Report on Single Index Model

Ergan Shang

USTC

June 7, 2022

Overview

1. Single Index Model
   - Under Gaussian Space
   - Least Squares on Monotone Functions
   - Score Estimation on Monotone Functions
2. Sample Complexity of One-Hidden-Layer Neural Networks
3. What am I thinking about

Overview

1. Single Index Model
   - Under Gaussian Space
   - Least Squares on Monotone Functions
   - Score Estimation on Monotone Functions
2. Sample Complexity of One-Hidden-Layer Neural Networks
3. What am I thinking about

Background and Introduction

Suppose our link function has the form

$$f(\boldsymbol{x}) = g(\langle \boldsymbol{u}^*, \boldsymbol{x} \rangle) = \sum_{\ell=0}^{\infty} a_\ell^* H_\ell(\langle \boldsymbol{x}, \boldsymbol{u}^* \rangle)$$

with the function class defined as $\mathcal{F} = \{g : \mathbb{R} \to \mathbb{R} : g(z) = \sum_{i=0}^{L} a_i H_i(z)\}$.

Associated Population Loss

$$R_L(\boldsymbol{u}) = \min_{\boldsymbol{a} \in \mathbb{R}^{L+1}} \mathbb{E}\left[\left(y - \sum_{\ell=0}^{L} a_\ell H_\ell(\langle \boldsymbol{u}, \boldsymbol{x} \rangle)\right)^2\right] = \sigma^2 + \sum_{\ell=1}^{L} a_\ell^{*2} \left(1 - \langle \boldsymbol{u}, \boldsymbol{u}^* \rangle^{2\ell}\right)$$

$$F_\ell(\boldsymbol{u}) = \mathbb{E}[f(\boldsymbol{x}) H_\ell(\langle \boldsymbol{u}, \boldsymbol{x} \rangle)] = a_\ell^* \langle \boldsymbol{u}, \boldsymbol{u}^* \rangle^\ell$$

Therefore, we can define the goodness-of-fit statistic as
$\hat{F}_\ell(\boldsymbol{u}) = \frac{1}{n} \sum_{i=1}^{n} y_i H_\ell(\langle \boldsymbol{x}_i, \boldsymbol{u} \rangle)$ $\quad \hat{\boldsymbol{u}} = \arg\max_{\boldsymbol{u} \in \mathbb{S}^{p-1}} \hat{F}_\ell(\boldsymbol{u})$ for some value ℓ

Concentration Results

In the algorithms to come, we have to analyze $\nabla \hat{F}_\ell(\boldsymbol{u}; data) = \frac{1}{n}\sum_{i=1}^n \sqrt{\ell} y_i H_{\ell-1}(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\boldsymbol{x}_i$ which can be further decomposed into the form like $\propto \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i) H_\ell(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\boldsymbol{x}_i$ and $\frac{1}{n}\sum_{i=1}^n \epsilon_i H_\ell(\langle \boldsymbol{u}, \boldsymbol{x}_i\rangle)\boldsymbol{x}_i$

Main Concentration Results

With probability at least $1 - 2\delta - \frac{4}{n}$,

$$\left\| \frac{1}{n}\sum_{i=1}^n y_i H_\ell(\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle)\boldsymbol{x}_i - \mathbb{E}[y H_\ell(\langle \boldsymbol{u}, \boldsymbol{x}\rangle)\boldsymbol{x}] \right\|$$

$$\leq 100(\|h\|_\infty + 4\sigma)2^\ell \sqrt{\frac{max(p, log\frac{1}{\delta})(logn)^\ell}{n}}$$

Assumptions and a good ℓ

Let $Z \sim \mathcal{N}(0,1)$

1. (Normalization) $\mathbb{E}[g^2(z)] = 1$ i.e. $1 = \sum_{\ell=1}^{\infty} a_\ell^{*2}$
2. (Smoothness) $\mathbb{E}\left[\left(\frac{d^2 g(z)}{dz^2}\right)\right] \leq R^2$ i.e. $\sum_{i=2}^{\infty} i(i-1)a_i^{*2} \leq R^2$
3. (Minimum Signal Strength) $\mathbb{E}\left[\left(\frac{dg(z)}{dz}\right)^2\right] \geq \mu$ i.e. $\sum_{i=1}^{\infty} i a_i^{*2} \geq \mu$
4. (Bounded Link Function) $\|g\| < \infty$

A good $\ell_\#$

There exists a $\ell_\# \leq \frac{2R^2}{\mu}$ such that $\ell_\# |a_{\ell_\#}^*|^2 \geq \frac{\mu^2}{4R^2}$

Algorithms and Conclusions

Algorithm 1 Estimate-Index-Vector-From-Harmonic(S,ℓ)

input Data $S = \{\boldsymbol{x}_i, y_i\} \subset \mathbb{R}^p \times \mathbb{R}$; Degree of Harmonic $\ell \in \mathbb{N}$

output Index Estimate $\hat{\boldsymbol{u}}_\ell \in \mathbb{R}^p$

1:Split S into two equal parts:

$S_1 = \{(\boldsymbol{x}_i, y_i), i = 1, 2, \cdots, \frac{n}{2}\}, S_2 = \{(\boldsymbol{x}_i, y_i), i = \frac{n}{2} + 1, \cdots, n\}$

2:Define $\hat{F}_\ell(\boldsymbol{u}; S_1) = \frac{2}{n} \sum_{i=1}^{\frac{n}{2}} y_i H_\ell(\langle \boldsymbol{x}_i, \boldsymbol{u} \rangle)$ and

$\hat{F}_\ell(\boldsymbol{u}; S_2) = \frac{2}{n} \sum_{i=\frac{n}{2}+1}^{n} y_i H_\ell(\langle \boldsymbol{x}_i, \boldsymbol{u} \rangle)$

3:Random Initialization: $\boldsymbol{u}_0 \sim Uniform(\mathbb{S}^{p-1})$

4:Compute two steps of iterative process: $\boldsymbol{u}_1 = \frac{\nabla \hat{F}_\ell(\boldsymbol{u}_0; S_1)}{\|\nabla \hat{F}_\ell(\boldsymbol{u}_0; S_1)\|}$ and

$\boldsymbol{u}_2 = \frac{\nabla \hat{F}_\ell(\boldsymbol{u}_1; S_2)}{\|\nabla \hat{F}_\ell(\boldsymbol{u}_1; S_2)\|}$

5:**return** $\hat{\boldsymbol{u}}_\ell := \boldsymbol{u}_2$

Algorithm 2 Learn-single-index-Model $(S, R^2, \mu, \sigma^2, \|f\|_\infty, \delta)$

Input Data:$S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$;
Output Index Estimate $\hat{\boldsymbol{u}} \in \mathbb{R}^p$
1:Split S into S_{train} and S_{test} such that
$m := |S_{test}| = 256 \cdot 2^{\frac{4R^2}{\mu}} R^4(\sigma^2 + \|f\|_\infty^2)/(\delta\mu^3)$
2:Let $L = \frac{2R^2}{\mu}$
3:Let $\hat{\boldsymbol{u}}_\ell :=$ Estimate-Index-Vector-From-Harmonic(S_{train}, ℓ) for each $\ell \in \{1, 2, \cdots, L\}$
4:Compute the good-of-fitness $T_\ell = \sum_{i \in S_{test}} y_i H_\ell(\langle \boldsymbol{x}_i, \hat{\boldsymbol{u}}_\ell \rangle)/m$ for each $\ell \in \{1, 2, \cdots, L\}$.
5:Let $\ell_{best} := arg\max_{\ell \in [L]} |T_\ell|$.
6:**return** $\hat{\boldsymbol{u}} := \boldsymbol{u}_{\ell_{best}}$.

Algorithms and Conclusions

Convergence Rate

Given any $\epsilon, \delta \in (0, 1)$; with probability at least $1 - \frac{4R^2}{\mu} e^{-\frac{p}{32}} - \frac{12R^2}{\mu}\delta - \frac{16R^2}{n\mu}$, the estimate returned by Algorithm 2, $\hat{\boldsymbol{u}}$ satisfies

$$|\langle \boldsymbol{u}^*, \hat{\boldsymbol{u}} \rangle| \geq 1 - \frac{3200 \cdot 2^{\frac{4R^2}{\mu}} (\|f\|_\infty + 4\sigma) R^4}{\mu^2 \sqrt{\mu}} \sqrt{\frac{max(p, log(\frac{1}{\delta}))(logn)^{\frac{2R^2}{\mu}}}{n}}$$

provided that n satisfies

$$n \geq \frac{1024 \cdot 10^4 (\|f\|_\infty + 4\sigma)^2 R^4}{\mu^3} \frac{2^{\frac{4R^2}{\mu}}}{\delta^{\frac{4R^2}{\mu} - 2}} max(p, log(\frac{1}{\delta})) p^{\frac{2R^2}{\mu} - 1} (logn)^{\frac{2R^2}{\mu}}$$

Least Squares under Monotonicity

Suppose the link function is monotone and data are not from Guassian space, we have the problem

$$h_n(\psi, \alpha) = \sum_{i=1}^{n} (Y_i - \psi(\alpha^T \boldsymbol{X}_i))^2 \quad (\psi, \alpha) \in \mathcal{M} \times \mathcal{S}_{d-1}$$

where $\mathcal{M}$ denotes function class of non-decreasing functions. Using the knowledge of isotonic regression, we define $n_k = \sum_{i=1}^{n} \mathbb{I}\{\alpha^T \boldsymbol{X}_i = Z_k\}, t_k = \frac{1}{n_k} \sum_{i=1}^{n} Y_i \mathbb{I}\{\alpha^T \boldsymbol{X}_i = Z_k\}$, we have $h_n(\psi, \alpha) = \sum_{k=1}^{m} n_k (t_k - \psi(Z_k))^2 + \sum_{i=1}^{n} Y_i^2 - \sum_{k=1}^{m} n_k t_k^2$. Thus, we set $\eta_k = \psi(Z_k)$ leading to

$$min \sum_{k=1}^{m} n_k (t_k - \eta_k)^2 \text{ over } \eta_1 \leq \cdots \leq \eta_m$$

Solution for Least Squares

Let P^X be the set of all permutations π on $\{1, \cdots, m\}$ such that $\exists \alpha \in \mathcal{S}_{d-1}$ that linearly induces π in the sense that $\alpha^T x_{\pi(1)} < \cdots < \alpha^T x_{\pi(m)}$ which is available for $\alpha \in S^X = \{\alpha \in \mathcal{S}_{d-1} : \alpha^T \boldsymbol{X}_i \neq \alpha^T \boldsymbol{X}_j \text{ for all } i \neq j \text{ such that } \boldsymbol{X}_i \neq \boldsymbol{X}_j\}$

Definition

$$\tilde{n}_k = \sum_{i=1}^{n} \mathbb{I}\{\boldsymbol{X}_i = \boldsymbol{x}_k\} \qquad \tilde{y}_k = \frac{1}{\tilde{n}_k} \sum_{i=1}^{n} Y_i \mathbb{I}\{\boldsymbol{X}_i = \boldsymbol{x}_k\}$$

We denote by $d_1^\pi \leq \cdots \leq d_m^\pi$ the left derivatives of the greatest convex minorant of the cumulative sum diagram define by the set of points

$$\left\{ (0,0), \left(\sum_{j=1}^{k} \tilde{n}_{\pi(j)}, \sum_{j=1}^{k} \tilde{n}_{\pi(j)} \tilde{y}_{\pi(j)} \right), k = 1, \cdots, m \right\}$$

Solutions for Least Squares

Theoretical Solution

The infimum of $(\psi, \alpha) \mapsto h_n(\psi, \alpha)$ over $\mathcal{M} \times \mathcal{S}_{d-1}$ is achieved. Moreover if $(\hat{\psi}_n, \hat{\alpha}_n)$ satisfies the following conditions, then it is a minimzer:

- $\hat{\alpha}_n \in S^X$ linearly induces $\hat{\pi}_n$ that minimizes $\pi \mapsto \tilde{h}_n(\pi) = \sum_{k=1}^{m} \tilde{n}_{\pi(k)}(\tilde{y}_{\pi(k)} - d_k^{\pi})^2$ over P^X
- $\hat{\psi}_n$ is monotone ,non-decreasing with $\hat{\psi}_n(\hat{\alpha}^T \boldsymbol{x}_{\hat{\pi}_{n(k)}}) = d_k^{\hat{\pi}_n}$

Solutions for Least Squares

Algorithm: Stochastic Search: Find optimal $(\hat{\alpha}_n, \hat{\psi}_n)$

1:Choose the total number(maximal iterations) N of stochastic searches to perform and set $k = 1$

2:Let $Z_k \sim \mathcal{N}(0, I_d)$ and $\alpha = \frac{Z_k}{\|Z_k\|}$ which is uniform on the sphere

3:Compute distinct values $t_1 \leq \cdots \leq t_L$ of $\alpha_k^T \boldsymbol{X}_i$ for $i \in [n]$
and also $n_\ell = \sum_{i=1}^n \mathbb{I}\{\alpha_k^T \boldsymbol{X}_i = t_\ell\}$,$y_\ell = \frac{1}{n_\ell} \sum_{i=1}^n Y_i \mathbb{I}\{\alpha_k^T \boldsymbol{X}_i = t_\ell\}$

4:Compute $d_1 \leq \cdots \leq d_L$, the left derivatives of the greatest convex minorant of
$\left\{(0,0), \left(\sum_{j=1}^{\ell} n_j, \sum_{j=1}^{\ell} n_j y_j\right), \ell = 1, \cdots, L\right\}$ using **PAVA**

5:Compute $A_k := \sum_{\ell=1}^L n_\ell (y_\ell - d_\ell)^2$ and set $k = k + 1$ and return to 2 when $k \leq N$

6:Compute $\hat{k}$ that minimizes A_k over $k \in [N]$

Return: $(\hat{\alpha}_n, \hat{\psi}_n) = (\alpha_{\hat{k}}, \psi_{\hat{k}})$ where $\psi_{\hat{k}}$ is piecewise constant: $\psi_{\hat{k}}(t_\ell) = d_\ell$

Entropy Results

Observing that $\frac{1}{n}\sum_{i=1}^{n}(Y_i - g(\boldsymbol{X}_i))^2 \propto -\frac{1}{n}\sum_{i=1}^{n}\left(Y_i g(\boldsymbol{X}_i) - \frac{g^2(\boldsymbol{X}_i)}{2}\right)$ and $\hat{g}_n$ maximizes $\mathbb{M}_n g := \frac{1}{n}\left(Y_i g(\boldsymbol{X}_i) - \frac{g^2(\boldsymbol{X}_i)}{2}\right)$ of form $g(\boldsymbol{x}) = \psi(\alpha^T \boldsymbol{x})$.
Similarly we define $\mathbb{M}g := \int_{\mathcal{X}\times\mathbb{R}}(yg(\boldsymbol{x}) - \frac{1}{2}g^2(\boldsymbol{x}))d\mathbb{P}(\boldsymbol{x}, y), \mathbb{Q}g := \int g d\mathbb{Q}$.
Also, we denote $\hat{f}_n(\boldsymbol{x}, y) = y\hat{g}_n(\boldsymbol{x}) - \frac{1}{2}\hat{g}_n^2(\boldsymbol{x}), f(\boldsymbol{x}, y) = yg(\boldsymbol{x}) - \frac{1}{2}g^2(\boldsymbol{x})$ and $\mathbb{M}_n g = \mathbb{P}_n f$ where $\mathbb{P}_n$ denotes empirical distribution.

Lemma 3.4.3

Let $\mathcal{F}$ be a class of measurable functions such that $\|f\|_{\mathbb{P},B} = \left(2\mathbb{P}(e^{|f|} - 1 - |f|)\right)^{\frac{1}{2}} \leq \delta$ for every $f \in \mathcal{F}$. Then

$$\mathbb{E}[\|G_n\|_{\mathcal{F}}] \lesssim J(\delta, \mathcal{F}, \|\cdot\|_{\mathbb{P},B})\left(1 + \frac{J(\delta, \mathcal{F}, \|\cdot\|_{\mathbb{P},B})}{\delta^2\sqrt{n}}\right) \tag{1}$$

where $J(\eta) = \int_0^\eta \sqrt{1 + H_B(\epsilon, \mathcal{F}, \|\cdot\|_{B,\mathbb{P}})}d\epsilon$ and $G_n = \sqrt{n}(\mathbb{P}_n - \mathbb{P}), \|G_n\|_{\mathcal{F}} = \sup_{g\in\mathcal{F}}|G_n g|$ and $\mathbb{P}_n$ denotes the empirical process.

Entropy Results

Lemma 3.4.2

Let $\mathcal{F}$ be a class of measurable functions such that $\|f\|_{\mathbb{P}} \leq \delta$ and $\|f\|_\infty \leq M$ for every f in $\mathcal{F}$. Then

$$\mathbb{E}_{\mathbb{P}}[\|G_n\|_{\mathcal{F}}] \lesssim J_n(\delta, \mathcal{F}, \|\cdot\|_{\mathbb{P}}) \left(1 + \frac{J_n(\delta, \mathcal{F}, \|\cdot\|_{\mathbb{P}})}{\sqrt{n}\delta^2} M\right)$$

where $G_n = \sqrt{n}(\mathbb{P}_n - \mathbb{P})$ and $J_n(\delta, \mathcal{F}, \|\cdot\|) = \int_0^\delta \sqrt{1 + H_B(\epsilon, \mathcal{F}, \|\cdot\|)} d\epsilon$.

Before using Lemma 3.4.2 and Lemma 3.4.3, there is a useful little trick: whenever we have a bound for $H_B \leq \frac{C}{\epsilon}$, we will write $J_n(e) = \int_0^e \left(1 + \frac{C}{\epsilon}\right)^{\frac{1}{2}} \leq e + 2\sqrt{c\epsilon}$, thus bounds for $\mathbb{E}[\|G_n\|]$

$$H_B \Rightarrow J_n \Rightarrow \mathbb{E} \overset{Markov}{\Rightarrow} \text{consistency}$$

Theorem for deriving Convergence Rates

Theorem 3.2.5

If for every θ in a neighborhood of θ_0

$$\mathbb{M}(\theta) - \mathbb{M}(\theta_0) \lesssim -d^2(\theta, \theta_0)$$

Suppose that, for every n and sufficiently small δ, the centered process $\mathbb{M}_n - \mathbb{M}$ satisfies

$$\mathbb{E}\left[\sup_{d(\theta,\theta_0)<\delta} \left|(\mathbb{M}_n - \mathbb{M})(\theta) - (\mathbb{M}_n - \mathbb{M})(\theta_0)\right|\right] \lesssim \frac{\phi_n(\delta)}{\sqrt{n}}$$

for functions $\phi_n(\delta)$ such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$. Let $r_n^2\phi_n\left(\frac{1}{r_n}\right) \leq \sqrt{n}$ hold and $\mathbb{M}_n(\hat{\theta}_n) \geq \mathbb{M}_n(\theta_0) - O_p(r_n^{-2})$. Then

$$r_n d(\hat{\theta}_n, \theta_0) = O_p(1) \tag{2}$$

Entropy for function class

Two useful conclusions

Under the assumption: $\exists a_0 > 0, M > 0$ such that $\forall s \geq 2, \boldsymbol{x} \in \mathcal{X}, \int |y|^s \mathbb{P}_{\boldsymbol{x}}(y) \leq a_0 s! M^{s-2}$, we have

$$\sup_{\boldsymbol{x} \in \mathcal{X}} |\hat{g}_n(\boldsymbol{x})| \leq O_p(logn) \qquad D(\hat{g}_n, g_0) = O_p(n^{-\frac{1}{3}}(logn)^{\frac{5}{3}})$$

where $D(g, g_0) = \left(\int_{\mathcal{X}} (g_0(\boldsymbol{x}) - g(\boldsymbol{x}))^2 d\mathbb{Q}(\boldsymbol{x})\right)^{\frac{1}{2}}$

By setting $K = Clogn, v = Cn^{-\frac{1}{3}}(logn)^2$ with proper constant C. We derive entropy bounds for these function class:

- $G_K = \{g(\boldsymbol{x}) = \psi(\alpha^T \boldsymbol{x}) : \alpha \in \mathcal{S}_{d-1}, \psi \in \mathcal{M}_K\}$
  $\mathcal{F}_K = \{f(\boldsymbol{x}, y) = yg(\boldsymbol{x}) - \frac{1}{2}g^2(\boldsymbol{x}) : g \in G_K\}$
- $G_{Kv} = \{g \in G_K : D(g, g_0) \leq v\}$
  $\mathcal{F}_{Kv} = \left\{f(\boldsymbol{x}, y) = yg(\boldsymbol{x}) - \frac{1}{2}g^2(\boldsymbol{x}) : g \in G_{Kv}\right\}$

Main Results: $O_p(n^{-\frac{1}{3}})$

With the entropy bounds at hand, combining with 15, we have

Consistency and Convergence

- $\left(\int_{\mathcal{X}}(\hat{g}_n(\boldsymbol{x}) - g_0(\boldsymbol{x}))^2 d\mathbb{Q}(\boldsymbol{x})\right)^{\frac{1}{2}} = O_p(n^{-\frac{1}{3}})$
- $\hat{\alpha}_n = \alpha_0 + o_p(1)$, particularly, $\|\hat{\alpha}_n - \alpha_0\| = O_p(n^{-\frac{1}{3}})$
- For all fixed continuity points t of ψ_0 in the interior of $C_{\alpha_0} = \{\alpha_0^T \boldsymbol{X} : \boldsymbol{X} \in \mathcal{X}\}$, $\psi_n(t) \xrightarrow{P} \psi_0(t)$; if ψ_0 is continuous, then $\sup_{t\in I}|\hat{\psi}_n(t) - \psi_0(t)| = o_p(1)$
- If moreover, ψ_0 has a derivative bounded from above on C_{α_0}, then $\left(\int_{\underline{c}+v_n}^{\bar{c}-v_n}(\hat{\psi}_n(t) - \psi_0(t))^2 dt\right)^{\frac{1}{2}} = O_p\left(n^{-\frac{1}{3}}\right)$ for all sequence v_n such that $n^{\frac{1}{3}} v_n \to \infty$ and $\underline{c} + v_n < \bar{c} - v_n$ with $\bar{c} = supC_{\alpha_0}, \underline{c} = infC_{\alpha_0}$.

Score Estimator

We define $S_n(\psi, \alpha) = \frac{1}{n}(Y_i - \psi(\alpha^T \boldsymbol{X}_i))^2$. For a fixed α, by the conclusion obtained via the left derivative of the greatest convex minorant of the cumulative sum diagram $\left\{(0,0), (\sum_{j=1}^{i} n_j^{\alpha}, \sum_{j=1}^{i} n_j^{\alpha} Y_j^{\alpha}), i = 1, \cdots, m\right\}$, the minimizer for a fixed α is denoted by $\hat{\psi}_{n\alpha}$. Now we consider the estimation for α:

$$\min_{\alpha} \frac{1}{n} \left(Y_i - \hat{\psi}_{n\alpha}(\alpha^T \boldsymbol{X}_i)\right)^2$$

Let $\mathbb{S} : \mathbb{R}^{d-1} \to \mathcal{S}_{d-1} \subset \mathbb{R}^d; \beta \to \alpha = \mathbb{S}(\beta)$ be a local parametrization, obtaining

$$\min_{\alpha} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{\psi}_{n\alpha}(\mathbb{S}(\beta)^T \boldsymbol{X}_i)\right)^2$$

Zero-Crossing Solution

By derivative, we get
$\frac{1}{n}\sum_{i=1}^{n} 2\left(Y_i - \hat{\psi}_{n\alpha}(\mathbb{S}(\beta)^T \boldsymbol{X}_i)\right) \cdot (-1)\frac{d\hat{\psi}_{n\alpha}(x)}{dx} \cdot (J_{\mathbb{S}}(\beta))^T \boldsymbol{X}_i = 0$, i.e.

$$\frac{1}{n}\sum_{i=1}^{n}(J_{\mathbb{S}}(\beta))^T \boldsymbol{X}_i\left(Y_i - \hat{\psi}_{n\alpha}(\mathbb{S}(\beta)^T \boldsymbol{X}_i)\right)$$

where $J_{\mathbb{S}}(\beta) = \left(\frac{\partial \mathbb{S}_i(\beta)}{\partial \beta_j}\right) \in \mathbb{R}^{d\times(d-1)}$
We cannot hope to find the exact solution for the equations, instead we can derive the zero-crossing of

$$\phi_n(\beta) = \int (J_{\mathbb{S}}(\beta))^T \boldsymbol{x}(y - \hat{\psi}_{n\alpha}\left(y - \hat{\psi}_{n\alpha}(\mathbb{S}(\beta)^T \boldsymbol{x})\right) d\mathbb{P}_n(\boldsymbol{x}, y)$$

and with population version

$$\phi(\beta) = \int (J_{\mathcal{S}}(\beta))^T \boldsymbol{x}(y - \psi_\alpha(\mathcal{S}(\beta)^T \boldsymbol{x}))dP_0(\boldsymbol{x}, y)$$

where $\psi_\alpha(u) = \mathbb{E}[\psi_0(\alpha_0^T \boldsymbol{X})|\alpha^T \boldsymbol{X} = u]$

Technical Lemmas

Link Function under L_2

The functional L_α given by $\psi \mapsto L_\alpha(\psi) = \int_{\mathcal{X}}(\psi_0(\alpha_0^T \boldsymbol{x}) - \psi(\alpha^T \boldsymbol{x}))^2 dG(\boldsymbol{x})$ admits a minimizer ψ^α over the set of non-decreasing functions, such that ψ^α is uniquely given by $\psi_\alpha = \mathbb{E}[\psi_0(\alpha_0^T \boldsymbol{X}) | \alpha^T \boldsymbol{X} = u]$

Distance and Bound

Similar to 16, we have when $\alpha \in B(\alpha_0, \delta_0)$

$$\max_\alpha \sup_{\boldsymbol{x} \in \mathcal{X}} |\hat{\psi}_{n\alpha}(\alpha^T \boldsymbol{x})| = O_p(logn) \quad \sup_\alpha \int \left(\hat{\psi}_{n\alpha}(\alpha^T \boldsymbol{x}) - \psi_\alpha(\alpha^T \boldsymbol{x})\right)^2 = O_p\left(\left(lo\right.\right.$$

$$\begin{array}{ccc} \text{The estimation } \hat{\psi}_{n\alpha} & \overset{distance}{\longleftrightarrow} & \psi_\alpha(u) \text{ link function} \\ \updownarrow & & \updownarrow \\ \phi_n(\beta) & \overset{\phi_n(\beta)=\phi(\beta)+o_p(1)}{\longleftrightarrow} & \phi(\beta) \end{array}$$

Asymptotic Property

- (Existence) A crossing of zero $\hat{\beta}_n$ of $\phi_n(\beta)$ exists with probability tending to 1
- (Consistency) $\hat{\alpha}_n \xrightarrow{P} \alpha_0$
- (Asymptotic normality) Define $A = \mathbb{E}\left[\psi_0'\left(\alpha_0^T \boldsymbol{X}\right) Cov\left(\boldsymbol{X}|\alpha_0^T \boldsymbol{X}\right)\right]$ and $\Sigma = \mathbb{E}\left[\left(Y - \psi_0(\alpha_0^T \boldsymbol{X})\right)^2 \left(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}|\alpha_0^T \boldsymbol{X}]\right)\left(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}|\alpha_0^T \boldsymbol{X}]\right)^T\right]$, then $\sqrt{n}(\hat{\alpha}_n - \alpha_0) \xrightarrow{d} \mathcal{N}_d(0, A^- \Sigma A^-)$

Overview

1. Single Index Model
   - Under Gaussian Space
   - Least Squares on Monotone Functions
   - Score Estimation on Monotone Functions
2. Sample Complexity of One-Hidden-Layer Neural Networks
3. What am I thinking about

Background and Introduction

Definition: Learnable algorithm

Suppose that F is a set of functions mapping from a domain X into the real interval $[0,1]$. A learning algorithm L for F is a function$L : \cup_{m=1}^{\infty}(X \times \mathbb{R})^m \to F$ with the following property:

- given any $\epsilon \in (0,1)$, $\delta \in (0,1)$, $B \geq 1$: there is an integer $m_0(\epsilon, \delta, B)$ such that if $m \geq m_0(\epsilon, \delta, B)$

then, for any probablity distribution P on $X \times [1-B, B]$,
if z is a training sample of length m, drawn randomly according to the product probability distribution P^m, then, with probability at least $1-\delta$, the function $L(z)$ output by L is such that

$$er_P(L(z)) < opt_P(F) + \epsilon$$

where $opt_P(F) = \inf_{f \in F} \mathbb{E}[(f(x)-y)^2]; er_P(f) = \mathbb{E}[(f(x)-y)^2]$
We say that F is **learnable** if there is a learning algorithm for F.

Fat shattering and Rademacher complexity

Lower bound for sample complexity through fat-shattering dimension

F is the function class: $X \to [0,1]$. Then for $B \geq 2, \epsilon \in (0,1)$ and $0 < \delta < \frac{1}{100}$, **any learning algorithm** L for F has sample complexity satisfying

$$m_L(\epsilon, \delta, B) \geq \frac{fat_F(\epsilon/\alpha) - 1}{16\alpha} \quad \forall \alpha \in (0, \frac{1}{4})$$

Rademacher complexity

$$R_m(\mathcal{F}) = \sup_{\{x_i\}_{i=1}^m \subset \mathcal{X}} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \epsilon_i f(x_i) \right]$$

Frankly speaking, although I read Chapter 13 in Martin's book, I still cannot figure out the relationship between Rademacher complexity and upper bound for sample complexity

Overview

1. Single Index Model
   - Under Gaussian Space
   - Least Squares on Monotone Functions
   - Score Estimation on Monotone Functions
2. Sample Complexity of One-Hidden-Layer Neural Networks
3. What am I thinking about

My idea

For single index model

$$f(\boldsymbol{x}) = \psi_0(\alpha_0^T \boldsymbol{x})$$

under i.i.d. observations $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, the problem is how to "test" $\alpha_0 = \mathbf{0}$

In those papers I read, they always assume $\alpha_0 \in \mathcal{S}_{d-1}$, with the condition ψ is monotone or noise ϵ_i and $\boldsymbol{x}_i$ are normal random variable. $\alpha_0 \in \mathcal{S}_{d-1}$ bypasses the problem of non-identifiability,i.e. if we define $\phi_0(t) = \psi_0(\|\alpha_0\| t)$. With $\beta_0 = \frac{\alpha_0}{\|\alpha_0\|}$, we have $\psi_0(\alpha_0^T \boldsymbol{x}) = \phi_0(\beta_0^T \boldsymbol{x})$.

Therefore, it may not work if we apply least squares methods to estimate $\alpha_0 \in \mathcal{S}_{d-1}$. Perhaps we have to solve the problem from a prospective of hypothesis testing.

Estimation or Testing?

- Estimation
  If we can tackle the problem of non-identifiability, then we can, under the conditions that the link is monotone or in Gaussian space, make an estimation for both ψ_0 and α_0. It will be clear to see whether $\alpha_0 = 0$.
- Testing

$$H_0 : Y = \psi_0(0) + \epsilon \sim \mathcal{N}(\mu, \sigma^2) \longleftrightarrow H_1 : Y = \psi_0(\alpha_0^T \boldsymbol{x}) + \epsilon \quad \alpha_0 \in \mathcal{S}_{d-1}$$

  where $\mu = \psi_0(0)$. If ψ_0 is linear, the testing problem can be done by ANOVA. Similarly, is it possible to derive a testing statistic similar to "RSS_1-RSS_2" and figure it out distribution?

Testing and Estimation

We can define $RSS_1 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ and $RSS_2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, where we can using the algorithm finding $\hat{\psi}_n, \hat{\alpha}_n$ and let $\hat{Y}_i = \hat{\psi}_n(\hat{\alpha}_n^T \boldsymbol{x}_i)$. And we can define F-statistic

$$F = \frac{(RSS_1 - RSS_2)/d.f.1}{RSS_2/d.f.2}$$

And another method is to define the likelihood ratio test and perhaps do minimax hypothesis testing. Observing that $\psi_0(\alpha_0^T \boldsymbol{x}) - \psi_0(0)$ is still a non-decreasing function, we can test

$$H_0 : \theta = 0 \longleftrightarrow H_1 : \theta = \psi_0(\alpha_0^T \boldsymbol{x})$$

where the null and alternative are both simple.

References

1 Balabdaoui F, Durot C, Jankowski H. Least squares estimation in the monotone single index model[J]. Bernoulli, 2019, 25(4B): 3276-3310.

2 Vardi G, Shamir O, Srebro N. The Sample Complexity of One-Hidden-Layer Neural Networks[J]. arXiv preprint arXiv:2202.06233, 2022.

3 Balabdaoui F , Piet Groeneboom. Score estimation in the monotone single index mode.

4 Dudeja R , Hsu D . Learning Single-Index Models in Gaussian Space. 2018.

5 Wainwright M J. High-dimensional statistics: A non-asymptotic viewpoint[M]. Cambridge University Press, 2019.

6 Anthony M, Bartlett P L, Bartlett P L. Neural network learning: Theoretical foundations[M]. Cambridge: cambridge university press, 1999.

7 Groeneboom P, Jongbloed G. Nonparametric estimation under shape constraints[M]. Cambridge University Press, 2014.

Thank you !