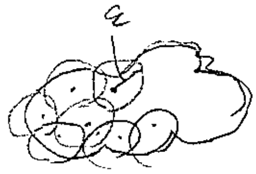
Rademacher複雑度の上界



Def (Covering number, 被覆数)

$\mathcal{F}$: $\mathbb{R}^d \to \mathbb{R}$ の関数の集合

d: $\mathcal{F}$ の距離

$\mathcal{F}$ の ε-被覆数: $N(\varepsilon, \mathcal{F}, d) := \min \{ n \geq 1 \mid \underbrace{\exists f_1, \ldots, f_n \in \mathcal{F}}_{\mathcal{F}\text{ の「}\varepsilon\text{-被覆」と言う}}$ s.t. $\forall f \in \mathcal{F}$ に対し $\exists f_j$ で $d(f, f_j) \leq \varepsilon \}$ //

データ点 $x_1, \ldots, x_n \in \mathbb{R}^d$ に対し、

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

とする。

Thm $\forall f \in \mathcal{F}$ が $\|f\|_\infty \leq R$ をみたすなら、ある $x_1, \ldots, x_n \in \mathbb{R}^d$ が与えられたもとで

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

Proof $f \in \mathcal{F}$ に対し $v[f]$ を f に一番近い ε-被覆の点とする ($\|f - v[f]\|_{n,1} \leq \alpha$)

$$\begin{aligned}
\hat{R}_n(\mathcal{F}) &= E_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i) \,\middle|\, x_1, \ldots, x_n \right] \\
&= E_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i (f(x_i) - v[f](x_i)) + \frac{1}{n} \sum_{i=1}^n \sigma_i v[f](x_i) \right] \\
&\leq E_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i (f(x_i) - v[f](x_i)) \right] + E_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i v[f](x_i) \right] \\
&\leq \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n |f(x_i) - v[f](x_i)| + E_\sigma \left[\max_{1 \leq j \leq N(\alpha, \mathcal{F}, \|\cdot\|_{n,1})} \frac{1}{n} \sum_{i=1}^n \sigma_i f_j(x_i) \right] \\
&\leq \alpha + R\sqrt{\frac{2\log N}{n}}
\end{aligned}$$

(∵ Massart の ~~定理~~ 不等式) //

- $p \leq q$ なら $N(\alpha, \mathcal{F}, \|\cdot\|_p) \leq N(\alpha, \mathcal{F}, \|\cdot\|_q)$ であることから、右辺の被覆数は $N(\alpha, \mathcal{F}, \|\cdot\|_p)$ $(p > 1)$ におきかえてもよい

Thm (Dudley のチェイニング)

ある $x_1, \ldots, x_n \in \mathbb{R}^d$ が与えられたもと。$\delta := \sup_{f \in \mathcal{F}} \|f\|_{n,2}$ とすると

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

が成立つ。 //

- $\mathcal{F}$ を有限個で近似し、近似精度を粗いものから密なものへつなげてゆく(チェインしてゆく)という操作に対応。
- 先のバウンドよりタイトになることが多い。

(証明は略)

ー パラメータサイズに基づくバウンド

スパースなNN.

$$\mathcal{F}'_{L,S,B} := \Big\{ (W^{(L)}\cdot + b^{(L)}) \circ \eta \circ (W^{(L-1)}\cdot + b^{(L-1)}) \circ \eta \circ \cdots \circ \eta \circ W^{(1)}x \;\Big|\; \sum_{\ell=1}^{L} \|W^{(\ell)}\|_0 + \sum_{\ell=1}^{L}\|b^{(\ell)}\|_0 \le S,\ \max_\ell \max\{\|W^{(\ell)}\|_\infty, \|b^{(\ell)}\|_\infty\} \le B,\ W^{(\ell)} \in \mathbb{R}^{m_\ell \times m_{\ell-1}},\ b_\ell \in \mathbb{R}^{m_\ell} \Big\}$$

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

ηはReLUとする.

$$\mathcal{F}_{L,S,B} = \Big\{ f(x) = \underbrace{\max\{-1, \min\{1, f(x)\}\}}_{\text{clipping}} \;\Big|\; f \in \mathcal{F}'_{L,S,B} \Big\}$$

clipping ±1の間におさめる.

<u>Lem</u> (Schmidt-Hieber, 2017)
$\max_\ell m_\ell \le D$ とすると.

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

・VC次元は $O(SL\log(S))$: Harvey, Liaw, Mehrabian, 2017.

これより.

$$\hat{R}_n(\mathcal{F}'_{L,S,B}) \le \inf_{\delta>0} \left\{ \delta + R\sqrt{\frac{2\log N(\delta, \mathcal{F}_{L,S,B}, \|\cdot\|_\infty)}{n}} \right\} \quad (\text{ただし. } R := \sup_{f\in\mathcal{F}} \|f\|_\infty)$$

$$\le 4R\sqrt{\frac{SL\log(L(B\vee 1)\cdot D \cdot n)}{n}} \quad (\delta = \tfrac{1}{n} \text{を代入})$$

※ Rのbound : $\forall \ell, j$ で $\|W^{(\ell)}_{j:}\|_1 \le B_1$; $\|x\|_\infty \le 1$ という条件を追加すると,

$R \le B_1^L$ を得る.

$\boxed{B_1^L \text{はLについて指数的に増大}}$

一方で、 $\hat{R}_n(\mathcal{F}_{L,S,B}) \le 4\sqrt{\dfrac{SL\log(L(B\vee 1)Dn)}{n}}$ ← clipすることで複雑さが大きく抑えられる.

↓

B_1^Lが出て来ない

- Batch-Normalization
- ResNet

の有効性

- Fast learning rate と局所 Rademacher 複雑さ

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

- 推定量 $\hat f$ は f^* の"近く"にいるはずなので、より"局所的"な上界を用いるべき
- モデルバイアスも考慮に入れた、バイアス-バリアンスのトレードオフを考えるべき

仮定

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

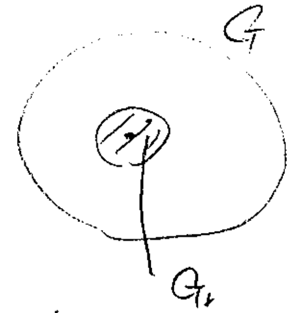
$$\frac{\alpha}{2}\|f-f^*\|^2_{L^2(P_X)} \le L(f)-L(f^*)$$

$\ell(y,f)=(y-f)^2$ で $\|f\|_\infty\le 1$, $\|f^*\|_\infty\le 1$ ($\forall f\in\mathcal{F}$) なら OK

- これまでの仮定がみたされている (ℓ は1-リプシッツ連続, $\|g\|_\infty\le M$)

$$G_r := \{ g=\ell\circ f-\ell\circ f^* \in G \mid Pg = L(f)-L(f^*) \le r,\ f\in\mathcal{F} \}$$ として、

$R_n(G_r)$ を局所Rademacher複雑さと言う



※ リプシッツ連続性と強凸性より

$$\frac{\alpha}{2}\|g\|^2_{L^2(P_Z)} \le \frac{\alpha}{2}\|f-f^*\|^2_{L^2(P_X)} \le Pg$$

なので、$Pg\le r$ なら $\|g\|_{L_2(P_Z)} \le \sqrt{\frac{2r}{\alpha}}$

$\|f-f^*\|^2_{L^2(P_X)} \le \frac{2r}{\alpha}$ ← 汎化誤差が小さければ最適解に"近い"

Thm (Peeling)

ある $\phi:[0,\infty)\to[0,\infty)$ なる関数が、ある $r^*\ge 0$ が存在し $\forall r>r^*$ で

$\phi(4r)\le 2\phi(r)$ かつ　　(r^* は モデルバイアスを表す)

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

ならば、$\forall r>r^*$ に対し

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

//

Excess risk $\le\varepsilon$ $\Rightarrow$ $\|\hat f-f^*\|^2 \le \varepsilon+\delta$

$\|f-f^\circ\|^2-\|f^*-f^\circ\|^2$

$= \|f-f^*+f^*-f^\circ\|^2 - \|f^*-f^\circ\|^2$

$= \|f-f^*\|^2 + 2\langle f-f^*, f^*-f^\circ\rangle$

$\ge \|f-f^*\|^2\cdot\frac{1}{2} - 2\|f^*-f^\circ\|^2$

Prf

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

両辺期待値とると、

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

$$\le \frac{\phi(r)}{r} + \frac{1}{r}\sum_{j=0}^{\infty}\frac{\phi(4^{j+1}r)}{4^j+1} \le \frac{\phi(r)}{r} + \frac{1}{r}\sum_{j=0}^{\infty}\frac{2^{j+1}\phi(r)}{4^j+1} \le \frac{4\phi(r)}{r}$$

//

Thm (Talagrandの集中不等式)

$(Z, \mathcal{A}, P)$: 測度空間

$\mathcal{G}$: $(Z,\mathcal{A})$上の可測関数の集合, $\forall g \in \mathcal{G}$, $\|g\|_\infty \leq B$, $E[g]=0$, $E[g^2] \leq \sigma^2$

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

すると $\forall t>0$ に対し

$$P_z\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n g(z_i) \geq 2 E_z\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n g(z_i)\right] + \sqrt{\frac{2t\sigma^2}{n}} + \frac{2tB}{n} \right]$$
$$\leq e^{-t}$$

//

この$\mathcal{G}$として, $\tilde{\mathcal{G}} = \left\{ \tilde{g} = \dfrac{Pg - g(z)}{Pg + r} \,\middle|\, g\in\mathcal{G} \right\}$ を用いると. ($r \geq r^*$ とする)

$$\|\tilde{g}\|_\infty = \left\|\frac{Pg-g}{Pg+r}\right\|_\infty \leq \frac{M}{r}, \quad \|\tilde{g}\|_{L_2}^2 = \frac{\|Pg-g\|_{L_2}^2}{(Pg+r)^2} \leq \frac{P(g^2)}{2r\cdot\frac{\alpha}{2}P(g^2)} = \frac{1}{\alpha r}$$

つまり $B = \dfrac{M}{r}$, $\sigma^2 = \dfrac{1}{\alpha r}$ とおける.

$\because \left(\frac{\alpha}{2}P(g^2) + r\right)^2 \geq$
$= r^2 + \alpha P(g^2) r + \cdots$

さらに、Peelingの議論より.

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

($\because$ ℓのリプシッツ性の縮約性による)

である. 以上より

$$P(\hat{g}) = L(\hat{f}) - L(f^\circ) \leq (P-P_n)(\underbrace{\ell\circ\hat{f} - \ell\circ f^\circ}_{\hat{g}\text{とおく}}) + \overbrace{\left|(P-P_n)(\ell\circ f^* - \ell\circ f^\circ)\right|}^{O_p\left(\sqrt{\frac{L(f^*)-L(f^\circ)}{n}}\right)}$$

$O_p\left(\sqrt{\dfrac{L(f^*)-L(f^\circ)}{n}}\right)$ と勝手せよ.

$\leq \sqrt{\dfrac{r^*}{n}} \leq r^* + \dfrac{1}{n}$

$$\leq (P\hat{g}+r)\underbrace{\left[\frac{16\phi(r)}{r} + \sqrt{\frac{2t}{\alpha r n}} + \frac{2tM}{rn}\right]}_{\psi_n(r)} + O\left(r^* + \frac{1}{n}\right)$$

w.p. $1-e^{-t}$

典型的な例に
$\phi(r) \lesssim \left(\dfrac{r^{\frac{1-p}{2}}}{\sqrt{n}}, n^{-\frac{1}{1+p}}\right)$ $r^*=0$

$\Longrightarrow \hat{r}^* = O\left(n^{-\frac{1}{1+p}}\right)$

例: ソボレフ空間 $W_2^m(\mathbb{R}^d)$, $p = \frac{d}{2m}$, $f^\circ \in W_2^m(\Omega)$

$n^{-\frac{2m}{2m+d}}$ (カーネル法)

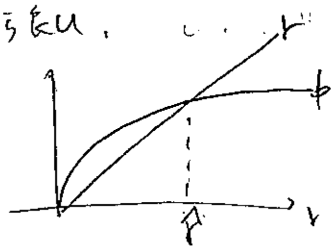
よって $\psi_n(r) \leq \frac{1}{2}$ なる任意の r に対し.

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

を得る. $\psi_n(r) \leq \frac{1}{2}$は

$$\max\left\{ \phi(r), \frac{t}{\alpha n}, \frac{Mt}{n}, r^* \right\} \lesssim r$$

ならよい.



φは凹関数であることが多い.

$\phi(r) = r$ なる $\hat{r}$ を求めればよい

(Fixed Point)

― 深層NNの局所Rademacher複雑さ.

先ほど定義したスパースなNNモデル $\mathcal{F}_{L,S,B}$ を考える.

Thm (informal)

$\max_{\ell} m_\ell \leq D$ とする.

$\mathcal{F} = \mathcal{F}_{L,S,B}$ に対するモデルバイアスを $r^* = \min_{f \in \mathcal{F}} L(f) - L(f^\circ)$ とする.

$^\forall r > r^*$ において

$$\underset{R_n(\mathcal{G}_r)}{\underbrace{\phi(r)}} \leq O\left(\sqrt{\frac{SLr}{n}\log(L(B \vee 1)\cdot D \cdot n)}\right)$$

(略証) $E_z[R_n(\mathcal{G}_r)] \leq E_z[R_n(\mathcal{F}_r)]$ （ただし $\mathcal{F}_r := \{f - f^\circ \mid L(f) - L(f^\circ) \leq r\}$）

contraction ineq.

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

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

(要) ここは本当はちょっと気を付ける必要あり $\|f\|_{n,2} \leq \|f\|_{L^2(P_X)}$

$$\lesssim \inf_{\alpha > 0}\left\{\alpha + \frac{1}{\sqrt{n}}\int_\alpha^{\sqrt{\frac{2r}{\alpha}}} \sqrt{2SL\log(L(B \vee 1)D\varepsilon^{-1})}\, d\varepsilon\right\}$$

$$\lesssim \sqrt{\frac{SLr}{n}\log(L(B \vee 1)\cdot D \cdot n)} \quad \left(\alpha = \frac{1}{n} を代入\right)$$

//

ノンパラメトリック回帰

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

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

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

とする. $\ell(y, f) = (y - f)^2$ とする.

※ $L(f) - L(f^\circ) = E[(Y - f(X))^2] - E[(Y - f^\circ(X))^2]$

$$= E[(f^\circ(X) + \varepsilon - f(X))^2] - E[\varepsilon^2]$$

$$= E[(f^\circ(X) - f(X))^2] = \|f^\circ - f\|_{L^2(P_X)}^2$$ に注意.

※ ℓ はリプシッツ連続ではないが $\|f\|_\infty, \|f^\circ\|_\infty \leq M$ と $|\varepsilon| \leq C$ なら $\|f\|_\infty \leq M$ の範囲でリプシッツ連続

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

$m = \lfloor \beta \rfloor$ = 最大の integer s.t. $m \leq \beta$

$C^\beta([0,1]^d) = \left\{ f \,\middle|\, \sum_{\alpha: |\alpha| < \beta} \|D^\alpha f\|_\infty + \sum_{|\alpha| = \lfloor \beta \rfloor} \sup \frac{|D^\alpha f(x) - D^\alpha f(y)|}{|x - y|_\infty^{\beta - \lfloor \beta \rfloor}} < \infty \right\}$

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

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

$$\inf_{f \in \mathcal{F}_{L,S,B}} \|f^\circ - f\|_{L^2(P_X)}^2 \leq O(N^{-\frac{2\beta}{d}})$$

深さが $O(\log(N))$ ないとこれは達成できない.

//

$N \gg 2$ に対し、L, S, B, D を次の Thm のように設定する

$\inf_{f \in \mathcal{F}} L(f) - L(f^\circ) \leq \underbrace{O(N^{-\frac{s}{d}})}_{= r^*}$ である

$\psi(r, \frac{1}{n}, r^*) \leq r$ を r について最小にする. $\hat{r}$ とおく.

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

よって、$N \asymp n^{-\frac{2s+d}{d}}$ とおけば (CVでせんたく)

$$\leq O\left(n^{-\frac{2s}{2s+d}} \log(n)^3 \right)$$

- ニューラルネットワークの最適化について.

- Over-parametrization で大域的最適解が見つかりやすくなる.

[Du et al. 2018; Allen-Zhu, Li & Song 2018; Li & Liang, 2018; Du et al. 2018]

Du, Zhai, Póczos, and Singh (2018) をベースに紹介する.

その他は講義資料pdfを参照

$$f(W,a,x) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \sigma(w_r^T x) \quad (a_r \in \mathbb{R}, w_r \in \mathbb{R}^d)$$

(mは十分大) → overparametrize.

基本的に.
$$\boxed{F(W) - F(W^*) \le \|\nabla F(W)\|^2}$$
を示す.

$$L(W) = \frac{1}{2}\sum_{i=1}^{n} (y_i - f(W,a,x_i))^2$$

$$W_{k+1} = W_k - \eta \left.\frac{\partial L(W)}{\partial W}\right|_{W=W_k} \quad \text{(Gradient Descent)}$$

初期化: $w_r \sim N(0, I)$, $a_r \sim p(a_r = 1) = p(a_r = -1) = \frac{1}{2}$ (i.i.d.)

Wのみを更新, aは止める.

- 時間を連続化して考える

$$\frac{dW(t)}{dt} = -\frac{\partial L(W(t))}{\partial W(t)}$$

$u(t) = (f(W(t), a, x_i))_{i=1}^{n} \in \mathbb{R}^n$ の動きを見る.

$$\frac{d}{dt}u_i(t) = \sum_{j=1}^{n} (y_j - u_j(t)) \cdot H_{ij}(t)$$

ただし. $H_{ij}(t) = \frac{1}{m} x_i^T x_j \sum_{r=1}^{m} \mathbb{1}\{x_i^T w_r(t) \ge 0, x_j^T w_r(t) \ge 0\}$

$$\frac{d}{dt}\|Y - u(t)\|^2 = -2(Y - u(t))^T H(t) (Y - u(t))$$

Fact

- 実は mを十分大きくとると. $H(0) \succeq \lambda_0 I$ $(\exists \lambda_0 > 0)$ とできる. (初期解)

 [$(x_i)_{i=1}^{n}$ が degenerate してなければ]

- $\|w_r(t) - w_r(0)\|_2 \lesssim \frac{\sqrt{n}\|Y - u(0)\|_2}{\sqrt{m}\lambda_0}$ ($w_r(t)$はあまり動かない)

- $H(t) \succeq \frac{\lambda_0}{2}$ (mが十分大なら)

$$\Rightarrow \|Y - u(t)\|^2 \le \exp(-\lambda_0 t) \|Y - u(0)\|^2$$

exponential convergence