修士論文 要旨

Analysis on Benign Overfitting of Two-layer Linear Neural Network

数理情報学専攻 48216219　鈴木 渓太
指導教員　鈴木 大慈 准教授

1　はじめに

深層学習は近年, 画像や自然言語処理をはじめとした多くの分野で高い予測精度を発揮している. しかし, 高い精度を実現しながらもモデルの複雑さゆえに学習データに対して過学習してしまうことが多く, これはモデルが汎化するためには複雑すぎてはならないとする従来の学習理論では説明できないものである. このような現象を良性過学習といい, 現在多くの研究者が高次元統計学の枠組みで深層学習の汎化現象を研究しているものの [1, 2, 3, 4], それらの研究は 1 出力の問題設定に限定されてしまっている. 多出力の問題は多クラス分類といった多くの重要な問題設定を含んでいるため, 深層学習の汎化現象を理解するには不可欠な問題設定である. 本研究では多出力線形回帰の問題設定で二層線形ニューラルネットワークにおける良性過学習の解析を行い, 特徴量学習したリッジ回帰 (二層線形ニューラルネットワーク) が特徴量学習をしていないリッジ回帰を優越すること (3 章), そして学習した二層線形ニューラルネットワークがベイズ最適性を満たすこと (4 章) を示した. これらの結果は多出力設定ならニューラルネットワークが高次元でも適切な特徴学習を実現できることを示している.

2　問題設定

本研究では出力が m 個の多出力の線形回帰問題, 特に教師データ $D_n = \left\{\left(x_1, \left(y_1^{(1)}, \ldots, y_1^{(m)}\right)\right), \ldots, \left(x_n, \left(y_n^{(1)}, \ldots, y_n^{(m)}\right)\right)\right\} \in \mathbb{R}^{d\times m}$ が

$$y_i^{(j)} = \beta_{*j}^\top x_i + \epsilon_i^{(j)}, \quad j = 1, \ldots, m$$

という関係で生成されている時, m 個のシグナル $\beta_{*i} \in \mathbb{R}^d$ $(i = 1, \ldots, m)$ を推定する問題を考える. 本研究では $X = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n\times d}$, 正則化パラメータ $\lambda > 0$ について β_{*i} の推定量として以下のものを考える:

$$\hat{\beta}_i(W) = WX^\top \left(XW^\top WX^\top + n\lambda I_n\right)^{-1} y^{(i)}.$$

但し, $W \in \mathbb{R}^{d\times d}$ は目的関数

$$R(W) = \frac{1}{m}\sum_{i=1}^m \min_{\beta_i \in \mathbb{R}^l} \frac{1}{n}\left\|y^{(i)} - XW^\top \beta_i\right\|^2 + \lambda\|\beta_i\|^2 + \frac{\sigma'^2}{n}\mathrm{Tr}\left(\hat{\Sigma}_{wx}\left(\hat{\Sigma}_{wx} + \lambda I_d\right)^{-1}\right)$$

の最小化元であるとする. 但し, σ'^2 はパラメータであり, $\hat{\Sigma}_{wx} := \frac{1}{n}WX^\top XW^\top$ である. この推定量は W を一層目のパラメータ, $\hat{\beta}_i(W)$ を二層目のパラメータとする二層線形ニューラルネットワークに対応することに注意されたい.

3　特徴量学習の効力

本章では二層線形ニューラルネットワークとリッジ回帰の比較によって特徴量学習が有益であることを示す. ここでは推定量 $\hat{\beta}_i(W)$ の性能の指標として以下に示す予測損失を考える:

$$\frac{1}{m}\sum_{i=1}^m \mathbb{E}_x\left[\left(x^\top \beta_{*i} - x^\top W^\top \hat{\beta}_i(W)\right)^2\right].$$

この予測損失は未来のテストデータに対する予測の誤差を表している. さらに予測損失は以下のようにバイアス・バリアンス分解することができる:

$$\frac{1}{m}\sum_{i=1}^m \mathbb{E}_x\left[\left(x^\top \beta_{*i} - x^\top W^\top \hat{\beta}_i(W)\right)^2\right] \lesssim B + V.$$

バイアスはシグナルと推定量全体の空間の “距離”, バリアンスはノイズによる予測のずれを表している. この時, バイアスとバリアンスは以下の定理 1 のように評価することができる.

定理 1. $R(W) - t\sigma^2 = o(1)$ を満たす $t > 0$ についてある $k = o(n)$ が存在し, この k, t がある定数 c_x について $t \in (1, n/c_x), \sqrt{k} + \sqrt{t} \le \sqrt{n}/c_x$ を満たす時, $2t\sigma^2 \le \sigma'^2, \sigma'^2 - t\sigma^2 = \Omega(1)$ なる σ'^2 について, ある定数 c_1, c_2 が存在して任意の $s \in (0, n)$ と $\delta = o(1)$ について確率 $1 - 20e^{-t/c_x} - 2e^{-c_1 k} - 2e^{-c_2 s} - \frac{\sigma^2 \max\|\beta_{*i}\|_{\Sigma_x}^2}{n\delta^2}$ 以上で以下のようにバイアスとバリアンスは評価できる.

$$B + V = O\left(\max\{R(W) - t\sigma^2, \delta\}\right).$$

定理 1 によって目的関数 $R(W)$ は予測損失の上界の推定量として機能することがわかる. 以下では定理

1 の上界に現れる $R(W) - t\sigma^2$ の評価を行う. $\Sigma_\beta := \frac{1}{m}\sum_{i=1}^m \beta_{*i}\beta_{*i}^\top, \sigma_i^2 := \mu_i(\Sigma_\beta)$ とし, $V \in \mathbb{R}^{d\times d}$ を各列に Σ_β の固有ベクトルを固有値の大き順に並べた行列とする. この時, 以下の定理が成立する.

定理 2. $t\sigma^2 < \sigma'^2$ とする. この時, $w_i^2 = \frac{n\lambda}{\sigma'^2 - t\sigma^2}\sigma_i^2$ なる w_i について $W = \mathrm{diag}(w_1, \ldots, w_d)V^\top$ とした時, $\mu_{k+1}\left(\Sigma_\beta^{\frac{1}{2}}\Sigma_x\Sigma_\beta^{\frac{1}{2}}\right) \le \frac{\sigma'^2 - t\sigma^2}{n} \le \mu_k\left(\Sigma_\beta^{\frac{1}{2}}\Sigma_x\Sigma_\beta^{\frac{1}{2}}\right)$ なる k が $k \le n$ を満たすなら高確率で以下が成立する.

$$R(W) - t\sigma^2 \lesssim \sum_{i=1}^d \min\left\{\frac{\sigma'^2 - t\sigma^2}{n}, \mu_i\left(\Sigma_\beta^{\frac{1}{2}}\Sigma_x\Sigma_\beta^{\frac{1}{2}}\right)\right\}.$$

定理 1 と定理 2 を組み合わせることで, 定理 2 の上界が十分小さければ $R(W)$ を最適化することで予測損失自体を十分小さくできることがわかる.

多出力の場合の通常のリッジ回帰の予測損失のバイアスとバリアンスの評価について, Σ_x の固有値を大きい順に $\lambda_1 \ge \cdots \ge \lambda_d$ と表記し, $k \le n$ について $\rho_k := \frac{1}{n\lambda_{k+1}}\left(n\lambda + \sum_{i>k}\lambda_i\right)$ とし, λ_i に対応する Σ_x の固有ベクトル $u_i \in \mathbb{R}^d$ について $\tilde{\sigma}_i^2 := u_i^\top \Sigma_\beta u_i$ を定義する. この時, リッジ回帰のバイアス B_R とバリアンス V_R は適切な仮定の下で [4] と同様に固有値の大きな k 成分と小さな $d-k$ 成分に分けて解析することで以下のように評価することができる:

$$B_R \approx \sum_{i=1}^k \lambda_i\tilde{\sigma}_i^2\frac{\rho_k^2\lambda_{k+1}^2}{\lambda_i^2} + \sum_{i=k+1}^d \lambda_i\tilde{\sigma}_i^2,$$
$$V_R \approx \frac{kt\sigma^2}{n} + \frac{t\sigma^2}{n}\sum_{i=k+1}^d \frac{\lambda_i^2}{\rho_k^2\lambda_{k+1}^2}.$$

この評価と定理 2 の上界を比較することで特徴量学習を行なった二層線形ニューラルネットワークが特徴量学習を行なっていない通常のリッジ回帰を以下の設定で優越することを示すことができる.

1. Σ_x の減衰が遅い時
2. x と β_{*i} の広がりを持つ方向が違う時
3. Σ_x の小さい固有ベクトルがある程度大きい時
4. 出力 $y^{(i)}$ が大きい時

これらの結果は特徴量学習によって Wx の分布がシグナル β_{*i} の分布と近くなることによる恩恵である.

4 ベイズ最適性

本章では二層線形ニューラルネットワークのバイアスとバリアンスがベイズ最適性を満たすことを示す. はじめに, 以下で定義される $X \in \mathbb{R}^{n\times d}$ と $y = X\beta_* + \epsilon \in \mathbb{R}^n$ を用いた推定量を考える:

$$\hat{\beta}_B(X, \beta_*, \epsilon) = \operatorname*{argmin}_{\hat{\beta}} \mathbb{E}_{x,\beta}\left[\left(\beta^\top x - \hat{\beta}^\top x\right)^2\right]$$

右辺は β_* の分布を事前分布であると解釈すればベイズリスクであると見なすことができ, ベイズリスクを最小化する推定量 $\hat{\beta}_B$ をベイズ推定量という. この推定量は $x \sim \mathcal{N}(0, \Sigma_x), \beta_* \sim \mathcal{N}(0, \Sigma_\beta)$, そして出力に加えられるノイズ ε が $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ である時, 以下のように解析的に書くことができる:

$$\hat{\beta}_B(X, y) = \left(X^\top X + \sigma^2\Sigma_\beta^{-1}\right)^{-1}X^\top y.$$

このベイズ推定量を用いることで, 以下の定理のように二層線形ニューラルネットワークのバイアスとバリアンスの下界を構成することができる.

定理 3. $2\sigma^2 \le 1$ かつ, $\sum_{i=k+1}^d \mu_i\left(\Sigma_\beta^{\frac{1}{2}}\Sigma_x\Sigma_\beta^{\frac{1}{2}}\right) = O\left(\min\left\{1, \sigma^2\right\}\right), \mu_{k+1}\left(\Sigma_\beta^{\frac{1}{2}}\Sigma_x\Sigma_\beta^{\frac{1}{2}}\right) = O\left(\frac{1}{n}\right)$ を満たす $k \le n$ が存在するとする. この時, $x \sim \mathcal{N}(0, \Sigma_x)$ であり, $\epsilon^{(i)}$ の各成分が平均 0, 分散 σ^2 の正規分布に *i.i.d.* に従うなら以下が高確率で成立する:

$$\begin{aligned}
&\min_W B + V \\
&= \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{x,\beta\sim\mathcal{N}(0,\Sigma_\beta)}\left[\left(\beta^\top x - \hat{\beta}(X, \beta_{*i}, \epsilon^{(i)})^\top x\right)^2\right] \\
&\gtrsim \sum_{i=1}^d \min\left\{\frac{\sigma^2}{n}, \mu_i\left(\Sigma_\beta^{\frac{1}{2}}\Sigma_x\Sigma_\beta^{\frac{1}{2}}\right)\right\}.
\end{aligned}$$

定理 2 と定理 3 を比較することで $\sigma^2 = \Theta(\sigma'^2 - t\sigma^2)$ であるならば定理 2 の上界はオーダーの意味でその下限を達成していることがわかる.

参考文献

[1] J. Ba, M. A. Erdogdu, T. Suzuki, Z. Wang, D. Wu, and G. Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, 2022.

[2] N. S. Chatterji, P. M. Long, and P. L. Bartlett. The interplay between implicit bias and benign overfitting in two-layer linear networks. *Journal of Machine Learning Research*, 23(263):1–48, 2022.

[3] S. Frei, N. S. Chatterji, and P. Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory*, pages 2668–2703. PMLR, 2022.

[4] A. Tsigler and P. L. Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.