Neural Networks and Singular Learning Theory

Sumio Watanabe
Tokyo Institute of Technology
Contents

1  Statistics and Learning
2  Regular Theory
3  Neural Networks are Singular
4  Algebraic Geometry
5  General Theory
6  Conclusion
1 Statistics and Learning
Learning Machine

\[ p(y|x,w) \]

\[ q(x,y) = q(x)q(y|x) \]

Unknown True and Learning Machine

Learning = Statistical Estimation

\[ p(y|x,w) \]
Notations

Datum : \((x,y)\) in \(\mathbb{R}^{MN}\)

Unknown True : \(q(x,y)\)

I.I.D. Sample \(D_n = (X^n, Y^n) = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}\)

Parameter \(w\) in \(W\) contained in \(\mathbb{R}^d\)

Learning Machine : \(p(y|x,w)\) \quad \text{Prior : } \varphi(w)\)
Learning Process

True $q(x) q(y|x)$

Sample

Generalization Error

Model and Prior

Predictive $p^*(y|x)$

Posterior

MLE

MAP
ML, MAP, and Bayes

Log Loss: \( L_n(w) = - \frac{1}{n} \sum_{i=1}^{n} \log p(Y_i|X_i,w) \)

Posterior \( p(w|D_n) = \frac{1}{Z_n} \varphi(w) \exp(-nL_n(w)) \)

How to determine the predictive: \( p^*(y|x) \)

**Maximum Likelihood:** \( p(y|x,w^*) \) where \( w^* \) minimizes \( L_n(w) \).

**MAP:** \( p(y|x,w^+) \) where \( w^+ \) maximizes \( p(w|D_n) \).

**Bayes:** \( p(y|x,D_n) = E_w[p(y|x,w)] \)

\( E_w[\cdot], V_w[\cdot] \): posterior mean and variance
The purpose of Statistical Learning Theory

Let $p^*(y|x)$ be a predictive by some method.

Training Loss: $T = -\frac{1}{n} \sum_{i=1}^{n} \log p^*(Y_i|X_i)$

Generalization Loss: $G = -E_{xy} \left[ \log p^*(Y|X) \right]$

Free energy: $F = -\log \int \phi(w) \exp(-nL_n(w)) \, dw$

Main Purpose of Statistical Learning Theory

In ML, MAP, and Bayes, clarify the distributions of $T$, $G$, and $F$. 
Summary 1

The purpose of the statistical learning theory is to clarify the distributions of training loss, Generalization loss, and free energy.
2 Regular Theory
Regular Case

\[ L(w) = - \mathbb{E}_{xy}[ \log p(Y|X,w) ] \]

\( w_0 \) is the parameter that minimizes \( L(w) \).

If \( L(w) \) can be approximated by a positive definite quadratic form in a neighborhood of \( w_0 \), then statistical estimation is called regular.
In regular cases, the distributions of MLE and MAP and the posterior distribution converge to \( w_0 \), when sample size \( n \) tends to infinity, resulting that the regular statistical theory was established in 1970.

**Definition.** Positive definite matrices \( I \) and \( J \) are defined.

\[
I = \mathbb{E}_{xy}[ \mathcal{J} \log p(Y|X,w_0) (\mathcal{J} \log p(Y|X,w_0))^T ]
\]

\[
J = - \mathbb{E}_{xy} [ \mathcal{J}^2 \log p(Y|X,w_0) ]
\]

**Remark.** If \( q(y|x) = p(y|x,w_0) \), then \( I = J \).
Theorem. In a regular case, the followings hold, where $d$ is the dimension of the parameter $w$.

**ML, MAP**

$$E[G] = L(w_0) + \frac{tr(IJ^{-1})}{2n} + o(1/n),$$
$$E[T] = L(w_0) - \frac{tr(IJ^{-1})}{2n} + o(1/n).$$

**Bayes**

$$E[G] = L(w_0) + \frac{d}{2n} + o(1/n),$$
$$E[T] = L(w_0) + \{d -2 \cdot tr(IJ^{-1})\} / (2n) + o(1/n).$$
Regular Case: Free Energy

Theorem. In a regular case, the following holds.

\[ F = nL_n(w_0) + (d/2) \log n + O_p(1). \]
Theorem. In a regular case, the followings hold.

(1) (Akaike 1974, Takeuchi 1976) In ML, MAP, and Bayes, 
\[ \text{AIC} = 2n T_n + 2 \text{tr}(IJ^{-1}) \]. Then 
\[ E[G] = E[\text{AIC}]/(2n) + o(1/n). \]

(2) (Shwarz, 1978) 
\[ \text{BIC} = 2n L_n(w^*) + d \log n \]. Then 
\[ F = \frac{\text{BIC}}{2} + O_p(1) \].

(3) (Stone, 1977) AIC is asymptotically equivalent to leave-one-out cross validation (LOOCV).
Summary 2

In regular cases, statistical learning theory was established in 1970.
3 Neural Networks are Singular
Singular Case

\[ L(w) = - \mathbb{E}_{xy}[ \log p(Y|X,w) ] \]

If \( L(w) \) cannot be approximated by any positive definite quadratic form in a neighborhood of the minimum of \( L(w) \), then statistical estimation is called singular.
In 1990, it is well known that neural networks are singular learning machines.


Neural Networks have many singularities in parameter space.
Hierarchical structure generates singularities.

Parameter space of a neural network.
Almost all learning machines are singular.
Neural Networks are singular.

(1) The map from a parameter to a probability distribution is not one-to-one.
(2) The likelihood function can not be approximated by any quadratic form.
(3) Neither MLE nor MAP has asymptotic normality.
(4) Bayes posterior distribution can not be approximated by any normal distribution.
Summary 3

Neural Networks are Singular.
4 Algebraic Geometry
The Load from Learning theory to Algebraic Geometry.

\[ K(w) = L(w) - L(w_0), \text{ where } w_0 \text{ minimizes } L(w). \]

The set \( \{w ; K(w) = 0 \} \) contains many singularities.

… There was no statistical theory.

… There was no probability theory.

… … … We need algebraic geometry.
Algebraic Geometry

Neural Networks
Neural Networks

- Hyperfunction
- D-module
- Zeta function
- Empirical Process
- Birational Invariants
- Resolution Theorem
Resolution of Singularities (Hironaka Theorem)

In each local coordinate, $K(g(u)) = u_1^{2k_1} u_2^{2k_2} \cdots u_d^{2k_d}$

For any $K(w) \geq 0$

Parameter Set $\mathbb{R}^d$

Manifold $M$

$w = g(u)$
Singular Likelihood is made to be a standard form.

By using \( K(g(u)) = u^{2k} = u_1^{2k_1} u_2^{2k_2} \cdots u_d^{2k_d} \),

\[
L_n(w) = -\frac{1}{n} \sum_{i=1}^{n} \log p(Y_i|X_i,w)
\]
is made to be

\[
nL_n(g(u)) - n L_n(w_0) = nu^{2k} - n^{1/2} u^k \xi_n(u),
\]

where \( \xi_n(u) \) converges to a Gaussian process in distribution.
Summary 4

By using resolution theorem, any singular log likelihood function can be made to be a common standard form on a manifold.
5 General Theory

General theory contains regular theory as a special case.
General Theory: MLE and MAP

Theorem. Assume that the parameter set is compact.

\[ \mu = E_\xi \left[ \max (0, \xi(u^*))^2 \right], \]

\[ u^* = \arg \max_u \left\{ (0, \xi_n(u))^2 / 4 + \log \varphi(g(u)) \right\}, \]

where \( \max_u \) shows the maximum value on \( \{u : K(g(u))=0\} \).
For MLE, \( \varphi(w) \) is set to be a constant.

Theorem. In ML and MAP,

\[ E[T] = L(w_0) - \mu / n + o(1/n), \]
\[ E[G] = L(w_0) + \mu / n + o(1/n) \]

General Theory : Bayes

Definition. Constants $\lambda$, $m$, $\nu$ are defined as follows.

- **RLCT**
  \[ \lambda = \min_{\text{Local}} \min_{j=1,2,\ldots,d} \frac{(h_j+1)}{(2k_j)} \]

- **Multiplicity**
  \[ m = \text{the number of } j \text{ that attain the maximum.} \]

- **Singular Fluctuation**
  \[ \nu = \mathbb{E}_{\xi}[<\xi(u)>] / 2 \]

Theorem. In Bayesian estimation,

\[ \mathbb{E}[G] = L(w_0) + \frac{\lambda}{n} + o(1/n) \]
\[ \mathbb{E}[T] = L(w_0) + \{\lambda - \nu\} / n + o(1/n) \]

General Theory : Free Energy

Theorem. The following holds.

\[ F = n \ln(w_0) + \lambda \log n + (m-1) \log \log n + O_p(1). \]


See also,
Theorem. Even in singular cases, the followings hold.

1. By the definition, $\text{WAIC} = T + \frac{1}{n} \sum_{i=1}^{n} V_{w}[\log p(X_i|w)]$, it follows that $E[G] = E[\text{WAIC}] + o(1/n)$.

2. By the definition, $\text{WBIC} = E_{w}^{(1/\log n)} [n \log(w)]$, it follows that $F = \text{WBIC} + o_{p}(\log n)$.

3. (Drton.et.al. 2017) By estimating $\lambda$

   $sBIC = n\log(w^*) + \lambda \log n$. Then $F = sBIC + o_{p}(\log n)$.

4. WAIC is asymptotically equivalent to LOOCV.
Information Criteria for Singular Models


(2) Sumio Watanabe, A widely applicable Bayesian information criterion. JMLR, 2013, pp.867-897.


An Experiment

Model: \[ p(y|x,a,b) = \left(\frac{1}{2\pi}\right)^{1/2} \exp\left(-\frac{1}{2}(y-a \tanh(bx))^2\right) \]

Prior: \( \varphi(a,b) \propto 1 \)

True: \( q(y|x)=p(y|x,0,0) \). True \( q(x) \) is the uniform on [-2,2].

In this case, \( \lambda = 1, m=2 \).

For \( n = 20, \ldots, 450 \),

BIC
WBIC
F
Theory
were compared.
Neural Networks

True input, hidden, output: 10, 5, 10

Candidates: 10, (1, 3, 5, 7, 9), 10

\[ n = 200 \]
\[ n_{\text{test}} = 1000 \]

Posterior was made by Langevin eq.
Experimental results for 10 trials

- Generalization Loss
- WAIC
- LOOCV
- AIC
Recent Advances of Singular Learning Theory

(5) Miki Aoyagi, Sumio Watanabe. Stochastic Complexities of Reduced Rank Regression in Bayesian Estimation, Neural Networks, No. 18, pp.924-933, 2005.
Summary 5

General theory which contains both singular and regular cases was established by using algebraic geometry.

Singularities make the generalization error very small. Neural Networks utilize singularities.
6 Conclusion

Generalization problem of neural networks was clarified by algebraic geometry.