A Singular Limit Theorem in Statistical Learning Theory

Sumio Watanabe
P&I Lab, Tokyo Institute of Technology

1. Statistical Learning

Let X be an $\mathbb{R}^N$ valued random variable which is subject to the probability distribution $q(x)dx$. Assume that $D_n = (X_1, X_2, ..., X_n)$ is a set of random variables which are independently subject to the same probability distribution as X. A statistical model $p(x|w)$ is defined as a probability density function of $x \in \mathbb{R}^N$ for a given parameter $w \in W \subset \mathbb{R}^d$. Let $\varphi(w)dw$ be a probability distribution on an open set W with compact support. The *a posteriori* distribution with the inverse temperature $\beta > 0$ is defined by

$$p(w|D_n)dw = \frac{1}{Z}\ \exp(-\beta H_n(w))\ \varphi(w)\ dw$$

where $H_n(w) = -\sum_{i=1}^n \log p(X_i|w)$ and Z is a normalizing constant. Let $E_w[\quad]$ be the expectation value using $p(w|D_n)dw$. The generalization error G and the training error T are respectively defined by

$$\begin{aligned} G &= -E_X\Big[\log E_w[p(X|w)]\Big], \\ T &= -\frac{1}{n}\sum_{i=1}^n \log E_w[p(X_i|w)]. \end{aligned}$$

In this report, we show that G and T are asymptotically determined by two birational invariants. Let $f(x,w) = \log(q(x)/p(x|w))$. Also let $S = -E_X[\log q(X)]$ and $S_n = -(1/n)\sum_i \log q(X_i)$. Then $K(w) = \int q(x)f(x,w)dx$ is a nonnegative function and

$$\begin{aligned} G &= S - E_X\Big[\log E_w[\exp(-f(X,w))]\Big], \\ T &= S_n - \frac{1}{n}\sum_{i=1}^n \log E_w[\exp(-f(X_i,w))]. \end{aligned}$$

Therefore asymptotic behaviors of G and T are given by the limit theorem of the average and empirical free energies. In statistical learning theory, the set $\{w \in W\ ;\ K(w) = 0\}$ is a nonempty analytic set with singularities in general, resulting that $\exp(-\beta H_n(w))$ cannot be approximated by any gaussian distribution.

2. Two Birational Invariants

Let $L^s(q)$ $(s \geq 2)$ be a real Banach space

$$L^s(q) = \{f(x)\ ;\ \int |f(x)|^s q(x)dx < \infty\}.$$

Assume that $w \mapsto f(x,w)$ is an $L^s(q)$-valued analytic function on W. By using resolution of singularities, there exist a manifold $\mathcal{M}$ and a real analytic map $g : \mathcal{M} \to W$ such that, in each local coordinate of $\mathcal{M}$,

$$\begin{aligned} K(g(u)) &= u^{2k} \equiv u_1^{2k_1}u_2^{2k_2}\cdots u_d^{2k_d}, \\ \varphi(g(u))|g'(u)| &= u^k\phi(u) \equiv u_1^{h_1}u_2^{h_2}\cdots u_d^{h_d}\phi(u), \end{aligned}$$

where $k = (k_1, k_2, ..., k_d)$ and $h = (h_1, h_2, ..., h_d)$ are sets of nonnegative integers, $|g'(u)|$ is the Jacobian determinant of $w = g(u)$, and $\phi(u) > 0$. Let $\{\alpha\}$ be a set of local coordinates of $\mathcal{M}$. The *log canonical threshold* λ is defined by

$$\lambda = \min_{\alpha} \min_{j=1}^{d} \Big(\frac{h_j + 1}{2k_j} \Big),$$

where we put $(h_j + 1)/k_j = \infty$ for $k_j = 0$. Let $\{\alpha^*\}$ be the set of all local coordinates in which the above minimum is attained. Since $f(x, g(u))$ is an analytic function on $\mathcal{M}$, there exists an $L^s(q)$-valued analytic function $a(x, u)$ such that $f(x, g(u)) = a(x, u)u^k$. Let $\xi(u)$ be a gaussian field on $\mathcal{M}$ which is uniquely determined by its expectation and covariance,

$$E_\xi[\xi(u)] = 0, \qquad E_\xi[\xi(u)\xi(v)] = E_X[a(X, u)a(X, v)] - E_X[a(X, u)]E_X[a(X, v)].$$

The *singular fluctuation* ν is defined by

$$\nu = \frac{\beta}{2} E_\xi E_X \Big[\langle a(X, u)^2 t \rangle - \langle a(X, u)\sqrt{t} \rangle^2 \Big],$$

where $\langle \ \rangle$ shows the expetation value over a renormalized *a posteriori* distribution,

$$\langle F(u, t) \rangle = \frac{\sum_{\alpha^*} \int dt \int du^* \ F(u, t) \ t^{\lambda - 1} \exp(-\beta t - \beta \sqrt{t} \xi(u))}{\sum_{\alpha^*} \int dt \int du^* \ t^{\lambda - 1} \exp(-\beta t - \beta \sqrt{t} \xi(u))},$$

where du^* is a measure whose support is contained in the set $\{u \in \mathcal{M}; K(g(u)) = 0\}$. Note that neither λ nor ν depends on the choice of desingularization $(\mathcal{M}, g)$, hence they are birational invariants.

Theorem. The following asymptotic expansions hold as $n \to \infty$,

$$\begin{aligned} E[G] &= S + \Big(\frac{\lambda - \nu}{\beta} + \nu \Big) \frac{1}{n} + o(\frac{1}{n}), \\ E[T] &= S + \Big(\frac{\lambda - \nu}{\beta} - \nu \Big) \frac{1}{n} + o(\frac{1}{n}). \end{aligned}$$

3. Application to statistics

The functional variance V is defined by

$$V = \sum_{i=1}^{n} \Big\{ E_w[(\log p(X_i|w))^2] - E_w[\log p(X_i|w)]^2 \Big\}.$$

Then $E[V] \to 2\nu/\beta$. Hence we can estimate $E[G]$ from $E[T]$ and $E[V]$ without any knowledge of $q(x)$, by *equation of state in statistical learning*,

$$E[G] = E[T] + \frac{\beta}{n} E[V] + o(\frac{1}{n}).$$

This equation holds for an arbitrary $(q(x), p(x|w), \varphi(w))$, which can be understood as the equation of state for Boltzmann distribution $p(w|D_n)$ with random Hamiltonian $H_n(w)$.

References

[1] S. Watanabe, "Algebraic geometry and statistical learning theory," Cambridge University Press, 2009.