



Title	Learning coefficients in semiregular models
Author(s)	車谷, 優樹
Citation	大阪大学, 2026, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/105548
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

<https://ir.library.osaka-u.ac.jp/>

The University of Osaka

Learning coefficients in semiregular models

YUKI KURUMADANI

MARCH 2026

Learning coefficients in semiregular models

A dissertation submitted to
THE GRADUATE SCHOOL OF ENGINEERING SCIENCE
The University of Osaka

in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY IN SCIENCE

BY
YUKI KURUMADANI

MARCH 2026

Abstract

This thesis discusses learning coefficients for singular models. In Bayesian inference, the asymptotic behavior of evaluation metrics for predictive accuracy, such as generalization loss and free energy, is determined by a model-specific rational number known as the learning coefficient (real log canonical threshold). Learning coefficients are already known for models that satisfy regularity conditions. However, for singular models that do not satisfy these conditions, specific values of learning coefficients have been provided for certain models such as reduced-rank regression. Nevertheless, a general formula for learning coefficients that broadly applies to singular models has not yet been established.

In Chapter 3, we derive a formula for the learning coefficients of statistical models that satisfy a condition slightly less stringent than the regularity condition, which we refer to as *semiregular models*. This formula reveals that a natural number m , determined by the partial derivatives of the log-likelihood ratio function f , significantly affects the learning coefficient. However, it can only be used when the Kullback-Leibler divergence $K(\theta)$ is analytic or when the set of realization parameters Θ_* is a single point. It is not applicable to more general cases.

In Chapter 4 and Chapter 5, we present two generalizations to overcome this limitation. First, we show that the formula for the learning coefficient still holds when $K(\theta)$ is not analytic. Second, we derive a formula that is applicable even when Θ_* is not a single point. As a result, we gain a systematic understanding of how the natural number m , determined by partial derivatives of f , and the dimension of Θ_* , a geometric quantity, influence the learning coefficient.

Our main results are applicable to a wide range of singular models. As a specific example, we provide upper bounds on the learning coefficients for neural networks and mixture distributions, which had not been elucidated before.

Contents

1	Introduction	1
1.1	Overview	3
1.2	Definitions and Assumptions	3
2	Background	5
2.1	Bayesian inference and generalization error	5
2.2	Real log canonical thresholds	6
2.3	Regularity	8
3	Semiregular models	9
3.1	Semiregularity	9
3.2	Derivation Method for m	11
3.3	Proof of Theorem 3.1	12
4	Extension1 : generalization to the nonanalytic case	19
5	Extension2 : generalization to the case where Θ_* has more than one point	21
5.1	Local learning coefficients at non-singular points	21
5.2	Derivation Method for m	28
5.3	Examples and Experiments	29
5.3.1	Neural Networks	29
5.3.2	Mixture Models	30
5.4	Proof of the Main Theorem	32
6	Conclusions and future work	39
A	Proof of Theorem 3.2	40
B	Derivation Method for m	46
C	Proof of Theorem 4.1	48
D	Details on Example 5.1(1)	53

List of Publications	57
Acknowledgements	58
References	59

List of Figures

3.1	Lowest-order terms in the Taylor expansion of $K(\theta)$	13
3.2	The procedure of coordinate transformations g performed in the proof of Theorem 3.1	14
5.1	Graph of learning coefficients of three-layer neural networks ($\sigma = \tanh$)	30
5.2	Graph of estimated learning coefficients of the binomial distribution	32
5.3	The procedure of coordinate transformation g performed in Example 5.5	36
5.4	The procedure of coordinate transformations g performed in the proof of the Main Theorem	37

List of Tables

5.1	Differences in the upper bound on the learning coefficient depending on the activation function	29
5.2	Learning coefficients of three-layer neural networks when $\sigma =$ $\tanh$	30
5.3	Estimated learning coefficients for the binomial distribution .	31
5.4	Normal crossings for each local coordinate	36

Chapter 1

Introduction

Hierarchical models such as neural networks, normal mixture models, and Boltzmann machines are referred to as singular learning models. These models are widely used in the analysis of real data, but their theoretical analysis has not been sufficiently developed. These models possess nonpositive definite Fisher information matrices and therefore cannot be analyzed by classic theory [28]. For example, model selection methods such as AIC [1], TIC [23], and BIC [22], which require the Fisher information matrix to be positive definite and the existence of a unique minimizer of the log loss function, cannot be applied to singular models.

Watanabe established a theory of learning coefficients using algebraic geometry for singular models [28]. However, only universal laws were presented, and further mathematical considerations are necessary to obtain specific learning coefficients.

Based on the free energy demonstrated by Watanabe [28], widely applicable Bayesian information criteria (WBIC) [30] and the singular Bayesian information criterion (sBIC) [12] have been developed. The sBIC effectively uses the learning coefficient λ and its multiplicity $\mathbf{m}$ with a fixed point equation system of marginal likelihood. Moreover, the learning coefficients λ and their multiplicities $\mathbf{m}$ provide mathematical indicators for analyzing the accuracy of model selection methods such as WAIC ([1], [24], [25], [26], [28], [29], [31]) and cross-validation in Bayesian estimation approaches.

The learning coefficient λ is equal to the real log canonical threshold of the Kullback-Leibler function in algebraic geometry, and its specific value can be obtained using the recursive blow-up process [14]. However, this result is an abstract existence theorem, and it is not clear which specific blow-ups should be performed. Furthermore, the theory of real log canonical thresholds has fewer achievements in the real number field compared to the complex number field or algebraically closed fields [16, 18]. Therefore, obtaining real log canonical thresholds is a matter of interest in various fields, including mathematics.

Aoyagi and Watanabe [11] and Aoyagi [9] obtained the specific values of the learning coefficients for reduced-rank regression models. Furthermore, Aoyagi [2, 5, 7, 8] obtained specific values or upper bounds of the learning coefficients for models such as three-layer neural networks and normal mixture models related to Vandermonde matrix-type singularities. In addition, Aoyagi [10] obtained learning coefficients for three-layer neural network models with ReLU units. Additionally, Aoyagi obtained specific values for restricted Boltzmann machines [6]. Moreover, Rusakov and Geiger [19, 20], Zwiernik [33] and Drton, Lin, Weihs, and Zwiernik [13] respectively obtained learning coefficients for naive Bayesian networks, directed tree models with hidden variables, and Gaussian latent tree and forest models. Additionally, Lau, Furman, Wang, Murfet, and Wei [17] empirically developed a method to obtain the local learning coefficient for deep linear networks.

While there is previous research on learning coefficients λ for each statistical model, a formula for learning coefficients λ that can be widely applied to singular models has not been identified.

In Chapter 3, we generalize the formula for the learning coefficient λ for regular models and obtain a formula for λ that can be applied to relatively simple singular models. Regular models refer to cases where the Fisher information matrix is positive definite and there is a formula $\lambda = d/2$ using the number of parameters d . We consider cases where the Fisher information matrix is not positive definite and its rank r satisfies $r < d$. Assuming that a certain type of linear independence holds, we derive the formula

$$\lambda = \frac{d - r + rm}{2m},$$

where m is a natural number determined by the order of differentiability of the log-likelihood ratio function. This generalizes the regular case. In particular, we clarify that this natural number m , along with the number of parameters d and the rank r of the Fisher information matrix, is a factor that determines the learning coefficient λ . We utilize this formula to obtain the learning coefficients for a specific three-layer neural networks with a fixed weight, which had previously remained unresolved. However, it can only be used when the Kullback-Leibler divergence $K(\theta)$ is analytic or when the set of realization parameters Θ_* is a single point.

In Chapter 4 and Chapter 5, we generalize this formula so that it covers a broader class of singular models. We overcome these constraints in two ways.

First, we prove that the learning coefficient formula still holds even when $K(\theta)$ is not analytic. This result makes the formula applicable to practically important nonanalytic models such as normal mixture models.

Second, we handle cases where Θ_* has a spread. From a geometric perspective, we usually know that Θ_* contains at least one nonsingular point. We show that if the statistical model satisfies certain assumptions at a non-

singular point P , then the local learning coefficient λ_P at P can be computed in the same form as the formula in Chapter 3.

$$\lambda_P = \frac{d_1 - r + rm}{2m}$$

where d_1 is determined by the codimension of Θ_* . Since this provides an upper bound on the learning coefficient, we can evaluate it accordingly.

Our main results apply to many models, including a certain class of three-layer neural networks and mixture distribution models. Specifically, the three-layer neural network $Y = \sum_{i=1}^H a_i \sigma(b_i X) + \mathcal{N}$, which were treated in Chapter 3, are models in which the weights a_i are fixed. In contrast, by applying the framework developed in Chapter 5, the weights a_i can be treated as parameters. As a result, our method can be applied to a broader class of models. Based on our main results, we clarified that the learning coefficient is significantly influenced by the integer r , which is the rank of the Fisher information matrix, the integer m , which is determined by the differentiability of f , and the integer d_1 , which is determined by the codimension of Θ_* .

1.1 Overview

The rest of this thesis is organized as follows. In Chapter 2, we outline the concepts of singularity resolution and real log canonical thresholds in algebraic geometry. In Chapter 3, we describe a formula for the learning coefficients of the semiregular model and discuss their applications. In Chapter 4, we show that this formula remains valid without the analyticity assumption on $K(\theta)$. In Chapter 5, we show that a formula of the same type holds at nonsingular points of Θ_* . We discuss the conclusions and future work in Chapter 6.

1.2 Definitions and Assumptions

Throughout this thesis, we consider statistical models $p(x|\theta)$ with continuous parameters $\theta = (\theta_1, \dots, \theta_d) \in \Theta \subset \mathbb{R}^d$ where $d \geq 1$, and we denote the true distribution by $q(x)$. We denote the set of all possible data x by χ . We assume that the statistical model is *realizable*, meaning that there exists a parameter θ_* such that $q(x) = p(x|\theta_*)$ almost surely. We refer to such parameters θ_* as realization parameters and denote the entire set of realization parameters by Θ_* . Without loss of generality, we assume $\theta_* = 0$. We assume that the prior distribution of the parameters $\varphi(\theta)$ satisfies $\varphi(\theta_*) > 0$ for any realization parameter θ_* . We denote a random variable following the true distribution q by X , and we use $\mathbb{E}_X[\cdot]$ to represent the expectation over the random variable X . In this thesis, we assume that the operations

of taking expectations and taking partial derivatives with respect to θ are interchangeable.

We define the log-likelihood ratio function f and the Kullback-Leibler divergence $K(\theta)$ respectively as

$$f(x|\theta) := \log \frac{q(x)}{p(x|\theta)}, \quad K(\theta) := \mathbb{E}_X [f(X|\theta)].$$

Furthermore, we define

$$I = \text{Cov} \left(\left. \frac{\partial f(X|\theta)}{\partial \theta_i} \right|_{\theta=0}, \left. \frac{\partial f(X|\theta)}{\partial \theta_j} \right|_{\theta=0} \right)_{i,j=1,\dots,d},$$

$$J = \left(\left. \frac{\partial^2 K(\theta)}{\partial \theta_i \partial \theta_j} \right|_{\theta=0} \right)_{i,j=1,\dots,d}.$$

The matrix I is the Fisher information matrix at $\theta = 0$. Generally, I and J do not coincide, but under the conditions of this thesis, they do coincide (see Remark A.1). We denote the rank of this matrix by r .

In this thesis, we also define the following function $H(\theta)$ as a measure of the distance between $p(x|\theta)$ and $q(x)$, which we use alongside $K(\theta)$:

$$H(\theta) := \frac{1}{2} \int_{\mathcal{X}} \{p(x|\theta) - q(x)\}^2 dx.$$

Chapter 2

Background

In this chapter, we collect the mathematical background needed in the rest of the thesis. We first recall the basic Bayesian setting and the definition of the learning coefficient. We then review resolution of singularities and real log canonical thresholds, and explain how they arise in singular learning theory.

2.1 Bayesian inference and generalization error

We consider a statistical model

$$\{p(x|\theta) \mid \theta \in \Theta \subset \mathbb{R}^d\},$$

where x takes values in a measurable space χ , and Θ is an open subset of $\mathbb{R}^d$. The true distribution of the data is denoted by $q(x)$. $K(\theta)$ is the Kullback-Leibler divergence,

$$K(\theta) := \mathbb{E}_X \left[\log \frac{q(X)}{p(X|\theta)} \right].$$

where the expectation is taken with respect to the true distribution q .

Given a prior density $\varphi(\theta)$ on Θ , the zeta function is defined by

$$\zeta(z) := \int_{\Theta} K(\theta)^z \varphi(\theta) \, d\theta$$

Definition 2.1 (learning coefficient). [17, 28] Let $-\lambda$ and $\mathfrak{m}$ be the largest pole of the meromorphic function $\zeta(z)$ and its order. Then $\lambda > 0$ is called a *learning coefficient*. If we restrict the domain of integration of the zeta function to a neighborhood of a realization parameter P , then the resulting learning coefficient is called the *local learning coefficient* and is denoted by λ_P , and its multiplicity is denoted by $\mathfrak{m}_P$.

If we obtain the learning coefficient, then we can predict the stochastic complexity and the Bayes generalization error theoretically.

We assume that the statistical model is realizable. Let $X^n = (X_1, \dots, X_n)$ be an i.i.d. sample from q . The posterior density is

$$\pi_n(\theta|X^n) = \frac{\{\prod_{i=1}^n p(X_i|\theta)\} \varphi(\theta)}{\int_{\Theta} \{\prod_{i=1}^n p(X_i|\theta)\} \varphi(\theta) d\theta},$$

and the corresponding predictive density for a new observation X is

$$p_n(x) := \int p(x|\theta) \pi_n(\theta|X^n) d\theta.$$

The *Bayesian generalization error* is defined as

$$G_n := \mathbb{E}_X \left[\log \frac{q(X)}{p_n(X)} \right].$$

Under regularity, $\mathbb{E}_{X^n}[G_n]$ behaves as $\frac{d}{2n} + o(1/n)$, but in singular models the coefficient $\frac{d}{2}$ is replaced by the learning coefficient λ , leading to

$$\mathbb{E}_{X^n}[G_n] = \frac{\lambda}{n} + o(1/n).$$

The *free energy* F_n is defined to be the negative log of the marginal likelihood:

$$F_n := -\log \left(\int_{\Theta} \left\{ \prod_{i=1}^n p(X_i|\theta) \right\} \varphi(\theta) d\theta \right).$$

The asymptotic expansion in n is given by

$$F_n = -\sum_{i=1}^n \log q(x_i) + \lambda \log n - (\mathfrak{m} - 1) \log \log n + O_p(1).$$

Moreover, if we restrict the parameter space to a neighborhood of the realization parameter P , then the expectation is governed by the local learning coefficient λ_P , and we obtain

$$\mathbb{E}_{X^n}[G_n] = \frac{\lambda_P}{n} + o(1/n)$$

$$F_n = -\sum_{i=1}^n \log q(x_i) + \lambda_P \log n - (\mathfrak{m}_P - 1) \log \log n + O_p(1).$$

2.2 Real log canonical thresholds

It is known that learning coefficients correspond to the concept of *real log canonical thresholds* in algebraic geometry. Real log canonical thresholds are defined using a resolution of singularities. Here, the resolution of singularities refers to transforming the analytic function F into a normal crossing, as specified by Theorem 2.1.

Theorem 2.1 (Resolution of Singularities). [14, 28] Let $F(x)$ be a real analytic function defined near the origin in $\mathbb{R}^d$ that is not identically zero, with $F(0) = 0$. Then, there exists an open set $W \subset \mathbb{R}^d$ containing the origin, a real analytic manifold U , and a proper analytic map $g : U \rightarrow W$ satisfying:

- (1) Let $W_0 := F^{-1}(0)$ and $U_0 := g^{-1}(W_0)$; then, $g : U \setminus U_0 \rightarrow W \setminus W_0$ is an analytic isomorphism.
- (2) At any point Q in U_0 , taking local coordinates $u = (u_1, \dots, u_d)$ of U at Q as the origin, it can be expressed as:

$$\begin{aligned} F(g(u)) &= a(u)u_1^{k_1}u_2^{k_2}\cdots u_d^{k_d}, \\ |g'(u)| &= |b(u)u_1^{h_1}u_2^{h_2}\cdots u_d^{h_d}|, \end{aligned} \tag{2.1}$$

where k_i, h_i ($i = 1, \dots, d$) are nonnegative integers and $a(u), b(u)$ are real analytic functions with $a(u) \neq 0, b(u) \neq 0$.

Expressions such as (2.1) are referred to as *normal crossings*.

Definition 2.2 (Real Log Canonical Threshold). [28] Let F be a real analytic function defined on an open set O in $\mathbb{R}^d$, and let $C \subset O$ be a compact set. For each point P in C where $F(P) = 0$, we perform a translation so that point P coincides with the origin in $\mathbb{R}^d$, and Theorem 2.1 applies, fixing (W, U, g) as guaranteed by Theorem 2.1(2). We denote the nonnegative integers h_i, k_i given by Theorem 2.1(2) near any point $Q \in U_0$ as $h_i^{(Q)}, k_i^{(Q)}$.

- (1) The *real log canonical threshold* λ_P at point P , and the *real log canonical threshold* λ for the compact set C , are defined respectively as:

$$\lambda_P := \inf_{Q \in U_0} \left\{ \min_{i=1, \dots, d} \frac{h_i^{(Q)} + 1}{k_i^{(Q)}} \right\}, \quad \lambda := \inf_{P \in C} \lambda_P,$$

with $(h_i^{(Q)} + 1)/k_i^{(Q)} = \infty$ defined for cases where $k_i^{(Q)} = 0$ [28].

- (2) For the point $P \in C$ giving the minimum in (1), the maximum number of i satisfying $\lambda_P = (h_i^{(Q)} + 1)/k_i^{(Q)}$ is called the *multiplicity* $\mathbf{m}$. If there are multiple such points P , the multiplicity is defined as the maximum number among them.

We consider a case where the analytic function F corresponds to the Kullback-Leibler divergence $K(\theta)$. Under the condition that the prior distribution $\varphi(\theta_*) > 0$, it is known that the *learning coefficient* λ is equal to the *real log canonical threshold on the compact set* $\Theta_* = \{\theta \in \Theta \mid K(\theta) = 0\}$ [28, Theorem 6.6, Definition 6.4]. This λ_P is equal to the local learning coefficient at P .

Theorem 2.1 ensures the existence of the function g , which can be constructed by a finite number of repetitions of an operation known as blow-up [28]. This result is demonstrated in [14] and [28, Theorem 3.6]. However, the existence of g is only proven abstractly, and it is usually difficult to find a specific construction method for g . In this thesis, we construct the coordinate transformation g through a concrete sequence of blow-ups to calculate the learning coefficient λ .

2.3 Regularity

Definition 2.3 (regularity). [28] $q(x)$ is said to be *regular* for a statistical model $p(x|\theta)$ if the following conditions are satisfied:

- The set of realizing parameters Θ_* contains only the single element θ_* and is contained within the interior of Θ .
- The Hessian matrix $\nabla^2 K(\theta_*)$ as θ_* defined by

$$(\nabla^2 K(\theta_*))_{i,j} = \left. \frac{\partial^2 K}{\partial \theta_i \partial \theta_j} \right|_{\theta=\theta_*}$$

is positive definite.

Proposition 2.1 (Formula for the regular model). [28]

We assume that the true distribution is regular and realizable with respect to the statistical model. If $\varphi(\theta_) > 0$, then we have $\lambda = d/2$ and $\mathbf{m} = 1$.*

Proposition 2.2. [28] *If there exists an open set $U \subset \Theta$ such that*

$$\{\theta \in \Theta \mid K(\theta) = 0, \varphi(\theta) > 0\}$$

is not the empty set, then

$$\lambda \leq \frac{d}{2}.$$

Proposition 2.3. [28]

Assume that a parameter is represented by $(\theta, \tau) \in \Theta$ ($\theta \in \mathbb{R}^{d_1}, \tau \in \mathbb{R}^{d_2}$) and that $K(\theta, \tau)$ and $\varphi(\theta, \tau)$ satisfy the following conditions.

- (1) *For arbitrary τ , $K(\theta_0, \tau) = 0$.*
- (2) *There exists an open set $V \subset \mathbb{R}^{d_2}$ such that $\varphi(\theta_0, \tau) > 0$ ($\tau \in V$).*

Then

$$\lambda \leq \frac{d_1}{2}.$$

Chapter 3

Semiregular models

3.1 Semiregularity

Definition 3.1 (Semiregularity). A statistical model $p(x|\theta)$ is said to be *semiregular*¹ at a parameter θ_* (which realizes the true model $q(x)$) if the following conditions are satisfied:

- The set of realizing parameters Θ_* contains only the single element θ_* and is contained within the interior of Θ .
- If $r = d$, then the r random variables

$$\left. \frac{\partial f(X|\theta)}{\partial \theta_1} \right|_{\theta=0}, \dots, \left. \frac{\partial f(X|\theta)}{\partial \theta_r} \right|_{\theta=0} \quad (3.1)$$

are linearly independent². If $r < d$, then the model satisfies Assumption 1.

Assumption 1. Divide the d parameters $\theta = (\theta_1, \dots, \theta_d)$ into $\theta_1, \dots, \theta_r$ and $\theta_{r+1}, \dots, \theta_d$, and assume:

- (1) There exists a maximal integer m (with $2 \leq m < \infty$) such that for all multiple indices $(i_{r+1}, \dots, i_d)$ satisfying $1 \leq i_{r+1} + \dots + i_d \leq m - 1$, the following condition holds:

$$\left. \frac{\partial^{i_{r+1} + \dots + i_d} f(X|\theta)}{\partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_d^{i_d}} \right|_{\theta=0} = 0 \quad \text{a.s.}$$

- (2) For all $(s_{r+1}, \dots, s_d) \neq 0$, the function

¹This term is specific to this thesis and is not generally established.

²The random variables being linearly independent means that they are so with probability one.

$$F_m(X|s_{r+1}, \dots, s_d) := \sum_{\substack{i_{r+1} + \dots + i_d = m \\ i_{r+1}, \dots, i_d \in \mathbb{Z}_{\geq 0}}} \frac{1}{i_{r+1}! \dots i_d!} \left. \frac{\partial^m f(x|\theta)}{\partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_d^{i_d}} \right|_{\theta=0} s_{r+1}^{i_{r+1}} \dots s_d^{i_d}.$$

and the r random variables in (3.1) are linearly independent. When $r = 0$, this condition is understood to mean that

$$F_m(X|s_{r+1}, \dots, s_d) \neq 0 \quad \text{a.s.}$$

When $r = d$, this definition reduces to the usual notion of regularity, so it indeed generalizes the classical concept. Therefore, we exclude such cases from our discussion, and we assume $r \leq d - 1$ throughout this chapter.

Remark 3.1. We now explain the meaning of the assumption of semiregularity. The assumptions other than Assumption 1 are the same as those in the regularity condition, and Assumption 1 generalizes the regularity condition that the Fisher information matrix is positive definite.

As for Assumption 1(1), when $1 \leq r \leq d - 1$, the first derivatives of f with respect to $\theta_{r+1}, \dots, \theta_d$ are linearly dependent on the r random variables in (3.1). As explained in Section 3.2, by applying an appropriate coordinate transformation, all such linearly dependent derivatives can be made zero. Thus, Assumption 1(1) can always be satisfied.

As for Assumption 1(2), it is a condition necessary to guarantee normal crossings (see the proof of Theorem 3.1). Unlike (1), there exist statistical models that do not satisfy this condition.

Example 3.1. Let $\mathcal{N}$ be a random variable that is subject to the standard normal distribution. The probability distribution $q(x)$ of X is assumed to have compact support. Consider the following statistical model in which a random variable Y for a given X is defined by

$$Y = A_1 \sigma(\theta_1 X) + A_2 \sigma(\theta_2 X) + \mathcal{N},$$

where A_1, A_2 are nonzero constants. This is a layered neural network with one input unit, two hidden units, and one output unit. The true distribution of Y is assumed to be

$$Y = 0 + \mathcal{N}.$$

The activation function $\sigma(x)$ is an analytic function satisfying $\sigma(0) = 0, \sigma'(0) \neq 0$, and the smallest natural number $m(\geq 2)$ such that $\sigma^{(m)}(0) \neq 0$ is defined as m . For instance, $m = 2$ for $\sigma(x) = e^x - 1$, $m = 3$ for $\sigma(x) = \sin(x), \arctan(x), \tanh(x)$, and m for $\sigma(x) = x^m + x$.

We assume that the constants A_1, A_2 satisfy $A_1^{m-1} + (-1)^m A_2^{m-1} \neq 0$ and that the random variables X and X^m are linearly independent. Under

these conditions, the statistical model satisfies Assumption 1 for the realized parameters $(\theta_1, \theta_2) = (0, 0)$ when $(d, r, m) = (2, 1, m)$.

Specifically, the log-likelihood ratio function f is given by

$$f(x, y|\theta) = -\{A_1\sigma(\theta_1x) + A_2\sigma(\theta_2x)\}y + \frac{1}{2}\{A_1\sigma(\theta_1x) + A_2\sigma(\theta_2x)\}^2.$$

By performing the coordinate transformation $\theta'_1 = \theta_1 + (A_2/A_1)\theta_2$, we obtain

$$\begin{aligned} \left. \frac{\partial f}{\partial \theta'_1} \right|_{(\theta'_1, \theta_2)=0} &= -A_1\sigma'(0)xy, \\ \left. \frac{\partial f}{\partial \theta_2^n} \right|_{(\theta'_1, \theta_2)=0} &= 0 \quad (1 \leq n < m), \\ \left. \frac{\partial f}{\partial \theta_2^m} \right|_{(\theta'_1, \theta_2)=0} &= -\frac{A_2}{A_1^{m-1}} \{A_1^{m-1} + (-1)^m A_2^{m-1}\} \sigma^{(m)}(0)x^m y. \end{aligned}$$

Since $x^m y$ and xy are linearly independent by assumption, the model satisfies Assumption 1 for $(d, r, m) = (2, 1, m)$.

Theorem 3.1. *[Learning coefficient for a semiregular model] Assume that $K(\theta)$ is analytic. Then the learning coefficient for a semiregular model is given by the following formula (with multiplicity $\mathbf{m} = 1$):*

$$\lambda = \frac{d - r + rm}{2m} \quad (3.2)$$

Remark 3.2. In the case of regular models, substituting $r = d$ into (3.2) results in

$$\lambda = \frac{d - d + d \times m}{2 \times m} = \frac{d}{2},$$

which matches the known result for regular models. Thus, this formula generalizes the regular case. Previous research has generally shown that $\lambda \leq d/2$ (Proposition 2.2), and

$$\frac{d - r + rm}{2m} \leq \frac{d}{2},$$

so (3.2) does not contradict this finding.

3.2 Derivation Method for m

Throughout this section, we assume that $r > 0$. To satisfy Assumption 1(1), we can explicitly construct the variable transformation as follows. Given the semiregularity, we have r parameters $\theta_1, \dots, \theta_r$ for which the derivatives

$$\left. \frac{\partial f}{\partial \theta_1} \right|_{\theta=0}, \dots, \left. \frac{\partial f}{\partial \theta_r} \right|_{\theta=0} \quad (3.3)$$

are linearly independent. For the remaining parameters $\theta_{r+1}, \dots, \theta_d$, starting from $l = 1$, if the derivative can be expressed as

$$\left. \frac{\partial^l f}{\partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_d^{i_d}} \right|_{\theta=0} = \sum_{k=1}^r a_k \left. \frac{\partial f}{\partial \theta_k} \right|_{\theta=0} \quad \text{a.s.,} \quad a_k \in \mathbb{R},$$

then we perform the coordinate transformation

$$\begin{cases} \theta'_1 := \theta_1 + \frac{a_1}{i_{r+1}! \dots i_d!} \theta_{r+1}^{i_{r+1}} \dots \theta_d^{i_d} \\ \vdots \\ \theta'_r := \theta_r + \frac{a_r}{i_{r+1}! \dots i_d!} \theta_{r+1}^{i_{r+1}} \dots \theta_d^{i_d} \end{cases} \quad \begin{cases} \theta'_{r+1} := \theta_{r+1} \\ \vdots \\ \theta'_d := \theta_d \end{cases}$$

which makes

$$\left. \frac{\partial^l f}{\partial \theta_{r+1}'^{i_{r+1}} \dots \partial \theta_d'^{i_d}} \right|_{\theta'=0} = 0 \quad \text{a.s.} \quad (3.4)$$

without changing other derivatives of order l or lower (this fact is proven in Proposition B.1). We repeat this process until we first encounter a derivative that is linearly independent from (3.3); the order of this derivative is m .

Remark 3.3. We must have $m < \infty$. If $m = \infty$, that is, if for any $i_{r+1}, \dots, i_d$, condition (3.4) is satisfied; then, in the Taylor expansion of f around $\theta = 0$, there would be no terms consisting solely of $\theta_{r+1}, \dots, \theta_d$. This means that for sufficiently small arbitrary values of $(\theta_{r+1}, \dots, \theta_d)$,

$$f(X|\theta_1 = 0, \dots, \theta_r = 0, \forall \theta_{r+1}, \dots, \forall \theta_d) = 0 \quad \text{a.s.,}$$

resulting in the statistical model $p(x|\theta)$ realizing the true distribution $q(x)$ at multiple points near the origin. This contradicts the assumption of semiregularity, which presumes realization at only one point.

3.3 Proof of Theorem 3.1

In the proof of Theorem 3.1, we use the following theorem (for the proof, see Appendix A):

Theorem 3.2. *Suppose that the statistical model $p(x|\theta)$ satisfies Assumption 1 and that $K(\theta)$ is analytic. Then the Taylor expansion of $K(\theta)$ at $\theta = 0$ can be expressed as:*

$$\begin{aligned} K(\theta) = & \frac{1}{2} \mathbb{E}_X \left[\{F_1(X|\theta_1, \dots, \theta_r) + F_m(X|\theta_{r+1}, \dots, \theta_d)\}^2 \right] \\ & + (\text{higher order terms}) \end{aligned}$$

Here, we define

$$F_1(X|s_1, \dots, s_r) := \sum_{k=1}^r \frac{\partial f}{\partial \theta_k} \Big|_{\theta=0} s_k,$$

and the higher order terms do not include the following:

- Terms of degree $2m$ or less consisting only of $\theta_{r+1}, \dots, \theta_d$.
- Terms of first degree in $\theta_1, \dots, \theta_r$ and up to m -th degree in $\theta_{r+1}, \dots, \theta_d$.
- Terms of second degree consisting only of $\theta_1, \dots, \theta_r$.

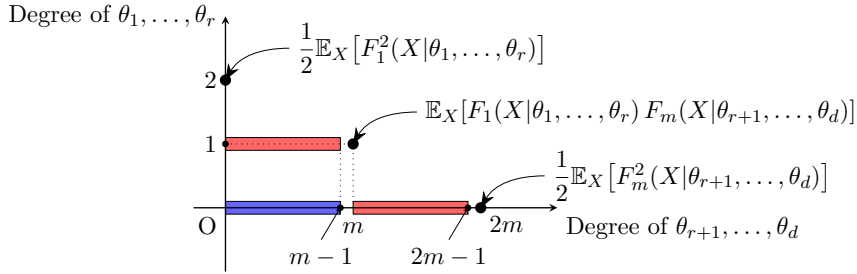


Figure 3.1: Lowest-order terms in the Taylor expansion of $K(\theta)$

Note. The horizontal axis represents the total degree with respect to $\theta_{r+1}, \dots, \theta_d$, and the vertical axis represents the total degree with respect to $\theta_1, \dots, \theta_r$. The blue region (including its boundary) corresponds to the degrees of the partial derivatives of the log-likelihood ratio function that are assumed to be zero almost surely. Under this assumption, the derivatives corresponding to the red region (including its boundary) have zero expectation; equivalently, all Taylor coefficients of $K(\theta)$ associated with these degrees vanish.

Proof of Theorem 3.1. Using Theorem 3.2, the Taylor expansion of $K(\theta)$ at $(\theta_1, \dots, \theta_d) = 0$ can be expressed as:

$$K(\theta) = \frac{1}{2} \mathbb{E}_X \left[\{F_1(X|\theta_1, \dots, \theta_r) + F_m(X|\theta_{r+1}, \dots, \theta_d)\}^2 \right] + (\text{higher order terms})$$

where the higher order terms are specifically expressed as the sum of the following four terms:

- $f_0(\theta_{r+1}, \dots, \theta_d)$: terms of degree $2m + 1$ and higher
- $f_1(\theta_1, \dots, \theta_d)$: terms of first degree homogeneous in $\theta_1, \dots, \theta_r$ and of degree $m + 1$ and higher in $\theta_{r+1}, \dots, \theta_d$
- $f_2(\theta_1, \dots, \theta_d)$: terms of second degree and higher in $\theta_1, \dots, \theta_r$ and of first degree and higher in $\theta_{r+1}, \dots, \theta_d$

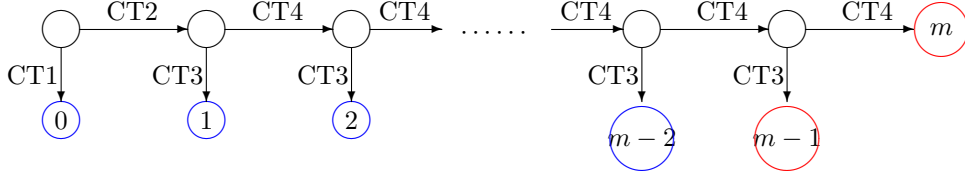


Figure 3.2: The procedure of coordinate transformations g performed in the proof of Theorem 3.1

Note. Local coordinates displayed in red represent the minimal real log canonical thresholds.

- $f_3(\theta_1, \dots, \theta_r)$: terms of third degree and higher.

We now consider these terms at their lowest degrees, namely:

- $f_0(\theta_{r+1}, \dots, \theta_d)$: terms of degree $2m + 1$ homogeneous
- $f_1(\theta_1, \dots, \theta_d)$: terms of first degree homogeneous in $\theta_1, \dots, \theta_r$ and $m+1$ degree homogeneous in $\theta_{r+1}, \dots, \theta_d$
- $f_2(\theta_1, \dots, \theta_d)$: terms of second degree homogeneous in $\theta_1, \dots, \theta_r$ and first degree homogeneous in $\theta_{r+1}, \dots, \theta_d$
- $f_3(\theta_1, \dots, \theta_r)$: terms of third degree homogeneous.

The proof for the general case proceeds similarly.

We perform the following four types of coordinate transformations, CT1–CT4 (see Figure 3.2) ³; in the case $r = 0$, we use only CT2. We demonstrate that normal crossings are achieved in all the resulting local coordinates.

CT1 For $1 \leq s \leq r$, $\{\theta_j \rightarrow \theta_s \theta'_j \mid 1 \leq j \leq d, j \neq s\}$

CT2 For $r + 1 \leq t \leq d$, $\{\theta_j \rightarrow \theta_t \theta'_j \mid 1 \leq j \leq d, j \neq t\}$

CT3 For the t selected in CT2 and for $1 \leq s \leq r$, $\{\theta'_j \rightarrow \theta_s \theta''_j, \theta_t \rightarrow \theta_s \theta'_t \mid 1 \leq j \leq r, j \neq s\}$

CT4 For the t selected in CT2, $\{\theta'_j \rightarrow \theta_t \theta''_j \mid 1 \leq j \leq r\}$

Initially, we perform CT1 and CT2. The first transformation achieves normal crossings, but the second does not. Therefore, we proceed to perform further iterations involving CT3 and CT4. In these iterations, CT3 achieves normal crossings, but CT4 does not. By repeating this process up to $m - 1$ times (resulting in a total of m iterations), we eventually achieve normal crossings. The details of this procedure are provided below.

³In terms of algebraic geometry, CT1 and CT2 correspond to “blow-ups centered at the origin of $\mathbb{R}^d$ ”. CT3 and CT4 correspond to “blow-ups centered at the subvariety $\{(\theta_1, \dots, \theta_d) \mid \theta_1 = \dots = \theta_r = \theta_t = 0\}$ when the exceptional surface from CT1 and CT2 is $\{\theta_t = 0\}$ ”.

Performing CT1 Once

We can assume $s = 1$ without loss of generality. In CT1, we perform $g = \{\theta_j = \theta_1 \theta'_j \mid 2 \leq j \leq d\}$.

$$\begin{aligned} K(\theta) &= \frac{1}{2} \mathbb{E}_X \left[\{F_1(X|\theta_1, \dots, \theta_r) + F_m(X|\theta_{r+1}, \dots, \theta_d)\}^2 \right] \\ &\quad + f_0(\theta_{r+1}, \dots, \theta_d) + f_1(\theta_1, \dots, \theta_d) + f_2(\theta_1, \dots, \theta_d) + f_3(\theta_1, \dots, \theta_r) \\ &= \frac{1}{2} \theta_1^2 \left\{ \mathbb{E}_X \left[\{F_1(X|1, \theta'_2, \dots, \theta'_r) + \theta_1^{m-1} F_m(X|\theta'_{r+1}, \dots, \theta'_d)\}^2 \right] \right. \\ &\quad + \theta_1^{2m-1} f_0(\theta'_{r+1}, \dots, \theta'_d) + \theta_1^m f_1(1, \theta'_2, \dots, \theta'_d) \\ &\quad \left. + \theta_1 f_2(1, \theta'_2, \dots, \theta'_d) + \theta_1 f_3(1, \theta'_2, \dots, \theta'_r) \right\} = \frac{1}{2} \theta_1^2 a(\theta_1, \theta'_2, \dots, \theta'_d) \end{aligned}$$

where a is an analytic function. Considering points on $U_0 = g^{-1}(O)$, the condition $\theta_1 = 0$ is satisfied. From Assumption 1(2), we have

$$a(0, \theta'_2, \dots, \theta'_d) = \mathbb{E}_X [F_1^2(X|1, \theta'_2, \dots, \theta'_r)] > 0,$$

and thus, a normal crossing is achieved.

The Jacobian of this coordinate transformation is θ_1^{d-1} , so

$$\inf_{Q \in U_0} \min_{j=1, \dots, d} \left\{ \frac{h_j^{(Q)} + 1}{k_j^{(Q)}} \right\} = \frac{(d-1) + 1}{2} = \frac{d}{2}$$

(with multiplicity $\mathbf{m} = 1$). The symbols $k_j^{(Q)}$ and $h_j^{(Q)}$ refer to Definition 2.2.

Sequence of Transformations: CT2 Once, CT4 $i \leq m-2$ Times, and CT3

Assume without loss of generality that $t = d$. First, perform CT2: $g = \{\theta_j \rightarrow \theta_d \theta'_j \mid 1 \leq j \leq d-1\}$, which transforms $K(\theta)$ as follows:

$$\begin{aligned} K(\theta) &= \frac{1}{2} \theta_d^2 \left\{ \mathbb{E}_X \left[\{F_1(X|\theta'_1, \dots, \theta'_r) + \theta_d^{m-1} F_m(X|\theta'_{r+1}, \dots, \theta'_{d-1}, 1)\}^2 \right] \right. \\ &\quad + \theta_d^{2m-1} f_0(\theta'_{r+1}, \dots, \theta'_{d-1}, 1) + \theta_d^m f_1(\theta'_1, \dots, \theta'_{d-1}, 1) \\ &\quad \left. + \theta_d f_2(\theta'_1, \dots, \theta'_{d-1}, 1) + \theta_d f_3(\theta'_1, \dots, \theta'_r) \right\} = \frac{1}{2} \theta_d^2 a(\theta'_1, \dots, \theta'_{d-1}, \theta_d), \end{aligned}$$

where a is an analytic function. Considering points on $U_0 = g^{-1}(O)$, the condition $\theta_d = 0$ is satisfied. We have

$$a(\theta'_1, \dots, \theta'_{d-1}, 0) = \mathbb{E}_X [F_1^2(X|\theta'_1, \dots, \theta'_r)].$$

For any point Q on U_0 with $(\theta'_1, \dots, \theta'_r) \neq 0$, the normal crossing of $K(\theta)$ is achieved, and

$$\inf_Q \min_{j=1, \dots, d} \frac{h_j^{(Q)} + 1}{k_j^{(Q)}} = \frac{d}{2}$$

(with multiplicity $\mathbf{m} = 1$). Therefore, we only need to consider points on U_0 where $(\theta'_1, \dots, \theta'_r) = 0$.

After applying CT4: $\{\theta'_j \rightarrow \theta_d \theta''_j \mid 1 \leq j \leq r\}$ i times, $K(\theta)$ can be expressed as:

$$\begin{aligned} K(\theta) = & \frac{1}{2} \theta_d^{2(i+1)} \left\{ \mathbb{E}_X \left[\left\{ F_1(X|\theta''_1, \dots, \theta''_r) \right. \right. \right. \\ & \left. \left. \left. + \theta_d^{m-i-1} F_m(X|\theta'_{r+1}, \dots, \theta'_{d-1}, 1) \right\}^2 \right] \right. \\ & + \theta_d^{2(m-i)-1} f_0(\theta'_{r+1}, \dots, \theta'_{d-1}, 1) + \theta_d^{m-i} f_1(\theta''_1, \dots, \theta'_{d-1}, 1) \\ & \left. \left. + \theta_d f_2(\theta''_1, \dots, \theta'_{d-1}, 1) + \theta_d^{i+1} f_3(\theta''_1, \dots, \theta''_r) \right\}. \end{aligned}$$

Additionally, we only need to consider the point where $(\theta''_1, \dots, \theta''_r, \theta_d) = (0, \dots, 0, 0)$. We denote $\theta''_1, \dots, \theta''_r$ as $\theta'_1, \dots, \theta'_r$.

Next, perform CT3: $\{\theta'_j \rightarrow \theta'_1 \theta'_j, \theta_d \rightarrow \theta'_1 \theta'_d \mid 2 \leq j \leq r\}$, which transforms $K(\theta)$ into:

$$\begin{aligned} K(\theta) = & \frac{1}{2} \theta_1'^{2(i+2)} \theta_d'^{2(i+1)} \times \\ & \left\{ \mathbb{E}_X \left[\left\{ F_1(X|1, \theta''_2, \dots, \theta''_r) \right. \right. \right. \\ & \left. \left. \left. + \theta_1'^{m-i-2} \theta_d'^{m-i-1} F_m(X|\theta'_{r+1}, \dots, \theta'_{d-1}, 1) \right\}^2 \right] \right. \\ & + \theta_1'^{2(m-i)-3} \theta_d'^{2(m-i)-1} f_0(\theta'_{r+1}, \dots, \theta'_{d-1}, 1) \\ & + \theta_1'^{m-i-1} \theta_d'^{m-i} f_1(1, \theta''_2, \dots, \theta'_{d-1}, 1) \\ & \left. \left. + \theta_1' \theta_d' f_2(1, \theta''_2, \dots, \theta'_{d-1}, 1) + \theta_1'^{i+2} \theta_d'^{i+1} f_3(1, \theta''_2, \dots, \theta''_r) \right\} \right. \\ & = \frac{1}{2} \theta_1'^{2(i+2)} \theta_d'^{2(i+1)} a(\theta'_1, \theta''_2, \dots, \theta'_d), \end{aligned}$$

where a is an analytic function. Considering points on $U_0 = g^{-1}(O)$, the condition $\theta'_1 = 0$ is satisfied. We have

$$\begin{aligned} & a(0, \theta''_2, \dots, \theta'_d) \\ = & \begin{cases} \mathbb{E}_X \left[\left\{ F_1(X|1, \theta''_2, \dots, \theta''_r) \right. \right. \right. \\ \quad \left. \left. \left. + \theta_d'^{m-i-1} F_m(X|\theta'_{r+1}, \dots, \theta'_{d-1}, 1) \right\}^2 \right] & \text{if } i = m-2, \\ \mathbb{E}_X \left[F_1(X|1, \theta''_2, \dots, \theta''_r)^2 \right] & \text{if } i < m-2. \end{cases} \end{aligned}$$

According to Assumption 1(2), this expression is positive. Therefore, in this local coordinate, the normal crossing of $K(\theta)$ is achieved at any point Q on U_0 . The Jacobian of this coordinate transformation is $\theta_1'^{d+r(i+1)-1}\theta_d'^{d+ri-1}$, leading to:

$$\begin{aligned} \inf_{Q \in U_0} \min_{j=1, \dots, d} \frac{h_j^{(Q)} + 1}{k_j^{(Q)}} &= \min \left\{ \frac{d + r(i+1) - 1 + 1}{2(i+2)}, \frac{d + ri - 1 + 1}{2(i+1)} \right\} \\ &= \frac{d - r + r(i+2)}{2(i+2)} \end{aligned}$$

(with multiplicity $\mathbf{m} = 1$).

After Performing CT2 Once, Then Applying CT4 $m - 1$ Times

Following the above method, after performing the coordinate transformation g and using the analytic function a , $K(\theta)$ can be expressed as:

$$\begin{aligned} K(\theta) &= \frac{1}{2} \theta_d^{2m} \left\{ \mathbb{E}_X \left[\{ F_1(X|\theta_1'', \dots, \theta_r'') + F_m(X|\theta_{r+1}', \dots, \theta_{d-1}', 1) \}^2 \right] \right. \\ &\quad + \theta_d f_0(\theta_{r+1}', \dots, \theta_{d-1}', 1) + \theta_d f_1(\theta_1'', \dots, \theta_{d-1}', 1) \\ &\quad \left. + \theta_d f_2(\theta_1'', \dots, \theta_{d-1}', 1) + \theta_d^m f_3(\theta_1'', \dots, \theta_r'') \right\} \\ &= \frac{1}{2} \theta_d^{2m} a(\theta_1'', \dots, \theta_{d-1}', \theta_d). \end{aligned}$$

Considering points on $U_0 = g^{-1}(O)$, the condition $\theta_d = 0$ is satisfied. In accordance with Assumption 1(2), we have

$$a(\theta_1'', \dots, \theta_{d-1}', 0) = \mathbb{E}_X \left[\left\{ F_1(X|\theta_1'', \dots, \theta_r'') + F_m(X|\theta_{r+1}', \dots, \theta_{d-1}', 1) \right\}^2 \right] > 0.$$

This conclusion can be understood as follows. The condition $(\theta_1'', \dots, \theta_{d-1}', 0) = 0$ holds if and only if

$$\begin{aligned} &F_1(X|\theta_1'', \dots, \theta_r'') + F_m(X|\theta_{r+1}', \dots, \theta_{d-1}', 1) = 0 \quad \text{a.s.} \\ \Leftrightarrow &\sum_{k=1}^r \theta_k'' \cdot \frac{\partial f(X|\theta)}{\partial \theta_k} \Big|_{(\theta_1, \dots, \theta_d)=0} + F_m(X|\theta_{r+1}', \dots, \theta_{d-1}', 1) = 0 \quad \text{a.s..} \end{aligned}$$

However, this equivalence does not hold because, by Assumption 1(2), the random variable $F_m(X|\theta_{r+1}', \dots, \theta_{d-1}', 1)$ is linearly independent of the r random variables

$$\frac{\partial f(X|\theta)}{\partial \theta_1} \Big|_{(\theta_1, \dots, \theta_d)=0}, \dots, \frac{\partial f(X|\theta)}{\partial \theta_r} \Big|_{(\theta_1, \dots, \theta_d)=0}.$$

Thus, the normal crossing of $K(\theta)$ is achieved. The Jacobian of this coordinate transformation is $\theta_d^{d+r(m-1)-1}$, so

$$\inf_{Q \in U_0} \min_{j=1, \dots, d} \frac{h_j^{(Q)} + 1}{k_j^{(Q)}} = \frac{d + r(m-1) - 1 + 1}{2m} = \frac{d - r + rm}{2m}$$

(with multiplicity $\mathbf{m} = 1$). Therefore, we achieve normal crossings at all local coordinates, and we determine the real log canonical threshold as:

$$\min_{1 \leq m' \leq m} \left\{ \frac{d - r + rm'}{2m'} \right\} = \frac{d - r + rm}{2m}.$$

□

Remark 3.4. When considering the coordinate transformation g performed in Theorem 3.1, we can deduce that the only point in the neighborhood of the origin in $\mathbb{R}^d$ where $K(\theta) = 0$ is the origin itself. For example, after performing CT2 once and CT4 $m - 1$ times, in the local coordinates $u = (\theta''_1, \dots, \theta'_{d-1}, \theta_d)$, we have $K(\theta) = K(g(u)) = 0 \Rightarrow \theta_d = 0$ for all $\theta''_1, \dots, \theta'_{d-1}$, and it follows that $g(\theta''_1, \dots, \theta'_{d-1}, \theta_d = 0) = (0, \dots, 0)$.⁴ Therefore, even without assuming the condition that “ Θ_* contains only the single element” in the definition of semiregularity, it follows that Θ_* consists of only one point in a neighborhood of $\theta = 0$, provided that Assumption 1 is satisfied.

⁴In the terminology of algebraic geometry, this follows because the centers of the blow-ups conducted are all contained in the exceptional surfaces of the previous blow-ups.

Chapter 4

Extension1 : generalization to the nonanalytic case

In Chapter 3, we assumed that $K(\theta)$ is analytic, but there are cases in which $K(\theta)$ is not analytic, such as in normal mixture distributions.

Example 4.1. Let $\tilde{p}(x|\mu)$ be the probability density function of a normal distribution with mean μ and variance 1. Consider the statistical model

$$p(x|\theta) := (1 - \theta) \cdot \tilde{p}(x|0) + \theta \cdot \tilde{p}(x|2)$$

If the true distribution is $q(x) := \tilde{p}(x|0)$, then this statistical model realizes the true distribution at $\theta = 0$. In this case, $K(\theta)$ is nonanalytic at $\theta = 0$ [27], [28, Example 7.3].

In this chapter, even if analyticity is not guaranteed, we claim that the same result holds under the following assumption.

Assumption 2. For a statistical model $p(x|\theta)$, assume that either

- $K(\theta)$ is analytic, or
- $K(\theta)$ is nonanalytic, but the learning coefficient λ equals the real log canonical threshold of the following analytic function $H(\theta)$:

$$H(\theta) := \frac{1}{2} \int_{\mathcal{X}} \{p(x|\theta) - q(x)\}^2 dx. \quad (4.1)$$

Remark 4.1. The following conditions provide a sufficient criterion for Assumption 2 [27]:

- There exists a measurable function $M(x)$ such that

$$\sup_{\theta \in \Theta} \frac{-\log \frac{p(x|\theta)}{q(x)} + \frac{p(x|\theta)}{q(x)} - 1}{\left(\frac{p(x|\theta)}{q(x)} - 1\right)^2} \leq M(x).$$

- Define an inner product $(u, v) := \int u(x) v(x) M(x) q(x) dx$, which induces a Hilbert space $\mathcal{H}$. Then the map

$$\Theta \ni \theta \mapsto \frac{p(x|\theta)}{q(x)} - 1 \in \mathcal{H}$$

is analytic with respect to the topology of $\mathcal{H}$.

In other words, if $p(x|\theta)$ is analytic in θ , then even if $K(\theta)$ itself is nonanalytic, we can obtain results equivalent to those in the analytic case. [27], [21] and [32] verified this sufficient condition for normal mixture models, Poisson mixture models, and multinomial mixture models. Hence, each of these models satisfies Assumption 2.

Theorem 4.1 (Generalization of Theorem 3.1). *Suppose we have a semiregular model in which, instead of assuming $K(\theta)$ is analytic, we assume Assumption 2 holds. Then the learning coefficient is expressed as*

$$\lambda = \frac{d - r + rm}{2m}, \quad m = 1$$

See Appendix C for the proof. In the proof of Theorem 4.1, the following theorem is essential.

Theorem 4.2 (An analogue of Theorem 3.2 for $K(\theta)$). *Consider a statistical model for which the realization parameter is $\theta = 0$ and which satisfies Assumption 1(1). Then the Taylor expansion of $H(\theta)$ at $\theta = 0$ can be expressed as*

$$H(\theta) = \frac{1}{2} \mathbb{E}_X \left[\{F_1(X|\theta_1, \dots, \theta_r) + F_m(X|\theta_{r+1}, \dots, \theta_d)\}^2 q(X) \right] \\ + (\text{higher order terms})$$

where the higher order terms do not include the following:

- Terms of degree $2m$ or less consisting only of $\theta_{r+1}, \dots, \theta_d$.
- Terms of first degree in $\theta_1, \dots, \theta_r$ and up to m -th degree in $\theta_{r+1}, \dots, \theta_d$.
- Terms of second degree consisting only of $\theta_1, \dots, \theta_r$.

Chapter 5

Extension2 : generalization to the case where Θ_* has more than one point

5.1 Local learning coefficients at non-singular points

In this section, we determine the local learning coefficient at nonsingular points in the real analytic variety Θ_* .

Definition 5.1. [28, Definition 2.6] Let $d = d_1 + d_2$ with $d_1, d_2 \in \mathbb{Z}_{\geq 0}$. A point $\theta_* \in \Theta_* \subset \mathbb{R}^{d_1+d_2}$ is called a nonsingular point if there exist open sets $U_1, U_2 \subset \mathbb{R}^{d_1+d_2}$ and an analytic isomorphism $\varphi : U_1 \rightarrow U_2$ such that

$$\varphi(\Theta_* \cap U_1) = \{(0, \dots, 0, x_1, \dots, x_{d_2}) | x_i \in \mathbb{R}\} \cap U_2.$$

We call d_1 the codimension and d_2 the dimension of Θ_* as a manifold.

Remark 5.1. It is known that such nonsingular points always exist in Θ_* [28, Theorem 3.4]. Moreover, when Θ_* consists of a single point, the conditions of Definition 5.1 are satisfied in the case of $(d_1, d_2) = (d, 0)$.

Let d_1 and d_2 be positive integers, and consider a statistical model $p(x|\theta, \tau)$ with $d_1 + d_2$ parameters $(\theta_1, \dots, \theta_{d_1}, \tau_1, \dots, \tau_{d_2})$. We consider the case where the set of realization parameters Θ_* is spread out in the τ -direction in a neighborhood of a certain point $P : (\theta, \tau) = (0, T)$. That is, there exists a neighborhood $V \subset \mathbb{R}^{d_2}$ of $\tau = T$ such that

$$\forall \tau \in V, \quad f(X|\theta = 0, \tau) = 0 \quad \text{a.s.} \quad (5.1)$$

or, equivalently,

$$\forall \tau \in V, \quad K(\theta = 0, \tau) = 0. \quad (5.2)$$

The equivalence of (5.1) and (5.2) follows from the following inequality for realizable statistical models [28, Theorem 6.3]:

$$K(\theta, \tau) = \mathbb{E}_X [f(X|\theta, \tau)] \geq \frac{1}{2} \mathbb{E}_X [f^2(X|\theta, \tau)] .$$

Assumption 3. A point $P : (\theta, \tau) = (0, T)$ is an interior point of the parameter space Θ , and $p(x|\theta, \tau)$ satisfies (5.1). Moreover, we assume that the rank r of the Fisher information matrix at the point P satisfies $r \leq d_1$. In the case $r = d_1$, we assume that for any fixed $\tau \in V$, the r random variables

$$\left. \frac{\partial f(X|\theta, \tau)}{\partial \theta_1} \right|_{\theta=0}, \dots, \left. \frac{\partial f(X|\theta, \tau)}{\partial \theta_r} \right|_{\theta=0} \quad (5.3)$$

are linearly independent. In the case $r < d_1$, we assume the following:

- (1) There exists a maximal integer m (with $2 \leq m < \infty$) such that for all multiple indices $(i_{r+1}, \dots, i_{d_1})$ satisfying $1 \leq i_{r+1} + \dots + i_{d_1} \leq m - 1$ and for any $\tau \in V$, the following condition holds:

$$\left. \frac{\partial^{i_{r+1} + \dots + i_{d_1}} f(X|\theta, \tau)}{\partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_{d_1}^{i_{d_1}}} \right|_{\theta=0} = 0 \quad \text{a.s.}$$

- (2) For all $(s_{r+1}, \dots, s_{d_1}) \neq 0$ and each fixed $\tau \in V$,

$$\begin{aligned} & F_m(X|s_{r+1}, \dots, s_{d_1}) \\ &:= \sum_{\substack{i_{r+1} + \dots + i_{d_1} = m \\ i_{r+1}, \dots, i_{d_1} \in \mathbb{Z}_{\geq 0}}} \frac{1}{i_{r+1}! \dots i_{d_1}!} \left. \frac{\partial^m f(X|\theta, \tau)}{\partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_{d_1}^{i_{d_1}}} \right|_{\theta=0} s_{r+1}^{i_{r+1}} \dots s_{d_1}^{i_{d_1}} \end{aligned}$$

and the r random variables (5.3) are linearly independent. When $r = 0$, this condition is understood to mean that

$$F_m(X|s_{r+1}, \dots, s_{d_1}) \neq 0 \quad \text{a.s.}$$

Note that when $d_2 = 0$, Assumption 3(1)(2) and Assumption 1 are equivalent. When $d_2 > 0$, the set of realization parameters Θ_* of the statistical model satisfying Assumption 3 is not a single point due to (5.1), and thus this statistical model does not satisfy semiregularity.

Remark 5.2. We now explain the meaning of Assumption 3. It is a local condition in the neighborhood of P . Condition (5.1) means that all points around $P : (\theta, \tau) = (0, T)$ in the τ -direction are realization parameters. Furthermore, Assumption 3(1)(2) is a generalization of Assumption 1(1)(2) so that it holds for every $\tau \in V$. Assumption 3(1) is necessary for the Taylor expansion result in Theorem 5.1. Assumption 3(2) is a condition necessary to guarantee normal crossings.

Example 5.1. The following statistical models do not satisfy semiregularity, but they satisfy Assumption 3.

- (1) Let $\tilde{p}(b)$ represent the binomial distribution with $M(\geq 2)$ trials and success probability b , the Poisson distribution with mean b , or the normal distribution with mean b and variance 1. Suppose $H(\geq 1)$ is the number of mixture components, and $(a_1, \dots, a_H)$ are their mixture weights.

Consider the mixture model with $2H$ parameters $(a_1, \dots, a_H, b_1, \dots, b_H)$:

$$p(x|a, b) := \sum_{i=1}^H a_i \tilde{p}(x|b_i)$$

If the true distribution $q(x)$ has H_0 mixture components with $H_0 \leq \min\{H, M/2\}$, then this model does not satisfy semiregularity. However, for some point $P \in \Theta_*$, it satisfies Assumption 3 for $(d_1, r, m) = (H + H_0 - 1, 2H_0 - 1, 2)$. Details are provided in Appendix D.

- (2) Let $\mathcal{N}$ be a random variable that is subject to the standard normal distribution. The probability distribution $q(x)$ of X has compact support. Consider the following statistical model in which a random variable Y for a given X is defined by

$$Y = \sum_{i=1}^H a_i \sigma(b_i X) + \mathcal{N}.$$

This is a three-layer neural network with one input unit, one output unit, and the number of hidden units is H . The true distribution of Y is assumed to a hidden layer with H_0 units, where $1 \leq H_0 < H$:

$$Y = \sum_{i=1}^{H_0} a_{i*} \sigma(b_{i*} X) + \mathcal{N}.$$

Here, $\{b_{1*}, \dots, b_{H_0*}\}$ are distinct nonzero constants, and $\{a_{1*}, \dots, a_{H_0*}\}$ are nonzero constants. Assume that the random variables

$$\sigma(b_{1*}x), \dots, \sigma(b_{H_0*}x), \sigma'(b_{1*}x)x, \dots, \sigma'(b_{H_0*}x)x, \sigma''(b_{H_0*}x)x^2$$

are linearly independent. Then, although this statistical model does not satisfy semi-regularity, it satisfies Assumption 3 with $(d_1, r, m) = (H + H_0, 2H_0, 2)$ at any point P in the following set of realization parameters:

$$\left\{ (a, b) \left| \begin{array}{ll} a_i = a_{i*}, & b_i = b_{i*} \quad (i = 1, \dots, H_0 - 1), \\ b_i = b_{H_0*} & (i = H_0, \dots, H), \\ \sum_{i=H_0}^H a_i = a_{H_0*}, \\ a_i(a_{H_0*} - \sum_{k=H_0+1}^H a_k) > 0 & (i = H_0 + 1, \dots, H) \end{array} \right. \right\} \subset \Theta_*$$

- (3) We now consider the same three-layer neural network model as in (2), with particular focus on the case $H_0 = H - 2$ (≥ 1). Assume that the activation function σ is analytic and satisfies $\sigma(0) = 0$, $\sigma'(0) \neq 0$, and let $m(\geq 2)$ be the smallest positive integer such that $\sigma^{(m)}(0) \neq 0$. Assume that the following $2H - 2$ random variables are linearly independent:

$$\sigma(b_{1*}x), \dots, \sigma(b_{H-2*}x), \sigma'(b_{1*}x)x, \dots, \sigma'(b_{H-2*}x)x, x, x^m \quad (5.4)$$

Then this statistical model satisfies Assumption 3 with $(d_1, r, m) = (2H - 2, 2H - 3, m)$ at any point P in the following set of realization parameters:

$$\left\{ (a, b) \left| \begin{array}{l} a_i = a_{i*}, \quad b_i = b_{i*} \quad (i = 1, \dots, H - 2), \\ b_{H-1} = b_H = 0, \\ \forall (a_{H-1}, a_H) ; a_{H-1}a_H \neq 0, a_{H-1}^{m-1} \neq (-a_H)^{m-1} \end{array} \right. \right\} \subset \Theta_*$$

We verify this. The log-likelihood ratio function f is expressed as

$$f(x, y | \theta) = \left[\sum_{i=1}^{H-2} a_{i*} \sigma(b_{i*}x) - \sum_{i=1}^H a_i \sigma(b_i x) \right] \times \left[y - \frac{\sum_{i=1}^{H-2} a_{i*} \sigma(b_{i*}x) + \sum_{i=1}^H a_i \sigma(b_i x)}{2} \right].$$

We next perform the coordinate transformation

$$a'_i = a_i - a_{i*}, \quad b'_i = b_i - b_{i*} \quad (i = 1, \dots, H - 2), \quad b'_{H-1} = b_{H-1} + \frac{a_H b_H}{a_{H-1}},$$

and compute the partial derivatives of f at $(\{a'_i\}_{i=1}^{H-2}, \{b'_i\}_{i=1}^{H-1}, b_H) = 0$. We obtain the following results using $z := y - \sum_{i=1}^{H-2} a_{i*} \sigma(b_{i*}x)$:

$$\left. \frac{\partial f}{\partial a'_i} \right|_{(a'_1, \dots, a'_{H-2}, b'_1, \dots, b'_{H-1}, b_H)=0} = -\sigma(b_{i*}x)z \quad (i = 1, \dots, H - 2) \quad (5.5)$$

$$\left. \frac{\partial f}{\partial b'_i} \right|_{(a'_1, \dots, a'_{H-2}, b'_1, \dots, b'_{H-1}, b_H)=0} = -a_{i*} \sigma'(b_{i*}x)xz \quad (i = 1, \dots, H - 2) \quad (5.6)$$

$$\left. \frac{\partial f}{\partial b'_{H-1}} \right|_{(a'_1, \dots, a'_{H-2}, b'_1, \dots, b'_{H-1}, b_H)=0} = -a_{H-1} \sigma'(0)xz \quad (5.7)$$

$$\left. \frac{\partial f}{\partial b_H} \right|_{(a'_1, \dots, a'_{H-2}, b'_1, \dots, b'_{H-1}, b_H)=0} = 0$$

$$\left. \frac{\partial^m f}{\partial b_H^m} \right|_{(a'_1, \dots, a'_{H-2}, b'_1, \dots, b'_{H-1}, b_H)=0} = -\frac{a_H \{a_{H-1}^{m-1} - (-a_H)^{m-1}\}}{a_{H-1}^{m-1}} \sigma^{(m)}(0)x^m z \quad (5.8)$$

Since $z|x$ follows the standard normal distribution and the random variables (5.4) are linearly independent, it follows that (5.5), (5.6), (5.7), and (5.8) are also linearly independent. Therefore, we conclude that this statistical model satisfies Assumption 3 when $(d_1, r, m) = (2H - 2, 2H - 3, m)$.

The above argument also applies to the case $(H, H_0) = (2, 0)$. If the random variables x and x^m are linearly independent, then the model satisfies Assumption 3 with $(d_1, r, m) = (2, 1, m)$.

- (4) Let X be a discrete random variable taking $N(\leq d+1)$ distinct values. We assume that there is a point $P \in \Theta_*$ with $r \geq N - 1$, and that the statistical model $p(x|\theta)$ is analytic with respect to θ . Then, $p(x|\theta)$ satisfies Assumption 3 for $(d_1, r) = (N - 1, N - 1)$ at the point P .

We verify this. Without loss of generality, we assume that the point P is the origin in $\mathbb{R}^d$, and we assume that

$$\left. \frac{\partial f(X|\theta)}{\partial \theta_1} \right|_{(\theta_1, \dots, \theta_d)=0}, \dots, \left. \frac{\partial f(X|\theta)}{\partial \theta_{N-1}} \right|_{(\theta_1, \dots, \theta_d)=0}$$

are linearly independent. From the identity

$$\left. \frac{\partial f(X|\theta)}{\partial \theta_i} \right|_{(\theta_1, \dots, \theta_d)=0} = -\frac{1}{q(X)} \left. \frac{\partial p(X|\theta)}{\partial \theta_i} \right|_{(\theta_1, \dots, \theta_d)=0},$$

we see that the linear independence of the derivatives of $f(X|\theta)$ implies the linear independence of

$$\left. \frac{\partial p(X|\theta)}{\partial \theta_1} \right|_{(\theta_1, \dots, \theta_d)=0}, \dots, \left. \frac{\partial p(X|\theta)}{\partial \theta_{N-1}} \right|_{(\theta_1, \dots, \theta_d)=0}. \quad (5.9)$$

Furthermore,

$$\begin{pmatrix} \left. \frac{\partial p(x_1|\theta)}{\partial \theta_1} \right|_{(\theta_1, \dots, \theta_d)=0} & \dots & \left. \frac{\partial p(x_1|\theta)}{\partial \theta_{N-1}} \right|_{(\theta_1, \dots, \theta_d)=0} \\ \vdots & \ddots & \vdots \\ \left. \frac{\partial p(x_{N-1}|\theta)}{\partial \theta_1} \right|_{(\theta_1, \dots, \theta_d)=0} & \dots & \left. \frac{\partial p(x_{N-1}|\theta)}{\partial \theta_{N-1}} \right|_{(\theta_1, \dots, \theta_d)=0} \end{pmatrix} \in \mathbb{R}^{(N-1) \times (N-1)} \quad (5.10)$$

has $N - 1$ column vectors that are linearly independent. Suppose that for some real numbers $\{s_i\}$, the following holds:

$$\sum_{i=1}^{N-1} s_i \left. \frac{\partial p(x_j|\theta)}{\partial \theta_i} \right|_{(\theta_1, \dots, \theta_d)=0} = 0, \quad 1 \leq j \leq N - 1. \quad (5.11)$$

Using the identity

$$\sum_{j=1}^N \left. \frac{\partial p(x_j|\theta)}{\partial \theta_i} \right|_{(\theta_1, \dots, \theta_d)=0} = 0, \quad 1 \leq i \leq N - 1,$$

we have

$$\sum_{i=1}^{N-1} s_i \frac{\partial p(x_N|\theta)}{\partial \theta_i} \Big|_{(\theta_1, \dots, \theta_d)=0} = - \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} s_i \frac{\partial p(x_j|\theta)}{\partial \theta_i} \Big|_{(\theta_1, \dots, \theta_d)=0} = 0.$$

Hence, (5.11) is an identity in the random variable X , and since the random variables (5.9) are linearly independent, it follows that $s_i = 0$ ($1 \leq i \leq N-1$).

Because X is discrete,

$$\Theta_* = \left\{ \theta \in \mathbb{R}^d \mid p(x_i|\theta) - q(x_i) = 0, i = 1, \dots, N-1 \right\}.$$

Applying the implicit function theorem, there exist $N-1$ analytic functions $g_1, \dots, g_{N-1}$ on a neighborhood V of the origin in $\mathbb{R}^{d-N+1}$, such that in a neighborhood U of the point P , the set Θ_* can be expressed as

$$\Theta_* \cap U = \{(\theta_1, \dots, \theta_d) \mid \theta_i = g_i(\theta_N, \dots, \theta_d), i = 1, \dots, N-1\} \cap U.$$

Therefore, performing the coordinate transformation

$$\theta'_i := \theta_i - g_i(\theta_N, \dots, \theta_d)$$

yields

$$\forall (\theta_N, \dots, \theta_d) \in V, \quad f(X|\theta'_1 = 0, \dots, \theta'_{N-1} = 0, \theta_N, \dots, \theta_d) = 0,$$

which implies (5.1). Moreover,

$$\frac{\partial f(X|\theta)}{\partial \theta'_i} \Big|_{(\theta'_1, \dots, \theta_d)=0} = \frac{\partial f(X|\theta)}{\partial \theta_i} \Big|_{(\theta_1, \dots, \theta_d)=0} \quad (i = 1, \dots, N-1)$$

are linearly independent. Therefore, we conclude that $p(x|\theta)$ satisfies Assumption 3 for $(d_1, r) = (N-1, N-1)$.

- (5) Let $\tilde{p}(x|b)$ be the probability function of a binomial distribution $\text{Bin}(H, b)$ with $H(\geq 2)$ trials and success probability b . We consider a mixture distribution model with $2H$ parameters $(a_1, \dots, a_H, b_1, \dots, b_H)$, where H is the number of mixture components, and $(a_1, \dots, a_H)$ represents the mixture weights. We assume that the number of mixture components matches that of the true distribution:

$$p(x|a, b) := \sum_{i=1}^H a_i \tilde{p}(x|b_i)$$

$$q(x) := \sum_{i=1}^H a_{i*} \tilde{p}(x|b_{i*}), \quad a_{i*} \neq 0, \quad b_{i*} \neq b_{j*} (i \neq j)$$

In this case, $(a, b) = (a_*, b_*)$ is a realization parameter, and

$$\left. \frac{\partial f}{\partial b_i} \right|_{(a,b)=(a_*,b_*)} = -\frac{a_{i*}}{q(x)} \frac{\partial \tilde{p}}{\partial b}(x|b_{i*}) \quad (i = 1, \dots, H).$$

Since $\{b_{1*}, \dots, b_{H*}\}$ are distinct, the same argument as in Lemma D.1 applies, and we see that these H random variables are linearly independent.

Therefore, at the realization parameter $(a, b) = (a_*, b_*)$, it follows that $r \geq H$. By applying Example 5.1(4), we conclude that this model satisfies Assumption 3 for $(d_1, r) = (H, H)$.

Remark 5.3.

- (1) Examples 5.1(2)-(3) are similar to Example 3.1, but there is a key distinction. In Example 3.1, the weights A_i in the model $Y = A_1\sigma(\theta_1 X) + A_2\sigma(\theta_2 X) + \mathcal{N}$ are treated as constants, whereas in Examples 5.1(2)-(3), the corresponding weights a_i are treated as parameters. Hence, Examples 5.1(2)-(3) represent more general statistical models.
- (2) Assumption 3 fundamentally relies on differentiability. Therefore, activation functions such as ReLU, defined by $\sigma(x) := \max(0, x)$, which are not differentiable at certain points, fall outside the scope of the present analysis.
- (3) The statistical model considered in Example 5.1(5) does not satisfy condition (5.1) before the coordinate transformation. However, by applying the coordinate transformation $\theta'_i := \theta_i - g_i$ ($i = 1, \dots, d_1$) guaranteed by the implicit function theorem (i.e., the map φ in Definition 5.1), the model comes to satisfy condition (5.1). As a result, we separate the parameters into d_1 parameters $(\theta'_1, \dots, \theta'_{d_1})$ that locally define Θ_* , and the remaining d_2 redundant parameters.

Main Theorem. *Consider a statistical model $p(x|\theta, \tau)$ that satisfies Assumption 2 and Assumption 3. The local learning coefficient λ_P at the point P is given by the following formula:*

$$\lambda_P = \frac{d_1 - r + rm}{2m}, \quad \mathbf{m} = 1 \tag{5.12}$$

This provides an upper bound for the learning coefficient of this statistical model.

Remark 5.4. Since (5.2) holds, we can apply the previous result (Proposition 2.3) to derive $\lambda \leq d_1/2$. Furthermore, since

$$\frac{d_1 - r + rm}{2m} \leq \frac{d_1}{2},$$

it follows that the Main Theorem provides a tighter upper bound than this previous result.

Moreover, if $(d_1, d_2) = (d, 0)$, then it is consistent with Theorem 3.1 and Theorem 4.1.

Remark 5.5. By Remark 5.6, if the statistical model satisfies Assumption 3 at the point P , then P is a nonsingular point of Θ_* with codimension equal to d_1 . Therefore, the Main Theorem applies only when P is a nonsingular point of Θ_* .

5.2 Derivation Method for m

Throughout this section, we assume that the Fisher information matrix at P has rank $r > 0$. At the point P where $(\theta, \tau) = (0, T)$, it follows from (5.1) that

$$\left. \frac{\partial f}{\partial \tau_i} \right|_{(\theta, \tau) = (0, T)} = 0 \quad (i = 1, \dots, d_2).$$

Since $r > 0$, we can choose r parameters $\theta_1, \dots, \theta_r$ such that

$$\left. \frac{\partial f}{\partial \theta_1} \right|_{(\theta, \tau) = (0, T)}, \dots, \left. \frac{\partial f}{\partial \theta_r} \right|_{(\theta, \tau) = (0, T)}$$

are linearly independent. If necessary, we can restrict V so that for each fixed $\tau \in V$,

$$\left. \frac{\partial f}{\partial \theta_1} \right|_{\theta=0}, \dots, \left. \frac{\partial f}{\partial \theta_r} \right|_{\theta=0} \quad (5.13)$$

are linearly independent.

Next, consider the remaining parameters $\theta_{r+1}, \dots, \theta_{d_1}$. Fix an arbitrary constant $\tau \in V$. Starting with $l = 1$, if the derivative can be expressed as

$$\left. \frac{\partial^l f}{\partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_{d_1}^{i_{d_1}}} \right|_{(\theta_1, \dots, \theta_{d_1})=0} = \sum_{k=1}^r \varphi_k(\tau) \left. \frac{\partial f}{\partial \theta_k} \right|_{(\theta_1, \dots, \theta_{d_1})=0} \quad \text{a.s.} \quad (5.14)$$

where $\varphi_1(\tau), \dots, \varphi_r(\tau)$ are analytic functions on V . In that case, by applying a coordinate transformation

$$\begin{cases} \theta'_1 := \theta_1 + \frac{\varphi_1(\tau)}{i_{r+1}! \dots i_{d_1}!} \theta_{r+1}^{i_{r+1}} \dots \theta_{d_1}^{i_{d_1}} \\ \vdots \\ \theta'_r := \theta_r + \frac{\varphi_r(\tau)}{i_{r+1}! \dots i_{d_1}!} \theta_{r+1}^{i_{r+1}} \dots \theta_{d_1}^{i_{d_1}} \end{cases} \quad \begin{cases} \theta'_{r+1} := \theta_{r+1} \\ \vdots \\ \theta'_{d_1} := \theta_{d_1} \end{cases}$$

which is an analytic isomorphism in a neighborhood of the origin in $\mathbb{R}^{d_1}$, we obtain

$$\left. \frac{\partial^l f}{\partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_{d_1}^{i_{d_1}}} \right|_{\theta'=0} = 0 \quad \text{a.s.} \quad (5.15)$$

without changing any other derivatives of order l or lower. This fact can be proved by applying Proposition B.1, considering τ as a constant. Repeating this process, m is defined as the order of derivatives at which we first encounter a derivative that is linearly independent from (5.13).

5.3 Examples and Experiments

5.3.1 Neural Networks

In this section, we consider three-layer neural networks in which both the input and output layers consist of a single unit. Let H denote the number of hidden units of the statistical model, and H_0 that of the true model.

Example 5.2. The three-layer neural network with $(H, H_0) = (2, 0)$ satisfies Assumption 3 with $(d_1, r, m) = (2, 1, m)$ at a certain realization parameter, as shown in Example 5.1(3). Hence, by the Main Theorem, the local learning coefficient at P is given by

$$\lambda_P = \frac{1+m}{2m}, \quad m = 1 \quad (5.16)$$

In particular, this provides an upper bound on the learning coefficient of the model.

For example, if $\sigma(x) = e^x - 1$, then $m = 2$, and (5.16) yields $3/4$, which matches the specific value [28, Example 7.1]. For $\sigma(x) = \tanh(x)$, we have $m = 3$, and (5.16) becomes $2/3$, which also matches the specific value [28, Section 7.2]. Thus, this upper bound is tight (see Table 5.1).

Table 5.1: Differences in the upper bound on the learning coefficient depending on the activation function

Activation function $\sigma(x)$	m	Upper bound (5.16)
$e^x - 1, x/(1 + e^{-x})$	2	3/4
$\sin(x), \tanh(x), \arctan(x)$	3	2/3
$x^m + x$	m	$(m+1)/2m$

Example 5.3. For a three-layer neural network with $1 \leq H_0 < H$, Example 5.1(2) shows that there exists a realization parameter such that the model satisfies Assumption 3 with $(d_1, r, m) = (H + H_0, 2H_0, 2)$. Therefore, an upper bound on the learning coefficient of this model is given by

$$\lambda \leq \frac{H + 3H_0}{4} \quad (5.17)$$

For example, when $\sigma(x) = \tanh(x)$, the learning coefficient λ is known from [3]. When $H \leq H_0 + 9$, it is given by

$$\lambda = H_0 - \frac{1}{2} + \frac{k_1^2 + k_1 + 2(H - H_0 + 1)}{4(k_1 + 1)};$$

$$k_1 := \max\{i \in \mathbb{Z} \mid i^2 + i \leq 2(H - H_0 + 1)\}$$

We compare this value with the upper bound in (5.17) for $H_0 = 1, 2, 3$ in Table 5.2 and Figure 5.1. We observe that the bound matches λ when H is small, while the difference between them increases as H becomes larger.

Table 5.2: Learning coefficients of three-layer neural networks when $\sigma = \tanh$

		H								
$H_0 = 1$		2	3	4	5	6	7	8	9	10
upper bound (5.17)		$\frac{5}{4}$	$\frac{3}{2}$	$\frac{7}{4}$	2	$\frac{9}{4}$	$\frac{5}{2}$	$\frac{11}{4}$	3	$\frac{13}{4}$
λ		$\frac{5}{4}$	$\frac{3}{2}$	$\frac{5}{3}$	$\frac{11}{6}$	2	$\frac{17}{8}$	$\frac{9}{4}$	$\frac{19}{8}$	$\frac{5}{2}$

		H								
$H_0 = 2$		3	4	5	6	7	8	9	10	
upper bound (5.17)	-	$\frac{9}{4}$	$\frac{5}{2}$	$\frac{11}{4}$	3	$\frac{13}{4}$	$\frac{7}{2}$	$\frac{15}{4}$	4	
λ	-	$\frac{9}{4}$	$\frac{5}{2}$	$\frac{8}{3}$	$\frac{17}{6}$	3	$\frac{25}{8}$	$\frac{13}{4}$	$\frac{27}{8}$	

		H								
$H_0 = 3$		4	5	6	7	8	9	10		
upper bound (5.17)	-	-	$\frac{13}{4}$	$\frac{7}{2}$	$\frac{15}{4}$	4	$\frac{17}{4}$	$\frac{9}{2}$	$\frac{19}{4}$	
λ	-	-	$\frac{13}{4}$	$\frac{7}{2}$	$\frac{11}{3}$	$\frac{23}{6}$	4	$\frac{33}{8}$	$\frac{17}{4}$	

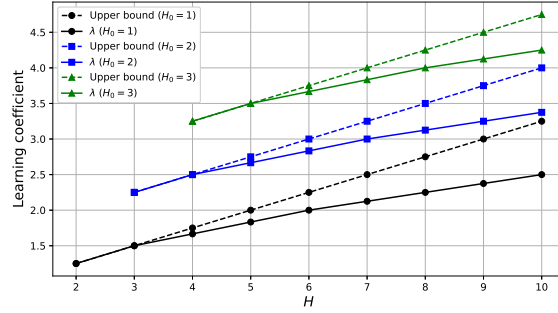


Figure 5.1: Graph of learning coefficients of three-layer neural networks ($\sigma = \tanh$)

5.3.2 Mixture Models

Example 5.4. The mixture model in Example 5.1(1) satisfies Assumption 2 as shown in Remark 4.1. For a point $P \in \Theta_*$, the model satisfies Assump-

tion 3 with $(d_1, r, m) = (H + H_0 - 1, 2H_0 - 1, 2)$. Hence, by the Main Theorem, the local learning coefficient is

$$\lambda_P = \frac{3H_0 + H - 2}{4}, \quad \mathbf{m} = 1.$$

For mixture Poisson distributions and mixture normal distributions with variance 1, the learning coefficient is known from [21], [8]. In each case, the value coincides with the upper bound derived here. Therefore, this upper bound is tight.

We verify these results through numerical experiments. We use the estimator $\widehat{\lambda}^{100}$ proposed by [15]:

$$\widehat{\lambda}^{100} := \frac{1}{100} \sum_{k=1}^{100} \widehat{\lambda}, \quad \widehat{\lambda} := t^2 \mathbb{V}_\theta^t \{\log p(X^n|\theta)\}, \quad t := \frac{1}{\log(n)}$$

$$\mathbb{E}_\theta^t \{h(\theta)\} := \frac{\int h(\theta) p(X^n|\theta)^t \varphi(\theta) d\theta}{\int p(X^n|\theta)^t \varphi(\theta) d\theta}, \quad \mathbb{V}_\theta^t \{h(\theta)\} := \mathbb{E}_\theta^t \{h(\theta)^2\} - (\mathbb{E}_\theta^t \{h(\theta)\})^2$$

where $p(X^n|\theta)$ represents the likelihood function for n independent and identically distributed samples following the true distribution q .

We conducted the numerical experiments with $n = 10000$, using the Hamiltonian Monte Carlo method implemented in the R package *RStan* to obtain the posterior distribution.

We let $M = 4$ and $H_0 \in \{1, 2\}$. We assume a prior distribution $\varphi(\theta)$ with $\theta \sim \text{Unif}(0, 1)$ for the binomial parameter. For $H_0 = 1$, the success probability of the true distribution is 0.5. For $H_0 = 2$, the true distribution is a mixture of two binomial distributions with success probabilities 0.1 and 0.9, each with equal mixture weights. Table 5.3 and Figure 5.2 show that the estimated values indeed lie below the theoretical upper bound from the Main Theorem.

Table 5.3: Estimated learning coefficients for the binomial distribution

	H							
$H_0 = 1$	1	2	3	4	5	6	7	8
Theoretical upper bound	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25
Estimated value	0.495	0.728	0.954	1.174	1.390	1.609	1.820	2.044
$H_0 = 2$								
Theoretical upper bound	-	1.50	1.75	2.00	2.25	2.50	2.75	3.00
Estimated value	-	1.502	1.694	1.879	2.044	2.214	2.389	2.550

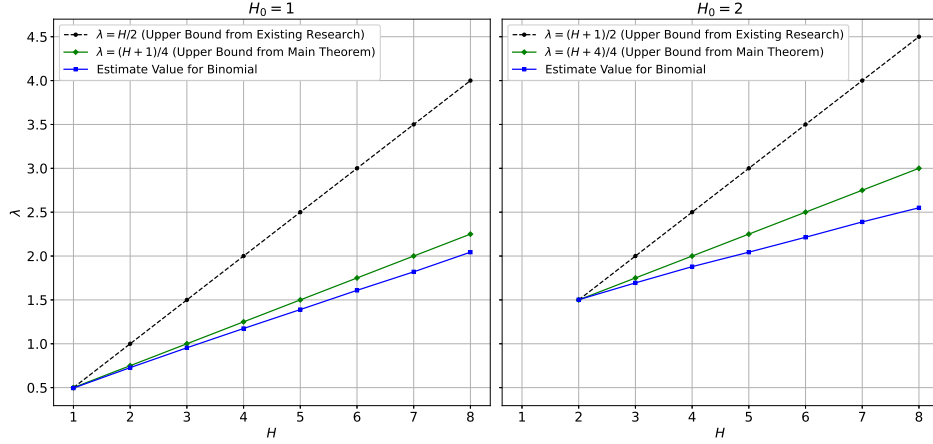


Figure 5.2: Graph of estimated learning coefficients of the binomial distribution

5.4 Proof of the Main Theorem

We use the following theorem for proving the Main Theorem.

Theorem 5.1. *Suppose that the statistical model $p(x|\theta, \tau)$ satisfies (5.1), Assumption 2, and Assumption 3(1). For any fixed $\tau \in V$, consider $K(\theta, \tau)$ and $H(\theta, \tau)$ as functions of θ . Then, the Taylor expansion at $\theta = 0$ can be expressed as*

$$\begin{aligned}
 K(\theta, \tau) &= \frac{1}{2} \mathbb{E}_X \left[\{F_1(X|\theta_1, \dots, \theta_r) + F_m(X|\theta_{r+1}, \dots, \theta_{d_1})\}^2 \right] \\
 &\quad + (\text{higher order terms}), \\
 H(\theta, \tau) &= \frac{1}{2} \mathbb{E}_X \left[\{F_1(X|\theta_1, \dots, \theta_r) + F_m(X|\theta_{r+1}, \dots, \theta_{d_1})\}^2 q(X) \right] \\
 &\quad + (\text{higher order terms}).
 \end{aligned}$$

The first equation is valid if K is analytic. Here, we define

$$F_1(X|s_1, \dots, s_r) := \sum_{k=1}^r \frac{\partial f}{\partial \theta_k} \Big|_{\theta=0} s_k,$$

and the higher order terms do not include the following:

- Terms of degree $2m$ or less consisting only of $\theta_{r+1}, \dots, \theta_{d_1}$.
- Terms of first degree in $\theta_1, \dots, \theta_r$ and up to m -th degree in $\theta_{r+1}, \dots, \theta_{d_1}$.
- Terms of second degree consisting only of $\theta_1, \dots, \theta_r$.

Proof. We fix any $\tau \in V$. From (5.1) and (5.2), there is no constant term in the Taylor expansions of $K(\theta, \tau)$ or $H(\theta, \tau)$ at $\theta = 0$. In other words, a term consisting only of τ does not exist. Moreover, by Assumption 3(1), for all indices $(i_{r+1}, \dots, i_{d_1})$ satisfying $1 \leq i_{r+1} + \dots + i_{d_1} \leq m - 1$, we have

$$\left. \frac{\partial^{i_{r+1} + \dots + i_{d_1}} f(X|\theta, \tau)}{\partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_{d_1}^{i_{d_1}}} \right|_{\theta=0} = 0 \quad \text{a.s..}$$

Therefore, we can apply Theorem 3.2 and Theorem 4.2, which lead to the following equation:

$$\begin{aligned} K(\theta, \tau) &= \frac{1}{2} \mathbb{E}_X \left[\{F_1(X|\theta_1, \dots, \theta_r) + F_m(X|\theta_{r+1}, \dots, \theta_{d_1})\}^2 \right] \\ &\quad + (\text{higher order terms}), \\ H(\theta, \tau) &= \frac{1}{2} \mathbb{E}_X \left[\{F_1(X|\theta_1, \dots, \theta_r) + F_m(X|\theta_{r+1}, \dots, \theta_{d_1})\}^2 q(X) \right] \\ &\quad + (\text{higher order terms}). \end{aligned}$$

This completes the proof. $\square$

Example 5.5. We verify the validity of Theorem 5.1 and the Main Theorem through explicit computations. Let X be a random variable following the binomial distribution $\text{Bin}(2, \theta)$ with parameter θ , where $0 < \theta < 1$:

$$\tilde{p}(X = x|\theta) = \binom{2}{x} \theta^x (1 - \theta)^{2-x} = \begin{cases} (1 - \theta)^2, & (x = 0) \\ 2\theta(1 - \theta), & (x = 1) \\ \theta^2, & (x = 2). \end{cases}$$

Consider the following mixture model with parameters $(\theta_1, \theta_2, \tau)$:

$$p(X|\theta, \tau) := \left(\frac{1}{2} - \tau\right) \cdot \tilde{p}\left(X \middle| \theta_1 - \frac{1+2\tau}{1-2\tau}\theta_2 + \frac{1}{2}\right) + \left(\tau + \frac{1}{2}\right) \cdot \tilde{p}\left(X \middle| \theta_2 + \frac{1}{2}\right). \quad (5.18)$$

Assuming the true distribution is $\tilde{p}(X|1/2)$, this model realizes the true distribution at parameters $(\theta_1, \theta_2, \tau) = (0, 0, 0)$. We determine the real log canonical threshold at the point $P : (\theta_1, \theta_2, \tau) = (0, 0, 0) \in \Theta_*$.

Because this model is a mixture of two binomial distributions, it satisfies Assumption 2 by Remark 4.1. Thus, we do not need to show that $K(\theta, \tau)$ is analytic, and it suffices to compute the real log canonical threshold of the analytic function $H(\theta, \tau)$. We fix any τ in a neighborhood V of 0. Since $f(x|\theta = 0, \tau) = 0$, condition (5.1) is satisfied. The following derivatives exist at $(\theta_1, \theta_2) = 0$:

$$\left. \frac{\partial f(X|\theta, \tau)}{\partial \theta_1} \right|_{(\theta_1, \theta_2)=0}, \left. \frac{\partial f(X|\theta, \tau)}{\partial \theta_2} \right|_{(\theta_1, \theta_2)=0}, \left. \frac{\partial^2 f(X|\theta, \tau)}{\partial \theta_2^2} \right|_{(\theta_1, \theta_2)=0}$$

resulting, respectively, in:

$$2(1-2\tau) \times \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \frac{8(1+2\tau)}{1-2\tau} \times \begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix}.$$

Each column vector corresponds to $x = 0$, $x = 1$, and $x = 2$. Since the first and third vectors are linearly independent, the model satisfies Assumption 3 with $(d_1, r, m) = (2, 1, 2)$.

$$\begin{aligned} F_1(X|s_1) + F_2(X|s_2) &= \frac{\partial f(X|\theta, \tau)}{\partial \theta_1} \Big|_{(\theta_1, \theta_2)=0} s_1 + \frac{1}{2} \frac{\partial^2 f(X|\theta, \tau)}{\partial \theta_2^2} \Big|_{(\theta_1, \theta_2)=0} s_2^2 \\ &= \begin{bmatrix} 2(1-2\tau)s_1 - \frac{4(1+2\tau)}{1-2\tau}s_2^2 \\ \frac{4(1+2\tau)}{1-2\tau}s_2^2 \\ -2(1-2\tau)s_1 - \frac{4(1+2\tau)}{1-2\tau}s_2^2 \end{bmatrix} \end{aligned}$$

Then,

$$\begin{aligned} &\frac{1}{2} \mathbb{E}_X [\{F_1(X|s_1) + F_2(X|s_2)\}^2 q(X)] \\ &= \frac{1}{2} \left\{ \left(2(1-2\tau)s_1 - \frac{4(1+2\tau)}{1-2\tau}s_2^2 \right)^2 \cdot \frac{1}{4^2} + \left(\frac{4(1+2\tau)}{1-2\tau}s_2^2 \right)^2 \cdot \frac{1}{2^2} \right. \\ &\quad \left. + \left(-2(1-2\tau)s_1 - \frac{4(1+2\tau)}{1-2\tau}s_2^2 \right)^2 \cdot \frac{1}{4^2} \right\} \\ &= \frac{(1-2\tau)^2}{4} s_1^2 + \frac{3(1+2\tau)^2}{(1-2\tau)^2} s_2^4. \end{aligned}$$

On the other hand, we obtain:

$$\begin{aligned} H(\theta, \tau) &= \frac{(1-2\tau)^2}{4} \theta_1^2 + \frac{3(1+2\tau)^2}{(1-2\tau)^2} \theta_2^4 + \frac{3(1-2\tau)^2}{4} \theta_1^4 - 3(1-4\tau^2) \theta_1^3 \theta_2 \\ &\quad + 6(1+\tau)(1+2\tau) \theta_1^2 \theta_2^2 - \frac{6(1+2\tau)^2}{1-2\tau} \theta_1 \theta_2^3 \\ &\quad + (\text{terms of order 5 or higher}). \end{aligned} \tag{5.19}$$

where terms of order five and higher are omitted. Therefore, as stated in Theorem 5.1, the lower order terms of $H(\theta, \tau)$ are indeed

$$\frac{1}{2} \mathbb{E}_X [\{F_1(X|\theta_1) + F_2(X|\theta_2)\}^2 q(X)] = \frac{(1-2\tau)^2}{4} \theta_1^2 + \frac{3(1+2\tau)^2}{(1-2\tau)^2} \theta_2^4.$$

Next, we verify that the Main Theorem holds by performing the coordinate transformation g , known as a "blow-up" (CT1)-(CT4) as illustrated in Figure 5.3. This transformation ensures that a normal crossing is obtained for each local coordinate. We define $W_0 := \{(\theta_1, \theta_2, \tau) \mid \theta_1 = 0, \theta_2 = 0, \forall \tau \in V\}$.

(CT1) Transforming with $\theta_2 = \theta_1 \theta'_2$ in (5.19) gives:

$$\begin{aligned} H(\theta, \tau) &= \theta_1^2 \left\{ \frac{(1-2\tau)^2}{4} + \frac{3(1+2\tau)^2}{(1-2\tau)^2} \theta_1^2 \theta_2'^4 + \theta_1 \times \dots \right\} \\ &= \theta_1^2 a_1(\theta_1, \theta'_2, \tau), \quad a_1(0, \theta'_2, \tau) = \frac{(1-2\tau)^2}{4} > 0. \end{aligned} \quad (5.20)$$

Considering points on $U_0 = g^{-1}(W_0)$, the condition $\theta_1 = 0$ ($\forall \tau \in V$) is satisfied. Therefore, a normal crossing is achieved in this local coordinate system $(\theta_1, \theta'_2, \tau)$.

(CT2) Transforming with $\theta_1 = \theta_2 \theta'_1$ in (5.19) results in:

$$\begin{aligned} H(\theta, \tau) &= \theta_2^2 \left\{ \frac{(1-2\tau)^2}{4} \theta_1'^2 + \frac{3(1+2\tau)^2}{(1-2\tau)^2} \theta_2^2 + \theta_2 \times \dots \right\} \\ &= \theta_2^2 a_2(\theta'_1, \theta_2, \tau), \quad a_2(\theta'_1, 0, \tau) = \frac{(1-2\tau)^2}{4} \theta_1'^2. \end{aligned} \quad (5.21)$$

Considering points on $U_0 = g^{-1}(W_0)$, the condition $\theta_2 = 0$ ($\forall \tau \in V$) is satisfied. Therefore, a normal crossing is achieved in this local coordinate system $(\theta'_1, \theta_2, \tau)$ except at the points $(\theta'_1, \theta_2) = 0$.

To determine the normal crossing at the points $(\theta'_1, \theta_2) = 0$, we perform another blow-up centered at these points. Under the transformations (CT3): $\theta_2 = \theta'_1 \theta''_2$ and (CT4): $\theta'_1 = \theta_2 \theta''_1$, we obtain:

$$\begin{aligned} H(\theta, \tau) &= \theta_1'^4 \theta_2''^2 \left\{ \frac{(1-2\tau)^2}{4} + \frac{3(1+2\tau)^2}{(1-2\tau)^2} \theta_2''^2 + \theta'_1 \times \dots \right\} \\ &= \theta_1'^4 \theta_2''^2 a_3(\theta'_1, \theta''_2, \tau), \end{aligned} \quad (5.22)$$

$$\forall \tau \in V, a_3(0, \theta''_2, \tau) = \frac{(1-2\tau)^2}{4} + \frac{3(1+2\tau)^2}{(1-2\tau)^2} \theta_2''^2 > 0,$$

$$\begin{aligned} H(\theta, \tau) &= \theta_2^4 \left\{ \frac{(1-2\tau)^2}{4} \theta_1''^2 + \frac{3(1+2\tau)^2}{(1-2\tau)^2} + \theta_2 \times \dots \right\} \\ &= \theta_2^4 a_4(\theta''_1, \theta_2, \tau), \end{aligned} \quad (5.23)$$

$$\forall \tau \in V, a_4(\theta''_1, 0, \tau) = \frac{(1-2\tau)^2}{4} \theta_1''^2 + \frac{3(1+2\tau)^2}{(1-2\tau)^2} > 0.$$

Therefore, normal crossings are achieved in these local coordinates.

We summarize the normal crossings for each local coordinate in Table 5.4, where the symbols $g, k_i^{(Q)}, h_i^{(Q)}$ refer to Definition 2.2. Thus, we conclude that the real log canonical threshold at P is $3/4$ with multiplicity $\mathbf{m} = 1$.

Table 5.4: Normal crossings for each local coordinate

	u	$g(u)$	$H(g(u))$	$ g'(u) $	$k_i^{(Q)}$	$h_i^{(Q)}$	λ
(5.20)	$(\theta_1, \theta'_2, \tau)$	$(\theta_1, \theta_1 \theta'_2, \tau)$	$\theta_1^2 a_1(\theta_1, \theta'_2, \tau)$	$ \theta_1 $	$(2, 0, 0)$	$(1, 0, 0)$	1
(5.22)	$(\theta'_1, \theta''_2, \tau)$	$(\theta_1'^2 \theta''_2, \theta'_1 \theta''_2, \tau)$	$\theta_1'^4 \theta''_2^2 a_3(\theta'_1, \theta''_2, \tau)$	$ \theta_1'^2 \theta''_2 $	$(4, 2, 0)$	$(2, 1, 0)$	3/4
(5.23)	$(\theta'_1, \theta_2, \tau)$	$(\theta_1' \theta_2^2, \theta_2, \tau)$	$\theta_2^4 a_4(\theta'_1, \theta_2, \tau)$	$ \theta_2^2 $	$(0, 4, 0)$	$(0, 2, 0)$	3/4

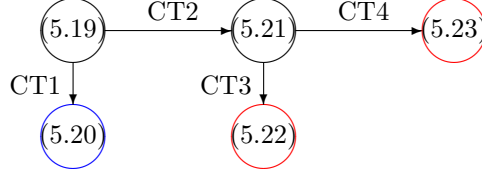


Figure 5.3: The procedure of coordinate transformation g performed in Example 5.5

Note. Local coordinates displayed in red represent the minimal real log canonical thresholds.

Proof of the Main Theorem. We restrict our proof to the case where $K(\theta, \tau)$ is analytic, but a similar argument can be used to prove the result in the nonanalytic case by replacing K with an analytic function H . (See the proof of Theorem 4.1.)

By applying Theorem 5.1, for each fixed $\tau \in V$, the Taylor expansion of $K(\theta, \tau)$ with respect to θ at $(\theta_1, \dots, \theta_{d_1}) = 0$ can be expressed as:

$$K(\theta, \tau) = \frac{1}{2} \mathbb{E}_X \left[\{F_1(X|\theta_1, \dots, \theta_r) + F_m(X|\theta_{r+1}, \dots, \theta_{d_1})\}^2 \right] + (\text{higher order terms}).$$

We perform the following four coordinate transformations, CT1-CT4 (see Figure 5.4).¹ ; in the case $r = 0$, we use only CT2. We demonstrate that normal crossings are achieved in all the resulting local coordinates.

CT1 For $1 \leq s \leq r$, $\{\theta_j \rightarrow \theta_s \theta'_j \mid 1 \leq j \leq d_1, j \neq s\}$.

CT2 For $r+1 \leq t \leq d_1$, $\{\theta_j \rightarrow \theta_t \theta'_j \mid 1 \leq j \leq d_1, j \neq t\}$.

CT3 For the t selected in CT2 and for $1 \leq s \leq r$, $\{\theta'_j \rightarrow \theta_s \theta''_j, \theta_t \rightarrow \theta_s \theta'_t \mid 1 \leq j \leq r, j \neq s\}$

CT4 For the t selected in CT2, $\{\theta'_j \rightarrow \theta_t \theta''_j \mid 1 \leq j \leq r\}$.

¹In terms of algebraic geometry, CT1 and CT2 correspond to “blow-ups centered at the subvariety $W_0 = \{(\theta, \tau) \mid \theta = 0, \tau \in V\}$ ”. CT3 and CT4 correspond to “blow-ups centered at the subvariety $\{(\theta, \tau) \mid \theta_1 = \dots = \theta_r = \theta_t = 0, \tau \in V\}$ when the exceptional surface from CT2 is $\{\theta_t = 0\}$ ”.

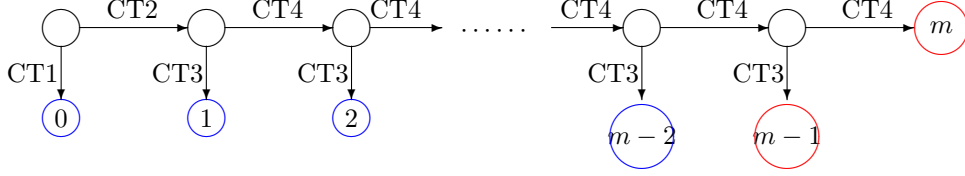


Figure 5.4: The procedure of coordinate transformations g performed in the proof of the Main Theorem

Note. Local coordinates displayed in red represent the minimal real log canonical thresholds.

Initially, we perform CT1 and CT2. The first transformation achieves normal crossings, but the second does not. Therefore, we proceed to perform further iterations involving CT3 and CT4. In these iterations, CT3 achieves normal crossings, but CT4 does not. By repeating this process up to $m - 1$ times (resulting in a total of m iterations), we eventually achieve normal crossings. For detailed calculations, see the proof of Theorem 3.1.

Here, we verify only the following three inequalities, which are crucial for achieving normal crossings:

$$\begin{aligned}
 a_1(s_2, \dots, s_{d_1}, \tau) &:= \mathbb{E}_X [F_1^2(X|1, s_2, \dots, s_r)] > 0, \\
 a_2(s_2, \dots, s_{d_1}, \tau) &:= \mathbb{E}_X \left[\left\{ F_1(X|1, s_2, \dots, s_r) + s_{d_1} F_m(X|s_{r+1}, \dots, s_{d_1-1}, 1) \right\}^2 \right] > 0, \\
 a_3(s_1, \dots, s_{d_1-1}, \tau) &:= \mathbb{E}_X \left[\left\{ F_1(X|s_1, \dots, s_r) + F_m(X|s_{r+1}, \dots, s_{d_1-1}, 1) \right\}^2 \right] > 0. \\
 (s_i \in \mathbb{R}, \tau \in V)
 \end{aligned}$$

For the first inequality, fixing any $\tau \in V$, $a_1(s_2, \dots, s_{d_1}, \tau) = 0$ implies

$$F_1(X|1, s_2, \dots, s_r) = 0 \text{ a.s.} \Leftrightarrow \frac{\partial f}{\partial \theta_1} \Big|_{\theta=0} + \sum_{k=2}^r s_k \cdot \frac{\partial f}{\partial \theta_k} \Big|_{\theta=0} = 0 \text{ a.s..}$$

Since

$$\frac{\partial f}{\partial \theta_1} \Big|_{\theta=0}, \dots, \frac{\partial f}{\partial \theta_r} \Big|_{\theta=0} \quad (5.24)$$

are linearly independent, there exists no $(s_2, \dots, s_{d_1}, \tau)$ that satisfies this equivalence. Therefore, we conclude that $a_1(s_2, \dots, s_{d_1}, \tau) > 0$ holds.

The second inequality is analogous to the third, so we verify only the

third. $a_3(s_1, \dots, s_{d_1-1}, \tau) = 0$ implies

$$\begin{aligned} & F_1(X|s_1, \dots, s_r) + F_m(X|s_{r+1}, \dots, s_{d_1-1}, 1) = 0 \quad \text{a.s.} \\ \Leftrightarrow & \sum_{k=1}^r s_k \cdot \left. \frac{\partial f}{\partial \theta_k} \right|_{\theta=0} + F_m(X|s_{r+1}, \dots, s_{d_1-1}, 1) = 0 \quad \text{a.s..} \end{aligned}$$

For each parameter $(s_{r+1}, \dots, s_{d_1-1}, 1) \neq 0$, Assumption 3(2) implies that the random variable $F_m(X|s_{r+1}, \dots, s_{d_1-1}, 1)$ and the r derivatives (5.24) are linearly independent for each fixed $\tau \in V$. Thus, there does not exist $(s_1, \dots, s_{d_1-1}, \tau)$ satisfying this equivalence relation, so we conclude that $a_3(s_1, \dots, s_{d_1-1}, \tau) > 0$ holds. $\square$

Remark 5.6. Remark 3.4 shows that the only point in the neighborhood of the origin $\theta = 0$ in $\mathbb{R}^d$ where $K(\theta) = 0$ is the origin itself. Applying the same argument to $K(\theta, \tau)$ implies that, in a neighborhood U_1 of the point P in $\mathbb{R}^{d_1+d_2}$, the points satisfying $K(\theta, \tau) = 0$ occur only when $\theta = 0$. Because the τ -direction remains arbitrary,

$$\Theta_* \cap U_1 = \{(\theta, \tau) \mid \theta = 0, \tau \in V\}.$$

Hence by Definition 5.1, P is a nonsingular point of Θ_* .

Chapter 6

Conclusions and future work

In this thesis, we defined semiregular models as a generalization of regular models and provided a formula for the learning coefficient λ for semiregular models (Theorem 3.1).

Furthermore, we generalized Theorem 3.1 in two directions. First, we proved that the same formula holds even when $K(\theta)$ is not analytic (Theorem 4.1). Second, in the Main Theorem we derived a formula for the local learning coefficient at nonsingular points when Θ_* is not a single point, and we provided an upper bound on the learning coefficient for a certain class of three-layer neural networks and mixture models.

The derived upper bound coincides with the learning coefficient for mixture models, but a gap remains in the case of three-layer neural networks when the number of hidden units is general. To narrow this gap, we need to consider the local learning coefficient at singular points of Θ_* . For future research, we intend to further generalize the methods used in this study and derive local learning coefficients at singular points of Θ_* .

Appendix A

Proof of Theorem 3.2

Let $i_1, \dots, i_k \in \{1, \dots, d\}$. We define

$$Y_{i_1, \dots, i_k} := -\frac{\partial^k f(x|\theta)}{\partial \theta_{i_1} \dots \partial \theta_{i_k}},$$

where we aim to identify patterns in the partial derivatives of the log-likelihood ratio function f . We refer to $Y_{i_1, \dots, i_k}$ as indeterminate symbols of weight k . Setting $\theta = (\theta_1, \dots, \theta_d)$, we obtain the following:

$$\begin{aligned} \frac{\partial f(x|\theta)}{\partial \theta_{i_1}} &= -\frac{1}{p(x|\theta)} \frac{\partial p(x|\theta)}{\partial \theta_{i_1}}, \\ \frac{\partial^2 f(x|\theta)}{\partial \theta_{i_1} \partial \theta_{i_2}} &= -\frac{1}{p(x|\theta)} \frac{\partial^2 p(x|\theta)}{\partial \theta_{i_1} \partial \theta_{i_2}} + Y_{i_1} Y_{i_2}, \\ \frac{\partial^3 f(x|\theta)}{\partial \theta_{i_1} \partial \theta_{i_2} \partial \theta_{i_3}} &= -\frac{1}{p(x|\theta)} \frac{\partial^3 p(x|\theta)}{\partial \theta_{i_1} \partial \theta_{i_2} \partial \theta_{i_3}} + Y_{i_1} Y_{i_2} Y_{i_3} \\ &\quad + Y_{i_1, i_2} Y_{i_3} + Y_{i_2, i_3} Y_{i_1} + Y_{i_3, i_1} Y_{i_2}, \\ \frac{\partial^4 f(x|\theta)}{\partial \theta_{i_1} \partial \theta_{i_2} \partial \theta_{i_3} \partial \theta_{i_4}} &= -\frac{1}{p(x|\theta)} \frac{\partial^4 p(x|\theta)}{\partial \theta_{i_1} \partial \theta_{i_2} \partial \theta_{i_3} \partial \theta_{i_4}} + Y_{i_1} Y_{i_2} Y_{i_3} Y_{i_4} \\ &\quad + Y_{i_1, i_2} Y_{i_3} Y_{i_4} + Y_{i_1, i_3} Y_{i_2} Y_{i_4} + Y_{i_1, i_4} Y_{i_2} Y_{i_3} \\ &\quad + Y_{i_2, i_3} Y_{i_1} Y_{i_4} + Y_{i_2, i_4} Y_{i_1} Y_{i_3} + Y_{i_3, i_4} Y_{i_1} Y_{i_2} \\ &\quad + Y_{i_1, i_2} Y_{i_3, i_4} + Y_{i_1, i_3} Y_{i_2, i_4} + Y_{i_1, i_4} Y_{i_2, i_3} \\ &\quad + Y_{i_1, i_2, i_3} Y_{i_4} + Y_{i_1, i_2, i_4} Y_{i_3} + Y_{i_1, i_3, i_4} Y_{i_2} + Y_{i_2, i_3, i_4} Y_{i_1}. \end{aligned} \tag{A.1}$$

Assuming we can interchange the order of partial differentiation and integration, the first term on the right-hand side becomes zero at $\theta = 0$ when taking the expectation with respect to X . Therefore, the partial derivatives of $K(\theta)$ at $\theta = 0$ depend primarily on the terms from the second term onward.

Remark A.1. By taking the expectation with respect to X in (A.1) and substituting $\theta = 0$, we find that the Fisher information matrices I and J coincide.

The subsequent terms on the right-hand side are sums of products of two or more indeterminate symbols Y . Each term has a coefficient of 1, and the total weight of each product equals the order of differentiation. This pattern holds in general.

Proposition A.1.

$$G_n(i_1, \dots, i_n) := \frac{\partial^n f(x|\theta)}{\partial \theta_{i_1} \cdots \partial \theta_{i_n}} + \frac{1}{p(x|\theta)} \frac{\partial^n p(x|\theta)}{\partial \theta_{i_1} \cdots \partial \theta_{i_n}}$$

is a polynomial consisting of indeterminate symbols $Y_{i_1, \dots, i_k}$ with weights less than n . Each term in the polynomial has a coefficient of 1, and each term is a product of two or more indeterminates, with a total weight of n . The polynomial is the sum of all such terms with these weights.

Proof. We prove this by induction on n . For $n = 1$, $G_1(i_1) = 0$ satisfies the condition since there are no terms. Assuming $G_n(i_1, \dots, i_n)$ is such a polynomial for some $n \geq 1$, we obtain

$$\begin{aligned} & \frac{\partial^{n+1} f(x|\theta)}{\partial \theta_{i_1} \cdots \partial \theta_{i_{n+1}}} \\ &= \frac{\partial}{\partial \theta_{i_{n+1}}} \left\{ -\frac{1}{p(x|\theta)} \frac{\partial^n p(x|\theta)}{\partial \theta_{i_1} \cdots \partial \theta_{i_n}} + G_n(i_1, \dots, i_n) \right\} \\ &= -\frac{1}{p(x|\theta)} \frac{\partial^{n+1} p(x|\theta)}{\partial \theta_{i_1} \cdots \partial \theta_{i_{n+1}}} + \frac{1}{p(x|\theta)} \frac{\partial^n p(x|\theta)}{\partial \theta_{i_1} \cdots \partial \theta_{i_n}} Y_{i_{n+1}} + \frac{\partial G_n(i_1, \dots, i_n)}{\partial \theta_{i_{n+1}}} \\ &= -\frac{1}{p(x|\theta)} \frac{\partial^{n+1} p(x|\theta)}{\partial \theta_{i_1} \cdots \partial \theta_{i_{n+1}}} + G_n(i_1, \dots, i_n) Y_{i_{n+1}} + Y_{i_1, \dots, i_n} Y_{i_{n+1}} \\ & \quad + \frac{\partial G_n(i_1, \dots, i_n)}{\partial \theta_{i_{n+1}}}. \end{aligned}$$

Thus,

$$G_{n+1}(i_1, \dots, i_{n+1}) = G_n(i_1, \dots, i_n) Y_{i_{n+1}} + Y_{i_1, \dots, i_n} Y_{i_{n+1}} + \frac{\partial G_n(i_1, \dots, i_n)}{\partial \theta_{i_{n+1}}}.$$

By the induction hypothesis, this is a polynomial consisting of indeterminate symbols Y with weights less than or equal to n , where each term has a coefficient of 1 and the sum of weights in each term equals $n + 1$. Moreover, the terms containing the indeterminate $Y_{i_{n+1}}$ are derived from the first and second terms on the right-hand side, while the terms that do not contain $Y_{i_{n+1}}$ are derived from the third term on the right-hand side. Thus, we have shown that each term is a product of two or more indeterminate symbols, summing to a total weight of $n + 1$. $\square$

Remark A.2. The statement of Proposition A.1 can be written explicitly as follows:

$$G_n(i_1, \dots, i_n) = \sum_{T \in S_{i_1, \dots, i_n}} \prod_{U \in T} Y_U, \quad (\text{A.2})$$

where $S_{i_1, \dots, i_n}$ is the set of all disjoint partitions of the set $\{i_1, \dots, i_n\}$. A disjoint partition satisfies

$$\bigcup_i U_i = \{i_1, \dots, i_n\}, \quad U_i \cap U_j = \{\} \text{ for } i \neq j, \quad U_i \neq \{\}, \{i_1, \dots, i_n\},$$

and each partition is denoted as $T = \{U_1, U_2, \dots\}$.

For example, when $n = 2$, the disjoint partition of $\{i_1, i_2\}$ consists of a single element, $T_1 = \{\{i_1\}, \{i_2\}\}$, so $S_{i_1, i_2} = \{T_1\}$. Substituting this into equation (A.2), we have

$$G_2(i_1, i_2) = \prod_{U \in T_1} Y_U = Y_{i_1} Y_{i_2}.$$

Similarly, when $n = 3$, the disjoint partitions of $\{i_1, i_2, i_3\}$ are given by $T_1 = \{\{i_1\}, \{i_2\}, \{i_3\}\}$, $T_2 = \{\{i_1, i_2\}, \{i_3\}\}$, $T_3 = \{\{i_2, i_3\}, \{i_1\}\}$, $T_4 = \{\{i_3, i_1\}, \{i_2\}\}$, so $S_{i_1, i_2, i_3} = \{T_1, T_2, T_3, T_4\}$. Substituting into equation (A.2), we have

$$\begin{aligned} G_3(i_1, i_2, i_3) &= \prod_{U \in T_1} Y_U + \prod_{U \in T_2} Y_U + \prod_{U \in T_3} Y_U + \prod_{U \in T_4} Y_U \\ &= Y_{i_1} Y_{i_2} Y_{i_3} + Y_{i_1, i_2} Y_{i_3} + Y_{i_2, i_3} Y_{i_1} + Y_{i_3, i_1} Y_{i_2}. \end{aligned}$$

Corollary A.1.

- (1) If all indeterminate symbols of weight $m - 1$ or less are zero almost surely at $\theta = 0$, then G_k (for $k \leq 2m - 1$) will also be zero almost surely at $\theta = 0$.
- (2) When the interchange of differentiation and integration is permissible,

$$\left. \frac{\partial^n K(\theta)}{\partial \theta_{i_1} \cdots \partial \theta_{i_n}} \right|_{\theta=0} = \mathbb{E}_X \left[\left. \frac{\partial^n f(X|\theta)}{\partial \theta_{i_1} \cdots \partial \theta_{i_n}} \right|_{\theta=0} \right]$$

equals the expectation over X of a polynomial. This polynomial is obtained by summing all terms where at least two indeterminates $Y_{i_1, \dots, i_k}$ are multiplied together, the total weight is n , and the coefficients of all terms are 1.

Proof. (1) follows because terms with a sum of weights up to $2m - 1$ are products of two or more terms, so at least one term will have a weight of $m - 1$ or less and thus will be zero almost surely at $\theta = 0$. (2) follows directly from the previous discussions. $\square$

Proof of Theorem 3.2. We need to prove the following five points about the Taylor expansion of $K(\theta)$ at $\theta = 0$:

- (1) Coefficients of terms of order zero in $(\theta_1, \dots, \theta_r)$ and of order $2m - 1$ or less in $(\theta_{r+1}, \dots, \theta_d)$ are zero.
- (2) Coefficients of terms of first order in $(\theta_1, \dots, \theta_r)$ and of order $m - 1$ or less in $(\theta_{r+1}, \dots, \theta_d)$ are zero.
- (3) Terms of first order in $(\theta_1, \dots, \theta_r)$ and of m -th order in $(\theta_{r+1}, \dots, \theta_d)$ are represented by:

$$\mathbb{E}_X [F_1(X|\theta_1, \dots, \theta_r) F_m(X|\theta_{r+1}, \dots, \theta_d)].$$

- (4) Terms of second order in $(\theta_1, \dots, \theta_r)$ and of zero order in $(\theta_{r+1}, \dots, \theta_d)$ are represented by:

$$\frac{1}{2} \mathbb{E}_X [F_1^2(X|\theta_1, \dots, \theta_r)].$$

- (5) Terms of zero order in $(\theta_1, \dots, \theta_r)$ and of $2m$ -th order in $(\theta_{r+1}, \dots, \theta_d)$ are represented by:

$$\frac{1}{2} \mathbb{E}_X [F_m^2(X|\theta_{r+1}, \dots, \theta_d)].$$

For (1), it suffices to show that for $i_{r+1} + \dots + i_d \leq 2m - 1$,

$$\left. \frac{\partial^{i_{r+1} + \dots + i_d} f(X|\theta)}{\partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_d^{i_d}} \right|_{\theta=0}$$

is a random variable with an expected value of zero. This has already been demonstrated in Corollary A.1(1). For (2), for $1 \leq j \leq r$ and $i_{r+1} + \dots + i_d \leq m - 1$,

$$\left. \frac{\partial^{1+i_{r+1} + \dots + i_d} f(X|\theta)}{\partial \theta_j \partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_d^{i_d}} \right|_{\theta=0}$$

is shown to have an expected value of zero. Among the products of two or more indeterminates claimed in Corollary A.1(2), those that do not include θ_j are zero with probability 1 by assumption. Therefore, their expected value is zero.

For (3), using a similar argument as in (2), consider a case where $1 \leq j \leq r$ and $i_{r+1} + \dots + i_d = m$. For

$$\left. \frac{\partial^{m+1} f(X|\theta)}{\partial \theta_j \partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_d^{i_d}} \right|_{\theta=0},$$

the nonzero products of two or more indeterminates claimed in Corollary A.1(2) are limited to

$$\left. \frac{\partial f(X|\theta)}{\partial \theta_j} \right|_{\theta=0} \left. \frac{\partial^m f(X|\theta)}{\partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_d^{i_d}} \right|_{\theta=0}.$$

Therefore,

$$\mathbb{E}_X \left[\frac{\partial^{m+1} f(X|\theta)}{\partial \theta_j \partial \theta_{r+1}^{i_{r+1}} \cdots \partial \theta_d^{i_d}} \Big|_{\theta=0} \right] = \mathbb{E}_X \left[\frac{\partial f(X|\theta)}{\partial \theta_j} \Big|_{\theta=0} \frac{\partial^m f(X|\theta)}{\partial \theta_{r+1}^{i_{r+1}} \cdots \partial \theta_d^{i_d}} \Big|_{\theta=0} \right],$$

validating the claim in (3).

Since (4) is a special case of (5), it suffices to demonstrate (5) alone. For $i_{r+1} + \cdots + i_d = 2m$,

$$\frac{\partial^{2m} f(X|\theta)}{\partial \theta_{r+1}^{i_{r+1}} \cdots \partial \theta_d^{i_d}} \Big|_{\theta=0}$$

is restricted to the sum over the set of nonzero products of two or more indeterminates claimed in Corollary A.1(2), given by

$$\left\{ \frac{\partial^m f(X|\theta)}{\partial \theta_{r+1}^{j_{r+1}} \cdots \partial \theta_d^{j_d}} \Big|_{\theta=0} \cdot \frac{\partial^m f(X|\theta)}{\partial \theta_{r+1}^{i_{r+1}-j_{r+1}} \cdots \partial \theta_d^{i_d-j_d}} \Big|_{\theta=0} \mid \begin{array}{l} j_{r+1} + \cdots + j_d = m, \\ i_h \geq j_h \geq 0 \end{array} \right\}.$$

Note that each element in this set appears as

$$\binom{i_{r+1}}{j_{r+1}} \cdots \binom{i_d}{j_d} \times \begin{cases} \frac{1}{2} & \text{if } (2j_{r+1}, \dots, 2j_d) = (i_{r+1}, \dots, i_d), \\ 1 & \text{otherwise.} \end{cases}$$

Thus, we can express

$$\begin{aligned} & \frac{\partial^{2m} K(\theta)}{\partial \theta_{r+1}^{i_{r+1}} \cdots \partial \theta_d^{i_d}} \Big|_{\theta=0} = \mathbb{E}_X \left[\frac{\partial^{2m} f(X|\theta)}{\partial \theta_{r+1}^{i_{r+1}} \cdots \partial \theta_d^{i_d}} \Big|_{\theta=0} \right] \\ &= \frac{1}{2} \sum_{\substack{j_{r+1} + \cdots + j_d = m \\ i_h \geq j_h \geq 0}} \binom{i_{r+1}}{j_{r+1}} \cdots \binom{i_d}{j_d} \\ & \quad \times \mathbb{E}_X \left[\frac{\partial^m f(X|\theta)}{\partial \theta_{r+1}^{j_{r+1}} \cdots \partial \theta_d^{j_d}} \Big|_{\theta=0} \cdot \frac{\partial^m f(X|\theta)}{\partial \theta_{r+1}^{i_{r+1}-j_{r+1}} \cdots \partial \theta_d^{i_d-j_d}} \Big|_{\theta=0} \right]. \end{aligned}$$

Transforming to the target expression, we find:

$$\begin{aligned} & \frac{1}{2} \mathbb{E}_X [F_m^2(X|\theta_{r+1}, \dots, \theta_d)] \\ &= \sum_{\substack{k_{r+1} + \cdots + k_d = m \\ k_h \geq 0}} \sum_{\substack{l_{r+1} + \cdots + l_d = m \\ l_h \geq 0}} \frac{\theta_{r+1}^{k_{r+1}+l_{r+1}} \cdots \theta_d^{k_d+l_d}}{2k_{r+1}! \cdots k_d! l_{r+1}! \cdots l_d!} \\ & \quad \times \mathbb{E}_X \left[\frac{\partial^m f(X|\theta)}{\partial \theta_{r+1}^{k_{r+1}} \cdots \partial \theta_d^{k_d}} \Big|_{\theta=0} \frac{\partial^m f(X|\theta)}{\partial \theta_{r+1}^{l_{r+1}} \cdots \partial \theta_d^{l_d}} \Big|_{\theta=0} \right] \\ &= \sum_{\substack{i_{r+1} + \cdots + i_d = 2m \\ i_h \geq 0}} \frac{1}{i_{r+1}! \cdots i_d!} \frac{\partial^{2m} K(\theta)}{\partial \theta_{r+1}^{i_{r+1}} \cdots \partial \theta_d^{i_d}} \Big|_{\theta=0} \theta_{r+1}^{i_{r+1}} \cdots \theta_d^{i_d}. \end{aligned}$$

The proof is completed using the variable substitutions $k_h + l_h = i_h$, $k_h = j_h$ for $h = r + 1, \dots, d$. Thus, the claim is demonstrated. $\square$

Appendix B

Derivation Method for m

Proposition B.1 (Properties of the Coordinate Transformation). *Fix non-negative integers $i_{r+1}, \dots, i_d$ with $i_{r+1} + \dots + i_d \geq 1$. For $k = 1, \dots, r$, consider the coordinate transformation $\varphi : (\theta_1, \dots, \theta_d) \mapsto (\theta'_1, \dots, \theta'_d)$ defined by*

$$a_k(\theta_{r+1}, \dots, \theta_d) := \frac{c_k}{i_{r+1}! \dots i_d!} \cdot \theta_{r+1}^{i_{r+1}} \dots \theta_d^{i_d}, \quad c_k \in \mathbb{R},$$

where

$$\begin{aligned} \theta'_1 &= \theta_1 + a_1(\theta_{r+1}, \dots, \theta_d), \quad \dots, \quad \theta'_r = \theta_r + a_r(\theta_{r+1}, \dots, \theta_d), \\ \theta'_{r+1} &= \theta_{r+1}, \quad \dots, \quad \theta'_d = \theta_d. \end{aligned}$$

This transformation satisfies the following conditions:

- (1) $\varphi(0) = 0$, $|\det \varphi'(0)| = 1$, and φ is bijective and analytic in a neighborhood of 0.
- (2) For $j = 1, \dots, r$, we have $\frac{\partial}{\partial \theta'_j} = \frac{\partial}{\partial \theta_j}$.
- (3) For any $(h_{r+1}, \dots, h_d) \in \mathbb{Z}_{\geq 0}^{d-r}$ with $h_{r+1} + \dots + h_d \leq i_{r+1} + \dots + i_d$, the following holds:

$$\begin{aligned} & \left. \frac{\partial^{h_{r+1} + \dots + h_d}}{\partial \theta_{r+1}^{h_{r+1}} \dots \partial \theta_d^{h_d}} \right|_{(\theta'_1, \dots, \theta'_d)=0} \\ &= \left. \frac{\partial^{h_{r+1} + \dots + h_d}}{\partial \theta_{r+1}^{h_{r+1}} \dots \partial \theta_d^{h_d}} \right|_{\theta=0} - \left\{ \sum_{k=1}^r c_k \frac{\partial}{\partial \theta_k} \right|_{\theta=0} \begin{array}{ll} \text{if } (h_{r+1}, \dots, h_d) \neq (i_{r+1}, \dots, i_d), \\ \text{if } (h_{r+1}, \dots, h_d) = (i_{r+1}, \dots, i_d). \end{array} \end{aligned}$$

Proof. (1) is straightforward. (2) follows from the relationship of partial derivatives before and after the transformation:

$$\begin{aligned}\frac{\partial}{\partial \theta'_j} &= \sum_{k=1}^d \frac{\partial}{\partial \theta_k} \cdot \frac{\partial \theta_k}{\partial \theta'_j} = \sum_{k=1}^r \frac{\partial}{\partial \theta_k} \cdot \frac{\partial \theta_k}{\partial \theta'_j} + \sum_{k=r+1}^d \frac{\partial}{\partial \theta_k} \cdot \delta_{k,j} \\ &= \frac{\partial}{\partial \theta_j} - \sum_{k=1}^r \frac{\partial a_k}{\partial \theta_j} \cdot \frac{\partial}{\partial \theta_k} \quad (j = r+1, \dots, d), \\ \frac{\partial}{\partial \theta'_j} &= \frac{\partial}{\partial \theta_j} \quad (j = 1, \dots, r).\end{aligned}$$

For (3), let $m := i_{r+1} + \dots + i_d \geq 1$ and $l := h_{r+1} + \dots + h_d \leq m$. We have:

$$\begin{aligned}\frac{\partial^l}{\partial \theta_{r+1}^{h_{r+1}} \dots \partial \theta_d^{h_d}} &= \prod_{j=r+1}^d \left(\frac{\partial}{\partial \theta_j} - \sum_{k=1}^r \frac{\partial a_k}{\partial \theta_j} \cdot \frac{\partial}{\partial \theta_k} \right)^{h_j} \\ &= \prod_{j=r+1}^d \left(\frac{\partial^{h_j}}{\partial \theta_j^{h_j}} - \sum_{k=1}^r \frac{\partial^{h_j} a_k}{\partial \theta_j^{h_j}} \cdot \frac{\partial}{\partial \theta_k} + \dots \right) \quad (\text{B.1})\end{aligned}$$

$$= \frac{\partial^l}{\partial \theta_{r+1}^{h_{r+1}} \dots \partial \theta_d^{h_d}} - \sum_{k=1}^r \frac{\partial^l a_k}{\partial \theta_{r+1}^{h_{r+1}} \dots \partial \theta_d^{h_d}} \cdot \frac{\partial}{\partial \theta_k} + \dots \quad (\text{B.2})$$

We focus on the terms indicated by “...” in equations (B.1) and (B.2), which involve derivatives of $a_k(\theta_{r+1}, \dots, \theta_d)$ of order less than or equal to $l-1$ with respect to $\theta_{r+1}, \dots, \theta_d$. Since $a_k(\theta_{r+1}, \dots, \theta_d)$ is a homogeneous polynomial of degree m in $\theta_{r+1}, \dots, \theta_d$, these terms evaluate to zero at $(\theta_1, \dots, \theta_d) = 0$. Therefore, we have:

$$\left. \frac{\partial^l a_k}{\partial \theta_{r+1}^{h_{r+1}} \dots \partial \theta_d^{h_d}} \right|_{\theta=0} = \begin{cases} 0 & \text{if } (h_{r+1}, \dots, h_d) \neq (i_{r+1}, \dots, i_d), \\ c_k & \text{if } (h_{r+1}, \dots, h_d) = (i_{r+1}, \dots, i_d). \end{cases}$$

Consequently,

$$\begin{aligned}&\left. \frac{\partial^l}{\partial \theta_{r+1}^{h_{r+1}} \dots \partial \theta_d^{h_d}} \right|_{(\theta'_1, \dots, \theta'_d)=0} \\ &= \left. \frac{\partial^l}{\partial \theta_{r+1}^{h_{r+1}} \dots \partial \theta_d^{h_d}} \right|_{\theta=0} - \begin{cases} 0 & \text{if } (h_{r+1}, \dots, h_d) \neq (i_{r+1}, \dots, i_d), \\ \sum_{k=1}^r c_k \frac{\partial}{\partial \theta_k} \Big|_{\theta=0} & \text{if } (h_{r+1}, \dots, h_d) = (i_{r+1}, \dots, i_d), \end{cases}\end{aligned}$$

as desired. $\square$

Appendix C

Proof of Theorem 4.1

We outline the proof as follows. First, we prove Theorem 4.2, which provides the Taylor expansion for the analytic function $H(\theta)$. This theorem asserts that the Taylor expansion of $H(\theta)$ takes the same form as in Theorem 3.2, which describes the Taylor expansion of $K(\theta)$. The proof is based on the argument in Appendix A. Here, we focus on the differences from Appendix A. Finally, using this result, we derive Theorem 4.1 by directly applying the proof technique in Section 3.3.

Let $i_1, \dots, i_k \in \{1, \dots, d\}$ and define

$$Z_{i_1, \dots, i_k} := \frac{\partial^k p(x|\theta)}{\partial \theta_{i_1} \dots \partial \theta_{i_k}}, \quad h(x|\theta) := \frac{1}{2} \{p(x|\theta) - q(x)\}^2,$$

where we aim to identify patterns in the partial derivatives of h . We refer to $Z_{i_1, \dots, i_k}$ as *indeterminate symbols of weight k* . Setting $\theta = (\theta_1, \dots, \theta_d)$, we obtain

$$\begin{aligned} \frac{\partial h(x|\theta)}{\partial \theta_{i_1}} &= \{p(x|\theta) - q(x)\} \cdot \frac{\partial p(x|\theta)}{\partial \theta_{i_1}} \\ \frac{\partial^2 h(x|\theta)}{\partial \theta_{i_1} \partial \theta_{i_2}} &= \{p(x|\theta) - q(x)\} \cdot \frac{\partial^2 p(x|\theta)}{\partial \theta_{i_1} \partial \theta_{i_2}} + Z_{i_1} Z_{i_2} \\ \frac{\partial^3 h(x|\theta)}{\partial \theta_{i_1} \partial \theta_{i_2} \partial \theta_{i_3}} &= \{p(x|\theta) - q(x)\} \cdot \frac{\partial^3 p(x|\theta)}{\partial \theta_{i_1} \partial \theta_{i_2} \partial \theta_{i_3}} \\ &\quad + Z_{i_1, i_2} Z_{i_3} + Z_{i_2, i_3} Z_{i_1} + Z_{i_3, i_1} Z_{i_2} \\ \frac{\partial^4 h(x|\theta)}{\partial \theta_{i_1} \partial \theta_{i_2} \partial \theta_{i_3} \partial \theta_{i_4}} &= \{p(x|\theta) - q(x)\} \cdot \frac{\partial^4 p(x|\theta)}{\partial \theta_{i_1} \partial \theta_{i_2} \partial \theta_{i_3} \partial \theta_{i_4}} \\ &\quad + Z_{i_1, i_2} Z_{i_3, i_4} + Z_{i_1, i_3} Z_{i_2, i_4} + Z_{i_1, i_4} Z_{i_2, i_3} \\ &\quad + Z_{i_1, i_2, i_3} Z_{i_4} + Z_{i_1, i_2, i_4} Z_{i_3} + Z_{i_1, i_3, i_4} Z_{i_2} + Z_{i_2, i_3, i_4} Z_{i_1}. \end{aligned}$$

In each of these, the first term on the right-hand side is zero at $\theta = 0$. Therefore, if we can interchange the order of partial differentiation and integration, the partial derivatives of $H(\theta)$ at $\theta = 0$ depend primarily on the terms from the second term onward.

In Appendix A, instead of using h , the partial derivatives of the log-likelihood ratio function f were examined. The second and subsequent terms on the right-hand sides of both expressions are similar. However, the partial derivatives of h consist of a *product of two* indeterminate symbols Z , whereas the partial derivatives of f consist of a *product of two or more* indeterminate symbols Y . Since terms consisting of a product of two indeterminate symbols play a crucial role in the following discussion, we can obtain the same conclusion for the partial derivatives of h .

Proposition C.1.

$$G_n(i_1, \dots, i_n) := \frac{\partial^n h(x|\theta)}{\partial \theta_{i_1} \dots \partial \theta_{i_n}} - \{p(x|\theta) - q(x)\} \cdot \frac{\partial^n p(x|\theta)}{\partial \theta_{i_1} \dots \partial \theta_{i_n}}$$

is a polynomial consisting of indeterminate symbols $Z_{i_1, \dots, i_k}$ with weights less than n . Each term in the polynomial has a coefficient of 1, and each term is a product of two indeterminate symbols, with a total weight of n . The polynomial is the sum of all such terms with these weights.

Proof. We prove this by induction on n . For $n = 1$, $G_1(i_1) = 0$ satisfies the condition since there are no terms. Assuming $G_n(i_1, \dots, i_n)$ is such a polynomial for some $n \geq 1$, we obtain

$$\begin{aligned} & \frac{\partial^{n+1} h(x|\theta)}{\partial \theta_{i_1} \dots \partial \theta_{i_{n+1}}} \\ &= \frac{\partial}{\partial \theta_{i_{n+1}}} \left\{ \{p(x|\theta) - q(x)\} \cdot \frac{\partial^n p(x|\theta)}{\partial \theta_{i_1} \dots \partial \theta_{i_n}} + G_n(i_1, \dots, i_n) \right\} \\ &= \{p(x|\theta) - q(x)\} \cdot \frac{\partial^{n+1} p(x|\theta)}{\partial \theta_{i_1} \dots \partial \theta_{i_{n+1}}} + Z_{i_1, \dots, i_n} Z_{i_{n+1}} + \frac{\partial G_n(i_1, \dots, i_n)}{\partial \theta_{i_{n+1}}} \end{aligned}$$

Thus,

$$G_{n+1}(i_1, \dots, i_{n+1}) = Z_{i_1, \dots, i_n} Z_{i_{n+1}} + \frac{\partial G_n(i_1, \dots, i_n)}{\partial \theta_{i_{n+1}}}.$$

By the induction hypothesis, this is a polynomial consisting of indeterminate symbols Z with weights less than or equal to n , where each term has a coefficient of 1 and the sum of weights in each term equals $n + 1$. Moreover, the terms containing the indeterminate symbol $Z_{i_{n+1}}$ are derived from the first and second terms on the right-hand side, while the terms that do not contain $Z_{i_{n+1}}$ are derived from the third term on the right-hand side. Thus, we have shown that each term is a product of two indeterminate symbols, summing to a total weight of $n + 1$. $\square$

Corollary C.1. (1) If all indeterminate symbols of weight $m - 1$ or less are zero almost surely at $\theta = 0$, then G_k (for $k \leq 2m - 1$) will also be zero almost surely at $\theta = 0$.

(2) If differentiation and integration can be interchanged,

$$\left. \frac{\partial^n H(\theta)}{\partial \theta_{i_1} \cdots \partial \theta_{i_n}} \right|_{\theta=0} = \int_{\chi} \left. \frac{\partial^n h(x|\theta)}{\partial \theta_{i_1} \cdots \partial \theta_{i_n}} \right|_{\theta=0} dx$$

equals the value obtained by substituting $\theta = 0$ and integrating over χ a polynomial. This polynomial is obtained by summing all terms where two indeterminate symbols $Z_{i_1, \dots, i_k}$ are multiplied together, the total weight is n , and the coefficients of all terms are 1.

Proof. (1) follows because terms with a sum of weights up to $2m - 1$ are products of two terms, so at least one term will have a weight of $m - 1$ or less and thus will be zero almost surely at $\theta = 0$. (2) follows directly from the previous discussions. $\square$

Proof of Theorem 4.2. We define

$$\begin{cases} P_1(x|s_1, \dots, s_r) := \sum_{j=1}^r \left. \frac{\partial p(x|\theta)}{\partial \theta_j} \right|_{\theta=0} s_j, \\ P_m(X|s_{r+1}, \dots, s_d) := \sum_{\substack{i_{r+1} + \dots + i_d = m \\ i_{r+1}, \dots, i_d \in \mathbb{Z}_{\geq 0}}} \frac{1}{i_{r+1}! \cdots i_d!} \left. \frac{\partial^m p(X|\theta)}{\partial \theta_{r+1}^{i_{r+1}} \cdots \partial \theta_d^{i_d}} \right|_{\theta=0} s_{r+1}^{i_{r+1}} \cdots s_d^{i_d}. \end{cases}$$

In the proof of Theorem 3.2, by replacing Y with Z and f with either h or p , we can show that the Taylor expansion of $H(\theta)$ at $(\theta_1, \dots, \theta_d) = 0$ is given by

$$H(\theta) = \frac{1}{2} \int_{\chi} \{P_1(x|\theta_1, \dots, \theta_r) + P_m(x|\theta_{r+1}, \dots, \theta_d)\}^2 dx + (\text{higher order terms}). \quad (\text{C.1})$$

As an example, we confirm that terms of first order in $(\theta_1, \dots, \theta_r)$ and of m -th order in $(\theta_{r+1}, \dots, \theta_d)$ are represented by

$$\int_{\chi} P_1(x|\theta_1, \dots, \theta_r) P_m(x|\theta_{r+1}, \dots, \theta_d) dx. \quad (\text{C.2})$$

Under Assumption 1(1), we deduce from Proposition A.1(1) that for all indices $(i_{r+1}, \dots, i_d)$ satisfying $1 \leq i_{r+1} + \dots + i_d \leq m - 1$,

$$\left. \frac{\partial^{i_{r+1} + \dots + i_d} p(X|\theta)}{\partial \theta_{r+1}^{i_{r+1}} \cdots \partial \theta_d^{i_d}} \right|_{\theta=0} = 0 \quad \text{a.s.}$$

holds. For

$$\left. \frac{\partial^{m+1} h(X|\theta)}{\partial \theta_j \partial \theta_{r+1}^{i_{r+1}} \cdots \partial \theta_d^{i_d}} \right|_{\theta=0} \quad (1 \leq j \leq r, i_{r+1} + \dots + i_d = m),$$

the nonzero products of two indeterminate symbols claimed in Corollary C.1(2) are limited to

$$\left. \frac{\partial p(X|\theta)}{\partial \theta_j} \right|_{\theta=0} \cdot \left. \frac{\partial^m p(X|\theta)}{\partial \theta_{r+1}^{i_{r+1}} \cdots \partial \theta_d^{i_d}} \right|_{\theta=0}.$$

Hence,

$$\int_{\chi} \left. \frac{\partial^{m+1} h(x|\theta)}{\partial \theta_j \partial \theta_{r+1}^{i_{r+1}} \cdots \partial \theta_d^{i_d}} \right|_{\theta=0} dx = \int_{\chi} \left. \frac{\partial p(x|\theta)}{\partial \theta_j} \right|_{\theta=0} \cdot \left. \frac{\partial^m p(x|\theta)}{\partial \theta_{r+1}^{i_{r+1}} \cdots \partial \theta_d^{i_d}} \right|_{\theta=0} dx$$

validates the claim in (C.2).

Finally, we verify that the equation to be proven follows from (C.1). Since

$$\left. \frac{\partial f}{\partial \theta_j} \right|_{\theta=0} = -\frac{1}{q(x)} \left. \frac{\partial p}{\partial \theta_j} \right|_{\theta=0},$$

we get $P_1(x|s_1, \dots, s_r) = -F_1(x|s_1, \dots, s_r)q(x)$. Under Assumption 1(1), we deduce from Proposition A.1(1) that

$$\left. \frac{\partial^m f(x|\theta)}{\partial \theta_{r+1}^{i_{r+1}} \cdots \partial \theta_d^{i_d}} \right|_{\theta=0} = -\frac{1}{q(x)} \left. \frac{\partial^m p(x|\theta)}{\partial \theta_{r+1}^{i_{r+1}} \cdots \partial \theta_d^{i_d}} \right|_{\theta=0}.$$

Thus, we obtain $P_m(x|s_{r+1}, \dots, s_d) = -F_m(x|s_{r+1}, \dots, s_d)q(x)$. Therefore,

$$\begin{aligned} & \int_{\chi} \{P_1(x|\theta_1, \dots, \theta_r) + P_m(x|\theta_{r+1}, \dots, \theta_d)\}^2 dx \\ &= \mathbb{E}_X \left[\{F_1(X|\theta_1, \dots, \theta_r) + F_m(X|\theta_{r+1}, \dots, \theta_d)\}^2 q(X) \right] \end{aligned}$$

is established. $\square$

Proof of Theorem 4.1. Since most of the argument parallels that in Section 3.3, we outline only the main steps here. By Theorem 4.2, the Taylor expansion of $H(\theta)$ at $(\theta_1, \dots, \theta_d) = 0$ can be expressed as

$$\begin{aligned} H(\theta) &= \frac{1}{2} \mathbb{E}_X \left[\{F_1(X|\theta_1, \dots, \theta_r) + F_m(X|\theta_{r+1}, \dots, \theta_d)\}^2 q(X) \right] \\ &\quad + (\text{higher order terms}). \end{aligned}$$

We perform the following four coordinate transformations, CT1-CT4 (see Figure 3.2). In the case $r = 0$, we use only CT2. We demonstrate that normal crossings are achieved in all the resulting local coordinates.

CT1 For $1 \leq s \leq r$, $\{\theta_j \rightarrow \theta_s \theta'_j \mid 1 \leq j \leq d, j \neq s\}$

CT2 For $r+1 \leq t \leq d$, $\{\theta_j \rightarrow \theta_t \theta'_j \mid 1 \leq j \leq d, j \neq t\}$

CT3 For the t selected in CT2 and for $1 \leq s \leq r$, $\{\theta'_j \rightarrow \theta_s \theta''_j, \theta_t \rightarrow \theta_s \theta'_t \mid 1 \leq j \leq r, j \neq s\}$

CT4 For the t selected in CT2, $\{\theta'_j \rightarrow \theta_t \theta''_j \mid 1 \leq j \leq r\}$

Initially, we perform CT1 and CT2. The first transformation achieves normal crossings, but the second does not. Therefore, we proceed to perform further iterations involving CT3 and CT4. In these iterations, CT3 achieves normal crossings, but CT4 does not. By repeating this process up to $m - 1$ times (resulting in a total of m iterations), we eventually achieve normal crossings.

Here, we verify only the following three inequalities, which are essential for achieving normal crossings (each parameter s_i ranges over all real numbers):

$$\begin{aligned} a_1(s_2, \dots, s_d) &:= \mathbb{E}_X [F_1^2(X|1, s_2, \dots, s_r)q(X)] > 0 \\ a_2(s_2, \dots, s_d) &:= \mathbb{E}_X \left[\left\{ F_1(X|1, s_2, \dots, s_r) + s_d F_m(X|s_{r+1}, \dots, s_{d-1}, 1) \right\}^2 q(X) \right] > 0 \\ a_3(s_1, \dots, s_{d-1}) &:= \mathbb{E}_X \left[\left\{ F_1(X|s_1, \dots, s_r) + F_m(X|s_{r+1}, \dots, s_{d-1}, 1) \right\}^2 q(X) \right] > 0 \end{aligned}$$

For the first inequality, since $q(X) \neq 0$ a.s., $a_1(s_2, \dots, s_d) = 0$ holds if and only if

$$\begin{aligned} F_1(X|1, s_2, \dots, s_r) &= 0 \quad \text{a.s.} \\ \Leftrightarrow \frac{\partial f(X|\theta)}{\partial \theta_1} \Big|_{(\theta_1, \dots, \theta_d)=0} + \sum_{k=2}^r s_k \cdot \frac{\partial f(X|\theta)}{\partial \theta_k} \Big|_{(\theta_1, \dots, \theta_d)=0} &= 0 \quad \text{a.s..} \end{aligned}$$

By Assumption 1, the set of partial derivatives

$$\frac{\partial f(X|\theta)}{\partial \theta_1} \Big|_{(\theta_1, \dots, \theta_d)=0}, \dots, \frac{\partial f(X|\theta)}{\partial \theta_r} \Big|_{(\theta_1, \dots, \theta_d)=0} \quad (\text{C.3})$$

are linearly independent, so there do not exist $(s_2, \dots, s_d)$ satisfying the above equivalence and we conclude that $a_1(s_2, \dots, s_d) > 0$ holds.

The second inequality is analogous to the third, so we verify only the third. The equation $a_3(s_1, \dots, s_{d-1}) = 0$ implies

$$\begin{aligned} F_1(X|s_1, \dots, s_r) + F_m(X|s_{r+1}, \dots, s_{d-1}, 1) &= 0 \quad \text{a.s.} \\ \Leftrightarrow \sum_{k=1}^r s_k \cdot \frac{\partial f(X|\theta)}{\partial \theta_k} \Big|_{(\theta_1, \dots, \theta_d)=0} + F_m(X|s_{r+1}, \dots, s_{d-1}, 1) &= 0 \quad \text{a.s..} \end{aligned}$$

For a nonzero parameter $(s_{r+1}, \dots, s_{d-1}, 1) \neq 0$, Assumption 1(2) implies that the random variable $F_m(X|s_{r+1}, \dots, s_{d-1}, 1)$ and the r random variables in (C.3) are linearly independent. Therefore, there does not exist $(s_1, \dots, s_{d-1})$ that satisfies this equivalence relation. Thus, we conclude that $a_3(s_1, \dots, s_{d-1}) > 0$ holds. $\square$

Appendix D

Details on Example 5.1(1)

Let H be the number of mixture components with $H \geq 2$, and $(a_1, \dots, a_{H-1})$ be mixture weights. Consider the following mixture model with $2H - 1$ parameters $(a_1, \dots, a_{H-1}, b_1, \dots, b_H)$:

$$p(x|a, b) := \sum_{i=1}^{H-1} a_i \tilde{p}(x|b_i) + \left(1 - \sum_{i=1}^{H-1} a_i\right) \tilde{p}(x|b_H).$$

The true distribution $q(x)$ can be expressed using $H_0 - 1$ constants $\{a_{i*}\}_{i=1}^{H_0-1}$ satisfying $a_{i*} > 0$ and $\sum_{i=1}^{H_0-1} a_{i*} < 1$, along with H_0 distinct constants $b_{1*}, \dots, b_{H_0*}$, as follows:

$$q(x) = \begin{cases} \sum_{i=1}^{H_0-1} a_{i*} \cdot \tilde{p}(x|b_{i*}) + \left(1 - \sum_{i=1}^{H_0-1} a_{i*}\right) \cdot \tilde{p}(x|b_{H_0*}) & \text{if } H_0 > 1, \\ \tilde{p}(x|b_{1*}) & \text{if } H_0 = 1. \end{cases}$$

It is clear that points satisfying the following conditions are realization parameters:

$$\begin{cases} (a_1, \dots, a_{H_0-1}, a_{H_0}, \dots, a_{H-1}) &= (a_{1*}, \dots, a_{H_0-1*}, \forall a_{H_0}, \dots, \forall a_{H-1}) \\ (b_1, \dots, b_{H_0-1}, b_{H_0}, \dots, b_{H-1}, b_H) &= (b_{1*}, \dots, b_{H_0-1*}, b_{H_0*}, \dots, b_{H_0*}, b_{H_0*}). \end{cases}$$

We now consider the case where the following condition holds:

$$a_{H_0}, \dots, a_{H-1} > 0, \quad \sum_{k=H_0}^{H-1} a_k < 1 - \sum_{k=1}^{H_0-1} a_{k*}. \quad (\text{D.1})$$

Since the set of realization parameters is not a single point, this statistical model does not satisfy semiregularity. For any realization parameter P satisfying (D.1), we show that $p(x|a, b)$ satisfies (5.1) and Assumption 3 with $(d_1, r, m) = (H + H_0 - 1, 2H_0 - 1, 2)$.

By translating the coordinates, we can express

$$p(x|a, b) = \sum_{i=1}^{H_0-1} (a_i + a_{i*}) \tilde{p}(x|b_i + b_{i*}) + \sum_{i=H_0}^{H-1} a_i \tilde{p}(x|b_i + b_{H_0*}) \\ + \left(1 - \sum_{i=1}^{H_0-1} a_{i*} - \sum_{i=1}^{H-1} a_i \right) \tilde{p}(x|b_H + b_{H_0*}).$$

$$P \in \left\{ (a_1, \dots, a_{H_0-1}, b_1, \dots, b_H, a_{H_0}, \dots, a_{H-1}) \left| \begin{array}{l} a_1 = \dots = a_{H_0-1} = 0 \\ b_1 = \dots = b_H = 0 \end{array} \right. \right\}$$

It is clear that (5.1) holds. In the following, we arbitrarily fix the $H - H_0$ parameters $(a_{H_0}, \dots, a_{H-1})$ within the range satisfying (D.1). For the remaining $H + H_0 - 1$ parameters $a_1, \dots, a_{H_0-1}, b_1, \dots, b_H$, we perform the coordinate transformation

$$b'_H := b_H + \frac{\sum_{k=H_0}^{H-1} a_k b_k}{1 - \sum_{k=1}^{H_0-1} a_{k*} - \sum_{k=H_0}^{H-1} a_k}.$$

After performing this transformation, we examine the partial derivatives of $f(x|a, b)$ at the point P . Then,

$$\begin{aligned} \frac{\partial f}{\partial a_i} &= -\frac{1}{q(x)} \{ \tilde{p}(x|b_{i*}) - \tilde{p}(x|b_{H_0*}) \} & i = 1, \dots, H_0 - 1 \\ \frac{\partial f}{\partial b_i} &= -\frac{1}{q(x)} \begin{cases} a_{i*} \frac{\partial \tilde{p}}{\partial b}(x|b_{i*}) & i = 1, \dots, H_0 - 1 \\ 0 & i = H_0, \dots, H - 1 \end{cases} \\ \frac{\partial f}{\partial b'_H} &= -\frac{1}{q(x)} \left(1 - \sum_{k=1}^{H_0-1} a_{k*} - \sum_{k=H_0}^{H-1} a_k \right) \frac{\partial \tilde{p}}{\partial b}(x|b_{H_0*}). \end{aligned}$$

The second derivative with respect to $\mathbf{b} := (b_{H_0} \cdots b_{H-1})^\top \in \mathbb{R}^{H-H_0}$ is given by

$$F_2(x|\mathbf{b}) = \sum_{\substack{i_{H_0} + \dots + i_{H-1} = 2 \\ \{i_{H_0}, \dots, i_{H-1}\} \subset \mathbb{Z}_{\geq 0}}} \frac{b_{H_0}^{i_{H_0}} \cdots b_{H-1}^{i_{H-1}}}{i_{H_0}! \cdots i_{H-1}!} \frac{\partial^2 f}{\partial b_{H_0}^{i_{H_0}} \cdots \partial b_{H-1}^{i_{H-1}}} \Big|_{(a_1, \dots, a_{H_0-1}, b_1, \dots, b_H) = 0} \\ = \frac{1}{2(1 - \sum_{k=1}^{H_0-1} a_{k*} - \sum_{k=H_0}^{H-1} a_k)} \cdot \frac{-1}{q(x)} \frac{\partial^2 \tilde{p}}{\partial b^2}(x|b_{H_0*}) \cdot \mathbf{b}^\top \Sigma \mathbf{b}$$

where $\Sigma := (\sigma_{i,j})_{1 \leq i, j \leq H-H_0}$ is defined by

$$\sigma_{i,j} := \begin{cases} a_{i+H_0-1} (1 - \sum_{k=1}^{H_0-1} a_{k*} - \sum_{k=H_0}^{H-1} a_k + a_{i+H_0-1}) & i = j, \\ a_{i+H_0-1} \cdot a_{j+H_0-1} & i \neq j. \end{cases}$$

Under condition (D.1), it follows from Lemma D.2, which will be stated below, that Σ is a positive definite matrix, so for any $\mathbf{b} \neq 0$, we have $F_2(x|\mathbf{b}) \neq 0$ a.s..

From Lemma D.1 stated below,

$$\tilde{p}(x|b_{1*}), \dots, \tilde{p}(x|b_{H_0*}), \frac{\partial \tilde{p}}{\partial b}(x|b_{1*}), \dots, \frac{\partial \tilde{p}}{\partial b}(x|b_{H_0*}), \frac{\partial^2 \tilde{p}}{\partial b^2}(x|b_{H_0*}) \quad (\text{D.2})$$

are linearly independent, we see that for any $\mathbf{b}(\neq 0)$,

$$\frac{\partial f}{\partial a_1}, \dots, \frac{\partial f}{\partial a_{H_0-1}}, \frac{\partial f}{\partial b_1}, \dots, \frac{\partial f}{\partial b_{H_0-1}}, \frac{\partial f}{\partial b'_H}, F_2(x|\mathbf{b})$$

are linearly independent. Therefore, at point P , the model satisfies Assumption 3(1)(2) for $(d_1, r, m) = (H + H_0 - 1, 2H_0 - 1, 2)$, where we set $\theta := (a_1, \dots, a_{H_0-1}, b_1, \dots, b_{H-1}, b'_H)$ and $\tau := (a_{H_0}, \dots, a_{H-1})$.

For simplicity in the following lemma, we replace the notation H_0 with N and b_{i*} with b_i .

Lemma D.1. *Let $\tilde{p}(x|b)$ be one of the following distributions: the binomial distribution with $M(\geq 2)$ trials and success probability b , the Poisson distribution with mean b , or the normal distribution with mean b and variance 1. Let $N(\leq M/2)$ be an integer greater than or equal to 1. Then for N distinct parameters $\{b_1, \dots, b_N\}$, the following random variables are linearly independent:*

$$\tilde{p}(x|b_1), \dots, \tilde{p}(x|b_N), \frac{\partial \tilde{p}}{\partial b}(x|b_1), \dots, \frac{\partial \tilde{p}}{\partial b}(x|b_N), \frac{\partial^2 \tilde{p}}{\partial b^2}(x|b_N)$$

Proof. Let $\{s_i\}, \{t_i\}, u$ be real numbers, and suppose that

$$\sum_{i=1}^N s_i \tilde{p}(x|b_i) + \sum_{i=1}^N t_i \frac{\partial \tilde{p}}{\partial b}(x|b_i) + u \frac{\partial^2 \tilde{p}}{\partial b^2}(x|b_N) = 0 \quad \text{a.s..} \quad (\text{D.3})$$

We show $s_i = t_i = u = 0$ for all $i = 1, \dots, N$. Note that the n -th moment ($n \leq M$) of a random variable following the given probability distribution $\tilde{p}$ is a polynomial of degree n in b . First, summing (D.3) over $x \in \chi$ implies

$$\sum_{i=1}^N s_i = 0. \quad (\text{D.4})$$

Next, multiplying (D.3) by x and again summing over $x \in \chi$, then using (D.4), we obtain

$$\sum_{i=1}^N s_i b_i + \sum_{i=1}^N t_i = 0. \quad (\text{D.5})$$

Repeating this process for $n = 2, \dots, 2N (\leq M)$, multiplying (D.3) by x^n , summing over $x \in \chi$, and using (D.4)–(D.5), we get

$$\sum_{i=1}^N s_i b_i^n + {}_n P_1 \sum_{i=1}^N t_i b_i^{n-1} + {}_n P_2 u b_N^{n-2} = 0,$$

where for nonnegative integers $m \geq n$, we define ${}_m P_n := m!/(m-n)!$. Therefore, we have $W(s_1 \cdots s_N t_1 \cdots t_N u)^\top = \mathbf{0} \in \mathbb{R}^{2N+1}$,

$$W := \left(\begin{array}{ccc|ccc|c} 1 & \cdots & 1 & 0 & \cdots & 0 & 0 \\ b_1 & \cdots & b_N & 1 & \cdots & 1 & 0 \\ b_1^2 & \cdots & b_N^2 & {}_2 P_1 b_1 & \cdots & {}_2 P_1 b_N & {}_2 P_2 \\ b_1^3 & \cdots & b_N^3 & {}_3 P_1 b_1^2 & \cdots & {}_3 P_1 b_N^2 & {}_3 P_2 b_N \\ b_1^4 & \cdots & b_N^4 & {}_4 P_1 b_1^3 & \cdots & {}_4 P_1 b_N^3 & {}_4 P_2 b_N^2 \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ b_1^{2N} & \cdots & b_N^{2N} & {}_{2N} P_1 b_1^{2N-1} & \cdots & {}_{2N} P_1 b_N^{2N-1} & {}_{2N} P_2 b_N^{2N-2} \end{array} \right).$$

Thus, we show that $\det W \neq 0$. A straightforward computation gives

$$\det W \propto \prod_{1 \leq i < j \leq N-1} (b_i - b_j)^4 \times \prod_{1 \leq i \leq N-1} (b_i - b_N)^6.$$

Because $\{b_1, \dots, b_N\}$ are distinct, we conclude that $\det W \neq 0$. $\square$

Lemma D.2. *Let N be a positive integer, and let $A, T_1, \dots, T_N$ be variables. Define the symmetric matrix $\Sigma_N := (\sigma_{i,j})_{1 \leq i,j \leq N} \in \mathbb{R}^{N \times N}$ by*

$$\sigma_{i,j} := \begin{cases} T_i(A + T_i) & \text{if } i = j, \\ T_i \cdot T_j & \text{if } i \neq j \end{cases}$$

Then, for any $(A, T_1, \dots, T_N)$ satisfying $AT_i > 0$ for all $i = 1, \dots, N$, the matrix Σ_N is positive definite.

Proof.

$$(s_1, \dots, s_N) \Sigma_N (s_1, \dots, s_N)^\top = \left(\sum_{i=1}^N s_i T_i \right)^2 + \sum_{i=1}^N AT_i s_i^2 \geq 0$$

$\square$

List of Publications

- [1] Kurumadani, Y. (2025). Learning coefficients in semiregular models I: properties. *Japanese Journal of Statistics and Data Science*, 8, 1051-1079.
- [2] Kurumadani, Y. (2025). Learning coefficients in semiregular models II: extensions. *Japanese Journal of Statistics and Data Science*.
- [3] Kurumadani, Y., & Iwasawa, H. (2025). Improving Withdrawal Rate Estimation Using Generalized Additive Models. *Japanese Journal of Risk and Insurance* (in Japanese).

Acknowledgements

First, I would like to express my deepest gratitude to Prof. Joe Suzuki, my thesis supervisor, for their unwavering support and guidance throughout this journey. Your expertise and insights have been invaluable in shaping both my work and my professional growth.

Additionally, I am deeply grateful to my family for their constant love and encouragement, and for believing in my abilities. In particular, I would like to express my heartfelt gratitude to my wife, who has taken care of our young child and supported our family while I was preoccupied with my research and unable to spend sufficient time with them. The bond we share as a family and the emotional support in my academic pursuits are of immeasurable value to me. Without you, I would not be where I am today.

References

- [1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.
- [2] Aoyagi, M. (2006). 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.
- [3] Aoyagi, M. (2009). Log canonical threshold of vandermonde matrix type singularities and generalization error of a three-layered neural network in bayesian estimation. *International Journal of Pure and Applied Mathematics*, 52, 177–204.
- [4] Aoyagi, M. (2010). A Bayesian Learning Coefficient of Generalization Error and Vandermonde Matrix-Type Singularities. *Communications in Statistics – Theory and Methods*, 39(15), 2667–2687.
- [5] Aoyagi, M. (2013a). Consideration on singularities in learning theory and the learning coefficient. *Entropy*, 15(9), 3714–3733.
- [6] Aoyagi, M. (2013). Learning coefficient in Bayesian estimation of restricted Boltzmann machine. *Journal of Algebraic Statistics*, 4(1), 30–57.
- [7] Aoyagi, M. (2019a). Learning coefficient of vandermonde matrix-type singularities in model selection. *Entropy (Information Theory, Probability and Statistics)*, 21(6–561), 1–12.
- [8] Aoyagi, M. (2019b). Learning coefficients and information criteria. *Frontiers in Artificial Intelligence and Applications*, 351–362.
- [9] Aoyagi, M. (2024). Consideration on the learning efficiency of multiple-layered neural networks with linear units. *Neural Networks*, 172:106132.
- [10] Aoyagi, M. (2025). Singular learning coefficients and efficiency in learning theory. *arXiv:2501.12747*.
- [11] Aoyagi, M., & Watanabe, S. (2005). Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Networks*, 18(7):924–933.

- [12] Drton, M., Plummer, M. (2017). A Bayesian information criterion for singular models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(2), 323–380.
- [13] Drton, M., Lin, S., Weihs, L., & Zwiernik, P. (2017). Marginal likelihood and model selection for Gaussian latent tree and forest models. *Bernoulli*, 23(2), 1202–1232.
- [14] Hironaka, H. (1964). Resolution of Singularities of an Algebraic Variety Over a Field of Characteristic Zero. *Annals of Mathematics*, 79:109–326.
- [15] Imai, T. (2019). Estimating real log canonical thresholds. [arXiv:1906.01341](https://arxiv.org/abs/1906.01341).
- [16] Kashiwara, M. (1976). B-functions and holonomic systems. *Inventions Mathematics*, 38, 33–53.
- [17] Lau, E., Furman, Z., Wang, G., Murfet, D., & Wei, S. (2025). The local learning coefficient: A singularity-aware complexity measure. *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS 2025)*, *Proceedings of Machine Learning Research*, 258, 244–252.
- [18] Mustata, M. (2002). Singularities of pairs via jet schemes. *Journal of the American Mathematical Society*, 15, 599–615.
- [19] Rusakov, D., & Geiger, D. (2002). Asymptotic model selection for naive Bayesian networks. *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, 438–445.
- [20] Rusakov, D., & Geiger, D. (2005). Asymptotic model selection for naive Bayesian networks. *Journal of Machine Learning Research*, 6, 1–35.
- [21] Sato, K., & Watanabe, S. (2019). Bayesian Generalization Error of Poisson Mixture and Simplex Vandermonde Matrix Type Singularity. [arXiv:1912.13289](https://arxiv.org/abs/1912.13289).
- [22] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
- [23] Takeuchi, K. (1976). Distribution of an information statistic and the criterion for the optimal model. *Mathematical Science*, 153, 12–18.
- [24] Watanabe, S. (2001a). Algebraic analysis for nonidentifiable learning machines. *Neural Computation*, 13(4), 899–933.
- [25] Watanabe, S. (2001b). Algebraic geometrical methods for hierarchical learning machines. *Neural Networks*, 14(8), 1049–1060.

- [26] Watanabe, S. (2001c). Algebraic geometry of learning machines with singularities and their prior distributions. *Journal of Japanese Society of Artificial Intelligence*, 16(2), 308–315.
- [27] Watanabe, S., Yamazaki, K., & Aoyagi, M. (2004). Kullback information of normal mixture is not an analytic function. Technical report of IEICE (in Japanese).
- [28] Watanabe, S. (2009). *Algebraic Geometry and Statistical Learning Theory*, volume 25. Cambridge University Press, New York, USA.
- [29] Watanabe, S. (2010). Equations of states in singular statistical estimation. *Neural Networks*, 23(1), 20–34.
- [30] Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, (14), 867–897.
- [31] Watanabe, S. (2018). *Mathematical theory of Bayesian statistics*. CRC Press.
- [32] Watanabe, T., & Watanabe, S. (2022). Asymptotic Behavior of Bayesian Generalization Error in Multinomial Mixtures. [arXiv:2203.06884](https://arxiv.org/abs/2203.06884).
- [33] Zwiernik, P. (2011). An asymptotic behavior of the marginal likelihood for general Markov models. *Journal of Machine Learning Research*, 12, 3283–3310.