arXiv:2303.09154v1 [stat.ML] 16 Mar 2023

具有概念瓶颈结构和多任务形式的线性神经网络中的贝叶斯泛化误差

Naoki Hayashi*[1] 和 Yoshihide Sawada†[1]

[1]爱信株式会社东京研究中心

2023/03/16

摘要

概念瓶颈模型（CBM）是一种广泛使用的方法，它利用概念来解释神经网络。在 CBM 中，概念作为可观测值插入输出层与最后一个中间层之间。这有助于理解神经网络生成输出的原因：即从最后一个隐藏层到输出层、与这些概念相对应的权重。然而，由于神经网络通常属于奇异统计模型，目前尚无法理解 CBM 中泛化误差的行为。当模型是奇异模型时，无法建立从参数到概率分布的一一映射。这种不可辨识性使泛化性能难以分析。在本研究中，我们从数学上阐明了当 CBM 的架构为三层线性神经网络时，其贝叶斯泛化误差和自由能。我们还考虑了一种多任务问题，其中神经网络不仅输出原始输出，还输出概念。结果表明，与标准版本相比，CBM 会显著改变三层线性神经网络中参数区域和贝叶斯泛化误差的行为，而多任务形式则不会。

1 引言

自多层感知机问世以来，人工神经网络不断发展，并已广泛应用于许多领域 [17, 12]。然而，由于大多数神经网络都是黑箱，因此有必要解释神经网络的输出。为此，人们提出了各种方法来提高输出可解释性 [40]。用于解释神经网络行为的一种网络架构是概念瓶颈模型（CBM） [31, 32, 28]。CBM采用一种新颖的结构，称为概念瓶颈结构：将概念作为观测值插入输出层与最后一个中间层之间，并且从概念到输出的最后一层连接是线性的。因此，人们有望能够解读最后一层连接的权重，将其视为指定概念对输出的影响，这与线性回归的系数类似。例如，按照 [28], ，当我们使用X射线图像和CBM预测患者的膝关节炎等级时，会将医生校正过的临床发现设为概念，从而通过观察最后一层连接中学到的权重，了解临床发现如何基于相关性影响预测等级。基于概念的解释应用于国际象棋的知识发现 [39], 、视频表征 [44], 、医学影像 [25], 、临床风险预测 [45],、计算机辅助诊断 [27], 以及其他医疗保健领域问题 [10]。CBM是这些应用的重要基础，并且已有基于CBM提出的先进方法 [49, 44, 27] 。因此，阐明CBM的理论行为十分重要。

*naoki.hayashi05@aisin.co.jp
†yoshihide.sawada@aisin.co.jp

为了阐明与CBM相比的性能差异，也需要采用多任务形式（多任务） [61] ，因为前者可以输出一个拼接了原始输出和概念的向量，而不是将概念插入中间层；CBM和多任务使用相似类型的数据（输入、概念和输出）。两者的解释方式也相似：CBM根据概念与输出之间的回归得出解释，而多任务则根据它们的共现关系得出解释。

尽管已有研究探讨了 CBM 的一些局限性，但除了一项在 [28] 中使用最小二乘法对三层线性独立 CBM 进行的简单分析外，CBM 的泛化误差尚未得到阐明。多任务模型的泛化误差也仍然未知。这是因为神经网络通常不可识别，即无法从参数映射到表示模型的概率分布。此类模型称为奇异统计模型[55, 58], ，因为任何正态分布都无法逼近其似然和后验分布，且其费舍尔信息矩阵具有非正特征值。对于奇异统计模型，已有证明表明，就泛化性能而言，贝叶斯推断优于最大似然或后验估计 [55, 58]。因此，以下主要讨论贝叶斯推断。

正则统计模型定义为其参数到概率密度函数的映射是单射的。此时称模型是正则的，否则称模型是奇异的。令d 为参数维度， n 为样本量。在正则统计模型中，其期望泛化误差渐近等于 $d/2n + o(1/n)$，其中，泛化误差是数据生成分布与预测分布之间的库尔贝克-莱布勒散度 [1, 3, 2]。此外，其负对数边际似然（亦称自由能）的渐近展开式表示为 $nS_n+(d/2)\log n+O_p(1)$，其中 S_n 为经验熵 [50]。在一般情形下，也就是模型可能为奇异模型的情形，Watanabe 已证明其泛化误差 G_n 和自由能 F_n 的渐近形式如下[53, 54, 55]:

$$\mathbb{E}_n[G_n] = \frac{\lambda}{n} - \frac{m-1}{n\log n} + o\left(\frac{1}{n\log n}\right), \tag{1}$$

$$F_n = nS_n + \lambda\log n - (m-1)\log\log n + O_p(1), \tag{2}$$

其中， λ 是正有理数， m 是正整数，而 $\mathbb{E}_n[\cdot]$ 分别表示对整个数据集取期望的期望算子。常数 λ 称为学习系数，因为它是式（1）和（2）中首项的主导项；这两个式子分别表示 $\mathbb{E}_n[G_n]$-n 学习曲线和 F_n-n 学习曲线。上述形式不仅适用于模型为正则的情形，也适用于模型为奇异的情形。泛化损失定义为数据生成分布与统计模型之间的交叉熵，且等于 $S+G_n$，其中 S 是数据生成分布的熵。Watanabe 开创了这一理论，并提出了两种模型评估方法：WAIC [56] 和 WBIC [57], ，它们分别能够根据模型和数据估计正则模型与奇异模型的 $S+G_n$ 和 F_n。

令 $K:\mathrm{W} \to \mathbb{R}$, $w \mapsto K(w)$表示数据生成分布与统计模型之间的 Kullback–Leibler（KL）散度，其中 $\mathrm{W} \subset \mathbb{R}^d$ 是参数集，且 $w \in \mathrm{W}$。假设 W 是一个足够大的紧集， $K(w)$是解析的，且其取值非负。常数 λ 和 m 由 KL 散度零点集中的奇异性刻画： $K^{-1}(0)$，这是一个解析集（亦称代数簇）。它们可通过奇点消解 [24, 7]计算； λ 称为实对数典范阈值（RLCT）， m 在代数几何中称为重数。请注意，它们是双有理不变量； λ 和 m 不受奇点消解方式的影响。这里，假设先验在 $K^{-1}(0)$上为正且有界。在正则情形下，我们可以推导出 $\lambda = d/2$ 和 $m = 1$。此外，预测分布既可以是贝叶斯后验预测分布，也可以是参数采用最大似然估计量或后验估计量的模型。然而，在奇异情形下，RLCT λ 和重数m 取决于模型，贝叶斯推断与最大似然估计或后验估计有显著差异。CBM 和多任务中都可能出现这些情况。

确定实对数典范阈值对于估计充分样本量、构建学习流程和选择模型都很重要。事实上，许多奇异模型的实对数典范阈值已得到研究：混合模型[63, 65, 47, 37, 60], 玻尔兹曼机 [67, 4, 5], 非负矩阵分解 [22, 21, 18], 潜在类别分析 [13], 潜在狄利克雷分配 [23, 19], 朴素贝叶斯网络 [46], 贝叶斯网络 [64],

马尔可夫模型 [69], 隐马尔可夫模型 [66], 线性动力系统 [42], 高斯潜在树模型和森林模型 [14], 以及激活函数为线性函数 [6], 解析奇函数（如双曲正切函数） [54], 和 Swish [52] 的三层神经网络。此外，Drton 和 Plummer 提出了一种名为 sBIC 的模型选择方法，该方法使用统计模型的实对数典范阈值 [15]。此外，Drton 和 Imai 通过实证证明，当实对数典范阈值得到精确明确的确定或其紧界可用时，在选择正确模型方面，sBIC 比 WBIC 更准确 [15, 14, 26]。另外，Imai 提出了一种根据数据和模型估计实对数典范阈值的方法，并扩展了 sBIC [26]。实对数典范阈值的另一项应用，是 Nagata 提出的交换蒙特卡洛方法中交换概率的设计流程 [41]。

没有概念瓶颈结构的神经网络，其 RLCT 被称为 [28], 中的标准模型；对于三层线性神经网络情形，该值已在 [6] 中得到确切阐明。然而，CBM 和多任务模型的 RLCT 仍然未知。换言之，即使模型结构为三层线性，CBM 和多任务模型的贝叶斯泛化误差与边际似然也尚未得到阐明。此外，由于我们研究的是可解释机器学习领域中的模型，我们的结果表明，奇异学习理论有助于为负责任的人工智能／计算智能奠定基础。这是因为，解释方法往往可以被视为对模型参数空间的限制。参数约束可能会改变模型属于正则还是奇异情形 [16]。有一些研究视角侧重于参数限制，以分析奇异统计模型 [16, 20]。例如，在非负矩阵分解（NMF） [43, 33, 9]，中，参数是因子分解矩阵的元素，并被限制在非负区域，以提高分解结果的可解释性，例如从物品—用户表中获得的客户购买因素，以及他们对各项产品的兴趣程度 [29, 30]。如果这些元素可以为负，由于正负元素会相互抵消，就难以估计产品受欢迎程度和客户的潜在需求。NMF 的贝叶斯泛化误差和自由能与不受限制的矩阵分解表现不同 [22, 21, 18, 6]。因此，出于解释方法的考虑，对参数空间的限制会实质性地影响模型的学习行为，而通过研究奇异学习理论可以对这些行为进行分析。

本研究通过求出采用 CBM 的神经网络在结构为三层线性时的实对数典范阈值（RLCT），从数学上推导出贝叶斯泛化误差和自由能的精确渐近形式。我们还阐明了该情形下多任务模型的实对数典范阈值，并比较了 CBM 与多任务模型及标准模型的理论行为。本文其余部分分为四部分。第 2 节介绍贝叶斯推断的框架，以及在数据生成分布未知时如何验证模型。第 3 节陈述主要定理。第 4 节将主要定理扩展到分类变量的情形。第 5 节讨论这一理论结果。第 6 节总结全文。此外，本文还包含两个附录。附录 A 介绍数据生成分布未知时贝叶斯推断的数学理论。这一理论是本研究的基础，称为奇异学习理论。附录 B 证明主要定理及其扩展结果，并给出一个用于比较 CBM 与多任务模型实对数典范阈值的命题。

2 贝叶斯推断框架

设 $X^n = (X_1, \ldots\ldots, X_n)$ 是一组从数据生成分布中独立同分布抽取的随机变量。X_i 的函数值属于 X，其中 X 是有限维实欧几里得空间或离散空间的子集。本文将这组变量 X^n 称为数据集或样本，其元素 X_i 称为第 i 个数据。此外，设 $q : \mathrm{X} \to \mathbb{R}$, $x \mapsto q(x)$、$p(\cdot|w) : \mathrm{X} \to \mathbb{R}$, $x \mapsto p(x|w)$ 和 $\varphi : \mathrm{W} \to \mathbb{R}$, $w \mapsto \varphi(w)$ 分别是数据生成分布、统计模型和先验分布的概率密度。请注意，参数 w 及其集合 W 的定义与上一节相同。

我们定义一个 后验分布是指密度由以下函数给出的分布，定义在 $\mathcal{W}$:

$$\varphi^*(w|X^n) = \frac{1}{Z_n}\varphi(w)\prod_{i=1}^{n} p(X_i|w), \tag{3}$$

其中， Z_n 是用于满足条件 $\int \varphi^*(w|X^n)dw = 1$ 的归一化常数：

$$Z_n = \int dw\varphi(w) \prod_{i=1}^{n} p(X_i|w). \tag{4}$$

这称为边际似然或配分函数。其负对数值称为自由能$F_n = -\log Z_n$。注意，边际似然是数据集的概率密度函数。就数据集生成过程而言，自由能出现在数据生成分布与模型之差的首项中。此外，预测分布由 X 上的以下密度函数定义：

$$p^*(x|X^n) = \int dw\varphi^*(w|X^n)p(x|w). \tag{5}$$

这是新数据的概率分布。对于统计学和机器学习而言，从新数据生成过程的角度评估真实分布与模型之间的差异也很重要。

在此，我们说明贝叶斯推断的评估准则。数据生成分布与统计模型之间的 KL 散度记为

$$K(w) = \int dxq(x) \log \frac{q(x)}{p(x|w)}. \tag{6}$$

作为技术假设，我们假设参数集 W $\subset \mathbb{R}^d$ 足够宽且紧致，并且先验在其上为正且有界

$$K^{-1}(0) := \{w \in \mathcal{W} \mid K(w) = 0\}, \tag{7}$$

即，对于任意 $w \in K^{-1}(0)$， $0 < \varphi(w) < \infty$。此外，我们假设 $\varphi(w)$ 是 W 上的 C^∞-函数，而 $K(w)$ 是 W 上的解析函数。 $q(x)$ 的熵及其经验熵分别记为

$$S = -\int dxq(x) \log q(x), \tag{8}$$

$$S_n = -\frac{1}{n} \sum_{i=1}^{n} \log q(X_i). \tag{9}$$

根据定义， $X_i \sim q(x)$ 和 $X^n \sim \prod_{i=1}^n q(x_i)$ 成立；因此，令 $\mathbb{E}_n[\cdot]$ 为定义在整个数据集上的期望算子：

$$\mathbb{E}_n[\cdot] = \int dx^n \prod_{i=1}^{n} q(x_i)[\cdot], \tag{10}$$

其中 $dx^n = dx_1 \dots dx_n$。于是，我们得到以下 KL 散度

$$\int dx^n \prod_{i=1}^{n} q(x_i) \log \frac{\prod_{i=1}^n q(x_i)}{Z_n} = -\mathbb{E}_n\left[nS_n\right] - \mathbb{E}_n[\log Z_n] \tag{11}$$

$$= -nS + \mathbb{E}_n[F_n], \tag{12}$$

其中 $\mathbb{E}_n[S_n]$ 是 S。期望自由能是唯一取决于模型和先验的项。因此，自由能被用作选择模型的判据。另一方面，贝叶斯泛化误差 G_n 定义为数据生成分布与预测分布之间的 KL 散度：

$$G_n = \int dxq(x) \log \frac{q(x)}{p^*(x|X^n)}. \tag{13}$$

这里，贝叶斯推断定义为推断数据生成分布可能就是预测分布。对于任意有限的 n，根据边际似然（4）和预测分布的定义，

(5)，我们有

$$p^*(X_{n+1}|X^n) = \frac{1}{Z_n}\int dw\varphi(w)\prod_{i=1}^{n} p(X_i|w)p(X_{n+1}|w) \tag{14}$$

$$= \frac{1}{Z_n}\int dw\varphi(w)\prod_{i=1}^{n+1} p(X_i|w) \tag{15}$$

$$= \frac{Z_{n+1}}{Z_n}. \tag{16}$$

对两边取负对数的期望值，根据 [55], ，我们得到

$$\mathbb{E}_{n+1}[-\log p^*(X_{n+1}|X^n)] = \mathbb{E}_{n+1}[-\log Z_{n+1} - (-\log Z_n)] \tag{17}$$

$$\mathbb{E}_n[G_n] + S = \mathbb{E}_{n+1}[F_{n+1}] - \mathbb{E}_n[F_n]. \tag{18}$$

因此，当数据生成过程未知时， G_n 和 F_n 是贝叶斯推断中的重要随机变量。 $q(x)$未知的情况被视为一般情形 [38, 59]。此外，一般而言，当模型具有层次结构或潜变量 [55, 58]时，模型可能是奇异的。因此，我们研究它们在 CBM 和多任务情形下的渐近行为。为了从理论上处理模型奇异的情形，需要用到代数几何中的奇点消解。附录 A 对这一理论作了简要介绍。

3 主要定理

本节陈述主要定理：CBM 和多任务模型的实对数典范阈值的精确值。令 n 为样本量， N 为输入维度， K 为概念数量， M 为输出维度。为简单起见，我们考虑回归情形下的 CBM，且概念为实数：输入数据 x 是一个 N维向量，概念 c 是一个 K维向量，输出数据 y是一个 M维向量。我们考虑 c或 y包含分类变量的情形，如第 4 节所述。令 $A = (a_{ik})_{i=1}^{M\ K}, , _{k=1}$和 $B = (b_{kj})_{k=1}^{K\ N}, , _{j=1}$分别为 $M \times K$和 $K \times N$矩阵。它们是 CBM 的连接权重： $y = ABx$ 和 $c = Bx$。与 CBM 类似，我们考虑多任务模型。令 H 为该模型中间层的单元数。矩阵$U = (u_{ik})_{i=1\ k=1}^{M+K\ H}, ,$ 和 $V = (v_{kj})_{k=1}^{H\ N}, , _{j=1}$表示多任务模型的连接权重： $[y;\ c] = UVx$，其中$[y;\ c]$是通过将 y和 c拼接为列向量而构造的一个 $(M+K)$维向量，即令$z = [y; c]$，我们有

$$z = (z_h)_{h=1}^{M+K},\ z_h = \begin{cases} y_h & 1 \leqq h \leqq M, \\ c_{h-M+1} & M+1 \leqq h \leqq M+K. \end{cases} \tag{19}$$

符号说明另请参见表1。下文中，我们将 $[\cdot;\cdot]$ 算子应用于矩阵和向量，以与上文相同的方式将它们纵向拼接。

下文将定义 CBM 和多任务的实对数典范阈值。我们考虑三层线性神经网络。首先，我们说明这些模型结构。

定义3.1 (CBM)。 设 $q_1(y, c|x)$和 $p_1(y, c|A, B, x)$ 是给定$(y, c) \in \mathbb{R}^M \times \mathbb{R}^K$ 时 $x \in \mathbb{R}^N$ 的条件概率密度函数，如下所示：

$$q_1(y, c|x) = p_1(y, c|A_0, B_0, x), \tag{20}$$

$$p_1(y, c|A, B, x) \propto \exp\left(-\frac{1}{2}\|y - ABx\|^2\right)\exp\left(-\frac{\gamma}{2}\|c - Bx\|^2\right), \tag{21}$$

表1：变量描述

变量	描述	索引
$b_j = (b_{kj}) \in \mathbb{R}^K$ $a_k = (a_{ik}) \in \mathbb{R}^M$	从 x 到 c的连接权重 从 c 到 y的连接权重	对于 $k = 1, \ldots, K$ 对于 $i = 1, \ldots, M$
$v_j = (v_{kj}) \in \mathbb{R}^H$ $u_k = (u_{ik}) \in \mathbb{R}^{M+K}$	从 x 到中间层的连接权重 从中间层到 z的连接权重	对于 $k = 1, \ldots, H$ 对于 $i = 1, \ldots, M + K$
$x = (x_j) \in \mathbb{R}^N$ $c = (c_k) \in \mathbb{R}^K$ $y = (y_i) \in \mathbb{R}^M$ $z = (z_h) \in \mathbb{R}^{M+K}$	第j个输入是 x_j 第k个概念是 c_k 第i个输出是 y_i 多任务的第h个输出是 z_h	对于 $j = 1, \ldots, N$ 对于 $k = 1, \ldots, K$ 对于 $i = 1, \ldots, M$ 对于 $h = 1, \ldots, M + K$
$*_0$ 和 $*^0$	与 $*$对应的最优变量或真实变量	-

其中， $A_0 = (a^0_{ik})^{M\ K}_{i=1\ ,k=1}$ 和 $B_0 = (b^0_{kj})^{K\ N}_{k=1\ ,j=1}$ 是真实参数， $\gamma > 0$ 是用于控制任务与解释之间权衡的正常数 *[28]*。先验密度函数记为 $\varphi_1(A, B)$。*CBM* 的数据生成分布及其统计模型分别定义为 $q_1(y, c|x)$ 和 $p_1(y, c|A, B, x)$。

这些分布基于联合概念瓶颈模型的损失函数，该模型具有最高的分类性能 [28]。第 5 节将讨论其他类型的 CBM。该损失由 y 与 ABx 之间的损失和 c 与 Bx之间的损失的线性组合定义。对于回归，这些损失是欧几里得距离平方，它们的线性组合等价于负对数似然$(-\log \prod_{l=1}^n p_1(y^l, c^l|A, B, x^l))$，其中数据为 $(x^l, c^l, y^l)_{l=1}^n \in (\mathbb{R}^N \times \mathbb{R}^K \times \mathbb{R}^M)^n$。由于 CBM 假设概念是可观测的， c 服从数据生成分布，因此 A_0 的列数和 B_0 的行数均等于 K，即概念数量。

设一个密度为 $p_{12}(y|A, B, x) \propto \exp\left(-\frac{1}{2}\|y - ABx\|^2\right)$。那么，标准模型的统计模型为$p_{12}(y|A, B, x)$，标准模型的数据生成分布可表示为 $q_{12}(y|x) := p_{12}(y|A_0, B_0, x)$；然而，只考虑标准模型时， A_0B_0 的秩可能小于 K，也就是说，存在一对矩阵 (A'_0, B'_0)，使得 $A_0B_0 = A'_0B'_0$ 且秩$A'_0B'_0 < K$。换言之，如果我们无法观测概念，就会出现模型选择问题：我们应如何设计中间单元的数量，才能找到数据生成分布，或找到能够实现高泛化性能的预测分布？当中间层单元的数量超过真实秩（即这些单元的真实数量）时，多任务模型中也会出现这一问题： $H > H_0$。多任务模型的分布定义如下。

定义 3.2 (多任务)。令 $z = [y; c]$。令 $q_2(z|x)$ 和 $p_2(z|U, V, x)$ 为 $z \in \mathbb{R}^{M+K}$ 给定 $x \in \mathbb{R}^N$ 时的条件概率密度函数，如下所示：

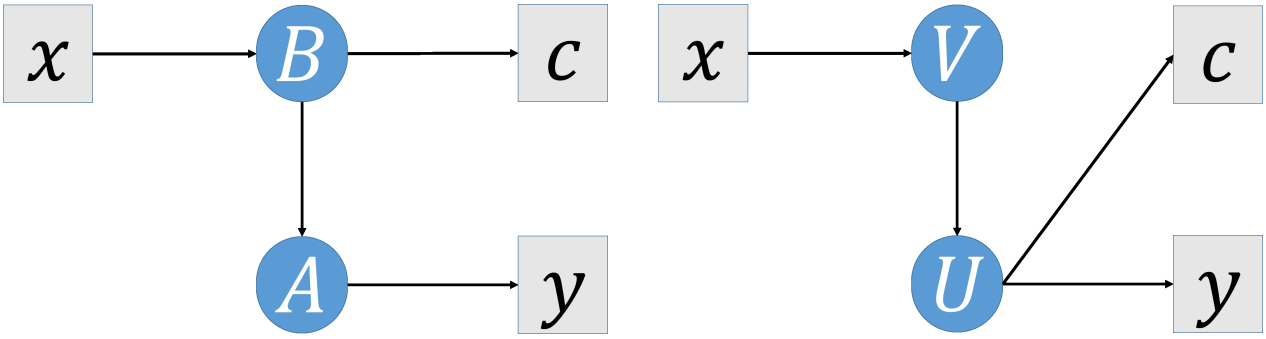
$$q_2(z|x) = p_2(z|U_0, V_0, x), \tag{22}$$

$$p_2(z|U, V, x) \propto \exp\left(-\frac{1}{2}\|z - UVx\|^2\right), \tag{23}$$

其中， $U_0 = (v^0_{ik})^{M+K\ H_0}_{i=1\ k=1\ ,}$和 $V_0 = (v^0_{kj})^{H_0\ N}_{k=1\ ,j=1}$ 为真实参数， H_0 是 U_0V_0 的秩。先验密度函数记为 $\varphi_2(U, V)$。多任务的数据生成分布及其统计模型分别定义为 $q_2(z|x)$ 和 $p_2(z|U, V, x)$。

标准模型的数据生成分布及统计模型定义为多任务在 $K = 0$ 时的 $q_2(y|x)$ 和 $p_2(y|U, V, x)$。

我们在图1a和图1b中将 CBM 和多任务表示为图模型。在 CBM（尤其是联合概念瓶颈模型）中，概念作为观测值插入最后一个中间层与输出层之间，但从 c 到 y 的连接权重 A 是根据关系 $y = ABx$学习的。概念插入被表示为模型的另一部分： $c = Bx$。然而，在多任务中，概念与输出拼接在一起，连接权重（U, V）作为 $[y$进行训练； $c] = UVx$。



(a) CBM 的图模型　　(b) 多任务的图模型

图 1：(a) 本图展示了神经网络为三层线性时 CBM（特别是联合概念瓶颈模型）的图模型。方块（x、 c、 y）表示观测数据，圆圈（A、 B）表示可学习参数（线性神经网络的连接权重）。箭头表示我们从左侧变量到右侧变量设定条件概率模型： $p_1(y, c|A, B, x) = p_{12}(y|A, B, x)p_{11}(c|x, B)$，其中$p_{12}(y|A, B, x)$是标准模型的统计模型， $p_{11}(c|x, B)$是满足$p_{11}(c|B, x) \propto \exp\left(-\frac{\gamma}{2}\|c - Bx\|^2\right)$的密度函数。

（b）该图展示了神经网络为三层线性结构时多任务的图模型。与上文相同，方形（x， c， y）表示观测值，圆形（U， V）表示可学习权重，箭头则分别对应条件概率模型。为便于与 CBM 比较，我们将 c 和 y 也画成方形；但实际上，由于它们被拼接成一个向量： p_2（$z|U$， V， x），因此它们被视为输出，其中 $z = [y;\ c]$。

我们将 CBM 和多任务的实对数典范阈值定义如下。

定义 3.3 （CBM 的实对数典范阈值）。 令 $K_1(A, B)$ 为 q_1 与 p_1之间的 *KL* 散度：

$$K_1(A, B) = \iiint dydcdxq'(x)q_1(y, c|x) \log \frac{q_1(y, c|x)}{p_1(y, c|A, B, x)}, \tag{24}$$

其中， $q'(x)$ 是输入的数据生成分布。 $q'(x)$ 不可观测，并假设其为正且有界。假设 $\varphi_1(A, B) > 0$ 在 $K_1^{-1}(0) \ni (A_0, B_0)$上为正且有界。那么，*CBM* 学习理论中的 ζ 函数是单变量复数z $(\mathrm{Re}(z) > 0)$的全纯函数

$$\zeta_1(z) = \iint K_1(A, B)^z dAdB \tag{25}$$

并且可以解析延拓为整个复平面 $\mathbb{C}$ 上的唯一亚纯函数，其所有极点都是负有理数。*CBM* 的实对数典范阈值定义为 λ_1，其中 $\zeta_1(z)$ 的最大极点为 $(-\lambda_1)$。其重数 m_1 定义为最大极点的阶数。

定义 3.4 （多任务的实对数典范阈值）。 令 $K_2(U, V)$ 为 q_2 与 p_2之间的 *KL* 散度 ：

$$K_2(U, V) = \iiint dydcdxq'(x)q_2(y, c|x) \log \frac{q_2(y, c|x)}{p_2(y, c|U, V, x)}, \tag{26}$$

其中 $q'(x)$ 与定义 *3.3* 相同。假设 $\varphi_2(U, V) > 0$ 在 $K_2^{-1}(0) \ni (U_0, V_0)$上为正且有界。与 *CBM* 情形一样，多任务的学习理论 ζ 函数是以下关于 z $(\mathrm{Re}(z) > 0)$的全纯函数

$$\zeta_2(z) = \iint K_2(U, V)^z dUdV \tag{27}$$

多任务的实对数典范阈值定义为 λ_2，其中 $\zeta_2(z)$ 的最大极点为 $(-\lambda_2)$，其重数 m_2定义为最大极点的阶数。

将 $N \times N$ m 阵 $\mathscr{X} = \left(\int x_i x_j q'(x) dx\right)^N$, $i=1$ ^{N}, $j=1$ · 那么，我们的主要结果如下定理所示 .

定理 3.1 (主定理1)。 假设 $\mathscr{X}$ 为正定，且(A, B) 属于某个紧集。当网络架构为三层线性时，*CBM* 是正则统计模型。因此，利用输入维度 N、概念数量 K和输出维度 M，*CBM* 的实对数典范阈值 λ_1及其重数 m_1 如下：

$$\lambda_1 = \frac{1}{2}(M+N)K,\ m_1 = 1. \tag{28}$$

定理 3.2 (主定理2)。 假设 $\mathscr{X}$为正定，且(U, V) 属于某个紧集。令 λ_2 为三层线性神经网络多任务形式的实对数典范阈值， m_2 为其重数。利用输入维度 N、概念数量 K和中间单元 H、输出维度 M以及真实秩 H_0， λ_2 和 m_2 可表示如下：

1. 在 $M+K+H_0 \leqq N+H$、 $N+H_0 \leqq M+K+H$ 和 $H+H_0 \leqq N+M+K$ 的情形下，

(a) 并且如果 $N+M+K+H+H_0$ 是偶数，那么

$$\lambda_2 = \frac{1}{8}\{2(H+H_0)(N+M+K) - (N-M-K)^2 - (H+H_0)^2\},\ m_2 = 1.$$

(b) 并且如果 $N+M+K+H+H_0$ 是奇数，那么

$$\lambda_2 = \frac{1}{8}\{2(H+H_0)(N+M+K) - (N-M-K)^2 - (H+H_0)^2 + 1\},\ m_2 = 2.$$

2. 在 $N+H < M+K+H_0$ 的情形下，那么

$$\lambda_2 = \frac{1}{2}\{HN + H_0(M+K-H)\},\ m_2 = 1.$$

3. 在 $M+K+H < N+H_0$ 的情形下，那么

$$\lambda_2 = \frac{1}{2}\{H(M+K) + H_0(N-H)\},\ m_2 = 1.$$

4. 否则（即 $N+M+K < H+H_0$），那么

$$\lambda_2 = \frac{1}{2}N(M+K),\ m_2 = 1.$$

这些定理给出了式 (1) 和 (2) 中贝叶斯期望泛化误差与自由能的精确渐近形式。其证明见 B。

定理 3.1 表明，当神经网络的网络架构为三层线性时，概念瓶颈结构会使神经网络成为正则模型。下面给出证明概要。利用三层神经网络的实对数典范阈值，可以立即证明定理 3.2 [6],，因为在模型 $p_2(z|A,\ B,\ x)$ 中，输入维度、中间层单元数、输出维度以及真实参数矩阵乘积的秩分别为 N、 H、 $M+K$ 和 H_0。

定理 *3.1* 的证明概要。设二元关系 $\sim$ 表示两侧的实对数典范阈值和重数相等。从 q_1 到 p_1 的 KL 散度可展开为

$$K_1(A,B) \propto \iiint dydcdxq'(x)q(y,c|x)\left(-\|y-A_0B_0x\|^2 - \gamma\|c-B_0x\|^2\right. \tag{29}$$

$$\left.+\|y-ABx\|^2 + \gamma\|c-Bx\|^2\right) \tag{30}$$

$$\sim \|AB-A_0B_0\|^2 + \|B-B_0\|^2. \tag{31}$$

为计算 λ_1和 m_1，我们求出 K_1^{-1}(0)。令 k$AB - A_0B_0$k^2 +k$B - B_0$k，则有 $(A,\ B) = (A_0,\ B_0)$。这意味着 K_1^{-1}(0) 可视为一个单点集；因此，三层线性情形下的 CBM 是正则的。

□

4 主要定理的展开

我们将观测噪声服从高斯分布时CBM的实对数典范阈值定义为.....（参见定义3.1和3.3）。这对应于原始CBM研究中从X射线图像到关节炎等级的回归[28]。不过，我们通常也可以将CBM视为分类器，并将概念视为分类变量。例如，在 [28], 中，Koh等人展示了一个利用鸟类属性概念进行鸟类物种分类的任务。综上，我们有以下四种情形：

1. $p_{12}(y|A,\ B,\ x)$和 $p_{11}(c|B,\ x)$均服从高斯分布（以实数概念为变量的回归任务）。

2. $p_{12}(y|A,\ B,\ x)$服从高斯分布，而 $p_{11}(c|B,\ x)$服从《伯努利》分布（以分类概念为变量的回归任务）。

3. $p_{12}(y|A,\ B,\ x)$ i 为类别变量，而 $p_{11}(c|B,\ x)$服从高斯分布（以实数概念为变量的分类任务 ）.

4. $p_{12}(y|A, B, x)$ 是类别型的，而 $p_{11}(c|B, x)$ 服从《伯努利》分布（概念为类别型的分类任务）。

请注意，概念并非互斥；因此， $p_{11}(c|B, x)$ 的分布必须是《伯努利》分布（而非类别分布）。我们证明，在上述情形中，类似于定理 3.1 的结果同样成立。在展开主要定理之前，我们先定义 sigmoid 函数和 softmax 函数。分别令 $\sigma_{K'}$: $\mathbb{R}^{K'} \to [0,1]^{K'}$为 K'维多元 sigmoid 函数， $s_{M'}$: $\mathbb{R}^{M'} \to \Delta_{M'}$为 M'维 softmax 函数：

$$\sigma_{K'}(u) = \left(\frac{1}{1+\exp(-u_j)}\right)_{j=1}^{K'},\ u \in \mathbb{R}^{K'}, \tag{32}$$

$$s_{M'}(w) = \left(\frac{\exp(w_j)}{\sum_{j=1}^{M'}\exp(w_j)}\right)_{j=1}^{M'} w \in \mathbb{R}^{M'}, \tag{33}$$

whe其中 $\Delta_{M'}$ 是一个 M'维单纯形。然后，我们可以将各个分布定义如下 ws:

$$p_{12}^1(y|A,B,x) \propto \exp\left(-\frac{1}{2}\|y - ABx\|^2\right), \tag{34}$$

$$p_{12}^2(y|A,B,x) = \prod_{j=1}^{M}(s_M(ABx))_j^{y_j}, \tag{35}$$

$$p_{11}^1(c|B,x) \propto \exp\left(-\frac{\gamma}{2}\|c - Bx\|^2\right), \tag{36}$$

$$p_{11}^2(c|B,x) \propto \left(\prod_{k=1}^{K}(\sigma_K(Bx))_k^{c_k}(1-(\sigma_K(Bx))_k)^{1-c_k}\right)^{\gamma}, \tag{37}$$

其中，$(s_M(ABx))_j$和 $(\sigma_K(Bx))_k$分别是 $s_M(ABx)$ 和 $\sigma_K(Bx)$ 的第 j 个和第 k 个元素。数据生成分布分别记为 $q_{12}^1(y|x) = p_{12}^1(y|A_0, B_0, x)$、 $q_{12}^2(y|x) = p_{12}^2(y|A_0, B_0, x)$、$q_{11}^1(c|x) = p_{11}^1(c|B_0, x)$ 和 $q_{11}^2(c|x) = p_{11}^2($ $c|B_0, x)$。基于上述内容，我们解释密度函数 $\cdot_{kl}^{h}$ 中索引的含义如下。上标 $h \in \{1, 2\}$ 表示响应变量的类型：实值和类别。对于双下标 kl, $k \in \{1, 2\}$ 表示模型（CBM 和多任务），而 $l \in \{1, 2\}$ 表示响应变量（c 和 y）。对于双上标 ij，如定理 4.1 和 4.2 中所用， i 和 j 分别表示响应变量 y 和 c。然后，定理 3.1 可展开如下：

定理 4.1（定理 3.1 的展开）。设

$$p_1^{ij}(y,c|A,B,x) = p_{12}^{i}(y|A,B,x)p_{11}^{j}(c|B,x),\ i=1,2,\ j=1,2, \tag{38}$$

$$q_1^{ij}(y,c|x) = p_1^{ij}(y,c|A_0,B_0,x),\ i=1,2,\ j=1,2. \tag{39}$$

若将 *KL* 散度写作

$$K_1^{ij}(A,B) = \iiint dydcdxq'(x)q_1^{ij}(y,c|x)\log\frac{q_1^{ij}(y,c|x)}{p_1^{ij}(y,c|A,B,x)},\ i=1,2,\ j=1,2, \tag{40}$$

其中 $q'(x)$ 是输入生成分布（与定义 *3.3* 相同）。假设 $\mathscr{X}$ 是正定的，且（*A, B*）属于一个紧集。那么，ζ 函数的最大极点（$-\lambda_1^{ij}$）及其阶数 m_1^{ij}

$$\zeta_1^{ij}(x) = \iint K_1^{ij}(A,B)^z dAdB \tag{41}$$

如下：对于 $i=1,2$ 和 $j=1,2$，

$$\lambda_1^{1j} = \frac{1}{2}(M+N)K, \tag{42}$$

$$\lambda_1^{2j} = \frac{1}{2}(M+N-1)K, \tag{43}$$

$$m_1^{ij} = 1. \tag{44}$$

此外，我们将主定理 3.2 推广到多任务情形。一般来说，多任务情形也有两个任务和概念类型，与上述情况相同。UVx的维数为 $M+K$；因此，我们可以像处理 $z=[y;c]$那样，将前面的 M 维部分与其余（K 维）部分进行分解。我们将这一分解定义为 $UVx = [(UVx)^{\mathrm{y}};(UVx)^{\mathrm{c}}]$，其中 $(UVx)^{\mathrm{y}} = ((UVx)_h)_{h=1}^{M}$ 和 $(UVx)^{\mathrm{c}} = ((UVx)_h)_{h=M+1}^{M+K}$。由于很容易证明 $\mathbf{k}z - UVx\mathbf{k}^2 = \mathbf{k}y - (UVx)^{\mathrm{y}}\mathbf{k}^2 + \mathbf{k}c - (UVx)^{\mathrm{c}}\mathbf{k}^2$,，多任务模型 $p_2(z|U,V,x)$ 可以分解为

$$p_2(z|U,V,x) = p_{22}(y|U,V,x)p_{21}(c|U,V,x), \tag{45}$$

其中

$$p_{22}(y|U,V,x) \propto \exp\left(-\frac{1}{2}\|y-(UVx)^{\mathrm{y}}\|^2\right), \tag{46}$$

$$p_{21}(c|U,V,x) \propto \exp\left(-\frac{1}{2}\|c-(UVx)^{\mathrm{c}}\|^2\right). \tag{47}$$

与 CBM 的情形类似，我们将每个分布定义如下：

$$p_{22}^{1}(y|U,V,x) \propto \exp\left(-\frac{1}{2}\|y-(UVx)^{\mathrm{y}}\|^2\right), \tag{48}$$

$$p_{22}^{2}(y|U,V,x) = \prod_{j=1}^{M}(s_M((UVx)^{\mathrm{y}}))_j^{y_j}, \tag{49}$$

$$p_{21}^{1}(c|U,V,x) \propto \exp\left(-\frac{1}{2}\|c-(UVx)^{\mathrm{c}}\|^2\right), \tag{50}$$

$$p_{21}^{2}(c|U,V,x) = \prod_{k=1}^{K}(\sigma_K((UVx)^{\mathrm{c}}))_k^{c_k}(1-((\sigma_K(UVx)^{\mathrm{c}}))_k)^{1-c_k}. \tag{51}$$

这些数据生成分布分别记为 $q_{22}^{1}(y|x) = p_{22}^{1}(y|U_0,\ V_0,\ x)$、$q_{22}^{2}(y|x) = p_{22}^{2}(y|U_0,\ V_0,\ x)$、$q_{21}^{1}(c|x) = p_{21}^{1}(c|U_0,\ V_0,\ x)$ 和 $q_{21}^{2}(c|x) = p_{11}^{2}(c|U_0,\ V_0,\ x)$。那么，定理 3.2 可展开如下：

定理 4.2（定理 3.2 的展开）。与定理 *4.1* 相同，模型和数据生成分布可以表示为

$$p_2^{ij}(z|U,V,x) = p_{22}^i(y|U,V,x)p_{21}^j(c|U,V,x),\ i=1,2,\ j=1,2, \tag{52}$$

$$q_2^{ij}(z|x) = p_2^{ij}(y,c|U_0,V_0,x),\ i=1,2,\ j=1,2. \tag{53}$$

此外，*KL* 散度可以表示为

$$K_2^{ij}(U,V) = \iiint dydcdxq'(x)q_2^{ij}(y,c|x)\log\frac{q_2^{ij}(y,c|x)}{p_2^{ij}(y,c|U,V,x)},\ i=1,2,\ j=1,2, \tag{54}$$

其中，$q'(x)$ 是输入生成分布（与定义 *3.4* 相同）。假设 $\mathscr{X}$ 是正定的，且(U,V) 属于一个紧集。λ_2 和 m_2 表示定理 *3.2* 中(N, H, M, K, H_0) 的函数。那么，ζ 函数的最大极点$(-\lambda_2^{ij})$ 及其阶数 m_2^{ij} 可以写为

$$\zeta_2^{ij}(x) = \iint K_2^{ij}(U,V)^z dUdV, \tag{55}$$

对于 $i=1$，2和 $j=1$，2，有

$$\lambda_2^{1j} = \lambda_2(N,H,M,K,H_0), \tag{56}$$

$$\lambda_2^{2j} = \lambda_2(N,H,M-1,K,H_0), \tag{57}$$

$$m_2^{1j} = m_2(N,H,M,K,H_0), \tag{58}$$

$$m_2^{2j} = m_2(N,H,M-1,K,H_0). \tag{59}$$

我们在 B 中证明定理 4.1 和 4.2。此外，上述展开后的定理还导出以下推论，这些推论考虑了输出或概念由实数和分类变量共同构成的情形（复合情形）。设 y^{r} 和 y^{c} 分别为 M^{r} 维实向量和 M_{c} 维分类变量。它们是输出的观测变量。设 c^{r} 和 c^{c} 分别为 K^{r} 维实向量和 K^{c} 维分类变量。它们作为概念，用于描述由 N 维输入生成的输出。按照 $z=[y$ 的定义方式； $c]$，令 $y=[y^{\mathrm{r}};\ y^{\mathrm{c}}]$ 和$c=[c^{\mathrm{r}};\ c^{\mathrm{c}}]$。此外，设 $M=M_{\mathrm{r}}+M_{\mathrm{c}}$和 $K=K^{\mathrm{r}}+K^{\mathrm{c}}$，其中 $M_{\mathrm{r}}M^{\mathrm{c}}, K^{\mathrm{r}}, K^{\mathrm{c}} \geqq 1$。类似地，有$ABx=[(ABx)^{\mathrm{r}};\ (ABx)^{\mathrm{c}}]$ 和 $Bx=[(Bx)^{\mathrm{r}};\ (Bx)^{\mathrm{c}}]$，其中

$$(ABx)^{\mathrm{r}} = ((ABx)_h)_{h=1}^{M^{\mathrm{r}}},\ (ABx)^{\mathrm{c}} = ((ABx)_h)_{h=M^{\mathrm{r}}+1}^{M}, \tag{60}$$

$$(Bx)^{\mathrm{r}} = ((Bx)_h)_{h=1}^{K^{\mathrm{r}}},\ (Bx)^{\mathrm{c}} = ((Bx)_h)_{h=K^{\mathrm{r}}+1}^{K}, \tag{61}$$

并且 $(ABx)_h$ 和 $(Bx)_h$ 分别是其中第 h 个矩阵元素。即使输出和概念同时由实数和分类变量构成，利用定理 4.1 和 4.2，我们也可以立即推导出实对数典范阈值 λ_1^{com} 及其重数 m_1^{com}，如下所示：

推论 4.1 (复合情形下 CBM 的实对数典范阈值)。设 $p_1^{\mathrm{com}}(y, c|A, B, x)$ 为复合情形下 *CBM* 的统计模型，而 $p_{12}^{\mathrm{com}}(y|A, B, x)$ 和 $p_{11}^{\mathrm{com}}(c|B, x)$ 为以下概率分布：

$$p_{12}^{\mathrm{com}}(y|A,B,x) \propto \exp\left(-\frac{1}{2}\|y^{\mathrm{r}}-(ABx)^{\mathrm{r}}\|^2\right) \times \prod_{j=1}^{M^{\mathrm{c}}}(s_{M^{\mathrm{c}}}((ABx)^{\mathrm{c}}))_j^{y_j}, \tag{62}$$

$$p_{11}^{\mathrm{com}}(c|B,x) \propto \exp\left(-\frac{\gamma}{2}\|c^{\mathrm{r}}-(Bx)^{\mathrm{r}}\|^2\right) \times \left(\prod_{k=1}^{K^{\mathrm{c}}}(\sigma_{K^{\mathrm{c}}}((Bx)^{\mathrm{c}}))_k^{c_k}(1-(\sigma_{K^{\mathrm{c}}}((Bx)^{\mathrm{c}}))_k)^{1-c_k}\right)^{\gamma}. \tag{63}$$

数据生成分布记为

$$q_1^{\mathrm{com}}(y,c|x) = p_{12}^{\mathrm{com}}(y|A_0,B_0,x)p_{11}^{\mathrm{com}}(c|B_0,x). \tag{64}$$

KL 散度可表示为

$$K_1^{\mathrm{com}}(A,B)=\iiint dydcdxq'(x)q_1^{\mathrm{com}}(y,c|x)\log\frac{q_1^{\mathrm{com}}(y,c|x)}{p_1^{\mathrm{com}}(y,c|A,B,x)}, \tag{65}$$

其中 $q'(x)$ 为输入生成分布（与定义 *3.3* 相同）。假设 $\mathscr{X}$ 为正定，且(A,B) 属于某个紧集。那么，K_1^{com}的实对数典范阈值 λ_1^{com}及其重数 m_1^{com}可表示如下：

$$\lambda_1^{\mathrm{com}}=\frac{1}{2}(M^{\mathrm{r}}+M^{\mathrm{c}}+N-1)(K^{\mathrm{r}}+K^{\mathrm{c}}), \tag{66}$$

$$m_1^{\mathrm{com}}=1. \tag{67}$$

这是因为这些概念按照与 $z=[y;\ c]$ 和 $UVx=[(UVx)^{\mathrm{y}};\ (UVx)^{\mathrm{c}}]$之间相同的对应关系，分解为实数部分和类别部分。多任务的复合情形也很容易确定，如下所示。注意，$(UVx)^{\mathrm{r}}$ 和$(UVx)^{\mathrm{c}}$ 的定义方式与 ABx相同。

推论4.2（多任务复合情形的实对数典范阈值）。 设 $p_2^{\mathrm{com}}(y,c|U,V,x)$ 为多任务复合情形下的统计模型，且 $p_{22}^{\mathrm{com}}(y|U,V,x)$ 和 $p_{21}^{\mathrm{com}}(c|U,V,x)$ 为以下概率分布：

$$p_{22}^{\mathrm{com}}(y|U,V,x)\propto\exp\left(-\frac{1}{2}\|y^{\mathrm{r}}-(UVx)^{\mathrm{r}}\|^2\right)\times\prod_{j=1}^{M^{\mathrm{c}}}(s_{M^{\mathrm{c}}}((UVx)^{\mathrm{c}}))_j^{y_j}, \tag{68}$$

$$p_{21}^{\mathrm{com}}(c|U,V,x)\propto\exp\left(-\frac{\gamma}{2}\|c^{\mathrm{r}}-(UVx)^{\mathrm{r}}\|^2\right)\times\prod_{k=1}^{K^{\mathrm{c}}}(\sigma_{K^{\mathrm{c}}}((UVx)^{\mathrm{c}}))_k^{c_k}(1-(\sigma_{K^{\mathrm{c}}}((UVx)^{\mathrm{c}}))_k)^{1-c_k}. \tag{69}$$

数据生成分布记为

$$q_2^{\mathrm{com}}(y,c|x)=p_{22}^{\mathrm{com}}(y|U_0,V_0,x)p_{21}^{\mathrm{com}}(c|U_0,V_0,x). \tag{70}$$

将 *KL* 散度记为

$$K_2^{\mathrm{com}}(U,V)=\iiint dydcdxq'(x)q_2^{\mathrm{com}}(y,c|x)\log\frac{q_2^{\mathrm{com}}(y,c|x)}{p_2^{\mathrm{com}}(y,c|U,V,x)}, \tag{71}$$

其中 $q'(x)$ 是输入生成分布（与定义 *3.4* 相同）。假设 $\mathscr{X}$ 是正定的，且(U,V) 属于某个紧集。那么， K_2^{com}的实对数典范阈值 λ_2^{com}及其重数 m_2^{com}如下：

$$\lambda_2^{\mathrm{com}}=\lambda_2(N,H,M^{\mathrm{r}}+M^{\mathrm{c}}-1,K^{\mathrm{r}}+K^{\mathrm{c}},H_0), \tag{72}$$

$$m_2^{\mathrm{com}}=m_2(N,H,M^{\mathrm{r}}+M^{\mathrm{c}}-1,K^{\mathrm{r}}+K^{\mathrm{c}},H_0). \tag{73}$$

5 讨论

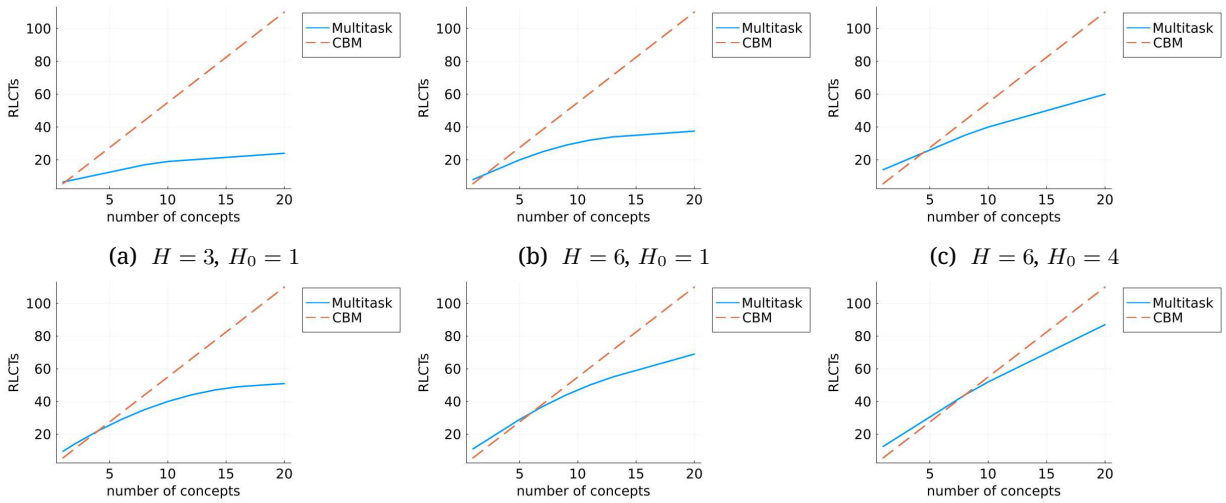
本文阐述了在三层线性神经网络情形下，如何确定 CBM 和多任务的实对数典范阈值。利用这些实对数典范阈值以及式（1）和式（2），我们还明确了这些模型的贝叶斯泛化误差和边际似然的精确渐近形式。

本研究有两项局限。第一，本文研究的是三层神经网络。如果输入是一个高精度神经网络的中间层，那么在学习最后一个全连接线性层时，该模型会被冻结。因此，我们的结果不仅为学习三层神经网络奠定了基础，也对迁移学习具有价值。事实上，从特征提取的角度来看，可以将最先进的神经网络的中间层用作另一个模型的输入，而非使用原始输入 [11, 51, 68]。

第二项局限是，我们用于贝叶斯推断的 CBM 公式基于联合概念瓶颈模型。CBM 还有另外两种类型：独立概念瓶颈模型和序贯概念瓶颈模型 [28]。在独立概念瓶颈模型中，函数 $x \mapsto c$ 和 $c \mapsto y$ 是独立学习的。当神经网络为三层线性网络时，学习独立概念瓶颈模型等价于仅估计两个独立的线性变换 $c = Bx$和 $y = Ac$。独立概念瓶颈模型的图模型为 $x \to B \to c \to A \to y$。显然，$(A,\ B)$ 可识别，且该模型是正则的。相比之下，序贯概念瓶颈模型采用两步估计。首先，将 B ^ ^估计为 $c = Bx$。然后，将 A 学习为 $y = A\hat{c}$，其中 B 是 B 的估计量，且$\hat{c} = Bx$。由于$\hat{c}$ 服从在给定 x条件下的 c 预测分布，其图模型与联合概念瓶颈模型相同（图 1a）。着眼于两步估计这一特点，对于 A的贝叶斯推断，我们将 B 的先验设为由 $c = Bx$推断得到的 B 后验分布，即先验分布依赖于数据。若忽略两步估计这一特点，序贯概念瓶颈模型的贝叶斯推断就与联合概念瓶颈模型相同。具有数据依赖先验分布的奇异学习理论颇具挑战，因为该理论将先验作为积分的测度，用以刻画实对数典范阈值及其重数（见命题 A.1）。为解决这一问题，必须建立一种用于分析贝叶斯泛化误差和自由能的新方法。

尽管存在上述一些局限性，但通过本研究的贡献，我们可以获得一个新的视角：CBM 是一种参数受限模型。根据定理 3.1 的证明，概念瓶颈结构 $p_{11}(c|B,\ x)$使神经网络成为正则模型，而标准模型则是奇异的。换言之，为了求得实对数典范阈值，我们应考虑概念瓶颈结构为解析集 $K_1^{-1}(0)$施加的约束条件 $B = B_0$。这一结构被添加到标准模型中，以提升其可解释性。因此，在可解释模型的奇异学习理论中，定理 3.1 给出了非平凡的结果：对解释施加的参数约束会影响泛化行为；在这种情况下，为实现可解释性而施加的限制会使模型从奇异变为正则。

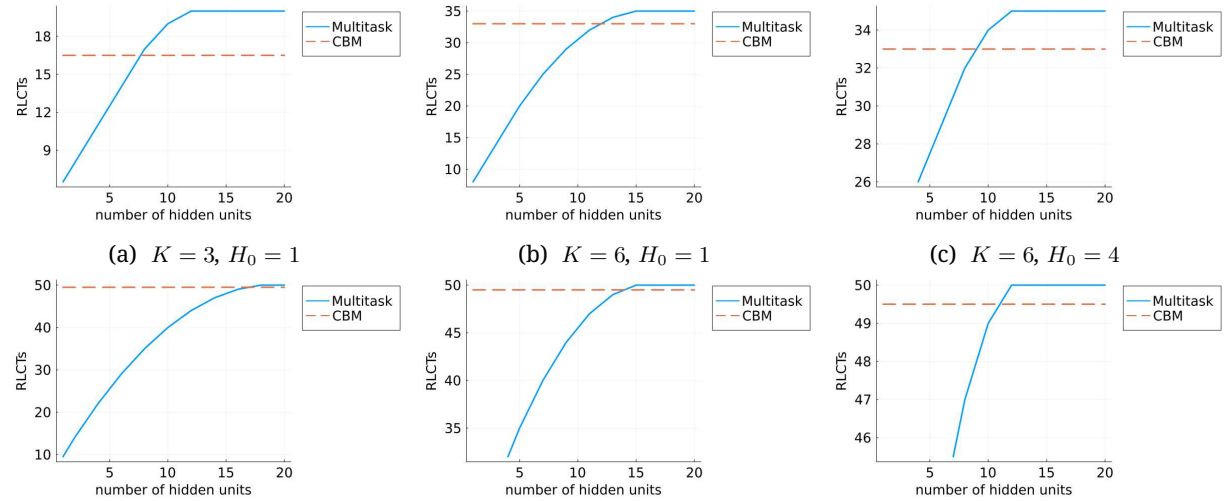
最后，我们讨论 CBM 和多任务模型的模型选择过程。这两种模型都使用由输入、概念和输出构成的类似数据集。此外，它们还利用观测到的概念来解释预测结果背后的原因。在这两种方法中，监督学习都是从输入到概念和输出进行的。不过，它们的模型结构有所不同：CBM 将概念用作中间层单元，而多任务模型则将其用于输出。这种差异会如何影响泛化性能和知识发现的准确性？我们从贝叶斯泛化误差和自由能（负对数边际似然）的角度探讨了这一问题。图 2a–2f 展示了 CBM 和多任务模型中的实对数典范阈值随概念数量（即 K）增加而变化的情况。此外，图 3a–3f 展示了中间层单元数量 H 增加时的情形。在这两组图中，CBM 的实对数典范阈值呈直线，而多任务模型的实对数典范阈值则呈等高线（分段线性）形状。对于 CBM，如定义3.1所述，其特征由 $(M,\ N,\ K)$ 表示，并且网络架构中的中间层单元与概念数量相同。由此必然得出，CBM 的实对数典范阈值不依赖于H，即使它使用与多任务模型相同的一对 $(y,\ c,\ x)$。对于多任务模型，根据 [6],，标准模型的实对数典范阈值图也类似于一个以 H 为自变量、以实对数典范阈值为因变量的函数图。这种相似性可直接由定理 3.2 和 [6] 推出（另请参见定理 3.2 的证明）。此外，我们可以确定 CBM 和多任务模型的实对数典范阈值曲线的交点。实对数典范阈值决定了贝叶斯泛化误差和自由能的渐近形式。如果实对数典范阈值较大，贝叶斯泛化误差和自由能也会增大。因此，只要厘清它们的理论行为，就能明确与数据分析方法选择相关的问题。因此，对于重视准确性的研究人员和从业者而言，比较 CBM 和多任务模型十分重要。命题 5.1 给出



(a) $H = 3$, $H_0 = 1$ (b) $H = 6$, $H_0 = 1$ (c) $H = 6$, $H_0 = 4$

(d) $H = 9$, $H_0 = 1$ (e) $H = 9$, $H_0 = 4$ (f) $H = 9$, $H_0 = 7$

图 2：CBM 和多任务的实对数典范阈值随 K 变化的行为。纵轴表示实对数典范阈值的值，横轴表示概念数量 K。实对数典范阈值的变化通过关于 K 的函数图像来呈现，其中 $M = 10$ 和 $N = 1$ 固定，H 和 H_0 作为子图标题。CBM 的实对数典范阈值用虚线表示，多任务的实对数典范阈值用实线表示。两者存在显著差异，因为一个是线性的，而另一个是非线性的（分段线性）。这是因为多任务的实对数典范阈值取决于 H 和 H_0，而 CBM 不受其影响。



(a) $K = 3$, $H_0 = 1$ (b) $K = 6$, $H_0 = 1$ (c) $K = 6$, $H_0 = 4$

(d) $K = 9$, $H_0 = 1$ (e) $K = 9$, $H_0 = 4$ (f) $K = 9$, $H_0 = 7$

图 3：CBM 和多任务的实对数典范阈值随 H 变化的行为。纵轴表示实对数典范阈值的值，横轴表示中间层单元（隐藏单元）的数量 H。实对数典范阈值随 H 变化的行为以函数图的形式呈现，其中 $M = 10$ 和 $N = 1$ 固定，K 和 H_0 作为子图标题。CBM 的实对数典范阈值以虚线表示，多任务的则以实线表示。CBM 的实对数典范阈值不依赖于 H，因此为常数。多任务的实对数典范阈值依赖于 H，标准模型的实对数典范阈值也如此，详见 [6]。

关于该问题的分界点：在给定的 $(M,\ H,\ N,\ K)$ 和假设 H_0 下，哪种方法性能更好？如果 $\lambda_1 > \lambda_2$，从贝叶斯泛化误差和自由能的角度来看，多任务优于 CBM。如果 $\lambda_1 \leqq \lambda_2$，则相反。命题 5.1 的证明见 B。

命题 5.1 (CBM 与多任务的实对数典范阈值比较)。结合定理 *3.2* 中的条件分支，*CBM* 的实对数典范阈值大小 λ_1 和多任务的实对数典范阈值大小 λ_2 按如下方式变化。

1. 在情形 $M + K + H_0 \leqq N + H$ 和 $N + H_0 \leqq M + K + H$ 和 $H + H_0 \leqq N + M + K$，

(a) 并且如果 $N + M + K + H + H_0$ 是偶数，则

$$\begin{cases} \lambda_1 > \lambda_2 & (K > H + H_0 - (\sqrt{M} - \sqrt{N})^2), \\ \lambda_1 \leqq \lambda_2 & (\text{otherwise}). \end{cases}$$

(b) 并且如果 $N + M + K + H + H_0$ 是奇数，则

$$\begin{cases} \lambda_1 > \lambda_2 & (K > H + H_0 - M - N + \sqrt{4MN + 1}), \\ \lambda_1 \leqq \lambda_2 & (otherwise). \end{cases}$$

2. 在情形 $N + H < M + K + H_0$ 下，

$$\begin{cases} \lambda_1 > \lambda_2 & ((M + N - H_0)K > (N - H_0)H + MH_0), \\ \lambda_1 \leqq \lambda_2 & (otherwise). \end{cases}$$

3. 在情形 $M + K + H < N + H_0$ 下，

$$\begin{cases} \lambda_1 > \lambda_2 & ((M + N - H_0)K > (N - H)H_0 + MH), \\ \lambda_1 \leqq \lambda_2 & (otherwise). \end{cases}$$

4. 否则（即 $N + M + K < H + H_0$），

$$\begin{cases} \lambda_1 > \lambda_2 & (K > N), \\ \lambda_1 \leqq \lambda_2 & (otherwise). \end{cases}$$

6 结论

我们得出了采用概念瓶颈模型和多任务形式的神经网络在三层线性结构下，贝叶斯泛化误差和自由能的精确渐近行为。这些行为是通过求出这些模型的实对数典范阈值得出的。结果表明，在三层线性情形下，概念瓶颈结构使神经网络成为正则（可识别）的模型。另一方面，三层线性网络的多任务形式仅涉及将概念添加到输出中；因此，贝叶斯泛化误差和自由能的行为与标准模型相似。未来的工作可对多层网络和非线性激活进行理论分析，也可基于奇异学习理论构建序贯概念瓶颈模型。阐明主要定理所呈现的数值行为，也可以成为未来的研究方向。

奇异学习理论

我们简要说明贝叶斯推断与代数几何之间的关系，换句话说，也就是为何需要进行奇点消解。这一理论称为奇异学习理论[55]。

以下奇点消解定理 [24] 的解析形式 [7] 有助于处理式 (6) 中的 $K(w)$ 及式 (7) 中的其零点 $K^{-1}(0)$。

定理 A.1 (Hironaka, Atiyah)。 设 K 是 $\mathcal{W} \subset \mathbb{R}^d$ 上的非负解析函数。假设 $K^{-1}(0)$ 非空。那么，存在一个开集 $\mathcal{W}'$、一个 d 维光滑流形 $\mathcal{M}$，以及一个解析映射 $g\colon \mathcal{M} \to \mathcal{W}'$，使得 $g\colon \mathcal{M} \setminus g^{-1}(K^{-1}(0)) \to \mathcal{W}' \setminus K^{-1}(0)$ 同构，并且

$$K(g(u)) = u_1^{2k_1} \dots u_d^{2k_d}, \tag{74}$$

$$|\det g'(u)| = b(u)|u_1^{h_1} \dots u_d^{h_d}| \tag{75}$$

对 $\mathcal{M}$ 的每个局部图册 $U \ni u$ 均成立，其中 k_j 和 h_j 对于 $j = 1, \dots\dots, d$ 均为非负整数，$\det g'(u)$ 是 g 的雅可比行列式，而 $b : \mathcal{M} \to \mathbb{R}$ 是严格正的解析函数：$b(u) > 0$。

Atiyah 推导出了这种形式，用于分析广义函数（又称超函数）的除法与局部型 ζ 函数之间的关系 [7]。利用定理 A.1，可证明如下结论 [7, 8, 48]。

定理 A.2 (Atiyah、Bernstein、Sato 和 Shintani)。 设 $K\colon \mathbb{R}^d \to \mathbb{R}$ 是变量 $w \in \mathcal{W}$ 的一个解析函数。 $a : \mathcal{W} \to \mathbb{R}$ 表示一个具有紧支集的 C^∞-函数 $\mathcal{W}$。以下单变量复函数

$$\zeta(z) = \int_{\mathcal{W}} |K(w)|^z a(w) dw \tag{76}$$

在 $\mathrm{Re}(z) > 0$ 上是全纯函数。此外，$\zeta(z)$ 可解析延拓为整个复平面 $\mathbb{C}$ 上唯一的亚纯函数。它的所有极点都是负有理数。

假设先验密度 $\varphi(w)$ 的紧支集为 W，且开集 W′ 满足 W $\subset$W′。根据定理 A.2，我们可以定义一个学习理论 ζ 函数。

定义 A.1 （学习理论 ζ 函数）。 设 $K(w) \geqq 0$ 为式（6）中提到的 *KL* 散度，且 $\varphi(w) \geqq 0$ 为满足上述假设的先验密度函数。学习理论 ζ 函数定义为以下单变量复函数

$$\zeta(z) = \int_{\mathcal{W}} K(w)^z \varphi(w) dw.$$

定义 A.2 （实对数典范阈值）。 设 $\zeta(z)$ 为定义 *A.1* 中表示的学习理论 ζ 函数。考虑定理 *A.2* 中 $\zeta(z)$ 的解析延拓。实对数典范阈值（*RLCT*） λ 定义为 $\zeta(z)$ 的负最大极点，其重数 m 定义为最大极点的阶：

$$\zeta(z) = \frac{C(z)}{(z+\lambda)^m} \frac{C_1(z)}{(z+\lambda_1)^{m_1}} \cdots \frac{C_D(z)}{(z+\lambda_D)^{m_D}} \cdots, \tag{77}$$

$$\lambda < \lambda_k \ (k = 1, \dots, D, \dots), \tag{78}$$

其中 $C(z)$ 和 $C_k(z)\,(k = 1, \dots, D, \dots)$ 是取值非零的复函数。

Watanabe 创立了奇异学习理论；他证明，实对数典范阈值 λ 和重数 m决定了渐近贝叶斯泛化误差和自由能 [53, 54, 55]：

定理 A.3 (Watanabe)。 $\zeta(z)$ 由定义 *A.1* 中的学习理论 ζ 函数表示。令 λ和 m分别为由 $\zeta(z)$定义的实对数典范阈值和重数。贝叶斯泛化误差 G_n 和自由能 $F_n = -\log Z_n$ 具有第 *1* 节中所示的渐近形式(1) 和(2) 。

该定理根植于定理 A.1。这就是为什么我们需要通过确定实对数典范阈值及其重数，利用奇点消解来阐明 G_n 和 F_n 的行为。

这里，我们说明如何确定与 $K(w)$ 对应的模型的实对数典范阈值 $\lambda > 0$ 。我们将定理 A.1 应用于学习理论 ζ 函数。由于我们假设参数空间是紧致的，

奇点消解中的流形也紧致。因此，该流形可以被每个局部坐标 U 对应的 $[0,1)^d$ 的并所覆盖。考虑到 $[0,1)^d$ 的单位分解，我们有

$$\zeta(z) = \int_U K(g(u))^z \varphi(g(u))|\det g'(u)|du \tag{79}$$

$$= \sum_\eta \int_U K(g(u))^z \varphi(g(u))|\det g'(u)|\phi_\eta(u)du \tag{80}$$

$$= \sum_\eta \int_{[0,1]^d} u_1^{2k_1 z+h_1} \dots u_d^{2k_d z+h_d} \varphi(g(u))b(u)\phi_\eta(u)du, \tag{81}$$

其中 ϕ_η 是单位分解：$\mathrm{supp}(\phi_\eta) = [0,1]^d$ 且 $\phi_\eta(u) > 0$ 在 $(0,1)^d$内。函数 $\varphi(g(u))$、$b(u)$ 和 ϕ_η 在 $(0,1)^d$内均严格为正；因此，我们应考虑其最大极点

$$\int_{[0,1]^d} u_1^{2k_1 z+h_1} \dots u_d^{2k_d z+h_d} du = \frac{1}{2k_1 z + h_1 + 1} \cdots \frac{1}{2k_d z + h_d + 1}. \tag{82}$$

允许重复后，极点集合可表示如下：

$$\left\{ \frac{h_j+1}{2k_j} \mid j = 1, \dots, d \right\}. \tag{83}$$

因此，我们可以在局部图册 U 中如下求出最大极点（$-\lambda_U$）

$$\lambda_U = \min_{j=1}^{d} \left\{ \frac{h_j+1}{2k_j} \right\}. \tag{84}$$

通过考虑指标的重复，我们还可以求出 U 中的重数，记为 m_U。因此，我们可以确定实对数典范阈值为 $\lambda = \min_U \lambda_U$，重数 m 为极点（$-\lambda$）的阶数，即 $m = m_{\underline{U}}$，其中 $\underline{U} = \mathrm{argmin}_U \lambda_U$。

此外，我们将如下说明实对数典范阈值的一个几何性质：体积维数[62, 55]的极限：

命题 A.1 **.** 设 $V\colon (0,\infty) \to (0,\infty)$、$t \mapsto V(t)$ 是由 $\varphi(w)dw$ 测得的 $K^{-1}((0,t))$ 体积：

$$V(t) = \int_{K(w)<t} \varphi(w)dw. \tag{85}$$

那么，实对数典范阈值 λ 满足以下条件：

$$\lambda = \lim_{t \to +0} \frac{\log V(t)}{\log t}. \tag{86}$$

实对数典范阈值及其重数是解析集 $K^{-1}(0)$ 的双有理不变量。由于它们是双有理不变量，因此不依赖于奇点消解。上述性质刻画了这一事实。

要确定 λ 和 m，我们应考虑与模型对应的具体簇的奇点消解 [24] 。为了明确奇异统计模型的学习系数，我们应计算一族函数的实对数典范阈值的理论值；然而，目前没有针对给定函数集合求实对数典范阈值的标准方法。因此，我们需要针对每个统计模型采用不同的实对数典范阈值求解流程。事实上，正如第 1 节所述，统计学和机器学习领域都已针对各种情形分析了若干模型的实对数典范阈值。我们关于 CBM 和多任务模型的研究为学习理论的知识体系作出了贡献：明确了奇异统计模型的实对数典范阈值。此类研究在实践层面的价值已在第 1 节中介绍。

B 证明

设 $\sim$ 为一个二项关系，其等式两边具有相同的实对数典范阈值和重数，且 $\mathrm{M}(M,\ N)$ 表示 $M \times N$ 实矩阵的集合。然后，我们定义如下实用工具。

定义 B.1（行提取器）。令 $(\cdot)_{<d} : \mathrm{M}(I, J) \to \mathrm{M}(d-1, J),\ \ 2 \leqq d \leqq I,\ 1 \leqq J$ 为一个矩阵算子，用于提取以下子矩阵：

$$(W)_{<d} = \begin{pmatrix} w_{11} & \dots & w_{1J} \\ \vdots & \ddots & \vdots \\ w_{(d-1)1} & \dots & w_{(d-1)J} \end{pmatrix},\ W = (w_{ij})_{i=1,j=1}^{I,J},\ W \in \mathrm{M}(I, J). \tag{87}$$

我们将此算子用于向量 $w \in \mathbb{R}^M$，记作 $\mathbb{R}^M \cong \mathrm{M}(M,1)$；我们称之为列向量。与上文相同，我们定义 $(\cdot)_{>d}$、$(\cdot)_{\leqq d}$ 和 $(\cdot)_{\geqq d}$，其中不等式对应于行索引。此外，$(\cdot)_{\neq d}$ 和 $(\cdot)_{=d}$ 分别定义为 $[(\cdot)_{<d}$；$(\cdot)_{>d}]$ 和 $((\cdot)_{\leqq d})_{\geqq d}$。我们可以立即证明，对于 $W \in \mathrm{M}(I, R)$ 和 $V \in \mathrm{M}(R, J)$，有 $(WV)_{<d} = (W)_{<d}V$，其中 $R \geqq 1$。此外，下标中的其他关系也满足上述规则；我们可以推导出 $(WV)_{>d} = (W)_{>d}V$、$(WV)_{\leqq d} = (W)_{\leqq d}V$、$(WV)_{\geqq d} = (W)_{\geqq d}V$、$(WV)_{\neq d} = (W)_{\neq d}V$ 和 $(WV)_{=d} = (W)_{=d}V$。此外，为了证明我们的定理，我们使用以下引理 [55, 6, 37]。

引理 B.1. 设 $f_1 : \mathcal{W} \to \mathbb{R}$ 和 $f_2 : \mathcal{W} \to \mathbb{R}$ 是非负解析函数。若存在正常数 $\alpha_1, \alpha_2 > 0$ 使得

$$\alpha_1 f_1(w) \leqq f_2(w) \leqq \alpha_2 f_1(w) \tag{88}$$

在 $f_2^{-1}(0)$ 的邻域上成立，则 $f_1 \sim f_2$。

引理 B.2. 设 $K : \mathrm{M}(I, J) \to \mathbb{R},\ W \mapsto K(W)$ 是

$$K(W) = \int dx q'(x) \|(W - W_0)x\|^2, \tag{89}$$

其中 $W_0 \in \mathrm{M}(I, J)$。令 $\Phi : \mathrm{M}(I, J) \to \mathbb{R},\ W \mapsto \Phi(W) = \|W - W_0\|^2$。对称矩阵 $J \times J$ 的 (i, j) 矩阵元素为 $\int x_i x_j q'(x) dx$，其中 $1 \leqq i \leqq N$ 且 $1 \leqq j \leqq N$，记作 $\mathscr{X}$。若 $\mathscr{X}$ 正定，则存在正常数 $\alpha_1, \alpha_2 > 0$，使得 $\alpha_1 \Phi(W) \leqq K(W) \leqq \alpha_2 \Phi(W)$ 在 $K^{-1}(0)$ 的某个邻域上成立。因此，$K \sim \Phi$。

引理 B.3. 设 $K : \Delta_d \to \mathbb{R},\ w \mapsto K(w)$ 为

$$K(w) = \sum_{y \in \mathrm{Onehot}(d)} \prod_{j=1}^{d} (w_j^0)^{y_j} \log \frac{\prod_{j=1}^{d} (w_j^0)^{y_j}}{\prod_{j=1}^{d} (w_j)^{y_j}}, \tag{90}$$

其中 w_0 Δ_d Δ_d d $w = w_j{}_{j=1}^{d} \in \Delta_d \Rightarrow \sum_{j=1}^{d} w_j =$ 位于 的内部，是 维 单纯形，即 $(), w_j \geqq 0)$，且 $\mathrm{Onehot}()$ 是 维 *onehot* 向量的集合

$$\mathrm{Onehot}(d) = \{y \in \{0,1\}^d \mid y_j = 1, y_l = 0\ (l \neq j), \text{for } j = 1, \dots, d\}. \tag{91}$$

令 $\Phi : \Delta_d \to \mathbb{R},\ w \mapsto \Phi(w) = \sum_{j=1}^{d-1} (w_j - w_j^0)^2$。存在正常数 $\alpha_1, \alpha_2 > 0$，使得$\alpha_1 \Phi(w) \leqq K(w) \leqq \alpha_2 \Phi(w)$ 在 $K^{-1}(0)$的某个邻域内成立。因此，$K \sim \Phi$。

引理 B.1 已在 [55] 中证明。引理 B.2 已在 [6] 中证明。引理 B.3 已在 [37] 中证明。此外，在情形 $d = 2$ 下使用引理 B.3，可以推出以下引理。

引理 B.4。令 $K\colon [0,1]^R \to \mathbb{R},\ w \mapsto K(w)$ 为

$$K(w) = \sum_{c=(c_k)_{k=1}^R \in \{0,1\}^R} \prod_{l=1}^{R} (w_l^0)^{c_l}(1-w_l^0)^{1-c_l} \log \frac{\prod_{k=1}^R (w_k^0)^{c_k}(1-w_k^0)^{1-c_k}}{\prod_{k=1}^R (w_k)^{c_k}(1-w_k)^{1-c_k}}, \tag{92}$$

其中 w_0 位于 $[0,1]^R$ 的内部。令 $\Phi\colon\ [0,1]^R \to \mathbb{R},\ w \mapsto \Phi(w) = \|w-w_0\|^2$。存在正常数 $\alpha_1, \alpha_2 > 0$，使得 $\alpha_1\Phi(w) \leqq K(w) \leqq \alpha_2\Phi(w)$ 在 $K^{-1}(0)$ 的某个邻域内成立。因此，$K \sim \Phi$。

引理 *B.4* 的证明。在 $R=1$ 的情形下，该引理等价于引理 B.3。在 $R \geqq 2$ 的情形下，展开 $K(w)$，我们得到

$$K(w) = \sum_{c=(c_k)_{k=1}^d \in \{0,1\}^R} \left\{ \prod_{l=1}^{R} (w_l^0)^{c_l}(1-w_l^0)^{1-c_l} \times \right. \tag{93}$$

$$\left. \sum_{k=1}^{R} \left(\log (w_k^0)^{c_k}(1-w_k^0)^{1-c_k} - \log (w_k)^{c_k}(1-w_k)^{1-c_k} \right) \right\} \tag{94}$$

$$= \sum_{c\in\{0,1\}^R} \prod_{l=1}^{R} (w_l^0)^{c_l}(1-w_l^0)^{1-c_l} \sum_{k=1}^{R} \log \frac{(w_k^0)^{c_k}(1-w_k^0)^{1-c_k}}{(w_k)^{c_k}(1-w_k)^{1-c_k}} \tag{95}$$

$$= \sum_{k=1}^{R} \sum_{c\in\{0,1\}^R} \prod_{l=1}^{R} (w_l^0)^{c_l}(1-w_l^0)^{1-c_l} \log \frac{(w_k^0)^{c_k}(1-w_k^0)^{1-c_k}}{(w_k)^{c_k}(1-w_k)^{1-c_k}}. \tag{96}$$

固定任意一个 $k \in \{1,\ldots,R\}$。如果 $l \neq k$，则由第 l 个伯努利分布 $(w_l^0)^{c_l}(1-w_l^0)^{1-c_l}$ 得到的期望不会影响第 k 个对数质量比：

$$\sum_{c_l=0}^{1} (w_l^0)^{c_l}(1-w_l^0)^{1-c_l} \log \frac{(w_k^0)^{c_k}(1-w_k^0)^{1-c_k}}{(w_k)^{c_k}(1-w_k)^{1-c_k}} = \log \frac{(w_k^0)^{c_k}(1-w_k^0)^{1-c_k}}{(w_k)^{c_k}(1-w_k)^{1-c_k}}. \tag{97}$$

由此得到如下结果：

$$K(w) = \sum_{k=1}^{R} \sum_{c_k=0}^{1} (w_k^0)^{c_k}(1-w_k^0)^{1-c_k} \log \frac{(w_k^0)^{c_k}(1-w_k^0)^{1-c_k}}{(w_k)^{c_k}(1-w_k)^{1-c_k}}. \tag{98}$$

现在，令 $v(k) = (w_k, 1-w_k) \in \Delta_2$、$v^0(k) = (w_k^0, 1-w_k^0) \in \Delta_2$，以及 $C(k) = (c_k, 1-c_k) \in \mathrm{Onehot}(2)$。那么，我们有

$$(w_k)^{c_k}(1-w_k)^{1-c_k} = \prod_{j=1}^{2} (v(k))_{=j}^{(C(k))_{=j}}, \tag{99}$$

$$(w_k^0)^{c_k}(1-w_k^0)^{1-c_k} = \prod_{j=1}^{2} (v^0(k))_{=j}^{(C(k))_{=j}}. \tag{100}$$

为简洁起见，我们分别将 $(v(k))_{=j}$ 和 $(C(k))_{=j}$ 记为 $v(k)_j$ 和 $C(k)_j$。由于 $C(k) \in \mathrm{Onehot}(2)$ 且 $v(k),\ v^0(k) \in \Delta_2$，我们得到

$$K(w) = \sum_{k=1}^{R} \sum_{c_k=0}^{1} \prod_{j=1}^{2} v^0(k)_j^{C(k)_j} \log \frac{\prod_{j=1}^2 v^0(k)_j^{C(k)_j}}{\prod_{j=1}^2 v(k)_j^{C(k)_j}} \tag{101}$$

$$= \sum_{k=1}^{R} \sum_{C(k)\in \mathrm{Onehot}(2)} \prod_{j=1}^{2} v^0(k)_j^{C(k)_j} \log \frac{\prod_{j=1}^2 v^0(k)_j^{C(k)_j}}{\prod_{j=1}^2 v(k)_j^{C(k)_j}}. \tag{102}$$

Put

$$\psi(v(k)) = \sum_{C(k)\in \text{Onehot}(2)} \prod_{j=1}^{2} v^0(k)_j^{C(k)_j} \log \frac{\prod_{j=1}^{2} v^0(k)_j^{C(k)_j}}{\prod_{j=1}^{2} v(k)_j^{C(k)_j}}. \tag{103}$$

应用引理 B.3，存在 R 维向量 $\alpha^1 = (\alpha_k^1)_{k=1}^R$ 和 $\alpha^2 = (\alpha_k^2)_{k=1}^R$，其各分量均为正常数，使得

$$\alpha_k^1 (w_k - w_k^0)^2 \leqq \psi(v(k)) \leqq \alpha_k^2 (w_k - w_k^0)^2, \tag{104}$$

在 $k = 1$、.....、R 的 $\psi^{-1}(0)$ 邻域内。将它们总结起来，可得

$$\sum_{k=1}^{R} \alpha_k^1 (w_k - w_k^0)^2 \leqq \sum_{k=1}^{R} \psi(v(k)) \leqq \sum_{k=1}^{R} \alpha_k^2 (w_k - w_k^0)^2. \tag{105}$$

由于 $\Phi(w) = \sum_{k=1}^{R} (w_k - w_k^0)^2$, ，我们有

$$\min_k \{\alpha_k^1\} \Phi(w) \leqq K(w) \leqq \max_k \{\alpha_k^2\} \Phi(w). \tag{106}$$

因此， $K \sim \Phi$。 □

上述引理表明，等价的差异具有相同的实对数典范阈值，此类差异的例子包括高斯分布、类别分布和《伯努利》分布之间的 KL 散度。

这里我们证明定理 3.1。

定理 *3.1* 的证明。通过使用

$$\|y - ABx\|^2 = \langle y - ABx, y - ABx \rangle \tag{107}$$

$$= \|y\|^2 - 2\langle y, ABx \rangle + \|ABx\|^2 \tag{108}$$

以及 $y - A_0 B_0 x$、 $c - Bx$ 和 $c - B_0 x$ 的对数，我们将 $\log q_1/p_1$ 展开如下：

$$\log \frac{q_1(y, c|x)}{p_1(y, c|A, B, x)} = \log \frac{\exp\left(-\frac{1}{2}\|y - A_0 B_0 x\|^2\right) \exp\left(-\frac{\gamma}{2}\|c - B_0 x\|^2\right)}{\exp\left(-\frac{1}{2}\|y - ABx\|^2\right) \exp\left(-\frac{\gamma}{2}\|c - Bx\|^2\right)} \tag{109}$$

$$= -\frac{1}{2}(\|y\|^2 - 2\langle y, A_0 B_0 x \rangle + \|A_0 B_0 x\|^2) - \frac{\gamma}{2}(\|c\|^2 - 2\langle c, B_0 x \rangle + \|B_0 x\|^2) \tag{110}$$

$$+ \frac{1}{2}(\|y\|^2 - 2\langle y, ABx \rangle + \|ABx\|^2) + \frac{\gamma}{2}(\|c\|^2 - 2\langle c, Bx \rangle + \|Bx\|^2) \tag{111}$$

$$= \frac{1}{2}(\|ABx\|^2 - 2\langle y, (AB - A_0 B_0)x \rangle - \|A_0 B_0 x\|^2) \tag{112}$$

$$+ \frac{\gamma}{2}(\|Bx\|^2 - 2\langle c, (B - B_0)x \rangle - \|B_0 x\|^2). \tag{113}$$

对 $q_1(y, c|x)$ 取平均，我们得到

$$\iint dc dy q_1(y, c|x) \log \frac{q_1(y, c|x)}{p_1(y, c|A, B, x)} = \frac{1}{2}(\|ABx\|^2 - 2\langle A_0 B_0 x, (AB - A_0 B_0)x \rangle - \|A_0 B_0 x\|^2) \tag{114}$$

$$+ \frac{\gamma}{2}(\|Bx\|^2 - 2\langle B_0 x, (B - B_0)x \rangle - \|B_0 x\|^2) \tag{115}$$

$$= \frac{1}{2}(\|ABx\|^2 - 2\langle A_0 B_0 x, ABx \rangle + \|A_0 B_0 x\|^2) \tag{116}$$

$$+ \frac{\gamma}{2}(\|Bx\|^2 - 2\langle B_0 x, Bx \rangle + \|B_0 x\|^2) \tag{117}$$

$$= \frac{1}{2}(\|(AB - A_0 B_0)x\|^2 + \gamma\|(B - B_0)x\|^2). \tag{118}$$

令 $\Psi_1(A,B) = (1/2)\int dxq'(x)\|(AB - A_0B_0)x\|^2$，且 $\Psi_2(B) = (1/2)\int dxq'(x)\|(B - B_0)x\|^2$。根据引理 B.2，存在正常数 $c_1, c_2, c_3, c_4 > 0$，使得

$$c_1\|AB - A_0B_0\|^2 \leqq \Psi_1(A,B) \leqq c_2\|AB - A_0B_0\|^2, \tag{119}$$

$$c_3\|B - B_0\|^2 \leqq \gamma\Psi_2(B) \leqq c_4\|B - B_0\|^2. \tag{120}$$

因此，将式 (119) 与式 (120) 相加，可得

$$c_1\|AB - A_0B_0\|^2 + c_3\|B - B_0\|^2 \leqq \Psi_1(A,B) + \gamma\Psi_2(B) \leqq c_2\|AB - A_0B_0\|^2 + c_4\|B - B_0\|^2. \tag{121}$$

分别令 C_1 和 C_2 为 $\min\{c_1, c_3\}$ 和 $\max\{c_2, c_4\}$。我们立即得到

$$C_1(\|AB - A_0B_0\|^2 + \|B - B_0\|^2) \leqq \Psi_1(A,B) + \gamma\Psi_2(B) \leqq C_2(\|AB - A_0B_0\|^2 + \|B - B_0\|^2). \tag{122}$$

将引理 B.1 应用于上述不等式，可得

$$\Psi_1(A,B) + \gamma\Psi_2(B) \sim \|AB - A_0B_0\|^2 + \|B - B_0\|^2. \tag{123}$$

因此，

$$K_1(A,B) = \iiint dxdcdyq'(x)q_1(y,c|x)\log\frac{q_1(y,c|x)}{p_1(y,c|A,B,x)} \tag{124}$$

$$= \Psi_1(A,B) + \gamma\Psi_2(B) \tag{125}$$

$$\sim \|AB - A_0B_0\|^2 + \|B - B_0\|^2. \tag{126}$$

确定　　实对数典范阈值 λ_1 及其重数 m_1时，我们应考虑以下解析集

$$\mathcal{V}_1 = \{(A,B) \mid \|AB - A_0B_0\|^2 + \|B - B_0\|^2 = 0, A \in \mathrm{M}(M,K) \text{ and } B \in \mathrm{M}(K,N)\}. \tag{127}$$

取 $\|AB - A_0B_0\|^2 + \|B - B_0\|^2 = 0$。由于 $\|AB - A_0B_0\|^2 \geqq 0$ 和 $\|B - B_0\|^2 \geqq 0$ 成立，我们有

$$\|AB - A_0B_0\|^2 + \|B - B_0\|^2 = 0 \Leftrightarrow \|AB - A_0B_0\|^2 = 0 \text{ and } \|B - B_0\|^2 = 0. \tag{128}$$

因此，$AB = A_0B_0$ 和 $B = B_0$，即 $(A,B) = (A_0,B_0)$。因此，$\mathrm{V}_1 = \{(A_0,B_0)\}$。这意味着不存在奇异性，即该模型是正则的。因此，实对数典范阈值等于参数维数的一半[55]。由于参数维数等于 $(M+N)K$，我们有

$$\lambda_1 = \frac{1}{2}(M+N)K, \ m_1 = 1. \tag{129}$$

□

接下来，我们证明定理 3.2。由青柳定理 [6] 可立即推出如下结论　　s.

定理 B.1 (Aoyagi 和 Watanabe)。假设 $\mathscr{X}$ 为正定。设 λ_3 为三层线性神经网络的标准的实对数典范阈值，m_3 为其重数。利用输入维度 N、中间单元数量 H、输出维度 M和真实秩 H_0，它们可表示如下：

1. 在情形 $M + H_0 \leqq N + H$ 和 $N + H_0 \leqq M + H$ 和 $H + H_0 \leqq N + M$下，

(a) 并且如果 $N + M + H + H_0$ 是偶数，那么

$$\lambda_3 = \frac{1}{8}\{2(H+H_0)(N+M) - (N-M)^2 - (H+H_0)^2\}, \ m_3 = 1.$$

(*b*) 并且如果 $N + M + K + H + H_0$ 是奇数，那么

$$\lambda_3 = \frac{1}{8}\{2(H+H_0)(N+M) - (N-M)^2 - (H+H_0)^2 + 1\}, \ m_3 = 2.$$

2. 在 $N+H<M+H_0$ 这一情形下，那么

$$\lambda_3=\frac{1}{2}\{HN+H_0(M-H)\},\ m_3=1.$$

3. 在情形 $M+H<N+H_0$ 下，则

$$\lambda_3=\frac{1}{2}\{HM+H_0(N-H)\},\ m_3=1.$$

4. 否则（即 $N+M<H+H_0$），则

$$\lambda_3=\frac{1}{2}NM,\ m_3=1.$$

定理 *3.2* 的证明。 在多任务情形下，输出维度从 M 扩展到 $M+K$，因为多任务会使输出与概念同时出现，从而推导出解释。从数学上说，这仅涉及增加维数。因此，将 $M+K$ 代入定理 B.1 中的 M，即可得到定理 3.2。 □

我们将主要结果扩展到第 4 节中提到的情形。

定理 *4.1* 的证明。 令 $u=(u_j)_{j=1}^M=s_M(ABx)$、 $u_0=(u_j^0)_{j=1}^M=s_M(A_0B_0x)$、 $v=(v_k)_{k=1}^K=\sigma_K(Bx)$ 和 $v_0=(v_k^0)_{k=1}^K=\sigma_K(B_0x)$。

(1) 当 $i=1$ 和 $j=1$ 时，这就是定理 3.1。

(2) 当 $i=1$ 和 $j=2$ 时，我们将 $\log q_1^{12}/p_1^{12}$ 展开如下：

$$\log\frac{q_1^{12}(y,c|x)}{p_1^{12}(y,c|A,B,x)}=\log\frac{\exp\left(-\frac{1}{2}\|y-A_0B_0x\|^2\right)\left(\prod_{k=1}^K(\sigma_K(B_0x))_k^{c_k}(1-(\sigma_K(B_0x))_k)^{1-c_k}\right)^\gamma}{\exp\left(-\frac{1}{2}\|y-ABx\|^2\right)\left(\prod_{k=1}^K(\sigma_K(Bx))_k^{c_k}(1-(\sigma_K(Bx))_k)^{1-c_k}\right)^\gamma} \tag{130}$$

$$=-\frac{1}{2}(\|y\|^2-2\langle y,A_0B_0x\rangle+\|A_0B_0x\|^2)+\gamma\log\prod_{k=1}^K(v_k^0)^{c_k}(1-v_k^0)^{1-c_k} \tag{131}$$

$$+\frac{1}{2}(\|y\|^2-2\langle y,ABx\rangle+\|ABx\|^2)-\gamma\log\prod_{k=1}^K(v_k)^{c_k}(1-v_k)^{1-c_k} \tag{132}$$

$$=\frac{1}{2}(\|ABx\|^2-2\langle y,(AB-A_0B_0)x\rangle-\|A_0B_0x\|^2) \tag{133}$$

$$+\gamma\log\frac{\prod_{k=1}^K(v_k^0)^{c_k}(1-v_k^0)^{1-c_k}}{\prod_{k=1}^K(v_k)^{c_k}(1-v_k)^{1-c_k}} \tag{134}$$

通过 $dydcq_1^{12}(y,\ c|x)$ 积分，我们得到

$$\iint dydcq_1^{12}(y,c|x)\log\frac{q_1^{12}(y,c|x)}{p_1^{12}(y,c|A,B,x)} \tag{135}$$

$$=\frac{1}{2}\|(AB-A_0B_0)x\|^2+\gamma\sum_{c=(c_k)_{k=1}^K\in\{0,1\}^K}\prod_{l=1}^K(v_l^0)^{c_l}(1-v_l^0)^{1-c_l}\log\frac{\prod_{k=1}^K(v_k^0)^{c_k}(1-v_k^0)^{1-c_k}}{\prod_{k=1}^K(v_k)^{c_k}(1-v_k)^{1-c_k}}. \tag{136}$$

根据引理 B.4，第二项 $\psi(v)$ 的计算结果如下所示为 $\mathbf{k}v-v_0\mathbf{k}^2$；存在正常数 $\alpha_1,\alpha_2>0$，使得

$$\alpha_1\|v-v_0\|^2\leqq\psi(v)\leqq\alpha_2\|v-v_0\|^2, \tag{137}$$

其中

$$\psi(v)=\gamma\sum_{c=(c_k)_{k=1}^K\in\{0,1\}^K}\prod_{k=1}^K(v_k^0)^{c_k}(1-v_k^0)^{1-c_k}\log\frac{\prod_{k=1}^K(v_k^0)^{c_k}(1-v_k^0)^{1-c_k}}{\prod_{k=1}^K(v_k)^{c_k}(1-v_k)^{1-c_k}}. \tag{138}$$

令 $\beta_1=\min\{1,\ \alpha_1\}$ 和 $\beta_2=\max\{1,\ \alpha_2\}$。由于 $\mathrm{k}(AB-A_0B_0)x\mathrm{k}^2/2\geqq 0$ 成立，将其加到等式两边，得到

$$\beta_1\left(\frac{1}{2}\|(AB-A_0B_0)x\|^2+\|v-v_0\|^2\right)\leqq\frac{1}{2}\|(AB-A_0B_0)x\|^2+\psi(v) \tag{139}$$

$$\leqq\beta_2\left(\frac{1}{2}\|(AB-A_0B_0)x\|^2+\|v-v_0\|^2\right), \tag{140}$$

因此，我们应当考 虑 $\mathrm{k}(AB-A_0B_0)x\mathrm{k}^2/2+\mathrm{k}v-v_0\mathrm{k}^2$。根据 $v=\sigma_K(Bx)$ 和 $v_0=\sigma_K(B_0x)$，我们有

$$\frac{1}{2}\|(AB-A_0B_0)x\|^2+\|v-v_0\|^2=\frac{1}{2}\|(AB-A_0B_0)x\|^2+\|\sigma_K(Bx)-\sigma_K(B_0x)\|^2. \tag{141}$$

根据引理 B.2，第一项关于 $dxq'(x)$ 的平均值等价于 $\mathrm{k}AB-A_0B_0\mathrm{k}^2$。另一方面，由于 σ_K 在其像上是解析同构的，且不含参数，第二项的平均值具有与线性回归 $Bx-B_0x$ 相同的实对数典范阈值（两个模型均为正则模型），即

$$\int dxq'(x)\|\sigma_K(Bx)-\sigma_K(B_0x)\|^2\sim\int dxq'(x)\|(B-B_0)x\|^2 \tag{142}$$

$$\sim\|B-B_0\|^2. \tag{143}$$

因此，

$$\iiint dxdydcq'(x)q_1^{12}(y,c|x)\log\frac{q_1^{12}(y,c|x)}{p_1^{12}(y,c|A,B,x)}\sim\|AB-A_0B_0\|^2+\|B-B_0\|^2 \tag{144}$$

成立，并由此得到定理 3.1。

(3) 当 $i=2$ 和 $j=1$ 时，与 $(i,\ j)=(1,\ 2)$ 的情形类似，因为

$$\log\frac{q_1^{21}(y,c|x)}{p_1^{21}(y,c|A,B,x)}=\log\frac{\prod_{j=1}^M(s_M(A_0B_0x))_j^{y_j}\exp\left(-\frac{\gamma}{2}\|c-B_0x\|^2\right)}{\prod_{j=1}^M(s_M(ABx))_j^{y_j}\exp\left(-\frac{\gamma}{2}\|c-Bx\|^2\right)} \tag{145}$$

$$=\log\prod_{j=1}^M(s_M(A_0B_0x))_j^{y_j}-\frac{\gamma}{2}(\|c\|^2-2\langle c,B_0x\rangle+\|B_0x\|^2) \tag{146}$$

$$-\log\prod_{j=1}^M(s_M(ABx))_j^{y_j}+\frac{\gamma}{2}(\|c\|^2-2\langle c,Bx\rangle+\|B_0x\|^2) \tag{147}$$

$$=\log\frac{\prod_{j=1}^M(s_M(A_0B_0x))_j^{y_j}}{\prod_{j=1}^M(s_M(ABx))_j^{y_j}}+\frac{\gamma}{2}(\|Bx\|^2-2\langle c,(B-B_0)x\rangle-\|B_0x\|^2), \tag{148}$$

我们有

$$\iint dydcq_1^{12}(y,c|x)\log\frac{q_1^{12}(y,c|x)}{p_1^{12}(y,c|A,B,x)} \tag{149}$$

$$=\sum_{y\in\mathrm{Onehot}(M)}\prod_{j=1}^M(u_j^0)^{y_j}\log\frac{\prod_{j=1}^M(u_j^0)^{y_j}}{\prod_{j=1}^M(u_j)^{y_j}}+\frac{\gamma}{2}(\|(B-B_0)x\|^2, \tag{150}$$

通过使用 $u = s_M(ABx)$ 和 $u_0 = s_M(A_0B_0x)$。根据引理 B.3 和 B.2，分别由 $dxq'(x)$ 求平均的第一项和第二项，与 $\sum_{j=1}^{M-1}(u_j - u_j^0)^2$ 和 $\mathrm{k}B - B_0\mathrm{k}^2$, 的平均值具有相同的实对数典范阈值。由于映射 $(w)_{<M} \mapsto (s_M(w))_{<M}$ 是解析的，且与其像同构，我们得到

$$\int dxq'(x) \sum_{j=1}^{M-1}(u_j - u_j^0)^2 = \int dxq'(x)\|(u)_{<M} - (u_0)_{<M}\|^2 \tag{151}$$

$$\sim \int dxq'(x)\|(ABx)_{<M} - (A_0B_0x)_{<M}\|^2 \tag{152}$$

$$= \int dxq'(x)\|((A)_{<M}B - (A_0)_{<M}B_0)x\|^2 \tag{153}$$

$$\sim \|(A)_{<M}B - (A_0)_{<M}B_0\|^2. \tag{154}$$

因此，我们有

$$\iiint dxdydcq'(x)q_1^{12}(y,c|x)\log\frac{q_1^{12}(y,c|x)}{p_1^{12}(y,c|A,B,x)} \sim \|(A)_{<M}B - (A_0)_{<M}B_0\|^2 + \|B - B_0\|^2. \tag{155}$$

与定理 3.1 的证明类似，上述函数的零点集为 $((A)_{<M},\ B) = ((A_0)_{<M},\ B_0)$。由此得到：

$$\lambda_1^{12} = \frac{1}{2}(M + N - 1)K, \tag{156}$$

$$m_1^{12} = 1. \tag{157}$$

(4) 当 $i = 2$ and $j = 2$ 时，KL 散度可以按照 $(i,\ j) = (1,\ 2)$ 和 $(2,\ 1)$ 情形中的相同方式展开。因此，我们有

$$K_1^{22}(A,B) = \int dxq'(x)\left(\sum_{y\in\mathrm{Onehot}(M)}\prod_{j=1}^{M}(u_j^0)^{y_j}\log\frac{\prod_{j=1}^{M}(u_j^0)^{y_j}}{\prod_{j=1}^{M}(u_j)^{y_j}}\right. \tag{158}$$

$$\left. +\gamma\sum_{c=(c_k)_{k=1}^K\in\{0,1\}^K}\prod_{k=1}^{K}(v_k^0)^{c_k}(1-v_k^0)^{1-c_k}\log\frac{\prod_{k=1}^{K}(v_k^0)^{c_k}(1-v_k^0)^{1-c_k}}{\prod_{k=1}^{K}(v_k)^{c_k}(1-v_k)^{1-c_k}}\right), \tag{159}$$

$$u = s_M(ABx),\ u_0 = s_M(A_0B_0x),\ v = \sigma_K(Bx),\ v_0 = \sigma_K(B_0x). \tag{160}$$

与.....类似

$$\sum_{y\in\mathrm{Onehot}(M)}\prod_{j=1}^{M}(u_j^0)^{y_j}\log\frac{\prod_{j=1}^{M}(u_j^0)^{y_j}}{\prod_{j=1}^{M}(u_j)^{y_j}} \sim \|(A)_{<M}B - (A_0)_{<M}B_0\|^2 \tag{161}$$

当 $i = 2$ and $j = 1$ 且

$$\gamma\sum_{c=(c_k)_{k=1}^K\in\{0,1\}^K}\prod_{k=1}^{K}(v_k^0)^{c_k}(1-v_k^0)^{1-c_k}\log\frac{\prod_{k=1}^{K}(v_k^0)^{c_k}(1-v_k^0)^{1-c_k}}{\prod_{k=1}^{K}(v_k)^{c_k}(1-v_k)^{1-c_k}} \sim \|B - B_0\|^2 \tag{162}$$

当 $i = 1$ and $j = 2$ 时，我们得到

$$K_1^{22}(A,B) \sim \|(A)_{<M}B - (A_0)_{<M}B_0\|^2 + \|B - B_0\|^2. \tag{163}$$

当 $i = 2$ 和 $j = 1$ 时也是如此

$$\lambda_1^{22} = \frac{1}{2}(M + N - 1)K, \tag{164}$$

$$m_1^{22} = 1. \tag{165}$$

根据上文所述的 (1)、(2)、(3) 和 (4)，定理得证。 □

定理 *4.2* 的证明。令 $u=(u_j)_{j=1}^{M}=s_M((UVx)^{\mathrm{y}})$、 $u_0=(u_j^0)_{j=1}^{M}=s_M((U_0V_0x)^{\mathrm{y}})$、 $v=(v_k)_{k=1}^{K}=\sigma_K((UVx)^{\mathrm{c}})$ 和 $v_0=(v_k^0)_{k=1}^{K}=\sigma_K((U_0V_0x)^{\mathrm{c}})$。另设 $\overline{U}=(U)_{\leqq M}$、 $\overline{U_0}=(U_0)_{\leqq M}$、 $\underline{U}=(U)_{>M}$ 和 $\underline{U}=(U_0)_{>M}$。

(1) 当 $i=1$ 和 $j=1$时，这与定理 3.2 相同。

(2) 当 $i=1$ 和 $j=2$时。使用定理 4.1 的证明中 $K_1^{12}(A,B)$ 所用的相同展开方法，我们得到

$$\log\frac{q_2^{12}(y,c|x)}{p_2^{12}(y,c|U,V,x)}=\frac{1}{2}(\|(UVx)^{\mathrm{y}}\|^2-2\langle y,(UVx)^{\mathrm{y}}-(U_0V_0x)^{\mathrm{y}}\rangle-\|(U_0V_0x)^{\mathrm{y}}\|^2) \tag{166}$$

$$+\gamma\log\frac{\prod_{k=1}^{H}(v_k^0)^{c_k}(1-v_k^0)^{1-c_k}}{\prod_{k=1}^{H}(v_k)^{c_k}(1-v_k)^{1-c_k}} \tag{167}$$

$$=\frac{1}{2}(\|(UVx)_{\leqq M}\|^2-2\langle y,(UVx)_{\leqq M}-(U_0V_0x)_{\leqq M}\rangle-\|(U_0V_0x)_{\leqq M}\|^2) \tag{168}$$

$$+\gamma\log\frac{\prod_{k=1}^{H}(v_k^0)^{c_k}(1-v_k^0)^{1-c_k}}{\prod_{k=1}^{H}(v_k)^{c_k}(1-v_k)^{1-c_k}} \tag{169}$$

$$=\frac{1}{2}(\|\overline{U}Vx\|^2-2\langle y,(\overline{U}V-\overline{U_0}V_0)x\rangle-\|\overline{U_0}V_0x\|^2) \tag{170}$$

$$+\gamma\log\frac{\prod_{k=1}^{H}(v_k^0)^{c_k}(1-v_k^0)^{1-c_k}}{\prod_{k=1}^{H}(v_k)^{c_k}(1-v_k)^{1-c_k}}. \tag{171}$$

对 si 求平均　　与当 $(i,j)=$(1,2) 时定理 4.1 的证明类似，并应用引理 B.4，我们得到　　n

$$K_2^{12}(U,V)=\int dxq'(x)\left(\|(\overline{U}V-\overline{U_0}V_0)x\|^2\right. \tag{172}$$

$$\left.+\gamma\sum_{c=(c_k)_{k=1}^{d}\in\{0,1\}^H}\prod_{k=1}^{H}(v_k^0)^{c_k}(1-v_k^0)^{1-c_k}\log\frac{\prod_{k=1}^{H}(v_k^0)^{c_k}(1-v_k^0)^{1-c_k}}{\prod_{k=1}^{H}(v_k)^{c_k}(1-v_k)^{1-c_k}}\right) \tag{173}$$

$$\sim\int dxq'(x)\left(\|(\overline{U}V-\overline{U_0}V_0)x\|^2+\|\sigma_K(\underline{U}Vx)-\sigma_K(\underline{U_0}V_0x)\|^2\right). \tag{174}$$

σ_K 是解析同构映射到其像上的，这就得到了以下结果:

$$K_2^{12}(U,V)\sim\int dxq'(x)\left(\|(\overline{U}V-\overline{U_0}V_0)x\|^2+\|\sigma_K(\underline{U}Vx)-\sigma_K(\underline{U_0}V_0x)\|^2\right) \tag{175}$$

$$\sim\|\overline{U}V-\overline{U_0}V_0\|^2+\|\underline{U}V-\underline{U_0}V_0\|^2 \tag{176}$$

$$=\|UV-U_0V_0\|^2. \tag{177}$$

因此，我们只需计算 $\mathrm{k}UV-U_0V_0\mathrm{k}^2$ 的实对数典范阈值。这与定理 3.2 相同。

(3) 在 $i=2$ 和 $j=1$ 的情形下。由于 $(w)_{<M}\to s_M(w)$， $w\in\mathbb{R}^{M-1}$ 解析同构地映射到其像上，采用与 $i=1$ 和 $j=2$ 的情形相同的方法，我们得到

$$K_2^{21}(U,V)\sim\int dxq'(x)\left(\|s_M(\overline{U}Vx)-s_M(\overline{U_0}V_0x)\|^2+\|(\underline{U}V-\underline{U_0}V_0)x\|^2\right) \tag{178}$$

$$\sim\|(U)_{<M}V-(U_0)_{<M}V_0\|^2+\|\underline{U}V-\underline{U_0}V_0\|^2 \tag{179}$$

$$=\|(U)_{<M}V-(U_0)_{<M}V_0\|^2+\|(U)_{>M}V-(U_0)_{>M}V_0\|^2 \tag{180}$$

$$=\mathrm{k}(U)_{\neq M}V-(U_0)_{\neq M}V_0\mathrm{k}^2. \tag{181}$$

$(U)_{\neq M}\in\mathrm{M}(M+K-1,\ H)$ 和 $(U_0)_{\neq M}\in\mathrm{M}(M+K-1,\ H_0)$ 成立。因此，$\mathrm{k}(U)_{\neq M}V-(U_0)_{\neq M}V_0\mathrm{k}^2$ 的实对数典范阈值可通过在定理 3.2 中将 $M-1$ 赋给 M 来计算。由此得到:

$$\lambda_2^{21}=\lambda_2(N,H,M-1,K,H_0), \tag{182}$$

$$m_2^{21}=m_2(N,H,M-1,K,H_0). \tag{183}$$

(4) 在 $i=2$ 和 $j=2$ 的情形下，结合上述方法，我们得到

$$K_2^{22}(U,V) \sim \int dxq'(x)\left(\|s_M(\overline{U}Vx) - s_M(\overline{U_0}V_0x)\|^2 + \|\sigma_K(\underline{U}Vx) - \sigma_K(\underline{U_0}V_0x)\|^2\right) \tag{184}$$

$$\sim \|(U)_{<M}V - (U_0)_{<M}V_0\|^2 + \|\underline{U}V - \underline{U_0}V_0\|^2 \tag{185}$$

$$=\mathrm{k}(U)_{\neq M}V - (U_0)_{\neq M}V_0\mathrm{k}^2. \tag{186}$$

这就得到了 $i=2$ and $j=1$ 时的情形。因此，我们得到

$$\lambda_2^{22} = \lambda_2(N,H,M-1,K,H_0), \tag{187}$$

$$m_2^{22} = m_2(N,H,M-1,K,H_0). \tag{188}$$

因此，根据 (1)、(2)、(3) 和 (4)，定理得证。 □

最后，我们证明命题 5.1。

命题 *5.1* 的证明。对每种情形构造 $\lambda_1-\lambda_2$ and 并求解 $\lambda_1>\lambda_2$。如果这些情形都得到解决，则可立即推出相反的情形 $\lambda_1 \leqq \lambda_2$。

(1) 在 $M+K+H_0 \leqq N+H$ and $N+H_0 \leqq M+K+H$ and $H+H_0 \leqq N+M+K$ 的情形下。当 $N+M+K+H+H_0$ 为偶数时，我们有

$$\lambda_1-\lambda_2 = \frac{1}{8}\{4MK+4NK-2(H+H_0)(N+M+K)+(N-M-K)^2+(H+H_0)^2\} \tag{189}$$

$$=\{\frac{1}{8}K^2+2(M+N-H-H_0)K+N^2+M^2+H^2+H_0^2 \tag{190}$$

$$-2HN-2H_0N-2MN-2HM-2H_0M+2HH_0\} \tag{191}$$

$$=\frac{1}{8}\{K^2+2(M+N-H-H_0)K+(M+N-H-H_0)^2-4MN\} \tag{192}$$

$$=\{\frac{1}{8}(K+M+N-H-H_0)^2-4MN\}. \tag{193}$$

上述方程最右侧是关于 K 的二次函数，且其最高次项的系数为正。根据这一假设，$H+H_0 \leqq M+N+K$，即 $K \geqq H+H_0-M-N$ 成立。因此，解 $\lambda_1>\lambda_2$，可得

$$K > H+H_0-M-N+2\sqrt{MN} \tag{194}$$

$$= H+H_0-(\sqrt{M}-\sqrt{N})^2. \tag{195}$$

反向结论也可按如下方式验证：

$$\lambda_1>\lambda_2 \Leftrightarrow K > H+H_0-M-N+(\sqrt{M}-\sqrt{N})^2. \tag{196}$$

当 $N+M+K+H+H_0$ 为偶数时，按照上述相同步骤，可得

$$\lambda_1-\lambda_2 = \frac{1}{8}\{(K+M+N-H-H_0)^2-4MN-1\}. \tag{197}$$

and

$$\lambda_1>\lambda_2 \Leftrightarrow K > H+H_0-M-N+\sqrt{4MN+1}. \tag{198}$$

(2) 在 $N+H<M+K+H_0$的情形下，我们有

$$\lambda_1-\lambda_2 = \frac{1}{2}\{MK+NK-HN-H_0(M+K-H)\} \tag{199}$$

$$= \frac{1}{2}\{(M+N-H_0)K-H(N-H_0)-MH_0\}. \tag{200}$$

因此，我们得到

$$\lambda_1 > \lambda_2 \Leftrightarrow (M + N - H_0)K > H(N - H_0) + MH_0. \tag{201}$$

(3) 当 $M + K + H < N + H_0$ 的情形下，我们有

$$\lambda_1 - \lambda_2 = \frac{1}{2}\{MK + NK - H(M + K) - H_0(N - H)\} \tag{202}$$

$$= \frac{1}{2}\{(M + N - H)K - H_0(N - H) - MH\}. \tag{203}$$

因此，我们得到

$$\lambda_1 > \lambda_2 \Leftrightarrow (M + N - H)K > H_0(N - H) + MH. \tag{204}$$

(4) 当 $N + M + K < H + H_0$ 的情形下，我们有

$$\lambda_1 - \lambda_2 = \frac{1}{2}(MK + NK - NM - NK) \tag{205}$$

$$= \frac{1}{2}M(K - N). \tag{206}$$

因此，我们得到

$$\lambda_1 > \lambda_2 \Leftrightarrow K > N. \tag{207}$$

因此，根据 (1)、(2)、(3) 和 (4)，可以说该定理成立。 □

参考文献

[1] Hirotogu Akaike. 信息论与最大似然原理的扩展。载于第二届国际信息论研讨会论文集，第1卷，第267–281页，1973年。[2] Hirotogu Akaike. 信息论与最大似然原理的扩展。载于赤池弘次论文选集，第199–213页。施普林格，1998年。[3] Hirotogu Akaike. 重新审视统计模型识别。*IEEE*自动控制汇刊，19(6):716–723，1974年。[4] 青柳美纪。贝叶斯估计中受限玻尔兹曼机的随机复杂度与泛化误差。机器学习研究期刊，11(Apr):1243–1272，2010年。[5] 青柳美纪。受限玻尔兹曼机贝叶斯估计中的学习系数。代数统计学杂志，4(1):30–57，2013年。[6] 青柳美纪和Sumio Watanabe。降秩回归贝叶斯估计中的随机复杂度。《神经网络》，18(7):924–933，2005年。[7] 迈克尔·弗朗西斯·阿蒂亚。奇点消解与广义函数的除法。纯粹与应用数学通讯，23(2):145–150，1970年。[8] 约瑟夫·伯恩斯坦。广义函数关于参数的解析延拓。泛函分析及其应用，6(4):26–40，1972年。[9] 阿里·T·杰姆吉尔。非负矩阵分解模型中的贝叶斯推断。计算智能与神经科学，2009(4):17，2009年。文章编号785152。

[10] Irene Y. Chen、Emma Pierson、Sherri Rose、Shalmali Joshi、Kadija Ferryman和Marzyeh Ghassemi。医疗保健领域的伦理机器学习。《生物医学数据科学年度评论》，4(1):123–144，2021年。[11] Jeff Donahue、Yangqing Jia、Oriol Vinyals、Judy Hoffman、Ning Zhang、Eric Tzeng和Trevor Darrell。Decaf：用于通用视觉识别的深度卷积激活特征。载于国际机器学习会议，第647–655页。PMLR，2014年。[12] Shi Dong、Ping Wang和Khushnood Abbas。深度学习及其应用综述。《计算机科学评论》，40:100379，2021年。[13] Mathias Drton。似然比检验与奇异性。《统计年刊》，37(2):979–1012，2009年。[14] Mathias Drton、Shaowei Lin、Luca Weihs、Piotr Zwiernik等。高斯潜在树模型和森林模型的边际似然与模型选择。《伯努利》，23(2):1202–1232，2017年。[15] Mathias Drton和Martyn Plummer。奇异模型的贝叶斯信息准则。《皇家统计学会杂志*B*辑》，79:323–380，2017年。附讨论。[16] Mathias Drton、Bernd Sturmfels和Seth Sullivant。代数统计学讲义，第39卷。施普林格科学与商业媒体，2008年。[17] Ian Goodfellow、Yoshua Bengio和Aaron Courville。《深度学习》。麻省理工学院出版社，2016年。[18] Naoki Hayashi。非负矩阵分解中的变分近似误差。《神经网络》，126:65–75，2020年。[19] Naoki Hayashi。潜在狄利克雷分配中贝叶斯泛化误差的精确渐近形式。《神经网络》，137:127–137，2021年。[20] Naoki Hayashi。参数受限奇异模型的统计学习理论。博士论文，东京工业大学，2021年。[21] Naoki Hayashi和Sumio Watanabe。非负矩阵分解实对数典范阈值的更紧上界及其在贝叶斯推断中的应用。载于*IEEE*计算智能研讨会系列（*IEEE SSCI*），第718–725页，2017年11月。[22] Naoki Hayashi和Sumio Watanabe。非负矩阵分解中贝叶斯泛化误差的上界。*Neurocomputing*，266C(11月29日):21–28，2017年。[23] Naoki Hayashi和Sumio Watanabe。潜在狄利克雷分配与随机矩阵分解中的贝叶斯泛化误差渐近形式。《*SN*计算机科学》，1(2):1–22，2020年。[24] Heisuke Hironaka。特征为零的域上代数簇的奇点消解。《数学年刊》，79:109–326，1964年。[25] Brian Hu、Bhavan Vasu和Anthony Hoogs。X-mir：可解释医学图像检索。载于*IEEE/CVF*冬季计算机视觉应用会议论文集，第440–450页，2022年。[26] Toru Imai。实对数典范阈值的估计。*arXiv*预印本 *arXiv:1906.01341*，2019年。[27] Ugne Klimiene、Ričards Marcinkevičs、Patricia Reis Wolfertstetter、Ece Özkan Elsen、Alyssia Paschke、David Niederberger、Sven Wellmann、Christian Knorr和Julia E Vogt。应用于儿童阑尾炎诊断的多视图概念瓶颈模型。载于第二届医疗保健可解释机器学习研讨会（*IMLH*），第1–15页。苏黎世联邦理工学院机器学习研究所，2022年。[28] Pang Wei Koh、Thao Nguyen、Yew Siang Tang、Stephen Mussmann、Emma Pierson、Been Kim和Percy Liang。概念瓶颈模型。载于Hal Daumé III和Aarti Singh编，第*37*届国际机器学习会议论文集，《机器学习研究论文集》第119卷，第5338–5348页。PMLR，2020年7月13–18日。

[29] Masahiro Kohjima、Tatsushi Matsubayashi 和 Hiroshi Sawada。概率非负不一致分辨率矩阵分解。载于*CIKM '15* 会议论文集：第*24*届 *ACM* 国际信息与知识管理会议，第1卷，第1855–1858页，2015年。[30] Masahiro Kohjima、Tatsushi Matsubayashi 和 Hiroshi Sawada。多重数据分析与非负矩阵／张量分解 [i]：多重数据分析及其进展。电子信息通信学会（*IEICE*）学报，99(6):543–550，2016年。日文。[31] Neeraj Kumar、Alexander C Berg、Peter N Belhumeur 和 Shree K Nayar。用于人脸验证的属性和明喻分类器。载于*2009*年 *IEEE* 第*12*届国际计算机视觉会议，第365–372页。IEEE，2009年。[32] Christoph H Lampert、Hannes Nickisch 和 Stefan Harmeling。通过类间属性迁移学习检测未见物体类别。载于*2009*年 *IEEE*计算机视觉与模式识别会议，第951–958页。IEEE，2009年。[33] Daniel D. Lee 和 H. Sebastian Seung。利用非负矩阵分解学习物体的组成部分。《自然》，401:788–791，1999年。[34] Joshua Lockhart、Nicolas Marchesotti、Daniele Magazzeni 和 Manuela Veloso。迈向利用概念瓶颈模型学习解释：缓解信息泄漏。载于与 *ICLR 2022*同期举行的社会责任机器学习（*SRML*）研讨会，第1卷，第1–6页，2022年。[35] Anita Mahinpei、Justin Clark、Isaac Lage、Finale Doshi-Velez 和 Weiwei Pan。黑箱概念学习模型的前景与陷阱。载于国际机器学习会议论文集：可解释人工智能理论基础、批判与应用趋势研讨会，第1卷，第1–13页，2021年。[36] Andrei Margeloiu、Matthew Ashman、Umang Bhatt、Yanzhi Chen、Mateja Jamnik 和 Adrian Weller。概念瓶颈模型是否按预期学习？载于表示学习会议论文集：负责任人工智能研讨会，第1卷，第1–8页，2021年5月。[37] Ken Matsuda 和 Sumio Watanabe。加权膨胀及其在多项分布混合中的应用。*IEICE Transactions*，J86-A(3):278–287，2003年。日文。[38] Richard McElreath。《统计学新思维：结合 *R* 和 *Stan* 示例的贝叶斯课程》。CRC出版社，第2版，2020年。[39] Thomas McGrath、Andrei Kapishnikov、Nenad Tomašev、Adam Pearce、Martin Wattenberg、Demis Hassabis、Been Kim、Ulrich Paquet 和 Vladimir Kramnik。在 AlphaZero 中习得国际象棋知识。《美国国家科学院院刊》，119(47):e2206625119，2022年。[40] Christoph Molnar。可解释机器学习。Lulu.com，2020年。[41] Kenji Nagata 和 Sumio Watanabe。交换蒙特卡洛方法中交换比率的渐近行为。《神经网络》，21(7):980–988，2008年。[42] Takuto Naito 和 Keisuke Yamazaki。线性动力系统的渐近边际似然。*IEICE*信息与系统汇刊，97(4):884–892，2014年。[43]Pentti Paatero 和 Unto Tapper。正矩阵分解：一种非负因子模型，能够最佳利用数据值的误差估计。《环境计量学》，5(2):111–126，1994年。doi:10.1002/env.3170050203。[44] Rui Qian、Shuangrui Ding、Xian Liu 和 Dahua Lin。用于自监督视频表征学习的静态与动态概念。载于欧洲计算机视觉会议，第145–164页。施普林格，2022年。

[45] Aniruddh Raghu、John Guttag、Katherine Young、Eugene Pomerantsev、Adrian V. Dalca 和 Collin M. Stultz。借助支持性证据进行预测：临床风险预测中的应用。载于健康、推理与学习会议论文集，第95—-104页，纽约，纽约州，美国，2021年。计算机协会。[46] Dmitry Rusakov 和 Dan Geiger。朴素贝叶斯网络的渐近模型选择。机器学习研究期刊，6(1月):1–35，2005年。[47] Kenichiro Sato 和 Sumio Watanabe。泊松混合的贝叶斯泛化误差与单纯形范德蒙矩阵型奇点。*arXiv* 预印本 *arXiv:1912.13289*，2019年。[48] Mikio Sato 和 Takuro Shintani。关于与预齐次向量空间相关的ζ函数。《数学年刊》，第131–170页，1974年。[49] Yoshihide Sawada 和 Keigo Nakamura。加入额外无监督概念的概念瓶颈模型。*IEEE Access*，10:41758–41765，2022年。[50] Gideon Schwarz。模型维数的估计。《统计年刊》，6(2):461–464，1978年。[51] Ali Sharif Razavian、Hossein Azizpour、Josephine Sullivan 和 Stefan Carlsson。开箱即用的 CNN 特征：出色的识别基线。载于*IEEE*计算机视觉与模式识别会议研讨会论文集，第806–813页，2014年。[52] Raiki Tanaka 和 Sumio Watanabe。采用 Swish激活函数的三层神经网络的实对数典范阈值。电子信息通信学会技术报告；*IEICE* 技术报告，119(360):9–15，2020年。[53] Sumio Watanabe。非正则学习机的代数分析。神经信息处理系统进展，12:356–362，2000年。美国丹佛。[54] Sumio Watanabe。层次学习机的代数几何方法。《神经网络》，13(4):1049–1060，2001年。[55] Sumio Watanabe。代数几何与统计学习理论。剑桥大学出版社，2009年。[56] Sumio Watanabe。奇异学习理论中贝叶斯交叉验证与广泛适用信息准则的渐近等价性。机器学习研究期刊，11(12月):3571–3594，2010年。[57] Sumio Watanabe。一种广泛适用的贝叶斯信息准则。机器学习研究期刊，14(3月):867–897，2013年。[58] Sumio Watanabe。贝叶斯统计的数学理论。CRC出版社，2018年。[59] Sumio Watanabe。面向未知信息源的贝叶斯统计数学理论。《皇家学会哲学汇刊 *A*辑》，第1–26页，2023年。即将发表。[60] Takumi Watanabe 和 Sumio Watanabe。多项分布混合中贝叶斯泛化误差的渐近行为。*arXiv* 预印本 *arXiv:2203.06884*，2022年。[61] Yiran Xu、Xiaoyin Yang、Lihang Gong、Hsuan-Chu Lin、Tz-Ying Wu、Yunsheng Li 和 Nuno Vasconcelos。面向自动驾驶车辆的可解释物体诱导动作决策。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第9523–9532页，2020年。[62] Keisuke Yamazaki。奇异学习机中随机复杂度的渐近展开。博士论文，东京工业大学，2003年。[63] Keisuke Yamazaki 和 Sumio Watanabe。混合模型中的奇异性与随机复杂度上界。《神经网络》，16(7):1029–1038，2003年。

[64] Keisuke Yamazaki 和 Sumio Watanabe。贝叶斯网络的随机复杂度。载于人工智能中的不确定性（*UAI'03*），第 592–599 页，2003年。[65] Keisuke Yamazaki 和 Sumio Watanabe。二项分布混合中的牛顿图与随机复杂度。载于国际算法学习理论会议，第 350–364 页。施普林格，2004年。[66]Keisuke Yamazaki 和 Sumio Watanabe。隐马尔可夫模型的代数几何与随机复杂度。*Neurocomputing*，69：62–84，2005年。第 1–3 期。[67] Keisuke Yamazaki 和 Sumio Watanabe。完全二部图型玻尔兹曼机中的奇异性及随机复杂度的上界。*IEEE*神经网络汇刊，16：312–324，2005年。第 2 期。[68] Jason Yosinski、Jeff Clune、Yoshua Bengio 和 Hod Lipson。深度神经网络中的特征可迁移性如何？载于 Z. Ghahramani、M. Welling、C. Cortes、N. Lawrence 和 K.Q. Weinberger 编，《style id='14'>神经信息处理系统进展》，第 27 卷，第 1–9 页。Curran Associates, Inc.，2014年。[69] Piotr Zwiernik。一般马尔可夫模型边际似然的渐近行为。机器学习研究期刊，12（11月）：3283–3310，2011年。