第3回 Math-iine learning Learning Functions: When Is Deep Better Than Shallow

1/28

Contents

1 Introduction

この論文では,one-hidden layer のニューラルネットワークと deep network を比較する.
この論文で定理とされているものを記載する.

Theorem 1.1. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be infinitely differentiable, and not a polynomial on any subinterval of $\mathbb{R}$.*

- *For $f \in W^{NN}_{r,d}$*

$$\mathrm{dist}(f, S_n) = O(n^{-r/d})$$

- *For $f \in W^{NN}_{H,r,2}$*

$$\mathrm{dist}(f, D_n) = O(n^{-2/d})$$

Theorem 1.2. *There exists a constant $c > 0$ depending on d alone with the following property. Let $\{C_m\}$ be a sequence of finite subsets with $\{C_m\} \subset [-cm, cm]^d$ with*

$$1/m \leq \max_{y \in K} \min_{x \in C} |x - y| \leq \eta(C_m)$$

If $\gamma > 0$ and $f \in W_{\gamma,d}$ then for integer $m \geq 1$ there exists $G \in N_{|Cm|,m}$ with centers at points in C_m such that

$$\|f - G\|_d \leq \frac{1}{m^\gamma} \|f\|_{\gamma,d}$$

Moreover, the coefficients of G can be chosen as linear combinations of the data $\{f(x) : x \in C_m\}$.

Theorem 1.3. *For each vinV, let $\{C_{m,v}\}$ be a sequence of finite subsets as described in Theorem 2. Let $\gamma > 0$ and $f \in TW_{\gamma,2}$. Then for integer $m \geq 1$, there exists $G \in TN_{\max|C_{m,v}|m}(\mathbb{R}^2)$ with centers of the constituent network G_v at vertex v at points in $C_{m,v}$ such that*

$$\|f - G\|_{\mathcal{T}} \leq \frac{1}{m^\gamma}\|f\|_{\mathcal{T},\gamma,2}$$

Moreover, the coefficients of each constituent G_v can be chosen as linear combinations of the data $\{f(x) : x \in C_{m,v}\}$.

Theorem 1.4. *(a) Let $\{C_m\}$ be a sequence of finite subsets of $\mathbb{R}^d$, such that for each integer $m \geq 1, C_m \subset C_{m+1}$, $|C_m| \leq c\exp exp(c_1 m^2)$, and $\eta(C_m) \geq 1/m$. Further, let $f \in C_0(\mathbb{R}^d)$, and for each $m \geq 1$, let G_m be a Gaussian network with centers among points in C_m, such that*

$$\sup_{m\geq 1} m^\gamma \|f - G_m\|_{\mathcal{T}} < \infty$$

Then $f \in W_{\gamma,d}$

(b) For each $v \in V$, let $\{C_{m,v}\}$ be a sequence of finite subsets of $\mathbb{R}^{d(v)}$, satisfying the conditions as described in part (a) above. Let $f \in \mathcal{T}C_0(\mathbb{R}^2), \gamma > 0$, and $\{G_m \in \mathcal{T}N_{n,m}\}$ be a sequence where, for each $v \in V$, the centers of the constitutent networks $G_{m,v}$ are among points in $C_{m,v}$, and such that

$$\sup_{m\geq 1} m^\gamma \|f - G_m\|_{\mathcal{T}} \geq \infty$$

Then $f \in \mathcal{T}W_{\gamma,2}$.

2 Previous Work

以前の仕事自体には興味が無いので，自分が疑問に思う点をここに記載する．全体の主張としては誤差が小さいものが存在するといってるだけ，誤差が $O(n^{-r/2})$ まで落とせると言っている．

- $\mathbb{Q}, \mathbb{Q}_p$ 上でうまく定義できるか
- n や d の関係を明確にして，その状況で問題設定を解決したい．
- 計算量に関する考察は何かできないか

3 Compositional functions

4 Main results

この章では，shallow network,deep network の 2 つの場合に近似定理を述べる．2 つとは，ReLU による deep network と deep Gaussian network である．*degree of approximation* は以下で定義される．

$$\mathrm{dist}(f, V_n) = \inf_{P\in V_n}\|f - P\| \tag{1}$$

Remark. V_n は関数の集合，実際にははニューラルネットワークとして定義される関数の集合として，使われていた．

4.1 Deep and shallow neural netwokrs

$I^d := [-1,1]^d, \mathbb{X} = C(I^d,\mathbb{R})$ とし，$\|f\| = \max_{x\in I^d}|f(x)|$ とする．S_n を n 個の *unit* を持つ *shallow netowork* のなす集合とする．すなわち，

$$S_n := \{f : \mathbb{R}^d \to \mathbb{R}| \text{ ある } w_k^i n\mathbb{R}^d, b_k.a_k \in \mathbb{R} \text{ が存在し}, f(x) = \sum_{k=1}^{n} a_k\sigma(w_k x + b_k)\}$$

この時，訓練パラメータが $(d+2)n$ 個存在する.*(メタ的で数学的ではない).* $W_{r,d}^{NN}$ で r 回連続偏微分可能であって, $\|f\| + \sum_{1\le|k|_1\le r}\|D^k f\| \le 1$ を満たすもの全体とする．また，$W_{H,r,2}^{NN}$ を以下で定義する．

$$W_{H,r,2}^{NN} := \{h|h = f_{11}\circ\cdots\circ f_{k2^k}(f_{ij}\in W_{r,2}^{NN})\}$$

$\mathcal{D}_n$ を S_n に属する関数の合成で書けるもの全体とする．上の書き方，かなりまずいけど，$f_11(f_21, f_22)$ で表せるもの？つまり，d が実質 2 のものということですかね．この時はパラメターの個数が $d = 2^m$ とした時に，$(d+2)m(1+2+\cdots+2^{m-1}) = (d+2)m(d-1)$ となる．

Theorem 4.1. $\sigma : \mathbb{R}\to\mathbb{R}$ を無限回微分可能であって，$\mathbb{R}$ の任意の開区間上で，多項式でないとする．この時以下が成り立つ．

1. 任意の $f \in W_{r,d}^{NN}$ に対し，
$$\mathrm{dist}(f, S_n) = O(n^{-r/d}) \tag{2}$$

2. 任意の $f \in W_{H,r,d}^{NN}$ に対し，
$$\mathrm{dist}(f, \mathcal{D}_n) = O(n^{-r/2}) \tag{3}$$

Proof. 1 つめの主張は他の論文にて示した．2 つめの主張を示す．f が無限回微分可能な時，特にリプシッツ連続である．よって，$f(g_1,g_2) - f(P_1,P_2) \le M|g_1 - P_1||g_2 - P_2|$ となる．これより，

$$\begin{aligned}|f(g_1,g_2) - P_0(P_1,P_2)| &\le |f(g_1,g_2) - f(P_1,P_2)| + |f(P_1,P_2) - P_0(P_1,P_2)|\\ &\le M|g_1 - P_1||g_2 - P_2| + \mathrm{dist}(f, S_n)\end{aligned}$$

となる．$|g_1 - P_1||g_2 - P_2| \le O(n^{-r})$ となるので．$f(g_1,g_2) - P_0(P_1,P_2) = O(n^{-r/2})$ となる．これを inductive に続けていけばよい． □

Remark. オーダとしてはこれが限界であることが示されている．