Learning coefficients for hierarchical learning models in Bayesian estimation

Miki Aoyagi, Keita Yamazaki and Tomoyasu Ohba
College of Science & Technology, Nihon University,
1–8–14, Surugadai, Kanda, Chiyoda-ku, Tokyo 101–8308 Japan,
aoyagi.miki@nihon-u.ac.jp

Abstract
Recently, singular learning theory has been analyzed using algebraic geometry as its basis. It is essential to determine the normal crossing divisors of learning machine singularities through a blowing-up process to observe the behaviors of state probability functions in learning theory. In this paper, we investigate learning coefficients for multi-layered neural networks with linear units, especially when dealing with a large number of layers in Bayesian estimation. We make use of the valuable results obtained in the paper [9], which provide the main terms for Bayesian generalization error and the average stochastic complexity (free energy). These terms are widely employed in numerical experiments, such as in information criteria.

Key Words: Resolution of singularities, learning coefficients, singular models, linear neural networks

1 Introduction

Let $q(x,y)$ be a true probability density function of variables, x, y, and let $(x,y)^n := \{(x_i,y_i)\}_{i=1}^n$ be n training samples selected independently and identically from $q(x,y)$. Consider a learning model that is written in probabilistic form as $p(x,y|w)$, where $w \in W \subset \mathbf{R}^d$ is a parameter.

Suppose that the purpose of the learning system is to estimate an unknown true density function $q(x,y)$ from $(x,y)^n$ using $p(x,y|w)$ in Bayesian estimation. Let $\psi(w)$ be an *a priori* probability density function on a parameter set W and $p(w|(x,y)^n)$ be the *a posteriori* probability density function,

$$p(w|(x,y)^n) = \frac{1}{Z_n(\beta)}\psi(w)\prod_{i=1}^n p(x_i,y_i|w)^\beta,$$

where

$$Z_n(\beta) = \int_W \psi(w) \prod_{i=1}^{n} p(x_i, y_i|w)^\beta \mathrm{d}w,$$

for an inverse temperature β; we typically set $\beta = 1$.

Define the average the average log loss function $L(w)$ by $L(w) = -E_{x,y}[\log p(x, y|w)]$ and the set of optimal parameters W_0 by

$$W_0 = \{w \in W | L(w) = \min_{w' \in W} L(w')\}.$$

Assume that its log likelihood function has relatively finite variance,

$$E_{x,y}[\log \frac{p(x, y|w_0)}{p(x, y|w)}] \geq cE_{x,y}[(\log \frac{p(x, y|w_0)}{p(x, y|w)})^2], \quad w_0 \in W_0, w \in W,$$

for a constant $c > 0$. Then, we have a unique probability density function $p_0(x, y) = p(x, y|w_0)$ for all $w_0 \in W_0$.

Define

$$E_w^\beta[g(w)] = \frac{\int dw g(w)\psi(w) \prod_{i=1}^n p(x_i, y_i|w)^\beta}{\int dw \psi(w) \prod_{i=1}^n p(x_i, y_i|w)^\beta},$$

and

$$V_w^\beta[g(w)] = E_w^\beta[g(w)^2] - E_w^\beta[g(w)]^2.$$

We then have the predictive density function $p(x, y|(x, y)^n) = E_w^\beta[p(x, y|w)]$, which is the average inference of the Bayesian density function. Let

$$f(x, y|w) = \log \frac{p_0(x, y)}{p(x, y|w)}$$

from which the Kullback function is defined as

$$K(w) = E_{x,y}[f(x, y|w)].$$

Applying Hironaka's Theorem [15] to the function $K(w)$, we obtain the proper analytic map π from manifold Y to neighborhood W,

$$K(\pi(u)) = u_1^{2k_1} u_2^{2k_2} \cdots u_d^{2k_d},$$

where $(u_1, \cdots, u_d)$ is a local analytic coordinate system on $U \subset Y$. Additionally, there exist analytic functions $a(x, y|u)$ and $b(u) \neq 0$ such that

$$f(x, y|\pi(u)) = u_1^{k_1} u_2^{k_2} \cdots u_d^{k_d} a(x, y|u),$$

and

$$\pi'(u)\psi(\pi(u)) = u_1^{h_1} u_2^{h_2} \cdots u_d^{h_d} b(u).$$

Let

$$\xi_n(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{u_1^{k_1} u_2^{k_2} \cdots u_d^{k_d} - a(x_i, y_i|\pi(u))\},$$

then, we have an empirical process $K_n(\pi(u))$ such that

$$nK_n(\pi(u)) = \sum_{i=1}^{n} f(x_i, y_i|\pi(u))$$

$$= nu_1^{2k_1} u_2^{2k_2} \cdots u_d^{2k_d} - \sqrt{n} u_1^{k_1} u_2^{k_2} \cdots u_d^{k_d} \xi_n(u).$$

We introduce learning coefficients

$$\lambda = \min_u \min_{1 \le j \le d} \frac{h_j + 1}{2k_j},$$

and its order

$$\theta = \max_u \mathrm{Card}(\{j : \frac{h_j + 1}{2k_j} = \lambda\}),$$

where $\mathrm{Card}(S)$ denotes the cardinality of a set S.

Without loss of generality, we can assume that

$$\lambda = \frac{h_1 + 1}{2k_1} = \frac{h_2 + 1}{2k_2} = \cdots = \frac{h_\theta + 1}{2k_\theta} < \frac{h_j + 1}{2k_j} \ (\theta + 1 \le j \le d).$$

Let

$$\Omega(w)dw = \frac{\psi(w) \prod_{i=1}^{n} p(x_i, y_i|w)^\beta}{\prod_{i=1}^{n} p_0(x_i, y_i)^\beta} dw.$$

Then, we have

$$\begin{aligned} \Omega(w)dw &= \frac{(\log n)^{\theta-1}}{n^\lambda} \int_0^\infty dt\, t^{\lambda-1} \exp(-t + \sqrt{\beta t}\xi_n(u)) du^* \\ &\quad + o_p\left(\frac{(\log n)^{\theta-1}}{n^\lambda}\right), \end{aligned}$$

where $\mu_j = -2\lambda k_j + h_j$,

$$du^* = \frac{\prod_{i=1}^{\theta} \delta(u_i) \prod_{j=\theta+1}^{d} u_j^{\mu_j}}{(\theta - 1)! \prod_{i=1}^{\theta} (2k_i)} b(u) du,$$

and $\delta(u)$ is Dirac's delta function.

Let ν be a singular fluctuation,

$$\nu = \frac{1}{2} E_\xi \left[\frac{\int_0^\infty dt \int du^* \xi(u) t^{\lambda - 1/2} \exp(-\beta t + \beta\sqrt{t}\xi(u))}{\int_0^\infty dt \int du^* t^{\lambda-1} \exp(-\beta t + \beta\sqrt{t}\xi(u))} \right],$$

where $\xi(u)$ is a convergence in distribution for $\xi_n(u)$ and a random variable of a Gaussian process with mean zero, and $E_\xi[\xi(w)\xi(u)] = E_{x,y}[a(x,y|w)a(x,y|u)]$ denotes the covariance.

Let G_n be the Bayes generalization loss,

$$G_n = -\int q(x,y)\log p(x,y|(x,y)^n)dx,$$

and T_n the Bayes training loss,

$$T_n = -\frac{1}{n}\sum_{i=1}^{n}\log p(x_i,y_i|(x,y)^n).$$

Watanabe [23, 24, 27, 29] proved the following relations,

$$\begin{aligned} E[G_n] &= L(w_0) + \frac{1}{n}\left(\frac{\lambda-\nu}{\beta}+\nu\right) + o(\frac{1}{n}), \\ E[T_n] &= L(w_0) + \frac{1}{n}\left(\frac{\lambda-\nu}{\beta}-\nu\right) + o(\frac{1}{n}). \end{aligned}$$

Using the above, we have in the Bayesian estimation approach model selection methods such as the widely-applicable information criterion (WAIC) [1, 23, 24, 25, 26, 27, 29] and cross-validation.

(1) WAIC [26]

$$W_n = T_n + \frac{\beta}{n}\sum_{i=1}^{n}\{E_w^\beta[(\log p(x_i,y_i|w))^2] - E_w^\beta[\log p(x_i,y_i|w)]^2\},$$

(2) Cross-validation

$$C_n = -\frac{1}{n}\sum_{i=1}^{n}\log p(x_i,y_i|(x,y)^n\backslash(x_i,y_i)) \quad (n \geq 2),$$

where

$$(x,y)^n\backslash(x_i,y_i) = \{(x_1,y_1),\cdots,(x_{i-1},y_{i-1}),(x_{i+1},y_{i+1}),\cdots,(x_n,y_n)\}.$$

Then, we have

$$E[W_n] = E[G_n] + o(\frac{1}{n}), \;\; E[C_n] = E[G_n] + o(\frac{1}{n}),$$

by using

$$\beta\sum_{i=1}^{n}\{E_w^\beta[(\log p(x_i,y_i|w))^2] - E_w^\beta[\log p(x_i,y_i|w)]^2\} \to 2\nu.$$

These relations show that the WAIC and cross-validation can estimate the Bayesian generalization loss G_n from data $(x, y)^n$ and learning model $p(x, y|w)$ without any knowledge of the true probability density function.

In addition, based on the free energy,

$$\begin{aligned} F_n(\beta) &= -\frac{1}{\beta}\log\int\prod_{i=1}^{n} p(x_i, y_i|w)^{\beta}\psi(w)dw \\ &= nL_n(w_0) + \frac{\lambda}{\beta}\log(n) - \frac{\theta-1}{\beta}\log\log(n) + o_p(1) \end{aligned}$$

which was shown by Watanabe [26], we have the two model-selection methods, namely, the "widely applicable Bayesian information criterion" (WBIC) [28] and "singular Bayesian information criterion" (sBIC) [12]. sBIC uses the learning coefficients very effectively with a fix point equation system of marginal likelihoods, whereas in practice WAIC, the cross-validation, and WBIC do not need these coefficients.

The learning coefficients are known as log canonical thresholds in algebraic geometry. Theoretically, their values are obtained using Hironaka's Theorem. However, these thresholds are studied mainly over the complex field or algebraically closed fields in algebraic geometry and algebraic analysis [17, 19, 16]. There are many differences for real and complex fields. For example, log canonical thresholds over the complex field are less than one, whereas those over the real field are not necessarily so. Obtaining these thresholds for learning models is difficult for several reasons, such as degeneration with respect to their Newton polyhedra and non-isolation of their singularities [14]. Therefore, it is of interest in various fields, even in mathematics, to obtain these thresholds.

Our purpose in this paper is to obtain λ and θ for deep-layered linear neural networks.

In recent studies, we obtained exact values or bounded values of the learning coefficients for Vandermonde matrix-type singularities, which are related to the three-layered neural networks and normal mixture models, among others [10, 2, 5, 7, 8]. We have also exact values for the restricted Boltzmann machine [6]. Additionally, Rusakov and Geiger [21, 22] and Zwiernik [30], respectively, obtained the learning coefficients for naive Bayesian networks and directed tree models with hidden variables. Drton et al. [13] considered these coefficients for the Gaussian latent tree and forest models.

2 Log canonical threshold

We denote constants by superscript $*$, for example, a^*, b^*, and w^*.

Definition 1 *Let f be an analytic function in neighborhood U of w^*, and ψ be a C^∞ function on U that is also analytic in a neighborhood of w^* with compact support. Define the log canonical threshold*

$$c_{w^*}(f, \psi) = \sup\{c : |f|^{-c} \textit{ is locally } L^2 \textit{ in a neighborhood of } w^*\}$$

over the complex field $\mathbf{C}$ *and*

$$c_{w^*}(f,\psi) = \sup\{c : |f|^{-c} \text{ is locally } L^1 \text{ in a neighborhood of } w^*\}$$

over the real field $\mathbf{R}$*. The value* $c_{w^*}(f,\psi)$ *is equal to the largest pole of the zeta function* $\int_U |f|^{kz}\psi(w)dw$ *for* $z \in \mathbf{C}$*, where* $k = 2$ *over the complex field and* $k = 1$ *over the real field. Let* $\theta_{w^*}(f,\psi)$ *be its order.*

If $\psi(w^*) \neq 0$*, then denote* $c_{w^*}(f) = c_{w^*}(f,\psi)$ *and* $\theta_{w^*}(f) = \theta_{w^*}(f,\psi)$ *because the log canonical threshold and its order are independent of* ψ*.*

For ideal I*, generated by real analytic functions* $f_1, \cdots, f_m$ *in a neighborhood of* w^**, define* $c_{w^*}(I) = c_{w^*}(f_1^2 + \cdots + f_m^2)$*.*

Here, $c_{w^*}(I)$ for ideal I is well-defined by Lemma 1.

Lemma 1 ([3, 4, 18]) *Let* U *be a neighborhood of* $w^* \in \mathbf{R}^d$*. Consider the ring of analytic functions on* U*. Let* $\mathcal{J}$ *be the ideal generated by* $f_1, \ldots, f_n$*, which are analytic functions defined on* U*.* (1) *If* $g_1^2 + \ldots + g_m^2 \leq f_1^2 + \cdots + f_n^2$*, then* $c_{w^*}(g_1^2 + \cdots + g_m^2) \leq c_{w^*}(f_1^2 + \cdots + f_n^2)$*.*

(2) *If* $g_1, \ldots, g_m \in \mathcal{J}$*, then* $c_{w^*}(g_1^2 + \cdots + g_m^2) \leq c_{w^*}(f_1^2 + \cdots + f_n^2)$*. In particular, if* $g_1, \ldots, g_m$ *generate ideal* $\mathcal{J}$*, then* $c_{w^*}(f_1^2 + \cdots + f_n^2) = c_{w^*}(g_1^2 + \cdots + g_m^2)$*.*

The following lemma is also used in the proofs.

Lemma 2 ([6]) *Let* $\mathcal{J}, \mathcal{J}'$ *be the ideals generated by* $f_1(w)$*, ...,* $f_n(w)$ *and* $g_1(w')$*, ...,* $g_m(w')$*, respectively. If* w *and* w' *are different variables, then*

$$c_{(w^*,w'^*)}(f_1^2 + \cdots + f_n^2 + g_1^2 + \cdots + g_m^2) = c_{w^*}(f_1^2 + \cdots + f_n^2) + c_{w'^*}(g_1^2 + \cdots + g_m^2).$$

The learning coefficient λ is the log canonical threshold of the Kullback function (relative entropy) over the real field.

Define the norm of a matrix $C = (c_{ij})$ as $||C|| = \sqrt{\sum_{i,j} |c_{ij}|^2}$.

Definition 2 *For a matrix* C*, let* $\langle C \rangle$ *be the ideal generated by all elements of* C*.*

3 Multiple-layered linear neural networks

In the paper [9], the learning coefficients for multiple-layered neural networks with linear units were obtained.

Define matrices $A^{(s)}$ of size $H^{(s)} \times H^{(s+1)}$ for $s = 1, \ldots, L$,

$$A^{(s)} = (a_{ij}^{(s)}), \qquad (1 \leq i \leq H^{(s)}, 1 \leq j \leq H^{(s+1)}).$$

Let W be the set of parameters

$$W = \{w = \{A^{(s)}\}_{1 \leq s \leq L} \mid \ A^{(s)} \text{ is an } H^{(s)} \times H^{(s+1)} \text{ matrix } \}.$$

Denote the input value by $x \in \mathbf{R}^{H^{(L+1)}}$ with probability density function $q(x)$ and output value $y \in \mathbf{R}^{H^{(1)}}$ for the multiple-layered neural network with linear units, which is given by

$$y = \prod_{s=1}^{L} A^{(s)} x + (\text{noise}),$$

with Gaussian noise. Consider the statistical model

$$p(y|x,w) = \frac{1}{(\sqrt{2\pi})^{H^{(1)}}} \exp(-\frac{1}{2}||y - \prod_{s=1}^{L} A^{(s)} x||^2), \quad p(x,y|w) = p(y|x,w)q(x).$$

The model has $H^{(1)}$ input units, $H^{(L+1)}$ output units, and $H^{(s)}$ hidden units in each hidden layer. Let

$$w^* = \{A^{*(s)}\}_{1 \leq s \leq L},$$

be the true parameter. Assume that the true density function

$$q(y|x) = \frac{1}{(\sqrt{2\pi})^{H^{(1)}}} \exp(-\frac{1}{2}||y - \prod_{s=1}^{L} A^{*(s)} x||^2), \quad q(x,y) = q(y|x)q(x),$$

which is included in the learning model. Moreover, assume that the *a priori* probability density function $\varphi(w)$ is a $C^\infty-$ function with compact support W, satisfying $\varphi(w^*) > 0$. Then, λ and θ for the model corresponding to the log canonical threshold $\lambda_{w^*}(||\prod_{s=1}^{L} A^{(s)} - \prod_{s=1}^{L} A^{*(s)}||^2)$ and its order θ are as follows.

Definition 3 *Let r be the rank of $\prod_{s=1}^{L} A^{*(s)}$ and $M^{(s)} = H^{(s)} - r$ for $s = 1, \ldots, L+1$. Define $\mathcal{M} \subset \{1, \ldots, L+1\}$ such that*

$$\begin{aligned}
&\ell = \mathrm{Card}(\mathcal{M}) - 1, \\
&\mathcal{M} = \{S_1, \ldots, S_{\ell+1}\}, \\
&M^{(S_j)} < M^{(s)} \textit{ for } S_j \in \mathcal{M} \textit{ and } s \notin \mathcal{M}, \\
&\sum_{k=1}^{\ell+1} M^{(S_k)} \geq \ell M^{(s)} \textit{ for } s \in \mathcal{M} \\
&\sum_{k=1}^{\ell+1} M^{(S_k)} < \ell M^{(s)} \textit{ for } s \notin \mathcal{M}.
\end{aligned}$$

Let M be the integer such that

$$M-1<\frac{\sum_{k=1}^{\ell+1}M^{(S_k)}}{\ell}\leq M,$$

and

$$a=\sum_{k=1}^{\ell+1}M^{(S_k)}-(M-1)\ell.$$

Theorem 1 ([9]) *We have*

$$\begin{aligned}\lambda &= \frac{-r^2+r(H^{(1)}+H^{(L+1)})}{2}+\frac{a(\ell-a)}{4\ell}\\ &\quad -\frac{\ell(\ell-1)}{4}(\frac{\sum_{j=1}^{\ell+1}M^{(S_j)}}{\ell})^2+\frac{1}{2}\sum_{1\leq i<j\leq \ell+1}M^{(S_i)}M^{(S_j)}\\ &= \frac{-r^2+r(H^{(1)}+H^{(L+1)})}{2}+\frac{a(\ell-a)}{4\ell}\\ &\quad +\frac{1}{4\ell}(\sum_{j=1}^{\ell+1}M^{(S_j)})^2-\frac{1}{4}\sum_{j=1}^{\ell+1}(M^{(S_j)})^2\\ &= \frac{-r^2+r(H^{(1)}+H^{(L+1)})}{2}+\frac{Ma+(M-1)\sum_{j=1}^{\ell+1}M^{(S_j)}}{4}\\ &\quad -\frac{1}{4}\sum_{j=1}^{\ell+1}(M^{(S_j)})^2\end{aligned}$$

and

$$\theta=a(\ell-a)+1.$$

Note that λ is a decreasing sequence with ℓ from the proof of Theorem [9].

Lemma 3 *Let $\tilde{M}$ be the integer such that $\tilde{M}\leq\frac{\sum_{k=1}^{\ell+1}M^{(S_k)}}{\ell+1}<\tilde{M}+1$. Fix $\tilde{M}$ and if ℓ is large enough to satisfy*

$$\frac{\sum_{k=1}^{\ell+1}M^{(S_k)}}{\ell+1}+\frac{\sum_{k=1}^{\ell+1}M^{(S_k)}}{\ell(\ell+1)}=\frac{\sum_{k=1}^{\ell+1}M^{(S_k)}}{\ell}<\tilde{M}+1,$$

then

$$M^{(S_j)}=\min_{1\leq s\leq L+1}M^{(s)}\ \textit{for all}\ S_j\in\mathcal{M}$$

and $\tilde{M}=M^{(S_1)}=\cdots=M^{(S_{\ell+1})}$.

(Proof)

Since $M^{(S_j)} \leq \frac{\sum_{k=1}^{\ell+1} M^{(S_k)}}{\ell} < \tilde{M} + 1$, we have

$$M^{(S_j)} \leq \tilde{M} \leq \frac{\sum_{k=1}^{\ell+1} M^{(S_k)}}{\ell+1}.$$

Therefore,

$$\frac{\sum_{k=1}^{\ell+1} M^{(S_k)}}{\ell+1} \leq \tilde{M} \leq \frac{\sum_{k=1}^{\ell+1} M^{(S_k)}}{\ell+1}.$$

That is, we have

$$M^{(S_j)} = \frac{\sum_{k=1}^{\ell+1} M^{(S_k)}}{\ell+1} = \tilde{M}.$$

The end of Proof

By Lemma 3, we have the followings.

Theorem 2 *Let* $M_{\min} = \min_{1 \leq s \leq L+1} M^{(s)}$. *Assume that* $M^{(S_1)} = M^{(S_2)} = \cdots = M^{(S_{\ell+1})} = M_{\min}$ *and* $\ell > M_{\min}$. *then we have* $a = M_{\min}$,

$$\lambda = \frac{-r^2 + r(H^{(1)} + H^{(L+1)})}{2} + \frac{M_{\min}^2 + M_{\min}}{4}$$

and

$$\theta = M_{\min}(\ell - M_{\min}) + 1.$$

4 Conclusions

In the paper [9], we have determined the precise values for the learning coefficients of multi-layered linear neural networks, thereby extending the results presented in the paper [11]. Utilizing these coefficients, we establish Theorem 2 for cases involving a large number of layers. This theorem demonstrates that the learning coefficient λ is exactly equal to

$$\frac{-r^2 + r(H^{(1)} + H^{(L+1)})}{2} + \frac{(H_{\min} - r)^2 + H_{\min} - r}{4},$$

where $H^{(s)}$ represents the number of perceptrons in each layer, r is the rank of its true probability density function, and $H_{\min}$ is the minimum among the values of $H^{(s)}$ for $1 \leq s \leq L+1$. Furthermore, this theorem reveals that when the number of layers exceeds $H_{\min} - r$, the value of λ remains constant and attains its minimum for a smaller number of layers than $H_{\min} - r$. This seems to explain phenomena like double descent [20] in machine learning.

Acknowledgments

This research was funded by Grants-in-Aid for Scientific Research - KAKENHI - under grant number 18K11479.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, 19:716–723, 1974.

[2] M. Aoyagi. The zeta function of learning theory and generalization error of three layered neural perceptron. *RIMS Kokyuroku, Recent Topics on Real and Complex Singularities*, 1501:153–167, 2006.

[3] M. Aoyagi. Log canonical threshold of Vandermonde matrix type singularities and generalization error of a three layered neural network. *International Journal of Pure and Applied Mathematics*, 52(2):177–204, 2009.

[4] M. Aoyagi. A Bayesian learning coefficient of generalization error and Vandermonde matrix-type singularities. *Communications in Statistics - Theory and Methods*, 39(15):2667–2687, 2010.

[5] M. Aoyagi. Consideration on singularities in learning theory and the learning coefficient. *Entropy*, 15(9):3714–3733, 2013.

[6] M. Aoyagi. Learning coefficient in Bayesian estimation of restricted Boltzmann machine. *Journal of Algebraic Statistics*, 4(1):30–57, 2013.

[7] M. Aoyagi. Learning coefficient of Vandermonde matrix-type singularities in model selection. *Entropy (Information Theory, Probability and Statistics)*, 21(6-561):1–12, 2019.

[8] M. Aoyagi. Learning coefficients and information criteria. *Frontiers in Artificial Intelligence and Applications*, pages 351–362, 2019.

[9] M. Aoyagi. Consideration on the learning efficiency of multiple-layered neural networks with linear units. *Preprint*, 2023.

[10] M. Aoyagi and S. Watanabe. Resolution of singularities and the generalization error with Bayesian estimation for layered neural network. *IEICE Trans. J88-D-II*, 10:2112–2124, 2005a.

[11] M. Aoyagi and S. Watanabe. Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Networks*, 18:924–933, 2005b.

[12] M. Drton. Conference lecture: Bayesian information criterion for singular models. *Algebraic Statistics 2012 in the Alleghenies at The Pennsylvania State University, http://jasonmorton.com/aspsu2012/*, 2012.

[13] M. Drton, S. Lin, L. Weihs, and P. Zwiernik. Marginal likelihood and model selection for Gaussian latent tree and forest models. *Bernoulli*, 23(2):1202–1232, 2017.

[14] W. Fulton. *Introduction to toric varieties, Annals of Mathematics Studies*. Princeton University Press, Princeton, NJ, USA, 1993.

[15] H. Hironaka. Resolution of singularities of an algebraic variety over a field of characteristic zero. *Annals of Math*, 79:109–326, 1964.

[16] M. Kashiwara. B-functions and holonomic systems. *Inventions Math.*, 38:33–53, 1976.

[17] J. Kollár. Singularities of pairs. *Algebraic geometry-Santa Cruz 1995, Proc. Symp. Pure Math., American Mathematical Society, Providence, RI,*, 62:221–287, 1997.

[18] S. Lin. Asymptotic approximation of marginal likelihood integrals. *(preprint)*, 2010.

[19] M. Mustata. Singularities of pairs via jet schemes. *J. Amer. Math. Soc.*, 15:599–615, 2002.

[20] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever. Deep double descent: Where bigger models and more data hurt. *ICLR2020, https://arxiv.org/pdf/1912.02292.pdf*, 2020.

[21] D. Rusakov and D. Geiger. Asymptotic model selection for naive Bayesian networks. *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 438–445, 2002.

[22] D. Rusakov and D. Geiger. Asymptotic model selection for naive Bayesian networks. *Journal of Machine Learning Research*, 6:1–35, 2005.

[23] S. Watanabe. Algebraic analysis for nonidentifiable learning machines. *Neural Computation*, 13(4):899–933, 2001a.

[24] S. Watanabe. Algebraic geometrical methods for hierarchical learning machines. *Neural Networks*, 14(8):1049–1060, 2001b.

[25] S. Watanabe. Algebraic geometry of learning machines with singularities and their prior distributions. *Journal of Japanese Society of Artificial Intelligence*, 16(2):308–315, 2001c.

[26] S. Watanabe. *Algebraic Geometry and Statistical Learning Theory*, volume 25. Cambridge University Press, New York, USA, 2009.

[27] S. Watanabe. Equations of states in singular statistical estimation. *Neural Networks*, 23(1):20–34, 2010.

[28] S. Watanabe. A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, (14):867–897, 2013.

[29] S. Watanabe. *Mathematical Theory of Bayesian Statistics*. Taylor and Francis, 2018.

[30] P. Zwiernik. An asymptotic behavior of the marginal likelihood for general Markov models. *Journal of Machine Learning Research*, 12:3283–3310, 2011.