma+\lambda I)^{-\frac{1}{2}}(\hat{g}-\hat{\Sigma}f^\circ)\right\|_{\mathcal{H}}^2$$

- $\left\|\sqrt{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}(\Sigma+\lambda I)^{\frac{1}{2}}\right\|_{\mathcal{L}(\mathcal{H})}$ の Bound

$$\text{左辺} \leqq \left\|(\Sigma+\lambda I)^{\frac{1}{2}}(\hat{\Sigma}+\lambda I)^{-1}(\Sigma+\lambda I)^{\frac{1}{2}}\right\|_{\mathcal{L}(\mathcal{H})}$$

$$= \left\|(\Sigma+\lambda I)^{\frac{1}{2}}(\hat{\Sigma}-\Sigma+\Sigma+\lambda I)^{-1}(\Sigma+\lambda I)^{\frac{1}{2}}\right\|_{\mathcal{L}(\mathcal{H})}$$

$$= \left\|\left[\underbrace{(\Sigma+\lambda I)^{-\frac{1}{2}}(\hat{\Sigma}-\Sigma)(\Sigma+\lambda I)^{-\frac{1}{2}}}_{} + I\right]^{-1}\right\|_{\mathcal{L}(\mathcal{H})}$$

↓

この大きさを見積もる
作用素 Bernstein の不等式

定理（作用素値確率変数のBernsteinの不等式）

- $\mathcal{H}$：可分なヒルベルト空間
- $X_1, \dots, X_n$：$\mathcal{H}$上の自己共役作用素に値をとるi.i.d.確率変数
  - $E[X_i] = 0$
  - $\|X_i\| \leqq M$ (a.s.)
  - $E[X_i^2] \leq S$ （Sは半正定値自己共役作用素）

このとき、$\forall \delta \in (0,1)$に対し、

$$P\left(\left\| \frac{1}{n}\sum_{i=1}^n X_i \right\| \leqq \sqrt{\frac{2\|S\|\beta}{n}} + \frac{2M\beta}{n} \right) \geqq 1-\delta,$$

ただし、$\beta = \log\left(\frac{2\mathrm{Tr}[S]}{\|S\|\delta} \right)$.

//

$$X_i = (\Sigma + \lambda I)^{-\frac{1}{2}} (k_{x_i} k_{x_i}^* - \Sigma)(\Sigma + \lambda I)^{-\frac{1}{2}}$$ とすれば、

$E[X_i] = 0$,
$$\begin{aligned}
\|X_i\| &\leqq \|(\Sigma+\lambda I)^{-\frac{1}{2}} \Sigma (\Sigma+\lambda I)^{-\frac{1}{2}}\| + \\
&\quad \|(\Sigma+\lambda I)^{-\frac{1}{2}} k_{x_i} k_{x_i}^T (\Sigma+\lambda I)^{-\frac{1}{2}}\| \\
&\leqq 1 + \sup_{x \in \mathrm{supp}(P_X)} \langle k_x, (\Sigma+\lambda I)^{-1} k_x \rangle_{\mathcal{H}} \\
&\leqq 1 + \sup_x \frac{1}{\lambda} \underbrace{\langle k_x, k_x \rangle_{\mathcal{H}}}_{= k(x,x)} \leqq 1 + \frac{1}{\lambda} =: M
\end{aligned}$$

$$\begin{aligned}
E[X_i^2] &\leq E_x\left[(\Sigma+\lambda I)^{-\frac{1}{2}} k_x k_x^* (\Sigma+\lambda I)^{-1} k_x k_x^* (\Sigma+\lambda I)^{-\frac{1}{2}} \right] \\
&\leqq \sup_x \langle k_x, (\Sigma+\lambda I)^{-1} k_x \rangle_{\mathcal{H}} E_x\left[(\Sigma+\lambda I)^{-\frac{1}{2}} k_x k_x^* (\Sigma+\lambda I)^{-\frac{1}{2}} \right] \\
&\leqq \frac{1}{\lambda} (\Sigma+\lambda I)^{-\frac{1}{2}} \Sigma (\Sigma+\lambda I)^{-\frac{1}{2}} = S
\end{aligned}$$

を用いる。$\|S\| \leqq \frac{1}{\lambda}$ を用いると、

$$P\left(\left\|(\Sigma+\lambda I)^{-\frac{1}{2}}(\hat{\Sigma}-\Sigma)(\Sigma+\lambda I)^{-\frac{1}{2}}\right\| \leq \sqrt{\frac{2\beta}{\lambda n}} + \frac{2(1+\frac{1}{\lambda})\beta}{n}\right) \geq 1-\delta,$$

$$\beta = \log\left(\frac{2\,\mathrm{Tr}[\Sigma(\Sigma+\lambda I)^{-1}]}{\|\Sigma(\Sigma+\lambda I)^{-1}\|\,\delta}\right)$$

である.

[十分小さな λ を選べば, $\|\Sigma(\Sigma+\lambda I)^{-1}\| \geq \frac{1}{2}$ である.
($\lambda \leq \|\Sigma\|$ とする)
$\Longrightarrow \quad \beta = \log\left(\frac{4N(\lambda)}{\delta}\right)$]

よって, $\frac{1}{n\lambda} \ll 1$ となるくらい n が十分大きければ,

$$\underline{\left\|(\Sigma+\lambda I)^{-\frac{1}{2}}(\hat{\Sigma}-\Sigma)(\Sigma+\lambda I)^{-\frac{1}{2}}\right\| \leq \frac{1}{2}}$$

が高い確率で成り立つ

$$\Longrightarrow \underline{\left\|\sqrt{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}(\Sigma+\lambda I)^{\frac{1}{2}}\right\|_{L(\mathcal{H})} \leq 2}$$

- $\|(\Sigma+\lambda I)^{-\frac{1}{2}}(\hat{g}-\hat{\Sigma}f^\circ)\|_{\mathcal{H}}^2$ のバウンド

$$(\Sigma+\lambda I)^{-\frac{1}{2}}(\hat{g}-\hat{\Sigma}f^\circ) = \frac{1}{n}\sum_{i=1}^n (\Sigma+\lambda I)^{-\frac{1}{2}}\left(k_{x_i}\overbrace{y_i}^{f^\circ(x_i)+\varepsilon_i} - k_{x_i}f^\circ(x_i)\right)$$

$$= \frac{1}{n}\sum_{i=1}^n \underbrace{(\Sigma+\lambda I)^{-\frac{1}{2}}k_{x_i}\varepsilon_i}_{= \xi(x_i,\varepsilon_i)}$$

$\xi(x_i,\varepsilon_i)$ とおく.

$\to \left\|\frac{1}{n}\sum_{i=1}^n \xi(x_i,\varepsilon_i)\right\|_{\mathcal{H}}^2$ を Bound する.

$E_{x,\varepsilon}[\xi(x,\varepsilon)] = 0$ に注意する.

<u>定理</u> (Bernstein の不等式)

- $\mathcal{H}$: 可分なヒルベルト空間
- ξ_i: $\mathcal{H}$ に値をとる i.i.d. 確率変数,

$E[\xi_i] = 0$ かつ. $E[\|\xi_i\|_{\mathcal{H}}^m] \leq \frac{1}{2}m!\, v\, M^{m-2}$ $(\forall m \geq 2)$

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^n \xi_i\right\|_{\mathcal{H}} \leq 2\left(\sqrt{\frac{v}{n}} + \frac{M}{n}\right)\log\left(\frac{2}{\delta}\right)\right) \geq 1-\delta$$
$(\forall \delta \in (0,1))$

ξ_i として、$\xi(x_i, \varepsilon_i)$ を代入する。

$$\mathbb{E}\left[\|\xi(x_i,\varepsilon_i)\|_{\mathcal{H}}^m\right]$$
$$\leq \sqrt{\mathbb{E}\left[\|\xi(x_i,\varepsilon_i)\|_{\mathcal{H}}^{2m}\right]}$$
$$\leq \sqrt{\mathbb{E}_{\varepsilon_i}\left[|\varepsilon_i|^{2m}\right]} \cdot \left(\int \|k_x(\Sigma+\lambda I)^{-\frac{1}{2}}\|_{\mathcal{H}}^{2m} \, dP_X(x)\right)^{\frac{1}{2}}$$
$$\leq \sqrt{(2m-1)!!\,\sigma^{2m}} \left(\underbrace{\int \langle k_x, (\Sigma+\lambda I)^{-1} k_x\rangle_{\mathcal{H}} \, dP_X(x)}_{N(\lambda)}\right)^{\frac{1}{2}} \times \left(\underbrace{\sup_x \langle k_x, (\Sigma+\lambda I)^{-1} k_x\rangle_{\mathcal{H}}}_{\leq \frac{1}{\lambda}}\right)^{\frac{m-2}{2}}$$
$$\leq \frac{1}{2} m! \underbrace{\left(\sigma\sqrt{N(\lambda)}\right)^2}_{= \nu} \underbrace{\left(\sqrt{\frac{2}{\lambda}}\right)}_{= H}^{m-2}$$

なので、

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\xi(x_i,\varepsilon_i)\right\|_{\mathcal{H}} \leq 2\left(\sqrt{\frac{\sigma^2 N(\lambda)}{n}} + \frac{1}{n}\sqrt{\frac{2}{\lambda}}\right)\log\left(\frac{2}{\delta}\right)\right) \geq 1-\delta$$

である。よって、

$$\|(\Sigma+\lambda I)^{-\frac{1}{2}}(\hat{g} - \hat{\Sigma} f^\circ)\|_{\mathcal{H}}^2 \leq 8\left(\underbrace{\frac{\sigma^2 N(\lambda)}{n}}_{\uparrow} + \frac{2}{n^2\lambda}\right)\log^2\left(\frac{2}{\delta}\right)$$

が確率 $1-\delta$ でみたされる。　ここに、自由度が現れる。

※キモ：

$$\mathbb{E}\left[\|\xi(x_i,\varepsilon_i)\|_{\mathcal{H}}^2\right]$$
$$= \underbrace{\mathbb{E}[\varepsilon_i^2]}_{= \sigma^2} \cdot \underbrace{\mathbb{E}_x\left[\|k_x(\Sigma+\lambda I)^{-\frac{1}{2}}\|_{\mathcal{H}}^2\right]}_{\shortparallel}$$
$$\mathbb{E}_x\left[\langle (\Sigma+\lambda I)^{-\frac{1}{2}} k_x, (\Sigma+\lambda I)^{-\frac{1}{2}} k_x\rangle_{\mathcal{H}}\right]$$
$$\shortparallel$$
$$\mathbb{E}_x\left\{\mathrm{Tr}\left[k_x k_x^* (\Sigma+\lambda I)^{-1}\right]\right\}$$
$$\shortparallel$$
$$\mathrm{Tr}\left[\Sigma(\Sigma+\lambda I)^{-1}\right] \text{：自由度}$$

以上をまとめると.

$$\|\hat{f} - f^\circ\|^2_{L^2(P_X)} \lesssim \underline{A(\lambda)} + \left(\underline{\frac{\sigma^2 N(\lambda)}{n}} + \frac{\lambda^{-1}A(\lambda)}{n} + \frac{1}{n^2}\left(\lambda^{-1}B(\lambda) + \lambda^{-1}\right)\right)\lg^2(\delta)$$

with prob. $1-\delta$.

この2項が主要項になりやすい.

$A(\lambda)$, $N(\lambda)$, $B(\lambda)$ 等を評価するために, いくつかの仮定を置く.
(カーネル法の理論では標準的)

$T_k f(x) := \int k(x,\cdot) f(x)\, dP_X(x)$
↑ データの分布を使う.

仮定

1. ある $r \geq \frac{1}{2}$ が存在して, $f^\circ \in T_k^r L^2(P_X)$.
   つまり $\exists h^\circ \in L^2(P_X)$ で $\underline{f^\circ = T_k^r h^\circ}$ ← $r \geq \frac{1}{2}$ より $f^\circ \in \mathcal{H}$ でもある.

2. P_X に対する k の Mercer 展開
   $$k(x,x') = \sum_{i=1}^{\infty} \mu_i e_i(x) e_i(x')$$
   $(\mu_1 \geq \mu_2 \geq \cdots \geq 0, \quad (e_i)_i$ は $L^2(P_X)$ 内の正規直交系$)$
   が可能であるとする. このとき, $\exists \alpha > 1$ で
   $$\mu_i \lesssim i^{-\alpha}$$

意味: 1. f° の"滑らかさ"を表す. r が大きいほど滑らか.
→ より RKHS の中に入る.

2. RKHS の"複雑さ"を表す. α が大きいほど単純.

⟨e⟩ α は P_X に依存

($\lambda \to 0$ を考える)

$$\Rightarrow \boxed{\begin{cases} \cdot\ A(\lambda) \leq \lambda^{2r} \|h^\circ\|^2_{L^2(P_X)} \\ \cdot\ N(\lambda) \lesssim \lambda^{-\frac{1}{\alpha}} \end{cases}}$$

← r が大きいと, バイアスが速く0へ収束

← α が大きいと, バリアンスが大きくなりにくい.

$$B(\lambda) \leq \lambda^{2r-1} \|h^\circ\|^2_{L^2(P_X)}$$

主要項だけ書き出すと,

$$\text{推定誤差} \lesssim \underbrace{\lambda^{2r}}_{\text{バイアス}} + \underbrace{\frac{\lambda^{-\frac{1}{\alpha}}}{n}}_{\text{バリアンス}} \xrightarrow{\text{バイアスとバリアンスのバランスをとる}} \boxed{n^{-\frac{2r\alpha}{2r\alpha+1}}}$$

$\lambda = n^{-\frac{\alpha}{2r\alpha+1}}$ とすれば

← (σ 項を無視している)

$$\|\hat{f} - f^\circ\|_{L^2}^2 \lesssim n^{-\frac{2r\alpha}{2r\alpha+1}}$$

は mini-max 最適レート と呼ばれる.

定理 (mini-max 最適レート)

$$\boxed{\inf_{\hat{f}:\text{推定量}} \sup_{\substack{f^\circ \in \mathcal{H}: \\ f^\circ = T_K^{\frac{r}{2}} h^\circ,\ \|h^\circ\|_{L^2(P_X)} \le 1}} \mathbb{E}\left[\|\hat{f} - f^\circ\|_{L^2(P_X)}^2\right] \gtrsim n^{-\frac{2r\alpha}{2r\alpha+1}}}$$

つまり、最悪誤差はどんな推定量を用いても $n^{-\frac{2r\alpha}{2r\alpha+1}}$ を超えられない.

$\longrightarrow$ リッジ回帰は (ほぼ) 最適

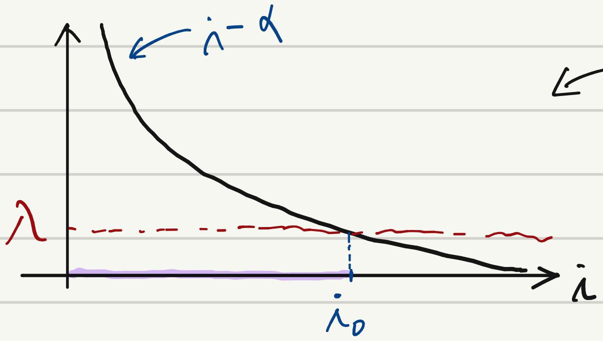
↳ 次ページに続く

◎ 自由度 $N(\lambda)$ の意味について（重要）

$$N(\lambda) = \mathrm{Tr}\left[\Sigma(\Sigma+\lambda I)^{-1}\right] = \sum_{i=1}^{\infty} \frac{\mu_i}{\mu_i+\lambda}$$

であることを思い出す。今 $\mu_i = O(i^{-\alpha})$ $(\alpha>1)$ ならば

$N(\lambda)$ は おおよそ $\displaystyle\sum_{i=1}^{\infty} \frac{i^{-\alpha}}{i^{-\alpha}+\lambda}$ でおさえられる（これが $O(\lambda^{-\frac{1}{\alpha}})$）



← このように、固有値を λ で切って、
λ 以上の固有値の個数 i_0 を数えると、

$$i_0^{-\alpha} \fallingdotseq \lambda \Longrightarrow i_0 \fallingdotseq \lambda^{-\frac{1}{\alpha}}$$

つまり、$i_0 \fallingdotseq N(\lambda)$

自由度は、λ 以下の固有値に対応する固有ベクトル空間を無視した有限次元空間の次元 ⟶ カーネルリッジ回帰はこの有限次元空間に制限して学習していることに相当する。

$$\mathcal{H} \ni f^{\circ} = \sum_{i=1}^{\infty} \alpha_i e_i(x) \xrightarrow{\text{近似}} \sum_{i=1}^{i_0} \alpha_i e_i(x) \quad (i_0 \fallingdotseq N(\lambda))$$

近似誤差 $A(\lambda)$: λ^{2r}

自由度 $N(\lambda)$ は λ の選択にも用いられる ⟶ Mallows' C_p 規準

AICと同様の効果：AICはモデル選択における次元の決定に使われる。
Mallows' C_p 規準では、AICで出てくる次元のペナルティの代わりに $N(\lambda)$ を用いる。
すると、λ の選択もできる。

☆
カーネル法が無限次元モデルにもかかわらず過学習せず適切な推定ができるのは、この $N(\lambda)$ が有限になるからである。
↑つまり、有限次元モデルで近似できる

- RKHSのcovering number

$\mathcal{B}(\mathcal{H})$: $\mathcal{H}$の単位球

$$\mu_i \le c\, i^{-\alpha} \quad \Longrightarrow \quad \begin{cases} \log N(\mathcal{B}(\mathcal{H}), \varepsilon, \|\cdot\|_{L^2(P_X)}) \lesssim \varepsilon^{-\frac{2}{\alpha}} \\ e_i(\mathcal{B}(\mathcal{H}), \|\cdot\|_{L^2(P_X)}) \lesssim i^{-\frac{\alpha}{2}} \end{cases}$$

$\lambda = \varepsilon^2$を代入すれば、 $\log N(\cdots) \lesssim \lambda^{-\frac{1}{\alpha}} \simeq N(\lambda)$

$\longrightarrow$ $\|f\|_{L^2(P_X)}^2 \le \lambda$は0と同一視すれば残りの次元は$N(\lambda)$

局所Rademacher複雑度を用いたfast rate

$$\Longrightarrow \quad \underline{n^{-\frac{1}{1+\frac{1}{\alpha}}} = n^{-\frac{\alpha}{\alpha+1}}} \qquad (r = \tfrac{1}{2}\text{のときのレート})$$

ちなみに、
$$\mathbb{E}\left[R_n\left(\{f \in \mathcal{F} \mid Pf^2 \le r, \|f\|_{\mathcal{H}} \le 1\}\right)\right]$$
$$\le \left(\frac{2}{n}\sum_{i=1}^{\infty} \min\{r, \mu_i\}\right)^{\frac{1}{2}}$$

である。 (Mendelson: Geometric parameters of kernel machines. COLT2002. pp. 29-43.)