◦ 汎化誤差解析と Rademacher 複雑度

$\mathcal{F}$: モデル　($f : x \in X \mapsto y \in \mathbb{R}$ なる関数の集合)

例: ニューラルネットワーク, 再生核ヒルベルト空間

$\ell : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$　を損失関数とする.

例: $\ell(y, f(x)) = (y - f(x))^2$: 2乗損失　　$(y \in \mathbb{R})$

$\ell(y, f(x)) = \log(1 + \exp(-yf(x)))$: ロジスティック損失　$(y \in \{\pm 1\})$

期待損失, 予測誤差 : $L(f) = E_{X,Y}[\ell(Y, f(X))]$

経験損失, 訓練誤差 : $\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i))$

データ : $Z_i = (x_i, y_i)$ $(i = 1, \dots, n)$ は i.i.d. に Z : ある分布 P に従うとする.

$Pg = E_{X,Y}[g(X,Y)]$, $P_n g = \frac{1}{n}\sum_{i=1}^{n} g(x_i, y_i)$　と書く.

($L(f) = P(\ell \circ f)$, $\hat{L}(f) = P_n(\ell \circ f)$ である)

← $Z = (X, Y)$ の分布と $Z_i = (x_i, y_i)$ の分布は同じである.

$f^\circ = \underset{f : \text{可測関数}}{\operatorname{argmin}} L(f)$, $f^* = \underset{f \in \mathcal{F}}{\operatorname{argmin}} L(f)$

$\hat{f}$: 何らかの推定量

(どちらも存在すると仮定)

- Excess risk (残余誤差, 余剰誤差)
  $L(\hat{f}) - L(f^\circ)$　または　$L(\hat{f}) - L(f^*)$
- Generalization gap (汎化ギャップ)
  $L(\hat{f}) - \hat{L}(\hat{f})$

← これを「汎化誤差」と言うことが多い.

Excess risk

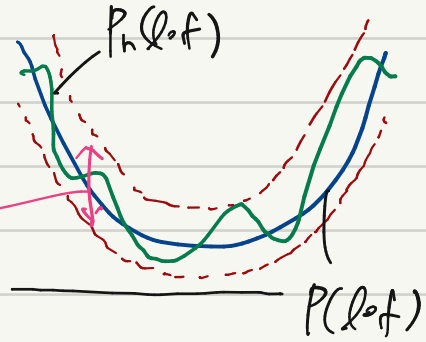
$$L(\hat{f}) - L(f^\circ) = \underbrace{L(\hat{f}) - \hat{L}(\hat{f})}_{\text{汎化ギャップ}} + \underbrace{\hat{L}(\hat{f}) - \hat{L}(f^*)}_{\substack{\hat{f} \text{ が訓練データを十分小さくしていれば} \\ 0 \text{ 以下}}} + \underbrace{\hat{L}(f^*) - L(f^*)}_{\substack{\text{大数の法則} \\ O_p(\frac{1}{\sqrt{n}})}} + \underbrace{L(f^*) - L(f^\circ)}_{\substack{\text{モデル} \\ \text{バイアス}}}$$

$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\ \hat{L}(f)$ なら

$$L(\hat{f}) - L(f^*) = \underbrace{(P - P_n)(\ell \circ \hat{f} - \ell \circ f^*)}_{\text{バリアンス}} + \underbrace{L(f^*) - L(f^\circ)}_{\text{(モデル)バイアス}}$$

- 汎化ギャップの Bound

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



(※) $\hat{f}$はデータに依存するため、単純な大数の法則は使えない。

$G := \{ g = \ell \circ f \mid f \in \mathcal{F} \}$ とおけば、

$\sup_{g \in G} (P - P_n) g$ を抑える問題になる → 経験過程

※ 以後 $\sup_{g \in G} P_n g$ や $\sup_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i g(x_i)$ は（↑は符号値）

可測である とする。可測でない場合は以下の議論は成り立たない。（Fubini の定理が成り立たない, $E_X E_\varepsilon [(\cdot)^*] \neq E_\varepsilon E_X [(\cdot)^*]$）

- van der Vaart & Wellner : Weak convergence and empirical processes. (1996) の Sec 2.3.
- Dudley : Uniform central limit theorems.
  → Suslin image admissible

◎ Rademacher complexity

以下. $\|g\|_\infty \leq M \quad (\forall g \in G)$ とする.

Def (Rademacher複雑度)

$\sigma_i \ (i=1,\dots,n)$: Rademacher変数 (i.i.d.)

$$P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$$

とする.

絶対値付けないバージョンもある.

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

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

//

Thm (対称化)

$\{z_i\}_{i=1}^n$ と書くのが大変なので z_i と書く.

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

//

(直接 sup を評価するより, $R_n(G)$ の方が評価しやすい)

(Proof)

z_i' と同じ分布に従う r.v.

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

$$\leq E_{z_i, z_i'}\left[\sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^n \left(g(z_i') - g(z_i) \right) \right| \right] \quad \text{(Jensenの不等式)}$$

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

(z_i' と z_i は同じ分布)

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

$$= 2 R_n(G) \quad (\because \text{Fubiniの定理})$$

//

↑ここで可測性仮定.
可積分性は有界性よりO.K.

Thm

$$P\left(\sup_{g\in G}|(P-P_n)g| \geq 2R_n(G) + M\sqrt{\frac{2\log(\frac{1}{\delta})}{n}}\right) \leq \delta \quad (\forall \delta > 0)$$

Proof

Lem (McDiamidの不等式)

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

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

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

が $\forall x_1, \dots, x_n, x_i' \in \mathcal{X}$ に対して成り立つ時,

$$P\left(f(X_1, \dots, X_n) - E[f(X_1, \dots, X_n)] \geq \varepsilon\right) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right) \quad (\forall \varepsilon > 0) \quad //$$

今, $\|g\|_\infty \leq M$ $(\forall g \in G)$ より

$$f(z_1, \dots, z_n) = \sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^n g(z_i) - Pg\right|$$

とおくと,

$$\begin{aligned} &|f(z_1, \dots, z_n) - f(z_1, \dots, z_i', \dots, z_n)| \\ &= \left|\sup_{g\in G}\left|\frac{1}{n}\sum_{j\neq i} g(z_j) + \frac{g(z_i)}{n} - \frac{g(z_i')}{n} + \frac{g(z_i')}{n} - Pg\right| - \sup_{g\in G}\left|\frac{1}{n}\sum_{j\neq i} g(z_j) + \frac{g(z_i')}{n} - Pg\right|\right| \\ &\leq \left|\sup_{g\in G}\left|\frac{1}{n}\sum_{j\neq i} g(z_j) + \frac{g(z_i)}{n} - Pg\right| - \sup_{g\in G}\left|\frac{1}{n}\sum_{j\neq i} g(z_j) + \frac{g(z_i')}{n} - Pg\right|\right| + \sup_{g\in G}\left|\frac{g(z_i)}{n} - \frac{g(z_i')}{n}\right| \\ &\leq \frac{2M}{n} \end{aligned}$$

同様にして, $f(z_1, \dots, z_i', \dots, z_n) - f(z_1, \dots, z_n) \leq \frac{2M}{n}$ も

絶対値は $\leq \frac{2M}{n}$ である

あとは McDiarmid の不等式と対称化より従う

Prop (Rademacher 複雑度の性質)

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

(2) $\phi_i : \mathbb{R} \to \mathbb{R}$ が $|\phi_i(x) - \phi_i(y)| \leq D|x-y|$ なら

$$E_{\sigma_i}\left[\sup_{g \in G} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i \phi_i(g(z_i))\right|\right] \leq 2D\hat{R}_n(G)$$

(contraction inequality)

($R_n(G)$ の定義で絶対値を使わない場合、左辺の $E[\cdots]$ の中の絶対値をはずすと、$2D$ を D にできる。)

(3) G の凸包を

$$\mathrm{conv}(G) = \left\{\sum_{i=1}^m \lambda_i g_i \;\middle|\; \sum_{i=1}^m \lambda_i = 1, \lambda_i \geq 0, g_i \in G, m = 1, 2, \cdots\right\}$$

とすると

$$R_n(G) = R_n(\mathrm{conv}(G))$$

(4) $R_n(c \cdot G) = |c| R_n(G) \quad (c \in \mathbb{R})$

Ex. (1) 線形モデル ← $x, \omega \in \mathbb{R}^d$ とする.

- $F = \{f(x) = x^T\omega \mid \|\omega\|_2 \leq 1\}$
- $E[\|x_i\|^2] \leq 1$

$$R_n(F) = E_{z_i, \sigma_i}\left[\sup_{\|\omega\| \leq 1} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i \omega^T x_i\right|\right]$$

$$\leq E_{z_i, \sigma_i}\left[\sup_{\|\omega\| \leq 1} \|\omega\| \cdot \left\|\frac{1}{n}\sum_{i=1}^n \sigma_i x_i\right\|\right]$$

$$\leq \sqrt{E_{z_i, \sigma_i}\left[\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \sigma_i \sigma_j x_i^T x_j\right]}$$

$$= \sqrt{\frac{1}{n^2}\sum_{i=1}^n E[\|x_i\|^2]} = \sqrt{\frac{1}{n}}$$

(2) $\mathcal{F}$ が有限集合のとき

$\mathcal{F} = \{f_1, \dots, f_N\}$, $\|f_i\|_\infty \le R$ なら

$$\hat{R}_n(\mathcal{F}) \le R\sqrt{\frac{2\log(2N)}{n}}$$

← R_n に $|\cdot|$ が入ってなければ $2N$ を N とできる.

(Massart の定理) ☆

← Hoeffding の不等式より示せる.

N はせいぜい $\log(N)$ でしか効かない

(Massart の定理の証明)

$t > 0$ に対し.

$$\exp\left(t\, E_\sigma\left[\max_k \left|\frac{1}{n}\sum_{i=1}^n \sigma_i f_k(z_i)\right|\right]\right) \le E_\sigma\left[\exp\left(t \max_k \left|\frac{1}{n}\sum_{i=1}^n \sigma_i f_k(z_i)\right|\right)\right]$$

$$\le E_\sigma\left[\sum_{1\le k\le N} \exp\left(t\frac{1}{n}\sum_{i=1}^n \sigma_i f_k(z_i)\right) + \sum_{1\le k\le N} \exp\left(-\frac{t}{n}\sum_{i=1}^n \sigma_i f_k(z_i)\right)\right]$$

$$= \sum_{1\le k\le N} 2\, E_\sigma\left[\exp\left(t\cdot\frac{1}{n}\sum_{i=1}^n \sigma_i f_k(z_i)\right)\right]$$

$$= \sum_{1\le k\le N} 2\, E_\sigma\left[\prod_{i=1}^n \exp\left(\frac{t}{n}\cdot\sigma_i f_k(z_i)\right)\right]$$

$$= 2\sum_{1\le k\le N} \prod_{i=1}^n E_{\sigma_i}\left[\exp\left(\frac{t}{n}\sigma_i f_k(z_i)\right)\right]$$

$$\le 2\sum_k \prod_{i=1}^n \exp\left(\frac{(\frac{t}{n})^2\cdot R^2}{2}\right) \quad (\because \text{Hoeffding の補題})$$

有界な r.v. は sub-Gaussian

$$\le 2N\cdot\exp\left(\frac{t^2R^2}{2n}\right)$$

(両辺 log とって, $\frac{1}{t}$ かけると)

$$\Rightarrow \hat{R}_n(\mathcal{F}) \le \frac{\log(2N)}{t} + \frac{tR^2}{2n} \quad (\forall t > 0)$$

$$\Rightarrow t = \sqrt{\frac{2n\log(2N)}{R^2}} \quad \because \text{右辺は最小.}$$

$$\Rightarrow \hat{R}_n(\mathcal{F}) \le R\cdot\sqrt{\frac{2\log(2N)}{n}}$$

※ R は $\hat{\sigma}^2 = \max_k \frac{1}{n}\sum_{i=1}^n f_k(z_i)^2$ におきかえても良い (証明が必要)

Contraction ineq. より. 損失関数がリプシッツ連続なら
汎化ギャップを $R_n(F)$ でおさえられる.

$|\ell(y,f) - \ell(y,f')| \leq L|f-f'|$ とする.

$$\begin{aligned}
\hat{L}(f) - L(f) &= (P_n - P)(\ell \circ f) \\
&\leq \sup_{g \in G} (P_n - P) g \\
&\leq 2R_n(G) + M\sqrt{\frac{2\log(\frac{1}{\delta})}{n}} \quad \text{with prob. } 1-\delta. \\
&\leq 4LR_n(F) + M\sqrt{\frac{2\log(\frac{1}{\delta})}{n}}
\end{aligned}$$

深層ニューラルネットワークの Rademacher 複雑度

ここでは やや特殊な条件のもと、深層 NN の Rademacher 複雑度を導出してみる。

$m_L = 1$ とする。

$$\mathcal{F} = \left\{ W^{(L)} \eta\left(W^{(L-1)} \eta\left(\cdots \eta\left(W^{(1)} x \right) \cdots \right) \cdots \right) \;\middle|\; W^{(\ell)} \in \mathbb{R}^{m_\ell \times m_{\ell-1}},\ \|W^{(\ell)}\|_{\infty,1} \le B_\ell \right\}$$

← このような条件強い

ただし、$\eta(x) = \max(x, 0)$ (要素ごと) (ReLU)

$$\|W^{(\ell)}\|_{\infty,1} := \max_{1 \le j \le m_\ell} \|W^{(\ell)}_{j,:}\|_1$$

とする。

$$\mathcal{F}_i = \left\{ f(x) = \sum_{j=1}^{m_{i-1}} w_j \eta(f_j(x)) \;\middle|\; f_j \in \mathcal{F}_{i-1},\ \|w\|_1 \le B_i \right\}$$

$$\mathcal{F}_1 = \left\{ f(x) = w^T x \;\middle|\; \|w\|_1 \le B_1 \right\}$$

とすれば、$\mathcal{F}_L = \mathcal{F}$ である。

Lem

$$\hat{\mathcal{R}}_n(\mathcal{F}_i) \le 2 B_i \hat{\mathcal{R}}_n(\mathcal{F}_{i-1})$$

Proof

$$\hat{\mathcal{R}}_n(\mathcal{F}_i) = \mathbb{E}_{\sigma_i}\left[\sup_{\substack{\|w\|_1 \le B_i \\ f_j \in \mathcal{F}_{i-1}}} \left| \frac{1}{n} \sum_{k=1}^{n} \sigma_k \sum_{j=1}^{m_i} w_j \eta(f_j(x_k)) \right| \right]$$

$$= \mathbb{E}_{\sigma_i}\left[\sup_{w, f_j} \left| \sum_{j=1}^{m_i} w_j \frac{1}{n} \sum_{k=1}^{n} \sigma_k \eta(f_j(x_k)) \right| \right]$$

$$\le \mathbb{E}_{\sigma_i}\left[\sup_{w, f_j} \underbrace{\|w\|_1}_{\le B_i} \max_{1 \le j \le m_i} \left| \frac{1}{n} \sum_{k=1}^{n} \sigma_k \eta(f_j(x_k)) \right| \right]$$

$$\le B_i \, \mathbb{E}_{\sigma_i}\left[\sup_{f \in \mathcal{F}_{i-1}} \left| \frac{1}{n} \sum_{k=1}^{n} \sigma_k \eta(f(x_k)) \right| \right]$$

← $\{\eta \circ f \mid f \in \mathcal{F}_{i-1}\}$ の Rademacher 複雑度

$\leq 2B_i R_n(\mathcal{F}_{i-1})$ (∵ contraction ineq.
η は 1-リプシッツ連続)

Lem

$\|x\|_\infty \leq 1$ (a.s.) なら

$$R_n(\mathcal{F}_1) \leq \sqrt{\frac{2\log(2d)}{n}}$$

(Proof)

$$\text{左辺} = E_\sigma\left[\sup_{\|w\|_1 \leq B_1} \|w\|_1 \max_{1\leq j\leq 1} \left|\frac{1}{n}\sum_{k=1}^n \sigma_k x_{k,j}\right|\right]$$

$$= B_1 E_\sigma\left[\max_{1\leq j\leq d} \left|\frac{1}{n}\sum_{k=1}^n \sigma_k x_{k,j}\right|\right]$$

$$\leq B_1 \sqrt{\frac{2\log(2d)}{n}} \quad (\because \text{Massartの定理})$$

Thm (DNNの Rademacher 複雑度)

$\|x\|_\infty \leq 1$ (a.s.) なら

$$R_n(\mathcal{F}) \leq 2^L \cdot \prod_{i=1}^L B_i \sqrt{\frac{2\log(2d)}{n}}$$

→ 各層のパラメータノルムが制御されれば次元は消える.
ネットワークのサイズが出てこない.

これ以外にも様々なバウンドが導出されている.
(2^L が出ないバウンドなど)

◦ Rademacher 複雑度の上界：カバリング・ナンバー，エントロピー・ナンバー

Def (Covering Number, 被覆数)

$\mathcal{F}$: 関数の集合

d: $\mathcal{F}$ の距離

$$N(\mathcal{F}, \varepsilon, d) := \min\{ n \geq 1 \mid \exists f_1, \dots, f_n \in \mathcal{F} \text{ s.t. } \forall f \in \mathcal{F} \text{ について} \exists f_j \text{ s.t. } d(f, f_j) \leq \varepsilon \}$$

Def (エントロピー数)

$$e_i(\mathcal{F}, d) := \inf\{ \varepsilon > 0 \mid N(\mathcal{F}, \varepsilon, d) \leq 2^{i-1} \}$$

データ点 $x_1, \dots, x_n \in \mathcal{X}$ に対して，

$$\|f\|_{n,p} := \left(\frac{1}{n} \sum_{i=1}^n |f(x_i)|^p \right)^{\frac{1}{p}} \quad (1 \leq p < \infty)$$

とする．

thm $\forall f \in \mathcal{F}$ が $\|f\|_\infty \leq R$ をみたすなら，ある $x_1, \dots, x_n \in \mathbb{R}^d$ が与えられたとき

$$\hat{R}_n(\mathcal{F}) \leq \inf_{\alpha \geq 0} \left\{ \alpha + R \sqrt{\frac{2 \log(2 N(\alpha, \mathcal{F}, \|\cdot\|_{n,1}))}{n}} \right\}$$

Proof $f \in \mathcal{F}$ に対して，$v[f]$ を f に一番近い ε-被覆の点とする．

$$\hat{R}_n(\mathcal{F}) = E_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i) \right| \middle| x_1, \dots, x_n \right]$$

以下略

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

$$\leq E_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \left(f(x_i) - v[f](x_i) \right) \right| \right] + E_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i v[f](x_i) \right| \right]$$

$$\leq \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n |f(x_i) - v[f](x_i)| + E_\sigma \left[\max_{1 \leq j \leq N(\mathcal{F}, \alpha, \|\cdot\|_{n,1})} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f_j(x_i) \right| \right]$$

ε-被覆の j 番目の元

$$\leq \alpha + R \sqrt{\frac{2 \log(2N)}{n}} \quad (\because \text{Massart の定理})$$

α ↑ 被覆の定義

・ $p \le q$ なら $\|\cdot\|_{n,p} \le \|\cdot\|_{n,q}$ であることから
$N(\varepsilon, \mathcal{F}, \|\cdot\|_{n,q}) \le N(\varepsilon, \mathcal{F}, \|\cdot\|_{n,p})$ であるので、左辺の被覆数は
$N(\varepsilon, \mathcal{F}, \|\cdot\|_{n,p})$ $(p>1)$ でおさえておく。 ($p=2$ は上く用いられる)

Thm (Dudley のチェイニング)
有る $x_1, \dots, x_n \in \mathbb{R}^d$ が与えられたとき、$\hat{\delta} := \sup_{f \in \mathcal{F}} \|f\|_{n,2} < \infty$ とおく

$$\hat{R}_n(\mathcal{F}) \le \inf_{\alpha > 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^{\hat{\delta}} \sqrt{\log 2N(\mathcal{F}, \varepsilon, \|\cdot\|_{n,2})}\, d\varepsilon \right\}$$

$$\hat{R}_n(\mathcal{F}) \le \frac{2}{\sqrt{n}} \left(\sum_{i=1}^{\infty} 2^{i/2} e_{2^i}(\mathcal{F} \cup \{0\}, \|\cdot\|_{n,2}) + \sup_{f \in \mathcal{F}} \|f\|_{n,2} \right)$$

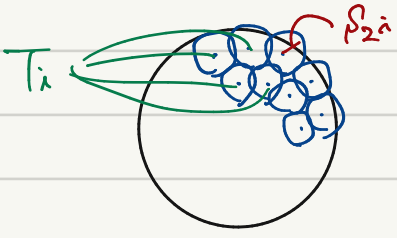
※ $\mathcal{F}$ を有限個の代表点で近似し、その近似精度を粗いものから密なものまで つなげてゆく という操作に対応.

Proof (エントロピー数の方だけ考えておく)

$$T = \{ (f(x_1), \dots, f(x_n)) \mid f \in \mathcal{F} \} \cup \{(0, \dots, 0)\} \subseteq \mathbb{R}^n$$
$$d(t, t') := \sqrt{\frac{1}{n} \sum_{i=1}^n (t_i - t_i')^2}$$ とおくと、(T, d) は可分な距離空間.

以後、$\forall i \ge 1$ $e_i(T, d)$ は有限であるとして一般性を失わない.
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
$$\delta_1 := \sup_{t \in T} d(t, 0)$$
$$\delta_{2^i} := (1+\varepsilon)\, e_{2^i}(T, d) \qquad (i = 1, 2, \dots)$$
とおく. また、$\forall i \ge 1$ に対し、δ_{2^i}-ネット T_i を $|T_i| \le 2^{2^i - 1}$ なるものとして取ってくる
($\bigcup_{t \in T_i} B_d(t, \delta_{2^i}) \supseteq T$ なるもの)
(エントロピー数の定義より存在)

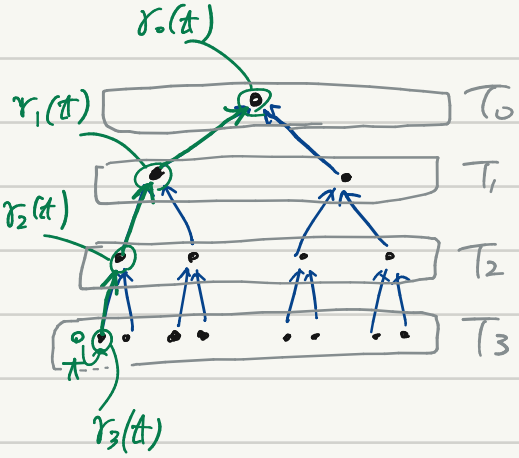


$T_0=\{0\}$ とする．すると，$e_j(T,d)\to 0$ $(j\to\infty)$ より，$U:=\bigcup_{j=0}^{\infty} T_j$ に対し，

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

が成り立つ．（U は T 内で稠密）よって，T の代わりに U を考えれば良い．

$j\geq 1$，$t\in T$ に対し，$\pi_j(t)\in T_j$ を，$d(t,\pi_j(t))\leq 2^{-j}$ なるものとして取ってまた特に，$t\in T_j$ なら，$\pi_j(t)=t$ としておく．



$j\geq 1$ を1つ固定して，γ_i $(i\leq j)$ を

$$\gamma_j(t)=\pi_j(t)$$
$$\gamma_{i-1}(t)=\pi_{i-1}\circ\gamma_i(t)\quad(i=1,\dots,j)$$

とする．

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

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

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

となる．よって，$U_k:=T_0\cup\cdots\cup T_k$ に対しても

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

が成り立つ．すると，有限集合の Rademacher 複雑度より

$$E_{\sigma_i}\left[\max_{t\in U_k}|h(t)|\right]\leq\sum_{j=1}^{k}E_{\sigma_i}\left[\max_{t\in T_j}|h(t)-h(\pi_{j-1}(t))|\right]$$
$$\leq\sum_{j=1}^{k}\sqrt{\frac{2\log(2\cdot 2^{2j}-1)}{n}}\cdot 2^{-j+1}$$

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

$$E_{\sigma_i}\left[\sup_{t\in U}h(t)\right]\leq\sum_{j=0}^{\infty}\sqrt{\frac{4\cdot 2^{j}}{n}}\,2^{-j}.$$

あとは $S_1=\sup_t d(t,0)$ に注意すれば…

◎深層NNのカバリングナンバー

スパースなNN

$$\mathcal{F}'_{L,S,B} := \Big\{ (W^{(L)}\cdot + b^{(L)}) \circ \eta(W^{(L-1)}\cdot + b^{(L-1)}) \circ \cdots \circ \eta(W^{(1)}x) \Big|$$

$$\sum_{\ell=1}^{L} \|W^{(\ell)}\|_0 + \sum_{\ell=1}^{L} \|b^{(\ell)}\|_0 \le S,$$

$$\max_{\ell} \max\{\|W^{(\ell)}\|_\infty, \|b^{(\ell)}\|_\infty\} \le B,$$

$$W^{(\ell)} \in \mathbb{R}^{m_\ell \times m_{\ell-1}},\ b_\ell \in \mathbb{R}^{m_\ell} \Big\}$$

(ただし $\|W\|_0 = W$ の非ゼロ要素の数, $\|W\|_\infty = \max_{i,j} |W_{ij}|$)

$$\mathcal{F}_{L,S,B} := \{ f(x) = \underline{\max\{-R, \min\{R, f(x)\}\}} \mid f \in \mathcal{F}'_{L,S,B} \}$$

↑ clipping $\pm R$ の間におさめる.

Prop (Schmidt-Hieber, 2017)

$\max_\ell m_\ell \le W$ とおくと

$$\log N(\varepsilon, \mathcal{F}_{L,S,B}, \|\cdot\|_\infty) \le 2SL \log(L(B\vee 1)W\cdot \varepsilon^{-1})$$ //

$$\Rightarrow \mathcal{R}_n(\mathcal{F}_{L,S,B}) \le 4R\sqrt{\frac{SL\log(L(B\vee 1)W\cdot n^{-1})}{n}}$$

$L, W, B = O(n)$ なら $\longrightarrow O\left(\sqrt{\frac{SL\log(n)}{n}}\right)$

(パラメータ数2乗スパースなら消化する)

② Fast learning rate と局所 Rademacher 複雑度

- 推定量 $\hat{f}$ は f^* の"近く"にいるはず → より"局所的"な上界を用いるべき

→ 結果的に 前より速い収束率が示せる.

仮定

- L は強凸：ある $\alpha > 0$ が存在して、$\forall f \in \mathcal{F}$ に対し
$$\frac{\alpha}{2}\|f - f^*\|^2_{L^2(P_X)} \leq L(f) - L(f^*)$$
- l は 1-リプシッツ連続，$\|l \circ f\|_\infty \leq M$ $(\forall f \in \mathcal{F})$

例　$l(y, f) = (y - f)^2$ で
$\|f\|_\infty \leq \frac{1}{2}$ $(\forall f \in \mathcal{F})$, $\|f^*\|_\infty \leq \frac{1}{2}$, $|y - f(x)| \leq \frac{1}{2}$ (a.s.)
なら成立.

← (定数は重要ではない. 適当にスケールすれば一般の定数に対し同様のオラクル不等式)

$$\mathcal{G}_r := \{g = l \circ f - l \circ f^* \mid Pg = L(f) - L(f^*) \leq r,\ f \in \mathcal{F}\}$$

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

$R_n(\mathcal{G}_r)$ を局所 Rademacher 複雑度 と言う.

※ リプシッツ連続性と強凸性より
$$\frac{\alpha}{2}\|g\|^2_{L^2(P_X)} \leq \frac{\alpha}{2}\|f - f^*\|^2_{L^2(P_X)} \leq Pg$$

なので. $Pg \leq r$ なら
$$\begin{cases} \|g\|_{L^2(P_X)} \leq \sqrt{\dfrac{2r}{\alpha}} \\ \|f - f^*\|^2_{L^2(P_X)} \leq \dfrac{2r}{\alpha} \end{cases}$$
← Excess risk が小さければ f は f^* に近い

thm (Peeling)

ある $\phi : [0,\infty) \to [0,\infty)$ なる関数が存在し、ある $r^* > 0$ に対し
$\forall r > r^*$ において

$$\phi(4r) \leq 2\phi(r)$$
$$R_n(G_r) \leq \phi(r)$$

が成り立つならば、$\forall r > r^*$ に対して

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

//

(Proof)

$$\text{左辺の } E[\cdot] \text{ の中身} \leq \sup_{g \in G_r} \frac{\frac{1}{n}\sum_{i=1}^n \sigma_i g(z_i)}{r} + \sum_{j=0}^{\infty} \sup_{g \in G_{r4^{j+1}} \setminus G_{r4^j}} \frac{\frac{1}{n}\sum_{i=1}^n \sigma_i g(z_i)}{r4^j + r}$$

両辺期待値を取ると、

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

//

★ Thm (Talagrandの集中不等式)

$(Z, \mathcal{A}, P)$：確率空間

$\mathcal{G}$：$(Z, \mathcal{A})$上の可測関数の集合，

$$E[g]=0,\ E[g^2]\le v,\ \|g\|_\infty \le B \quad (\forall g\in\mathcal{G})$$

かつ $\mathcal{G}$は$\|\cdot\|_\infty$に関して可分．

すると，$\forall t>0$ に対して，

$$P_Z\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n g(Z_i) \ge 2E_Z\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n g(Z_i)\right] + \sqrt{\frac{2tv}{n}} + \frac{2tB}{n}\right] \le e^{-t}$$

この$\mathcal{G}$について，$\tilde{\mathcal{G}}=\left\{\tilde{g}=\frac{Pg-g(z)}{Pg+r} \;\middle|\; g\in\mathcal{G}\right\}$ で考えると，

$$\|\tilde{g}\|_\infty = \left\|\frac{Pg-g}{Pg+r}\right\|_\infty \le \frac{M}{r} = B$$

$$\|\tilde{g}\|_{L^2}^2 = \frac{\|Pg-g\|_{L^2}^2}{(Pg+r)^2} \le \frac{Pg^2}{2r\,\frac{\alpha}{2}P(g^2)} \le \frac{1}{\alpha r} = v$$

（$\because \frac{\alpha}{2}P(g^2)\le Pg$）

および：Peelingの議論より

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

（$\phi(r)$は $R_n(\mathcal{G}_r)\le\phi(r)$ となるもの）

なので，

$\hat{f}$ をおく．

$$(P-P_n)\left(\frac{l_0\hat{f}-l_0f^0}{P\hat{g}+r}\right) \le \frac{16\phi(r)}{r} + \sqrt{\frac{2t}{\alpha r n}} + \frac{2tM}{rn} =: \gamma_n(r) \quad \text{w.p. } 1-e^{-t}$$

Ratio type Empirical process

$\hat{f} := \underset{f \in \mathcal{F}}{\operatorname{argmin}} \hat{L}(f)$ とおく.

このとき. $L(\hat{f}) - L(f^\circ) = L(\hat{f}) - \hat{L}(\hat{f}) + \underbrace{\hat{L}(\hat{f}) - \hat{L}(f^*)}_{\leq 0 \ (\because \hat{f}\text{のとり方})} + \hat{L}(f^*) - L(f^*)$

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

$= (P - P_n) g + P_n (\ell \circ f^* - \ell \circ f^\circ)$

$= (P - P_n) g + (P_n - P)(\ell \circ f^* - \ell \circ f^\circ) + P(\ell \circ f^* - \ell \circ f^\circ)$

$\leqq (P\hat{g} + r)\, \psi_n(r) + (P_n - P)(\ell \circ f^* - \ell \circ f^\circ) + L(f^*) - L(f^\circ)$

- $(P_n - P)(\ell \circ f^* - \ell \circ f^\circ)$ について.

  $r^* = L(f^*) - L(f^\circ)$ とすると. $\frac{\alpha}{2} P g^2 \leqq r^*$ $(g = \ell \circ f^* - \ell \circ f^\circ)$ と.

  Bernsteinの不等式より

  $(P_n - P) g = O_p\left(\sqrt{\frac{r^*}{n}} + \frac{1}{n}\right) \leqq O_p\left(\frac{1}{n}\right) + r^*$

  が成り立つ.

- $(P\hat{g} + r)\, \psi_n(r)$ について.

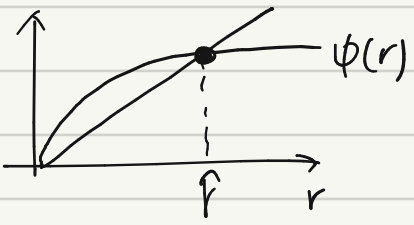
  $\psi_n(\hat{r}) \leqq \frac{1}{2}$ となるような $\hat{r}$ を選んでおく

$$P(\hat{g}) = L(\hat{f}) - L(f^\circ) \leqq 2\hat{r}\, \psi_n(\hat{r}) + 2r^* + O_p\left(\frac{1}{n}\right)$$

$$\leqq \hat{r} + 2r^* + O_p\left(\frac{1}{n}\right)$$

を得る. $\psi_n(\hat{r}) \leqq \frac{1}{2}$ は $\max\left\{\phi(\hat{r}), \frac{4\kappa}{\alpha n}, \frac{4\kappa}{n}, \hat{r}^*\right\} \leqq \hat{r}$ なら良い.

$\Longrightarrow$ $L(\hat{f}) - L(f^\circ) \leqq \hat{r} + 2r^* + O_p\left(\frac{1}{n}\right)$ w.p. $1 - e^{-t}$



ϕ は大体凹関数になる.

$\phi(r) \leqq r$ となる点を求めればよい

(fixed point)

例:

$\log N(\mathcal{F}, \varepsilon, \|\cdot\|_{n,2}) \lesssim \varepsilon^{-2\rho}$ (a.s.) のとき.

$$R_n(G_r) \lesssim \left(\frac{r^{\frac{1-\rho}{2}}}{\sqrt{n}}, n^{-\frac{1}{4+\rho}} \right)$$

$\Longrightarrow$ $R_n(G_r) \cong r$ iff $\hat{r} = n^{-\frac{1}{4+\rho}}$ で達成.

よって. $f^* = f^\circ$ のとき.

$$L(\hat{f}) - L(f^\circ) \lesssim n^{-\frac{1}{4+\rho}}$$

//

○ 深層NNの局所 Rademacher 複雑度

先の 深層 NN の カバリングナンバー Bound

$$\log N(\varepsilon, \mathcal{F}_{L,S,B}, \|\cdot\|_\infty) \leq 2SL \log\left(L(B\vee 1)\cdot W\cdot \varepsilon^{-1}\right)$$

より. $\forall r>r^*$ において. $\phi(r) \leq O\left(\sqrt{\frac{SLr}{n}\log\left(L(B\vee 1)W\cdot n\right)}\right)$

(略証)

$$E[\hat{R}_n(G_r)] \leq E_\xi[\hat{R}_n(\mathcal{F}_r)] \qquad (\text{ただし. } \mathcal{F}_r := \{f-f^\circ \mid L(f)-L(f^\circ)\leq r\})$$

↑ contraction ineq.

$$\leq E_\xi\left[\inf_{\alpha>0}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^{\hat{\delta}}\sqrt{\log 2N(\varepsilon,\mathcal{F}_r,\|\cdot\|_{n,2})}\,d\varepsilon\right\}\right]$$

ただし. $\hat{\delta} = \sup_{f\in\mathcal{F}_r}\|f\|_{n,2} \leq \sqrt{\frac{2r}{\alpha}}$

↑ ここはデータに依存するので.
この $\hat{\delta}$ のバウンドを本当は
局所 Rademacher を用いてバウンドする
必要がある.

$$\leq \frac{1}{n} + \frac{1}{\sqrt{n}}\int_{\frac{1}{n}}^{c\sqrt{\frac{r}{\alpha}}} \sqrt{2SL\log\left(L(B\vee 1)W\varepsilon^{-1}\right)+\log(2)}\,d\varepsilon$$

(定数 $c\sqrt{\frac{r}{\alpha}}$)

$$\lesssim \sqrt{\frac{SLr}{n}\log\left(L(B\vee 1)Wn\right)} \qquad //$$

ノンパラ回帰への応用:

$$y_i = f^\circ(x_i) + \xi_i$$

ξ_i は i.i.d. 雑音, 平均0, $|\xi_i| \leq C$ (a.s.)

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

↓ 次ページへ

Lem (Yarotsky, 2016; Schmidt-Hieber, 2017)

$f^\circ \in C^s([0,1]^d)$, $\|f^\circ\|_{C^s} \le 1$ のとき.

$N \gg 1$ なる整数に対し.
$\exists L = O(\log(N))$, $\exists S = O(N\log(N))$,
$\exists W = O(N)$, $\exists B = O(1)$ で、

$$\inf_{f \in \mathcal{F}_{L,S,B}} \|f^\circ - f\|_\infty \le O(N^{-\frac{s}{d}})$$

$C^s([0,1]^d)$：ヘルダークラス

$= \{f \mid \sum_{|\alpha| \le m} \|D^\alpha f\|_\infty + \sum_{|\alpha|=m} \sup \frac{\|D^\alpha f(x) - D^\alpha f(y)\|_\infty}{\|x-y\|_\infty^{s-m}} < \infty\}$

$(m = \lfloor s \rfloor)$

よって. $r^* = O(N^{-\frac{2s}{d}})$ ← バイアス

$\hat{r} = O\left(\frac{N \log(N)^2 (\log(N) + \log(n))}{n}\right)$ ← バリアンス

トレードオフ

となることが示せる.

$$\|\hat{f} - f^\circ\|^2_{L^2(P_X)} = O_p(\hat{r} + r^*)$$

である. $N \asymp n^{\frac{d}{2s+d}}$ とおけば、

$$\|\hat{f} - f^\circ\|^2_{L^2(P_X)} = O_p\left(n^{-\frac{2s}{2s+d}} \log(n)^3\right)$$

(注) 最近では非スパースなモデルで最適レートを達成する解析も現れている.

(注) 観測ノイズ ε_i が ガウス分布の場合, 有界性 $|\varepsilon_i| \le c$ は成り立たない. この場合は Talagrand の集中不等式 は使えない. 代わりに ガウス型集中不等式 を用いれば良い. (この授業では)

$(\xi_i)_{i=1}^n$: i.i.d. $N(0,\sigma^2)$, $(x_i)_{i=1}^n \subset \mathcal{X}$: 固定, $\mathcal{F}$: $\|\cdot\|_\infty$ について可分

$$P\left(\sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \xi_i f(x_i)\right| \ge E\left[\sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \xi_i f(x_i)\right|\right] + \sqrt{\frac{\sigma^2 \|\mathcal{F}\|_n^2 t}{n}}\right)$$
$$\le \exp\left(-\frac{t^2}{2}\right) \quad (\forall t > 0)$$

ただし. $\|\mathcal{F}\|_n^2 := \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(x_i)^2$.