Submitted to Statistical Science

Evidence bounds in singular models: probabilistic and variational perspectives

Anirban Bhattacharya, Debdeep Pati, Sean Plummer, and Yun Yang

Abstract. In Bayesian statistics, the marginal likelihood, a.k.a. the evidence, contains an intrinsic penalty accounting for larger model sizes and is of fundamental importance in Bayesian model comparison. Over the past two decades, there has been steadily increasing activity to understand the nature of this penalty in singular statistical models, building on pioneering works by Sumio Watanabe. Unlike regular models where the Bayesian information criterion (BIC) encapsulates a first-order expansion of the evidence, parameter counting gets trickier in singular models where a quantity called the real log-canonical threshold (RLCT) summarizes the effective model dimensionality. In this article, we offer a probabilistic treatment to recover non-asymptotic versions of established evidence bounds as well as prove a new result based on the Gibbs variational inequality. In particular, we show that mean-field variational inference correctly recovers the RLCT for any singular model in its standard form. We additionally exhibit sharpness of our bound empirically in dimension $d = 2$ and provide two conjectures concerning the asymptotics of the mean-field ELBO for singular models in standard form.

Key words and phrases: Bayesian model selection, Coordinate ascent, Gibbs variational inequality, Laplace approximation, Mean-field approximation, Real log-canonical threshold.

1. INTRODUCTION

The marginal likelihood (a.k.a. evidence) is a fundamental object in Bayesian model comparison [32], which encapsulates an intrinsic penalty for model complexity, and can be readily used to compare models with different parameter dimensions. However, barring conjugate settings, the marginal likelihood is rarely available in closed-form, necessitating approximate methods. A classical approach is to make analytic approximations, of which the Laplace approximation [34, 36, 22] is the most prominent. Applied to *regular* parametric models — the data generating distribution $q(x)$ is said to be *regular* for the model $\{p(x \mid \xi)\}_{\xi\in\Omega}$ if the minimum locus of the average log loss function $L(\xi) := -\int q(x) \log p(x \mid \xi) dx$ is a singleton $\{\xi_0\}$ and there exists an open neighborhood $U \subset \Omega$ containing ξ_0 such that the Hessian matrix $\nabla^2 L(\xi_0)$ is positive definite on U — the Laplace approximation yields an asymptotic expansion of the log-marginal likelihood as $\ell_n(\widehat{\xi}_n) - d/2 \cdot \log n + R_n$, where $\ell_n(\widehat{\xi}_n)$ is the log-likelihood evaluated at the maximum likelihood estimate $\widehat{\xi}_n$ based on the data, d is the parameter dimension, and the remainder term R_n is bounded in magnitude with high probability by a constant.[1]

In this article, our focus will be on *singular* statistical models. The data generating distribution $q(x)$ is said to be *singular* for the model $\{p(x \mid \xi)\}_{\xi\in\Omega}$ if either the minimum locus of $L(\xi)$ contains more than one point or there exists points in the minimum locus for which the Hessian matrix $\nabla^2 L$ fails to be positive definite. Singular models are commonly encountered in many modern applications such as artificial intelligence, robotics, and bioinformatics [49]. Examples of singular models include mixture models, factor models, hidden Markov models, latent class analysis, Bayesian networks, reduced rank regression, and neural networks; see [14]

Anirban Bhattacharya is Professor, Department of Statistics, Texas A&M University, College Station, TX, 77843, USA (e-mail: anirbanb@stat.tamu.edu). Debdeep Pati is Professor, Department of Statistics, University of Wisconson-Madison, Madison, WI, 53706, USA (e-mail: dpati2@wisc.edu). Sean Plummer is an Assistant Professor, Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR, 72701, USA (e-mail: seanp@uark.edu). Yun Yang is an Associate Professor, Department of Mathematics, University of Maryland, MD, 20742, USA (e-mail: yy84@umd.edu).

[1] See [22] for the necessary technical assumptions for the Laplace approximation.

for a more comprehensive list. As a simple concrete illustration, suppose the true data generating distribution is given by the standard matrix Gaussian distribution $q(X) = (2\pi)^{-2}\exp\{-\|X\|^2/2\}$ and consider a Gaussian matrix factorization model for the 2×2 real-valued matrix X, $p(X \mid A, B) = (2\pi)^{-2}\exp\{-\|X - AB\|^2/2\}$ with $A = (a, b)^{\mathrm{T}}$, $B = (c, d)$, and $(a, b, c, d) \in \mathbb{R}^4$. The map $(a, b, c, d) \mapsto p(\cdot \mid a, b, c, d)$ is clearly not one-to-one; the entire region $\Omega_0 := \{(a, b, c, d) \in \mathbb{R}^4 \,|\, (a, b) = (0, 0) \text{ or } (c, d) = (0, 0)\}$ inside the parameter space maps to the true data generating distribution $q(\cdot)$. In statistical terms, the Fisher information matrix is not positive definite on Ω_0.

While it may be apparent that developing a statistical theory for singular models is important, one is unable to utilize many of the existing tools from classical statistical theory which rely on the regularity assumptions. For example, the derivation of the Laplace approximation proceeds by localizing the log-marginal likelihood to a neighborhood of the maximum likelihood estimate (or the posterior mode) and subsequently applying a second-order Taylor series expansion of the log-likelihood around $\widehat{\theta}_n$ to reduce the marginal likelihood to a Gaussian integral. It should perhaps then be intuitive that this approximation will face difficulties for singular models where the Hessian matrix can be singular. This is indeed the case and can be verified via simulation in a straightforward manner; see, e.g., the instructive Example 1 of [14]. Hence, in general, the usual Laplace approximation no longer provides a correct approximation to the log-marginal likelihood.

The foundational groundwork for a general theory of singular models has been laid in a series of seminal contributions by Watanabe [46, 47, 48], with much of the subsequent development condensed into book-level treatments in [50, 52]. Watanabe shows that for a singular model which satisfies some mild technical conditions, the asymptotic behavior of the log-marginal likelihood can be characterized through

$$\ell_n(\xi^\star) - \lambda \log n + (m-1)\log(\log n) + R_n, \tag{1}$$

assuming the data is generated from $P^\star \equiv p(\cdot \mid \xi^\star)$, with the stochastic error term $R_n = O_{P^\star}(1)$. The quantity $\lambda \in (0, d/2]$ is called the *real log-canonical threshold* (RLCT) and the integer $m \geq 1$ its *multiplicity*. For a regular statistical model, we have $(\lambda, m) = (d/2, 1)$ and the expansion (1) reduces to the usual Laplace approximation. For more on model selection in singular settings, we refer the reader to [51, 14].

REMARK 1.1. The leading order term of Watanabe's expansion for the log-marginal likelihood is evaluated at the true parameter $\xi^\star$, while the leading order term for the BIC is the maximum likelihood estimator (MLE). Watanabe's theory of singular models is also capable of analyzing the behavior of MLEs if the set of parameters Ω is compact. If the set of parameters Ω is not compact, then an MLE may not exist. Watanabe shows that under similar conditions to Eq. 1, the log-marginal likelihood at the MLE can be asymptotically expanded about the true parameter $\xi^\star$,

$$\ell_n(\hat{\xi}_n) = \ell_n(\xi^\star) - \frac{1}{4n}\max\{0, W_n(\xi^\star)\}^2 + o_p\left(\frac{1}{n}\right).$$

where $W_n(\cdot)$ is a Gaussian process constructed from the statistical model; see Eq. (7). The details for this result can be found in section 6.4 in [52]. From the above expansion, we see that the log-marginal likelihood is also given by

$$\ell_n(\hat{\xi}_n) - \lambda \log n + (m-1)\log(\log n) + O_p(1).$$

See remark 1.17 in [50] and remark 50 in [52] for more information on why maximum likelihood estimation may not be appropriate for singular models.

Watanabe's original derivation of Eq. (1) is an asymptotic approximation to the marginal likelihood based on several deep results in modern mathematics which are less commonly known among the statistical community. In this article, we revisit the general problem of computing the evidence of a singular model. Our primary motivation behind this work is to provide a non-technical overview of Watanbe's original derivation of Eq. (1) and to explore the possibility of deriving Eq. (1) exclusively using probabilistic arguments, such as stochastic ordering and conditioning, readily accessible to the wider statistics and machine learning audience. As a by-product of the probabilistic treatment, all our results are non-asymptotic in nature. We carry out this program in § 3. We follow standard practice to first analyze a deterministic version of the problem, replacing the log-likelihood ratio with its expectation under the data generating model, and then proceed to handle the stochastic component. Interestingly, the RLCT and its multiplicity appear as the rate and shape parameters of a certain Gamma distribution in our analysis.

Analytic approximations are not the only tool which we can use to approximate the log-marginal likelihood. Variational approaches [25, 9, 39] have increasingly grown in popularity in Bayesian statistics as a different set of probabilistic tools to approximate the evidence. Variational Bayes (VB) aims to find the best approximation to the posterior distribution from a class of tractable probability distributions, with the approximation error most commonly measured in terms of Kullback–Leibler divergence; see [10] for an excellent recent survey of variational inference.

For the posterior distribution $\Pi(\xi \mid X)$ of a model $P(X \mid \xi)$ with prior $\varphi(\xi)$ and any probability measure ρ

on Ω with $\rho \ll \varphi$, the following well-known identity is easy to establish,

$$D\big(\rho \,\|\, \Pi(\cdot \mid X)\big) = \log P(X) + \left[- \int_\Omega \log P(X \mid \xi)\, \rho(\xi)\, d\xi + D(\rho \,\|\, \varphi) \right], \tag{2}$$

where $D(\mu \,\|\, \nu) := E_\mu(\log d\mu/d\nu)$ is the Kullback–Leibler divergence between μ and ν. An immediate upshot of this is the Gibb's variational inequality, which states that for any probability density $\rho \ll \varphi$ on Ω,

$$\log P(X) \geq \int \log P(X \mid \xi)\, \rho(\xi)\, d\xi - D(\rho \,\|\, \varphi), \tag{3}$$

with equality attained if and only if $\rho = \Pi(\xi \mid X)$. The Gibb's variational inequality is central to a variational approximation to the normalizing constant $P(X)$. The quantity in the right hand side of (3) is a lower bound to $\log P(X)$ for any $\rho \ll \varphi$. A variational lower bound to $\log P(X)$ is then obtained by optimizing the variational parameter ρ over a family of probability densities $\mathcal{F}$ on Ω,

$$\log P(X) \geq \mathrm{ELBO}(\mathcal{F}) := \sup_{\rho \in \mathcal{F}} \left\{ \int \log P(X \mid \xi)\, \rho(\xi)\, d\xi - D(\rho \,\|\, \varphi) \right\} \tag{4}$$

The notation ELBO here abbreviates *Evidence Lower Bound*, which is commonly used to designate the variational lower bound in Bayesian statistics. If the supremum in (4) is attained at some $\rho^\star \in \mathcal{F}$, the density $\rho^\star$ is called the optimal variational approximation. It follows from equation (2) that $\rho^\star$ is a best approximation to $\Pi(\xi \mid X)$ in terms of KL divergence from the class $\mathcal{F}$, i.e., $D\big(\rho^\star \,\|\, \Pi(\cdot \mid X)\big) = \inf_{\rho \in \mathcal{F}} D\big(\rho \,\|\, \Pi(\cdot \mid X)\big)$. The choice of the family $\mathcal{F}$ typically aims to balance computational tractability and expressiveness. A popular example is the mean-field family,

$$\mathcal{F}_{\text{MF}} := \big\{ \rho = \rho_1 \otimes \ldots \otimes \rho_d \,:\, \rho \ll \varphi \text{ a prob. measure on } \Omega \big\}, \tag{5}$$

where ρ is assumed to be a product-measure, with no further restriction on the constituent arms.

The statistical properties of VB have been studied for some specific examples of singular models such as various mixture models [42, 43, 44, 45], Bayesian graphical models [20, 21, 41, 28], and neural networks [30]. The primary focus of these works is deriving asymptotic, large n, bounds for the variational stochastic complexity, the negative ELBO, similar to the asymptotic expansion in Eq. (1). These asymptotic bounds allow us to measure the performance of mean-field variational inference in singular models as the gap between the two expansions bounds KL divergence between the optimal variational approximation and the posterior distribution.

In § 4 we show that mean-field variational inference correctly recovers the RLCT for models in standard form, even though the posterior distribution itself has strong dependence and is far from a product structure (see Figure 1 for a representative example). Furthermore, our analysis shows that this bound is sharp, i.e., the mean-field ELBO is not capable of recovering the $\log \log n$ term in (1). Our findings are supported by numerical computations.

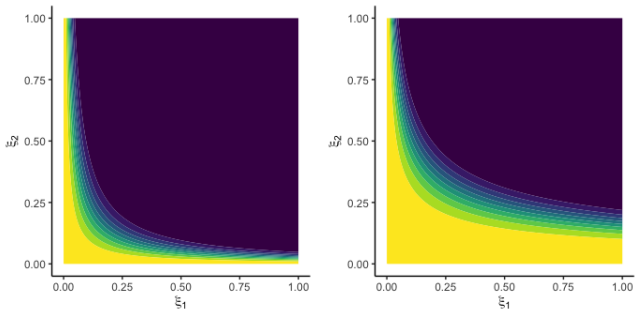

FIG 1. Contour plots of 2d target density proportional to $\exp(-nK(\xi_1, \xi_2))$ on $[0,1]^2$, with $n = 100$ and $K(\xi_1, \xi_2) = \xi_1^2 \xi_2^2$ (top) and $K(\xi_1, \xi_2) = \xi_1^2 \xi_2^4$ (bottom); see §2 for the relevance of such densities to singular models. The darker regions (red) represent smaller values close to 0 and the lighter regions (yellow) represent values close to 1. Notice that the shape of these distributions differ significantly from both the elliptical shape defined by the contour of a normal distribution or the rectangular shape of a typical product distribution.

2. A REVIEW OF SINGULAR MODEL THEORY

Our goal in this section is to provide a non-technical overview of Watanabe's original derivation of Eq. (1). In order to maintain a non-technical discussion of this derivation, many key results will be provided in a heuristic form together with motivation for the need and use of these tools as they arise in each step of the derivation. We will begin the section by introducing notation. We proceed by introducing three key tools, (a) Hironaka's resolution of singularities (page 3), (b) the standard form of the posterior (page 5), and (c) the state density function (page 5), that are needed for Watanabe's derivation. After introducing these tools we will sketch Watanabe's derivation of Eq. (1). We also refer the reader to Shaowei Lin's thesis [24] and the background section of [14] for lucid summaries of this theory.

We begin with introducing some notation. Let $X^{(n)} = (X_1, \ldots, X_n)^{\mathrm{T}}$ denote n independent and identically distributed observations from a probability density function q. A Bayesian analysis in this setting proceeds by setting up (i) a statistical model consisting of a family of probability distributions $\{p(\cdot \mid \xi) : \xi \in \Omega\}$ for the individual observations, indexed by a parameter ξ taking values in the

compact parameter space $\Omega \subseteq \mathbb{R}^d$, and (ii) a prior (probability) distribution $\varphi(\cdot)$ on Ω. The posterior distribution is given by

$$\Pi(\xi \mid X^{(n)}) = \frac{e^{\ell_n(\xi)}\varphi(\xi)}{m(X^{(n)})}, \tag{6}$$

$$\ell_n(\xi) := \sum_{i=1}^{n} \log p(X_i \mid \xi),\ m(X^{(n)}) = \int_\Omega e^{\ell_n(\xi)}\varphi(\xi)d\xi,$$

with $\ell_n(\xi)$ the log-likelihood function, and $m(X^{(n)})$ the marginal likelihood or evidence.

REMARK 2.1. The requirement that the parameter space Ω be compact is to reduce additional technical assumptions on the model. For example, without compactness one would need further tail conditions on the log-likelihood ratio of the model [52, CH 3] and other additional technical assumptions in order to establish convergence of the empirical log-likelihood ratio [52, CH 10].

To simplify the following discussion, we will assume that the true data generating distribution is *realizable*, i.e. $q(x) = p(x \mid \xi^\star)$ for some $\xi^\star \in \Omega$. We call $\xi^\star$ the true data generating parameter, and reserve the notations $\mathbb{E}^\star$ and $\mathbb{P}^\star$ to respectively denote expectation and probability under (the n-fold product of) $p(\cdot \mid \xi^\star)$. Let $K_n(\xi) = n^{-1}\left[\ell_n(\xi^\star) - \ell_n(\xi)\right]$ be the negative log-likelihood ratio scaled by a factor of n^{-1}, so that its $\mathbb{E}^\star$-expectation is the Kullback–Leibler divergence, $K(\xi) := \mathbb{E}^\star\big[K_n(\xi)\big] = D\big[p(\cdot \mid \xi^\star) \,\|\, p(\cdot \mid \xi)\big]$. Watanabe's analysis is based on the equivalent normalized evidence and its deterministic version, $\mathcal{Z}(n) = \int_\Omega e^{-nK_n(\xi)}\varphi(\xi)\,d\xi$ and $\mathcal{Z}_K(n) = \int_\Omega e^{-nK(\xi)}\varphi(\xi)\,d\xi$. It is immediate that $\log \mathcal{Z}(n) = \log m(X^{(n)}) - \ell_n(\xi^\star)$, and studying the asymptotic behavior of $\log m(X^{(n)})$ for large n is equivalent to studying that of $\log \mathcal{Z}(n)$. The deterministic quantity $\mathcal{Z}_K(n)$ is closely related to $\mathcal{Z}(n)$ as it is obtained by replacing the stochastic quantity $K_n(\xi)$ with its expectation $K(\xi)$ under the true distribution.

Our goal is to study the asymptotic behavior of the normalized evidence $\mathcal{Z}(n)$ and its deterministic counterpart $\mathcal{Z}_K(n)$ defined above. The integrals which define $\mathcal{Z}(n)$ and $\mathcal{Z}_K(n)$ are known as Laplace integrals. The asymptotic behavior of a Laplace integral as $n \to \infty$ concentrates on the minimum locus of $K(\xi)$, i.e., the set $\Omega_0 = \{\xi \in \Omega \,|\, K(\xi) = 0\}$; see Ch. 7 and Ch. 8 of [7] for more information on Laplace integrals and their asymptotic expansions. When $K(\xi)$ is a real analytic function, Ω_0 is called a *real analytic set*. For example, the asymptotic behavior of $\int_{[0,1]^2} e^{-n(\xi_2^2-\xi_1^3)^2}\,d\xi$ as $n \to \infty$ will be determined in a neighborhood of the set $\Omega_0 = \{\xi \,|\, \xi_2^2 - \xi_1^3 = 0\}$. The study of real analytic sets is a part of the field of mathematics known as algebraic geometry. [2]

2.1 Mathematical Issues in Singular Models

We briefly summarize some of the mathematical issues that arise when the regularity conditions are not met. The first issue is geometric in nature and arises from the fact that the set $\Omega_0 = \{\xi : K(\xi) = 0\}$ of global minimizers of $K(\cdot)$, or optimal parameters, may contain singular points. A point $\xi_0 \in \Omega_0$ is a *non-singular* point of Ω_0 if there exists open sets $U, V \subset \mathbb{R}^d$, $U \ni \xi_0$, and an analytic isomorphism $f : U \to V$, such that $f(\Omega_0 \cap U) = \{(x_1, x_2, \ldots, x_r, 0, 0, \ldots, 0) \,|\, x_i \in \mathbb{R}\} \cap V$, for some $r \leq d$ [50, Definition 2.6].[3] Informally, this definition says that Ω_0 can be expressed locally by a Euclidean coordinate system defined by real analytic functions. Singular points are defined as points which fail to be non-singular. No neighborhood of a singular point can be viewed as a real analytic coordinate transform of an open ball in some Euclidean space. This creates significant difficulties in analyzing the behavior of $\mathcal{Z}_K(\xi)$ on the set of nearly optimal parameters $\Omega_\varepsilon = \{\xi : K(\xi) < \varepsilon\}$. The second issue is probabilistic and arises while attempting to study the convergence of a stochastic process containing a singular point. This issue is illustrated by example 5.5 in Sec. 5.3 of [50]. Consider the function $f(a, b, X, Y) = aX + bY$, where $X, Y \overset{ind.}{\sim} N(0,1)$ and $\Omega = \{(a,b) \in [-1,1]^2\}$. For $X_i, Y_i \overset{ind.}{\sim} N(0,1)$ for $i \in [n]$, the function

$$f_n(a,b) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{f(a,b,X_i,Y_i)}{\sqrt{a^2+b^2}}$$

$$= \frac{a}{\sqrt{a^2+b^2}} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \right) + \frac{b}{\sqrt{a^2+b^2}} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \right)$$

defines an empirical process on $\Omega_1 = \{(a,b) = [-1,1]^2 : a^2 + b^2 \neq 0\}$. This empirical process is not well defined at the origin, $\lim_{(a,b)\to(0,0)} f_n(a,b)$ does not exist, despite $E[f_n(a,b)^2] = 1$, as the origin is a singularity of this process.

Both of these issues can be resolved by Hironaka's theorem on the resolution of singularities for real analytic functions [46]. Herein this result will be referred to as the resolution of singularities. As noted by Watanabe, the resolution of singularities guarantees the existence of a real analytic manifold $\mathcal{M}$ and real analytic transform

[2] Most literature on algebraic geometry focuses on zero-locus of polynomial systems. For introductions to algebraic geometry see [12, 17]. The geometry of real analytic sets is further discussed in [2].

[3] For an alternative definition of a singular point based on the rank of the Jacobian; see Theorem 3.4 of [50].

$g : \mathcal{M} \to \Omega$, $u \mapsto \xi$, such that the log-likelihood ratio after the transform can be represented as

$$nK_n(g(u)) = nu^{2\mathrm{k}} - \sqrt{n}\, u^{\mathrm{k}} W_n(u) \tag{7}$$
$$:= nu_1^{2k_1} u_2^{2k_2} \cdots u_d^{2k_d} - \sqrt{n}\, u_1^{k_1} u_2^{k_2} \cdots u_d^{k_d} W_n(u)$$

where $W_n(u)$ is a well-defined mean-zero stochastic process converging in distribution to a mean-zero Gaussian process $W(u)$ as $n \to \infty$. Surprisingly, all of the mathematical issues we discussed above are resolved in this transformed coordinate system.

The resolution of singularities guarantees that any neighborhood of a real analytic set can be understood as the image of normal crossing functions. A real analytic function $f : \mathbb{R}^d \to \mathbb{R}$ is *normal crossing* at a point x^* if there exists a neighborhood U of x^* such that $f(x) = a(x) \cdot \prod_j (x_j - x_j^*)^{k_j}$ for $x \in U$, where $a(\cdot)$ is a positive real analytic function and $k_1, k_2, \ldots, k_d$ are non-negative integers with at least one k_j positive. The importance of normal crossing functions derives from their ability to generalize the following result known as the factor theorem. If a one-dimensional real analytic function $f : \mathbb{R} \to \mathbb{R}$ satisfies $f(a) = 0$, then there exists a second real analytic function $g(x)$ for which $f(x) = g(x)(x - a)$. Unfortunately, this does not generalize to higher dimensions in general [50, Remark 2.9]. However, when zeros of a real analytic function $f(x)$ can be identified with the zeros of a normal crossing function $h(x)$, we have a multivariate analog of the factor theorem. If $\{x : f(x) = 0\} = \{x : h(x) = 0\}$ and $h(x)$ is normal crossing, then there exist another real analytic function $g(x)$ such that $f(x) = g(x)h(x)$. See Ch. 2, Sec. 5 of [50] for further details.

Watanabe calls the coordinate system $\mathcal{M}$ arising from the resolution of singularities the standard form of the model [52, Definition 9]. A statistical model is said to be in *standard form* if there exist real analytic functions $a(x, u)$ and $b(u) > 0$ such that: (i) $K(u) = u^{2\mathrm{k}} := u_1^{2k_1} \ldots u_d^{2k_d}$ is a monomial, where $\mathrm{k} = (k_1, \ldots, k_d)^{\mathrm{T}} \in \mathbb{N}^d$ is a multi-index having at least one positive entry; (ii) for the same multi-index k, $K_n(u) = n^{-1} \sum_i a(X_i, u)\, u^{\mathrm{k}}$ [52, Theorem 8]. This condition is used to prove the convergence of $W_n(u)$ to a centered Gaussian process $W(u)$; (iii) the prior density is $\varphi(u) = b(u)\, u^{\mathrm{h}}$, where $\mathrm{h} = (h_1, \ldots, h_d)^{\mathrm{T}} \in \mathbb{N}^d$ is another multi-index.[4] Moreover, the stochastic process $W_n(u) = n^{-1/2} \sum_i [u^{\mathrm{k}} - a(X_i, u)]$ is well-defined through Eq. (7) as it is now the ratio of the normal crossing functions.

For a given $K(\xi)$, none of the real analytic manifold $\mathcal{M}$, the real analytic map $g : \mathcal{M} \to \Omega$ from the resolution of singularities, or the multi-indices k and h are necessarily unique [52, Remark 46]. There are, however, numerical summaries of these objects which do not depend on the choice of $\mathcal{M}$ or g. These quantities are the real log-canonical threshold (RLCT) λ and the multiplicity m that arise in Eq. (1) and are determined using the local coordinates that compose $\mathcal{M}$. In each local coordinate system U_ℓ of $\mathcal{M}$, the local standard form of a model, $K(g_\ell(u)) = u^{2k_\ell}$ and $\varphi(g_\ell(u)) = b_\ell(u) u^{h\ell}$, can be used to determine the local RLCT λ_ℓ and the multiplicity m_ℓ via simple closed-form expressions, with

$$\lambda_\ell = \min_{j \in [d]} (h_{\ell,j} + 1)/(2k_{\ell,j}), \tag{8}$$
$$m_\ell = \#\big\{ i \in [d] : (h_{\ell,i} + 1)/(2k_{\ell,i}) = \lambda_\ell \big\}.$$

The RLCT λ and multiplicity m of the singular model are defined by

$$\lambda = \min_\ell \lambda_\ell, \quad m = \max\{m_\ell : \lambda_\ell = \lambda\}.$$

The RLCT has an important geometric interpretation as a measure of effective dimension of the set of optimal parameters Ω_0 since $\lambda = \lim_{\varepsilon \to 0} [\log \mathrm{Vol}(\varepsilon) / \log n]$, where $\mathrm{Vol}(\varepsilon) = \int_{K(\xi) < \varepsilon} \varphi(\xi) d\xi$.

REMARK 2.2. The resolution of singularities should be perceived as a purely theoretical result that only guarantees the existence of such a coordinate system. While this theorem provides a recursive algorithm to compute the real analytic manifold which puts the model into standard form, an application of this algorithm to specific singular models is far from being straightforward.

REMARK 2.3. Even for regular models the resolution of singularities is typically necessary to put the model in standard form. For example, the model $p(y \mid x, s, t) = \exp\{-(y_1 - s)^2 - (y_2 - t)^2\}/\pi$, for $W = \{(s, t) \mid s^2 + t^2 \leq 1\}$ and $y \in \mathbb{R}^2$. This model is regular if the true model is $q(x) = p(y \mid 1, 0)$, but the average log-density ratio $K(s, t) = s^2 + t^2 - 1$, which is not in standard form.

2.2 The State Density Function and the Asymptotic Expansion of the Evidence

The main analytic approximation that arises in Watanabe's derivation comes from the *state density function* of a statistical model. The combination of the state density function and the standard form, together with some additional tools from complex analysis allow Watanabe to derive an asymptotic form of the deterministic normalized evidence. Using the standard form of the model Watanabe connects the asymptotic form of the normalized evidence to the asymptotic expansion of the deterministic normalized evidence to the through convergence in distribution of stochastic process which defines their difference.

[4] The function $a(x, u)$ satisfies the following mathematical condition for arbitrary $s > 0$ and multi-index $k \geq 0$, $\int \sup_u |(\partial/\partial u)^k a(x, u)|^s p(x \mid u^\star) dx < \infty$

The state density function is the Schwartz distribution,

$$v(t) = \int_{\mathbb{R}^d} \delta(t - K(\xi))\, \varphi(\xi)\, d\xi. \tag{9}$$

See Appendix A for the formal definition of a Schwartz distribution. It is important to understand that the state density function is not actually an integral in the Lebesgue sense. The notation $\int F(x)\phi(x)dx$ is typically adopted when the explicit presence of the variable x is notationally helpful [15].

The importance of the state density function comes from is its action as a Schwartz distribution. As a Schwartz distribution, the state density function $v(t)$ acts as a change of variables in the integration,

$$\int_0^\infty F(t)v(t)dt = \int_\Omega F(K(\xi))\varphi(\xi)d\xi.$$

From a probabilistic perspective, the state density function arises as the density measure for the level set random variable $T = K(\xi)$ with $\xi \sim \varphi$, via the vector-to-scalar change of variables. The cumulative density function of T is $V(t) = \mathbb{P}[K(\xi) \leq t] = \int_{\{K(\xi)\leq t\}} \varphi(\xi)d\xi$, and v can be heuristically thought of as the a.e. derivative of V.

Using the identity in the above display, the marginal likelihood $\mathcal{Z}_K(n)$ can be viewed as the action of the state density function $v(t)$ on the function $F(t) = e^{-nt}$,

$$\mathcal{Z}_K(n) = \int_\Omega e^{-nK(\xi)}\, \varphi(\xi)\, d\xi = \int_0^\infty v(t)e^{-nt}dt.$$

Notice that the integral on the right hand side is the Laplace transform of the state density function $v(t)$ evaluated at n, $\mathcal{L}[v](n) = \int_0^\infty v(t)e^{-nt}dt$, where $\mathcal{L}[f](z) = \int_0^\infty f(t)e^{-zt}dt$ denotes the Laplace transform of a function f evaluated at the point $z \in \mathbb{C}$; The Laplace transform is defined for locally integrable functions satisfying the following two conditions: 1. The support of f is contained in $[0, \infty)$ and 2. There exists a $c \in \mathbb{R}$ such that $f(t)e^{-ct} \in L^1(\mathbb{R})$. The Laplace transform can also be extended to a well defined transform on the space of Schwartz distributions. For more details see [56, 11].

This view point allows Watanabe to derive an asymptotic expansion for the normalized evidence $\mathcal{Z}_K(n)$ by first deriving an asymptotic expansion for the corresponding state density function $v(t)$, and then pushing it back to the normalized evidence through the Laplace transform. In particular, the asymptotic behavior of the deterministic normalized evidence $\mathcal{Z}_K(n)$ as $n \to \infty$ corresponds to the asymptotic behavior of state density function $v(t)$ as $t \to +0$.

Due to the resolution of singularities, we only need to focus on the case when the model is in standard form and defined on the positive cone $\mathcal{K}_+ = \{\xi \in W \mid \xi_1, \ldots, \xi_d \geq 0\}$. In this case the state density function is given by,

$$v(t) = \int_{\mathbb{R}^d} \delta(t - \xi^{2\mathrm{k}})\, |\xi^{\mathrm{h}}|\, b(\xi)\chi(\xi)\, d\xi,$$

where $\chi(\xi)$ is the indicator function for the positive cone $\mathcal{K}_+$. Analysing the behavior of above state density function as $t \to +0$ requires a little technical finesse because the state density function is not technically a function, but a Schwartz distribution [16], which is not well-defined as $t \to +0$. Watanabe derives an asymptotic expansion for the state density function in terms of a well-defined Schwartz distribution $D(\xi)$ using the connection between the state density function and its Mellin transform, $\mathcal{M}[v](z) = \int_0^\infty t^z v(t)dt$, which can be computed in closed form when the model is written in standard form,

$$\delta(t - \xi^{2\mathrm{k}})\, |\xi^{\mathrm{h}}|\, b(\xi)\, \chi(\xi) \;\cong\; t^{\lambda-1}(-\log t)^{m-1} D(\xi), \tag{10}$$

here the real-log-canonical threshold (RLCT) λ and its multiplicity m are determined by the standard form of the statistical model; See Ch. 4 of [50], also Ch. 5 Theorem 9 and Remark 31 of [52].

Watanabe recovers an asymptotic expansion for the normalized evidence $\mathcal{Z}(n) = \int_\Omega e^{-nK_n(\xi)}\varphi(\xi)d\xi$ by connecting it back to the deterministic normalized evidence $\mathcal{Z}_K(n)$. For a statistical model in standard form, we have $nK_n(\xi) = n\xi^{2\mathrm{k}} - \sqrt{n}\xi^k W_n(\xi)$, where $W_n(\xi)$ is a stochastic process that can be shown to converge in distribution[5] to a Gaussian process $W(\xi)$. Under the same setting, we can expand the evidence as

$$\begin{aligned}\mathcal{Z}(n) &= \int_\Omega e^{-nK_n(\xi)}\varphi(\xi)d\xi \\ &= \int_\Omega e^{-n\xi^{2\mathrm{k}} + \sqrt{n}\xi^k W_n(\xi)}\, |\xi^h|\, b(\xi)\chi(\xi)d\xi.\end{aligned}$$

Performing a change of variables using the state density function,

$$\mathcal{Z}(n) = \int_\Omega \int_0^\infty e^{-ns + \sqrt{ns}W_n(\xi)} v(s)dsd\xi.$$

Finally, applying a change of variables $t = ns$ and rewriting the right hand side with the asymptotically equivalent $t^{\lambda-1}(-\log t)^{m-1}D(\xi)$, we obtain

$$\begin{aligned}\mathcal{Z}(n) = \frac{(\log n)^{m-1}}{n^\lambda} \int_\Omega D(\xi) \int_0^\infty t^{\lambda-1} e^{-t+\sqrt{t}W_n(\xi)} dtd\xi \\ + o_p\left(\frac{(\log n)^{m-1}}{n^\lambda}\right).\end{aligned} \tag{11}$$

Hence, the log-marginal likelihood has the following asymptotic expansion

$$\log \mathcal{Z}(n) \asymp -\lambda \log n + (m-1)\log\log n + O_p(1).$$

[5] The convergence in distribution follows from a technical condition on the log-likelihood ratio. See Chapter 5 of [50].

In summary, Watanabe analyzes the asymptotic behavior of normalized evidence $\mathcal{Z}(n)$ by first noticing the connection between the deterministic normalized evidence $\mathcal{Z}_K(n)$ for a model in standard form and the Laplace transform of the state density function Eq. (9). Watanabe then shows that the state density function corresponding to a model in standard form, where $K(\xi) = \xi^{2k}$, is asymptotically given by the well-defined Schwartz distribution $t^{\lambda-1}(-\log t)^{m-1}D(\xi)$, where λ is the RLCT of the model and m is the multiplicity. This allows Watanabe to properly study the asymptotics of $\mathcal{Z}_K(n)$ and hence $\mathcal{Z}(n)$ as $n \to \infty$ through the asymptotics of the state density function as $t \to 0+$. This yields the following asymptotic expansion for the normalized log-marginal likelihood,

$$\log \mathcal{Z}(n) \asymp -\lambda \log n + (m-1)\log\log n + \text{Const.}$$

The full details of this approach can be found in Chapter 5 of [52].

2.3 Watanabe's Original Derivation

In the above sections we have individually discussed the various tools Watanabe uses to establish an asymptotic expansion for the log-marginal likelihood. We will now discuss how to combine all of these tools together and provide a sketch of the proof of Watanabe's asymptotic expansion for the log-marginal likleihood of a general singular model. Further details can be found in Chapter of [52].

Unlike the previous sections, we do not assume that the singular model is in standard form. Given a singular model in some parameter coordinate system $w \in \mathbb{R}^d$, the first step in Watanabe's analysis is to decompose the normalized evidence $\mathcal{Z}(n)$ into the essential and non-essential parts, defined by

$$\mathcal{Z}_1(n) = \int_{\{w:K(w)<\varepsilon\}} e^{-nK_n(w)}\varphi(w)dw,$$
$$\mathcal{Z}_2(n) = \int_{\{w:K(w)\geq\varepsilon\}} e^{-nK_n(w)}\varphi(w)dw.$$

The essential set $\Omega_\varepsilon = \{w : K(w) < \varepsilon\}$ is the ε-neighborhood of the optimal parameter set Ω_0; the non-essential set is the complement of the essential set. Note, this step proceeds the application of the resolution of singularities.

For both singular and regular models, the non-essential part can be shown to exponentially decay in n of order $\mathcal{Z}_2(n) = o_p(\exp\{-n\varepsilon/3\})$. The major difference arises for the techniques used to handle the essential part. For regular models, the essential neighborhood is equivalent to $\{w : \|w - w^\star\| < \varepsilon\}$ and the Laplace expansion can now be invoked to show that $\mathcal{Z}_1(n) \propto n^{-d/2}$; see chapter 4 of [52] for details.

For singular models, we now invoke the resolution of singularites to express the essential set Ω_ε as the image of a finite number of local coordinate systems Ξ_j in which the model is in standard form. On each local coordinate system the discussion in Sec. 2.2 applies and we have that the evidence restricted to the local coordinate system asymptotically behaves like

$$\mathcal{Z}_j(n) = \frac{(\log n)^{m_j-1}}{n^{\lambda_j}} \int_{\Xi_j} D_j(\xi) \int_0^\infty t^{\lambda_j-1} e^{-t+\sqrt{t}W_n(\xi)} dt d\xi + o_p\left(\frac{(\log n)^{m_j-1}}{n^{\lambda_j}}\right),$$

and the RLCT λ_j, its multiplicity m_j, and the Schwartz distribution $D_j(\xi)$ will each be determined locally on the coordinate system Ξ_j. The leading order asymptotic behavior of the evidence is given by those local coordinate systems for which the local RLCT λ_j and its multiplicity m_j are equal to the global RLCT $\lambda = \min_j \lambda_j$ and its multiplicity $m = \#\{j \in [d]\ \lambda_j = \lambda\}$. Watanabe calls the local coordinate systems the essential local coordinates (ELC). The leading order asymptotic behavior of the evidence is given by

$$\mathcal{Z}(n) \asymp \frac{(\log n)^{m-1}}{n^{\lambda}} \int \sum_{j \in ELC} D_j(\xi) \int_0^\infty t^{\lambda-1} e^{-t+\sqrt{t}W_n(\xi)} dt d\xi + o_p\left(\frac{(\log n)^{m-1}}{n^{\lambda}}\right).$$

It follows that the log-marginal likelihood of a singular model is asymptotically

$$\log m(X^{(n)}) \asymp n\ell_n(\xi^\star) - \lambda \log n + (m-1)\log\log n + O_p(1).$$

It is important to realize that these results only hold asymptotically as the equivalence between the two Schwartz distributions in Eq. (10) only holds asymptotically. Furthermore, this expansion is unique up to an equivalence that arises in the blow-up of singularities in the resolution of singularities. It is possible to perform the blow-up procedure in multiple ways which result in different multi-indexes h, k. However, the RLCT and its multiplicity are birational invariants and will be the same for any properly constructed sequence of blow-ups of the singularities of the model.

This concludes our overview of Watanabe's asymptotic expansion of the log-marginal likelihood. Next we will present several examples of these calculations in action. In the next section we will present our original results, an exact expansion for the evidence using only probabilistic arguments.

2.4 Examples

We now present three simple examples to demonstrate some of the ideas introduced above. We have chosen these examples specifically because all of the necessary demonstrative computations are easily verified by straight forward computations.

In order to perform these computations by hand we will utilize an alternative, but equivalent, approach based on the Mellin transform of the state density function. The Mellin transform of the state density function for a general singular model not in standard form is given by the zeta function $\zeta(z) = \int K(\xi)^z \varphi(\xi) d\xi = \mathcal{M}[v](z)$. This function can be analytically continued to a meromorphic function defined on the entire complex plane whose poles $\{-\lambda_k\}$ are all real, negative, rational numbers with the order of the kth pole denoted by m_k. An asymptotic expansion of the state density function is then derived using the inverse Mellin transform of the Laurent expansion of this zeta function $\zeta(z)$, which shows $v(t) \asymp t^{\lambda-1}(-\log t)^{m-1}$ with $\lambda = \min_k \lambda_k$ and $m = m_k$. More details of this approach can be found in [50]. For a model in standard form, the smallest pole λ and its multiplicity m in this zeta function are the RLCT and its multiplicity defined in the expansion (1).

Note that the RLCT and its multiplicity are both intrinsic characteristics of a meromorphic function, and are therefore invariant to real analytic transformations.

REMARK 2.4. The connection with the poles of the zeta function of a statistical model allows an alternative route to computing the RLCT and its multiplicity. [5] utilizes this approach to compute the RLCT and its multiplicity for the general reduced rank regression model. Additional work in recent years has attempted to determine the RLCT and its multiplicity using this and alternative approaches [54, 33, 6, 3, 19, 13, 4].

2.4.1 Reduced Rank Regression Our first example is the 1-dimensional reduced rank regression on $\Omega = \{(a,b) \in [-1,1]^2\}$, $y_i \mid x_i, a, b \overset{ind.}{\sim} N(abx_i, 1)$ for $i \in [n]$, $x_i \overset{ind.}{\sim} \text{Unif}[-1,1]$ for $i \in [n]$, and $a, b \overset{ind.}{\sim} \text{Unif}[0,1]$. The true data generating distribution is given by $a^\star = b^\star = 0$. The set of optimal parameters is given by $\Omega_0 = \{(a,b) : a = 0 \text{ or } b = 0\}$. Straightforward computation shows that the log-likelihood ratio and the KL-divergence are respectively given by $nK_n(a,b) = ab(nS_{xx}ab - 2nS_{xy})/2$ and $K(a,b) = a^2b^2/6$. Notice that the model is in standard form in the original (a,b)-coordinate system. Hence we do not need to use the resolution of singularities. The asymptotic expansion of the evidence can be computed by first determining the RLCT and its multiplicity, and then applying the Laplace transform to the corresponding state-density function. In this example, we have $k = (1,1)$ and $h = (0,0)$, so $\lambda = 1/2$ and $m = 2$.

To obtain the state density function, we shall view it as the inverse Mellin transform of the zeta function of the statistical model. The zeta function for this model is given by $\zeta(z) = \int_0^1 \int_0^1 (a^2b^2)^z da db = (z+1/2)^{-2}/4$. The inverse Mellin transform of a function F of the form $F(z) = c_0(z+\lambda)^{-m}$, with c_0 being a constant, is given by $f(s) = c_0 s^{\lambda-1}(\log 1/s)^{m-1}/(m-1)!$. Hence, the state density function for this example is given by $v(t) = t^{\lambda-1}(-\log t)^{m-1}/4 = t^{-1/2}(-\log t)/4$, for $0 < t < 1$, and $v(t) = 0$ for $t = 0$. Finally, the evidence is the Laplace transform of the state density function,

$$\begin{aligned}
\mathcal{Z}_K(n) &= \int_0^1 v(t) e^{-nt} dt = n^{-1} \int_0^n v(t/n)\, e^{-t} dt \\
&= C_1 n^{-1/2} \log(n) - C_2/\sqrt{n} + C_3(n), \\
C_1 &= \int_0^\infty t^{-1/2} e^{-t} dt, C_2 = 1/4 \int_0^\infty t^{-1/2} \log(t) e^{-t} dt, \\
|C_3(n)| &\le C n^{-1/2} \log(n) \exp(-n/2),
\end{aligned}$$

with C a constant. Taking logarithms on both sides, we obtain $\log \mathcal{Z}_K(n) \asymp -1/2 \log n + \log\log(n) + o(1)$ as $n \to \infty$.

Analogous results for the general reduced rank regression can be found in [5]. The proof of the general reduced rank regression model requires the use of the resolution of singularities along with several approximation techniques that are necessary to derive the standard form of non-trivial singular models.

2.4.2 Normal Mixture Model Our second example is the normal mixture model on $\Omega = \{(s,t) : 0 \le s \le 1, -1 \le t \le 1\}$. Let $N(x)$ denote the density function of the univariate standard normal distribution. Our mixture model is given by $p(x \mid s,t) = (1-s)N(x) + sN(x-t)$ with prior $\varphi(s,t) = 1/2$, where the true data generating distribution is given by $N(x)$. The set of optimal parameters is given by $\Omega_0 = \{(s,t) : s = 0 \text{ or } t = 0\}$. A slightly tedious computation shows that the log-likelihood ratio has the form $f(x,s,t) = st \cdot a(x,s,t)$ and KL-divergence has the form $K(s,t) = s^2t^2K_0(s,t)$, with $K_0 \in C^1(\Omega)$, $K_0(s,t) > 0$. The model is rendered to a standard form by the change of variables $u = K_0(s,t)^{1/2}s$, $w = t$. It follows that the RLCT for this model is given by $\lambda = 1/2$ and its multiplicity is given by $m = 2$. Repeating the calculations from the previous example, we can reach $\log \mathcal{Z}_K(n) \asymp -1/2 \log n + \log\log(n) + o(1)$ as $n \to \infty$.

2.4.3 Layered Neural Network Neither of our previous examples required any significant algebraic reductions before applying the resolution of singularities to put the model into standard form. In general, there may be a significant amount of work necessary to express the set of optimal parameters in a form for which we can apply the resolution of singularities. We now present a simple

model which requires such techniques prior to the application of the resolution of singularities; this is example 7.1 from [50].

Consider the layered neural network model with one input, two hidden units, and activation function $\sigma(x) = e^x - 1$. Let $y_i \mid x_i \overset{ind.}{\sim} N(a\sigma(bx_i) + c\sigma(dx_i), 1)$ and $x_i \overset{ind.}{\sim} \text{Unif}(-1, 1)$. We assume the true distribution to be given by $y_i \overset{ind.}{\sim} N(0, 1)$. The log-likelihood ratio function is given by $f(x, a, b, c, d) = a\sigma(bx) + c\sigma(dx)$.

The set of optimal parameters is given by $\{(a, b, c, d) : K(a, b, c, d) = 1/2 \int (a\sigma(bx) + c\sigma(dx))^2 q(x)dx = 0\}$. In its current form we cannot directly apply the resolution of singularities with out first applying Hilbert's basis theorem for real polynomials to represent the set of optimal parameters with a finite number of polynomial equations. The Taylor expansion of f is given by $f(x, a, b, c, d) = \sum_k \frac{x^k}{k!} \cdot (ab^k + cd^k)$. Since $\{x^k\}$ is a set of linearly independent functions, $\{(a, b, c, d) \mid K(a, b, c, d) := \int (f(x, a, b, c, d))^2 \cdot q(x)dx = 0$ is equivalent to $g_k(a, b, c, d) := ab^k + cd^k = 0$ for all $k \in \mathbb{N}$. Notice that for any $k \in \mathbb{N}$, we have the identity

$$2g_{k+2}(a, b, c, d) = g_2(a, b, c, d)(b^k + d^k)$$
$$- g_1(a, b, c, d)(bd^k + db^k) + g_{k+1}(a, b, c, d)(b + d).$$

This shows that g_k for any $k \in \mathbb{N}$ can be generated[6] using g_1 and g_2. Hence $g_k(a, b, c, d) := ab^k + cd^k = 0$ for all $k \in \mathbb{N}$ is equivalent to $g_1(a, b, c, d) = g_2(a, b, c, d) = 0$. Computing the resolution of singularties allows us to construct the coordinate system which renders the model in standard form. This coordinate system[7] is given by $(a, b, c, d) = g(u)$, where $a = u_1$, $b = u_2u_4$, $c = u_1(u_2 - 1)u_2u_3u_4 - u_1u_2$, $d = u_4$, with Jacobian $|u_1(u_2 - 1)u_2u_4^2|$. Rewriting the KL divergence in the new coordinate system yields $K(u) = u_1^2u_2^2u_4^4K_0(u)$, with $K_0(u) \in C^1(\Omega)$, $K_0(u) > 0$. Using the change of variables[8] $w_1 = K_0^{1/2}u_1$, $w_j = u_j$ for $j > 1$, puts the system in standard form with $k = (1, 1, 0, 2)$ and $h = (1, 1, 0, 2)$. Thus the RLCT is given by $\lambda = 3/4$ with multiplicity $m = 1$. Repeating the zeta function calculations, we can obtain $\log \mathcal{Z}_K(n) \asymp -3/4 \log n + o(1) = -\lambda \log n + (m - 1) \log \log(n) + o(1)$ as $n \to \infty$.

3. NONASYMPTOTIC PROBABILISTIC BOUNDS FOR MODELS IN STANDARD FORM

In this section we present a non-asymptotic stochastic expansion of the evidence $\mathcal{Z}(n)$ using purely probabilistic arguments. Our proof proceeds in two parts. First, we provide a non-asymptotic two-sided bound to $\mathcal{Z}_K(n)$ for models in standard form based entirely on basic probabilistic arguments. Second, we utilize the previously established two-side bound, conditioning, and stochastic ordering to derive the non-asymptotic stochastic expansion of the evidence $\mathcal{Z}(n)$.

Connections with Watanabe's result: Our proof technique relies on simple probabilistic tools avoiding the highly technical nature of Watanabe's analytic approximation using the state density function and its Mellin transform. The main differences are as follows:

- The use of Dirac delta function as a generalized function is avoided by taking the conditional distribution of ξ with respect to ξ^{2k} and multiplying with the marginal density of ξ^{2k}.
- The conditioning neutralizes the effect of the singular part along with the stochastic component, while the marginal density provides the overall order.
- No approximation is made during the process. In contrast, while taking the inverse Mellin's transform, smaller order terms are dropped. Hence our representation of $\mathcal{Z}(n)$ in Theorem 3.4 is exact as opposed to Theorem 11 in [52]. Theorem 3.4 recovers Theorem 11 in [52] asymptotically as $n \to \infty$; see Appendix B.7.

3.1 The Deterministic Quantity $\mathcal{Z}_K(n)$

Our goal here is to provide a non-asymptotic two-sided bound to $\mathcal{Z}_K(n)$ for models in standard form based entirely on basic probabilistic arguments. Interestingly, the quantities λ and m turn out to be related to the rate and shape parameters of a collection of gamma densities, as we shall see below. In the first result, we assume $b(\xi) = 1$ and treat the general case as a corollary.

THEOREM 3.1. *Let $K(\xi) = \xi^{2k}$ for $\xi \in \Omega = [0, 1]^d$ and $\mathrm{k} = (k_1, \ldots, k_d)^{\mathrm{T}} \in \mathbb{N}^d$ with at least one positive entry, and let $\varphi(\cdot)$ be a probability density on Ω with $\varphi(\xi) \propto \xi^{\mathrm{h}}$, where $\mathrm{h} = (h_1, \ldots, h_d)^{\mathrm{T}} \in (0, \infty)^d$. Then, there exists positive constants C_1 and C_2 independent of n such that*

$$C_1 \frac{(\log n)^{m-1}}{n^\lambda} < \mathcal{Z}_K(n) < C_2 \frac{(\log n)^{m-1}}{n^\lambda},$$

where the RLCT λ and multiplicity m satisfy Eq. (8).

Below we provide an overview of the proof of the above theorem. The full proof can be found in Appendix B.1. Throughout, we use the convention that an $\text{Expo}(\beta)$ distribution has density $\beta e^{-\beta x} \mathbb{1}_{(0,\infty)}(x)$, that is, β denotes the rate parameter of the distribution. Let $\bar{d}$ denote the number of non-zero k_j in multi-index k. Define $\lambda_j := (h_j + 1)/(2k_j)$ and without loss of generality assume that these are sorted in non-decreasing order $\lambda_1 \leq \lambda_2 \ldots \leq \lambda_{\bar{d}}$.

[6] See CH 3 of [50] for more details.

[7] See pgs 102-103 in [50] for the derivation.

[8] This change of variables is well defined when $\{(s, t) \in \Omega \mid K_0(s, t)^{1/2} + s/2 \cdot K_0(s, t)^{-1/2} \cdot \partial K_0(s, t)/\partial s = 0\}$ is a set of measure zero. The K_0 for this example satisfies this condition.

Our analysis of the deterministic quantity $\mathcal{Z}_K(n)$ is an adaptation of Watanabe's proof using purely probabilistic tools. The main idea behind our proof is to exploit the natural representation of $\mathcal{Z}_K(n)$ as the expectation of a non-negative random variable with respect to the prior measure. Specifically, let $T = K(\xi)$, where $\xi \sim \varphi$ is a random variable distributed according to the prior measure. Then, it immediately follows that the non-negative random variable T takes values in the unit interval $[0,1]$ and $\mathcal{Z}_K(n) = \mathbb{E}[e^{-nT}]$. To obtain a handle on the distribution of T, we consider the random variable $Z := -\log T$ which can be expressed as a sum of $\bar{d}$ independent Expo(λ_j) random variables with $\lambda_j = (h_j+1)/(2k_j)$ as defined above; interestingly, observe the quantities $(h_j+1)/(2k_j)$s appear as the exponential rate parameters. Using a change of measure argument; see proof for details; we can express $\mathcal{Z}_K(n)$ in terms of the density of Z, denoted by g_Z,

$$\mathcal{Z}_K(n) = \int_0^n e^{-t}\,\frac{1}{t}\,g_Z\big(\log(n/t)\big)\,dt. \tag{12}$$

The intuition for the most general case of the proof follows from two special cases. First, consider the special case where $\lambda_j = \lambda$ for all $j = 1,\ldots,\bar{d}$ and $m = \bar{d}$. In this case Z is a Gamma(m,λ) random variable with density $\big(\lambda^m/\Gamma(m)\big)\,e^{-\lambda x}x^{m-1}\mathbb{1}_{(0,\infty)}(x)$. It follows that for any $t \in (0,n)$,

$$g_Z\big(\log(n/t)\big) = \frac{\lambda^m}{\Gamma(m)}\,n^{-\lambda}\,t^{\lambda}\big(\log(n/t)\big)^{m-1}.$$

It follows that

$$\mathcal{Z}_K(n) = Cn^{-\lambda}\int_0^n t^{\lambda-1}\,e^{-t}\,\big(\log(n/t)\big)^{m-1}\,dt \tag{13}$$

$$\asymp n^{-\lambda}\,(\log n)^{m-1}. \tag{14}$$

As a second special case, suppose $\lambda_1 < \ldots < \lambda_{\bar{d}}$, which implies that $\lambda = \lambda_1$ and $m = 1$. The distribution of Z is known as the Hypoexponential distribution [35, 23] and an analytic expression for its density is available in the literature as quoted below.

THEOREM 3.2 ([26, 8]). *Let $Z_k \overset{ind.}{\sim}$ Expo(λ_k) with density $g_k(z) = \lambda_k e^{-\lambda_k z}\mathbb{1}_{(0,\infty)}(z_k)$ for $k = 1,\ldots,K$, where $\lambda_1 < \ldots < \lambda_K$. Then, the density g_Z of $Z = \sum_{k=1}^K Z_k$ is*

$$g_Z(z) = \sum_{k=1}^K \underbrace{\left(\prod_{r\neq k}\frac{\lambda_r}{\lambda_r - \lambda_k}\right)}_{b_k} g_k(z).$$

The coefficients $\{b_k\}$ can be both positive and negative, and thus the above is not a mixture of exponential densities. However, the coefficient b_1 corresponding to the smallest rate parameter λ_1 is positive. We have, for any $t \in (0,n)$,

$$g_Z\big(\log(n/t)\big) = \sum_{j=1}^{\bar{d}} b_j\,\lambda_j\,n^{-\lambda_j}\,t^{\lambda_j}.$$

In this case we have

$$\mathcal{Z}_K(n) = \sum_{j=1}^{\bar{d}} b_j\,\lambda_j\,n^{-\lambda_j}\int_0^n e^{-t}t^{\lambda_j-1}dt$$

$$\asymp \sum_{j=1}^{\bar{d}} b_j n^{-\lambda_j} \asymp n^{-\lambda_1}.$$

This proves the theorem for this special case. The fact that $b_1 > 0$ has been crucially used to arrive at the last conclusion in the above display, along with the fact that $n^{-\alpha_1} > n^{-\alpha_2}$ for $\alpha_2 > \alpha_1 > 0$. This example carries the takeaway message that the exact form of the density g_Z is of secondary importance, and the focus should be on extracting the most significant contribution in terms of n. This is our strategy for the most general case.

In the general case, Z can be expressed as the sum of independent Gamma random variables. While there exist expressions for the density of sum of independent Gamma random variables [23, 26], they are much more cumbersome than the simpler case of exponentials in Theorem 3.2. Hence, we do not attempt to work with the density g_Z and instead aim to bound $\mathcal{Z}_K(n)$ from both sides. Using the idea of stochastic ordering of random variables we show that Z is stochastically bounded by two Gamma random variables whose expectations are of order $n^{-\lambda}(\log n)^{m-1}$.

We now state a corollary to Theorem 3.1 relaxing the assumption on the prior.

COROLLARY 3.3. *Assume the setup of Theorem 3.1. Let $b : U \to \mathbb{R}$ be an analytic function with $b(\mathbf{0}) \neq 0$, where U is any open subset of $\mathbb{R}^d$ containing Ω. Then,*

$$\int_\Omega b(\xi)\,e^{-nK(\xi)}\,\varphi(\xi)\,d\xi \asymp n^{-\lambda}(\log n)^{m-1}.$$

Corollary 3.3 shows that the assumption of a product prior in Theorem 3.1 can be relaxed to more general priors of the form $\widetilde{\varphi}(\xi) \propto b(\xi)\varphi(\xi)$, with the same asymptotic order of the normalizing constant as before.

3.2 The Stochastic Quantity $\mathcal{Z}(n)$

We now extend our probabilistic analysis from the previous subsection to analyze the stochastic quantity $\mathcal{Z}(n)$. Let us denote

$$Z_i(\xi) = \log\left\{\frac{p(X_i \mid \xi^\star)}{p(X_i \mid \xi)}\right\} - \mathbb{E}^\star \log\left\{\frac{p(X_i \mid \xi^\star)}{p(X_i \mid \xi)}\right\}, \tag{15}$$

$$i = 1,\ldots,n,$$

so that $n^{-1}\sum_{i=1}^n Z_i(\xi) = [K_n(\xi) - K(\xi)]$ characterizes the difference between K_n and K as an average of i.i.d. random variables. The log-marginal likelihood can be written as,

$$\mathcal{Z}(n) = \int_\Omega e^{-nK(\xi)-\sqrt{n}\xi^{\mathrm{k}}W_n(\xi)}\varphi(\xi)d\xi,$$

where

$$W_n(\xi) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi^{-\mathrm{k}} Z_i(\xi). \tag{16}$$

We make some simplifying assumptions to keep the presentation from getting notationally too heavy. We shall assume $\varphi(\xi) \propto \xi^{\mathrm{h}}$, and also that $\lambda_j = \lambda$ for all $j = 1, \ldots, \bar{d} < d$. Recall from § 3.1 that $m = \bar{d}$ in this case. Denote by I the set $\{1, \ldots, m\}$ and $J = \{m+1, \ldots, d\}$. Clearly, under $\varphi(\cdot)$, the common distribution of the independent random variables ξ_j for $j \in I$ is Beta$(h_j + 1, 1)$ and hence $\xi_j^{2k_j}$ is Beta$(\lambda, 1)$ distributed for $j \in I$.

We now state a stochastic approximation result for $\mathcal{Z}(n)$ in Theorem 3.4 that can be considered as a non-asymptotic version of Theorem 11 in [52]. The main idea lies in decoupling the effect of the singular part ξ_I controlled by $K(\xi)$ from the non-singular ξ_J part of ξ. As we shall see in Theorem 3.4, our proof relies heavily on the properties of the conditional density of ξ_I given $K(\xi)$.

Since the distribution of $Z = -\log T = -\sum_{j=1}^m 2k_j \log \xi_j$ is Gamma(m, λ), the conditional distribution

$$(-2k_1 \log \xi_1, \ldots, -2k_m \log \xi_m) := (Z_1, \ldots, Z_m) \mid Z$$

is given by $Z \times$ Dirichlet$(\mathbf{1}_m)$ and is expressed as

$$f_{Z_1,\ldots,Z_m|Z}(z_1, \ldots, z_m) = \frac{\Gamma(m)}{Z^{m-1}},$$
$$0 \le \sum_{i=1}^{m-1} z_i \le Z, \quad z_m = Z - \sum_{i=1}^{m-1} z_i.$$

Hence, the conditional density of

$$\xi \mid Z = (e^{-Z_1/(2k_1)}, \ldots, e^{-Z_m/(2k_m)}) \mid Z$$

is given by

$$\varphi_{\xi|Z}(\xi_I) = \frac{2^m \prod_{j=1}^m k_j}{\prod_{j=1}^m \xi_j}\frac{\Gamma(m)}{Z^{m-1}}, \tag{17}$$
$$e^{-Z} \le (\xi_{I_{-m}})^{2k_{-m}} \le 1, \quad \xi_I^{2k} = Z,$$

where $I_{-m} = I\backslash\{m\}, k_{-m} = k\backslash\{k_m\}$. Also, $\varphi_{\xi|Z=z}(\xi_I)$ is the same as the density $\varphi_{\xi|T=e^{-z}}(\xi_I)$. Define a sequence of stochastic processes $D_n(t, \xi)$ with index set $\mathbb{R}^+ \times [0, 1]^d$ as

$$D_n(t, \xi) = t^{\lambda-1}e^{-t-\sqrt{t}W_n(\xi)}\varphi_{\xi|Z=-\log(t/n)}(\xi_I)\varphi(\xi_J).$$

Further, define an integrated version of $D_n(t, \xi)$ as $D_n(t) = \int_\Omega D_n(t, \xi)\, d\xi$ for $t \in \mathbb{R}^+$.

THEOREM 3.4. *The following expression provides a non-asymptotic stochastic expansion of $\mathcal{Z}(n)$ with $n^{-\lambda}(\log n)^{m-1}$ as the leading term,*

$$\mathcal{Z}(n)\prod_{j=1}^d (h_j + 1) = \frac{n^{-\lambda}(\log n)^{m-1}\lambda^m}{\Gamma(m)}\int_0^n D_n(t)dt + R_n$$

where

$$R_n = \int_{t=0}^n r_n(t)D_n(t)dt,$$

$$r_n(t) = \frac{\lambda^m}{\Gamma(m)}n^{-\lambda}\sum_{j=1}^{m-2}\binom{m-1}{j}(\log n)^j(-\log t)^{m-1-j}.$$

Moreover, the remainder term R_n is smaller order in comparison with the dominating term. If the sequence of stochastic processes W_n satisfied $\|W_n\|_\infty = O_p(1)$, then

$$\frac{|R_n|}{n^{-\lambda}(\log n)^{m-1}} \to 0, \quad \text{as } n \to \infty$$

almost surely.

The proof follows in two parts. In the first part of the proof we use the conditional density of $\xi \mid T$, the change of variables $T = K(\xi)$ as in §3.1, and the identity $\varphi_K(t) = (1/t)g_Z\{\log(1/t)\}$, where $g_Z(\cdot)$ is the pdf of a Gamma(λ, m) to show that

$$\mathcal{Z}(n)\prod_{j=1}^d (h_j + 1)$$
$$= \frac{\lambda^m}{\Gamma(m)}n^{-\lambda}\left(\log n\right)^{m-1}\int_0^n\int D_n(t, \xi)d\xi dt + R_n.$$

The second part of the proof shows that

$$|R_n| = o_p\left(n^{-\lambda}(\log n)^{m-1}\right).$$

4. MEAN-FIELD APPROXIMATION FOR MODELS IN STANDARD FORM

In this section, our goal is to show that mean-field variational approximation always correctly recovers the RLCT for singular models in standard form, which therefore constitute an interesting class of statistical examples where the mean-field approximation is provably better than the Laplace approximation. While mean-field inference is known to produce meaningful parameter estimates in many statistical models [55, 31], the algorithmic landscape contains both positive and negative results [40, 57, 27, 18].

Previous work on variational Bayes in singular models focuses on deriving asymptotic bounds, as a function of the sample size n, for the ELBO in latent variable models such as Gaussian mixture models [43], hidden Markov

models [20], and other latent variable models [45, 28, 21]. These bounds take the form

$$-\nu_1 \log n + C_1 \leq \text{ELBO} \leq -\nu_2 \log n + C_2$$

where ν_1 and ν_2 are constants that are similar to the known upper bounds on the true RLCT of the corresponding models. The derivations of these bounds are based on specific inequalities which can be applied to the optimal variational posterior for these models, which is based on a block mean-field approximation to the posterior rather than a full mean-field factorization, and may not be applicable to other models. More interestingly this analysis proceeds in the original coordinate system rather than the resolved coordinate. This is due to the difficulty of determining the resolution mapping g, from the resolution of singularities, for any specific model; recall that the resolution of singularities only guarantees the existence of the transform. Given the key role the resolution of singularities provides in both Watanabe's and our analysis of singular models, it makes sense that it would also play a key role in the the study of variational inference for singular models.

We begin our investigation of variational inference in singular models by first determining the asymptotic behavior of the ELBO for the mean-field variation approximation to the posterior in the resolved coordinate system given by,

$$\gamma_K^{(n)}(\xi) = \frac{e^{-nK(\xi)}\,\varphi(\xi)}{\mathcal{Z}_K(n)}, \quad \xi \in \Omega, \tag{18}$$

with $K(\xi) = \xi^{2k}$ in standard form as in Theorem 3.1 and $\varphi(\xi) \propto b(\xi)\xi^{h}$ is a probability density on Ω (e.g. prior) with the analytic function $b(\cdot)$ as in Corollary 3.3. Clearly, $\mathcal{Z}_K(n)$ is then recognized as the normalizing constant of $\gamma_K^{(n)}(\cdot)$, which serves as a deterministic version of the posterior defined in Eq. (6). Recall the Gibb's variational inequality, Eq. (3), which states that for any probability density $\rho \ll \varphi$ on Ω,

$$\log \mathcal{Z}_K(n) \geq \Psi_n(\rho) := -\left[\int nK(\xi)\,\rho(d\xi) + D(\rho\,\|\,\varphi)\right], \tag{19}$$

with equality attained if and only if $\rho = \gamma_K^{(n)}$. The quantity $\Psi_n(\rho)$ on the right hand side of (19) is a lower bound to $\log \mathcal{Z}_K(n)$ for any $\rho \ll \varphi$. We are interested in computing an asymptotic expansion for the evidence lower bound (ELBO) of the mean-field family,

$$\text{ELBO}(\mathcal{F}_{\text{MF}}) := \sup_{\rho \in \mathcal{F}_{\text{MF}}} \Psi_n(\rho). \tag{20}$$

The aim of variational inference is to use the optimal variational approximation $\rho^\star$ as a surrogate for the true posterior $\gamma_K^{(n)}(\xi)$. We would like to understand how close the optimal variational approximation $\rho^\star$ from a variational family $\mathcal{F}$ approximates the true posterior. One way to quantify the quality of the approximation is by determining the discrepancy between the asymptotic behavior of $\log \mathcal{Z}_K(n)$ and the asymptotic behavior of $\text{ELBO}(\mathcal{F}_{\text{MF}})$. If the mean-field family is capable of providing a good approximation to $\gamma_K^{(n)}(\xi)$, then the asymptotic expansion of the $\text{ELBO}(\mathcal{F}_{\text{MF}})$ should properly capture some of the leading order terms of the asymptotic expansion of the log-marginal likelihood $\log \mathcal{Z}_K(n)$. Since $\gamma_K^{(n)}$ does not lie in $\mathcal{F}_{\text{MF}}$ for any n, it follows that the inequality in equation (20) is a strict one if we restrict $\mathcal{F}$ to the mean-field family. We, however, show below that the optimal mean-field approximation of Eq. (18) correctly recovers the leading order term of asymptotic expansion of $\log \mathcal{Z}_K(n)$.

THEOREM 4.1. *Consider a variational approximation (20) to $\log \mathcal{Z}_K(n)$ in equation (18), where the variational family $\mathcal{F}$ is taken to be the mean-field family $\mathcal{F}_{\text{MF}}$ defined in equation (5). Then, there exists constants C_1, C_2 independent of n such that*

$$-\lambda \log n - C_1 \leq \text{ELBO}(\mathcal{F}_{\text{MF}}) \leq -\lambda \log n - C_2,$$

where the RLCT λ and the multiplicity m satisfy Eq. (8).

Since $\log \mathcal{Z}_K(n) \asymp -\lambda \log n + (m-1)\log\log n$, it follows that the optimal mean-field approximation correctly recovers the asymptotic behavior of $\log \mathcal{Z}_K(n)$. This, in particular, implies that the relative error R_n due to the mean-field approximation

$$R_n := \frac{|\log \mathcal{Z}_K(n) - \text{ELBO}(\mathcal{F}_{\text{MF}})|}{|\log \mathcal{Z}_K(n)|} \to 0 \text{ as } n \to \infty.$$

This is rather interesting, as the density $\gamma_K^{(n)}$ clearly does not lie in $\mathcal{F}_{\text{MF}}$ for any finite n. Similar bounds have been derived using mean-field VI for specific singular models [29], but these results heavily rely on model specific equations and do not readily extend to singular models in general. Moreover, Theorem 4.1 shows that it is not possible for the mean-field approximation of Eq. (18) to recover the lower order $(m-1)\log\log n$ term in the asymptotic expansion of $\log \mathcal{Z}_K(n)$ in Eq. (1).

Our strategy is to first show that every mean-field distribution satisfying the stationary equations of Ψ_n, denoted by $\widetilde{\rho} = \otimes_{j=1}^{d} \widetilde{\rho}_j$, satisfies a two sided bound which is of the order $-\lambda \log n - C_1 \leq \Psi_n(\widetilde{\rho}) \leq -\lambda \log n - C_2$, for some constants C_1, C_2 free of n.

We introduce some notation before describing the $\widetilde{\rho}_j$s. For $k, h, \beta > 0$, define a density $f_{k,h,\beta}$ on $[0,1]$ given by

$$f_{k,h,\beta}(u) = \frac{u^h \exp(-\beta u^{2k})\mathbb{1}_{[0,1]}(u)}{B(k,h,\beta)}, \tag{21}$$

where $B(k,h,\beta) = \int_0^1 x^h \exp(-\beta x^{2k})dx$. We record two useful facts about $f_{k,h,\beta}$ in the Lemma below. We collect

some well-known facts first about the incomplete gamma function; see [1].

REMARK 4.2. For $x, a > 0$, denote the lower incomplete gamma function by $\gamma(a,x) = \Gamma(a)^{-1}\int_0^x t^{a-1}e^{-t}\,dt$. For any fixed $a > 0$, $\gamma(a,\cdot)$ takes values in $(0,1)$, with $\lim_{x\to\infty}\gamma(a,x) = 1$. Also, $\lim_{x\to 0}[\gamma(a,x)/x^a] = 1/\Gamma(a+1)$, and $\gamma(a+1,x) = \gamma(a,x) - x^a e^{-x}/\Gamma(a+1)$.

LEMMA 4.3. *Let the density $f_{k,h,\beta}$ be as in equation (21). Then,*

(i) The normalizing constant $B(k,h,\beta)$ is given by $B(k,h,\beta) = \beta^{-\lambda}\Gamma(\lambda)\gamma(\lambda,\beta)/(2k)$.
(ii) The quantity $\int_0^1 u^{2k} f_{k,h,\beta}(du)$ depends on k and h only through $\lambda := (h+1)/(2k)$. Call this expectation $G(\lambda,\beta)$, and we have $G(\lambda,\beta) := \int_0^1 u^{2k} f_{k,h,\beta}(u)du = \lambda\beta^{-1}\cdot\gamma(\lambda+1,\beta)/\gamma(\lambda,\beta)$.
(iii) We have, $\lim_{\beta\to\infty}\left(|\log B(k,h,\beta) - (-\lambda\log\beta)|\right)/(\lambda\log\beta) = 0$ and $\lim_{\beta\to\infty} G(\lambda,\beta)/(\lambda/\beta) = 1$. Thus, for large β, $\log B(k,h,\beta) \asymp -\lambda\log\beta$, and $G(\lambda,\beta) \asymp \lambda/\beta$.
(iv) We have, $\lim_{\beta\to 0}\log B(k,h,\beta) = -\log(2k\lambda)$, $G(\lambda,0) = \lambda/(\lambda+1)$. Thus, for small β, there exists constatnts C_1, C_2, C_3, C_4 such that and $C_1 \le G(\lambda,\beta) \le C_2$ and $C_3 \le \log B(k,h,\beta) \le C_4$.

REMARK 4.4. It follows from Lemma 4.3 that there exist constants C_1, $C_2 > 0$ and constants C_3, C_4,

$$\frac{C_1}{\max\{\beta,1\}} \le G(\lambda,\beta) \le \frac{C_2}{\max\{\beta,1\}}$$

$\beta \in [0,\infty)$ and, $C_3 - \lambda\log\beta \le \log B(k,h,\beta) \le C_4 - \lambda\log\beta$ for all $\beta \in [1,\infty)$.

Define $\lambda_j = (h_j+1)/(2k_j)$ for $1 \le j \le d$. Without loss of generality, we assume that $\lambda = \lambda_1 = \lambda_2 = \cdots = \lambda_m \le \lambda_{m+1} \le \cdots \le \lambda_d$, where λ is the RLCT and m its multiplicity. It follows from variational calculus that the optimal marginals of the mean-field approximation to $\gamma_K(\xi)$ are of the form

$$\rho_j(\xi_j) \propto \exp\{\mathbb{E}_{-j}[\log\gamma_K(\xi)]\},\ 1 \le j \le d,$$

where $\mathbb{E}_{-j}$ denotes the expectation with respect to

$$\rho_{-j}(\xi_{-j}) := \prod_{k\neq j}\rho_k(\xi_k).$$

A straightforward computation shows that the optimal marginals are of the form, $\rho_j(\xi_j) \propto \exp\{-\beta_j\xi_j^{2k_j}\}$, $\beta_j \in [0,\infty)$, for $j = 1,\ldots,d$. Hence, we consider $\widetilde{\rho}_j = f_{k_j,h_j,\beta_j}$ for $j = 1,\ldots,d$, with the $\{\beta_j\}$ terms to be specified at a later point. Let us now bound

$$\Psi_n(\widetilde{\rho}) = -\left[\int_\Omega nK(\xi)\widetilde{\rho}(\xi)d\xi + D(\widetilde{\rho}\,\|\,\varphi)\right]$$

by bounding its two parts. First, we have

$$\int_\Omega nK(\xi)\widetilde{\rho}(\xi)d\xi = n\prod_{j=1}^d G(\lambda_j,\beta_j)$$

where recall that $\lambda_j = (h_j+1)/(2k_j)$. From Remark 4.4 there exist constants $C_{1,K}$, $C_{2,K}$, free of n, such that

$$\frac{nC_{1,K}}{\prod_{j=1}^d\max\{\beta_j,1\}} \le \int_\Omega nK(\xi)\widetilde{\rho}(\xi)d\xi \le \frac{nC_{2,K}}{\prod_{j=1}^d\max\{\beta_j,1\}}.$$

Next, consider $D(\widetilde{\rho}\,\|\,\varphi)$. Let φ be the probability density on Ω with $\bar{\varphi}(\xi) \propto \xi^{\mathrm{h}}$. Write

$$D(\widetilde{\rho}\,\|\,\varphi) = D(\widetilde{\rho}\,\|\,\bar{\varphi}) + \int_\Omega \widetilde{\rho}\log\frac{\bar{\varphi}}{\varphi}.$$

The second term in the above display can be bounded away by constants independent of n on Ω. We bound $D(\widetilde{\rho}\,\|\,\bar{\varphi}) = \sum_{j=1}^d D(\widetilde{\rho}_j\,\|\,\bar{\varphi}_j)$, where $\bar{\varphi}_j$ is the jth marginal of $\bar{\varphi}$ with density $\bar{\varphi}_j(u) \propto u^{h_j}$ for $u \in [0,1]$, term-wise. Bounding each $D(\widetilde{\rho}_j\,\|\,\bar{\varphi}_j)$, for $j = 1,\ldots,d$, using Remark 4.4 and summing the bounds yields the two sided bound

$$A_1 + \sum_{j=1}^d \lambda_j\log\beta_j \le D(\widetilde{\rho}\,\|\,\bar{\varphi}) \le A_2 + \sum_{j=1}^d \lambda_j\log\beta_j,$$

for some constants A_1, A_2 free of n. Combining these two bounds produces the two-sided bound

$$-\widetilde{A}_2 + \frac{nC_{2,K}}{\prod_{j=1}^d\max\{\beta_j,1\}} - \sum_{j=1}^d\lambda_j\log\beta_j \le \Psi_n(\widetilde{\rho})$$

$$\le -\widetilde{A}_1 - \frac{nC_{1,K}}{\prod_{j=1}^d\max\{\beta_j,1\}} - \sum_{j=1}^d\lambda_j\log\beta_j.$$

for some constants $\widetilde{A}_1$, $\widetilde{A}_2$ free of n. Maximizing Ψ_n over the class of mean-field distributions $\mathcal{F}_{MF}$ is equivalent to minimizing the equations of the form

$$\frac{nC}{\prod_{j=1}^d\max\{\beta_j,1\}} + \sum_{j=1}^d\lambda_j\log\beta_j$$

over $\beta_1,\ldots,\beta_d \in [0,\infty)$. This equation is minimized by choosing $\beta_j = C_j n^{1/m}$, for $1 \le j \le m$ and $\beta_s = C_s$, for $m+1 \le s \le d$, where C_j and C_s are constants independent of n. Putting the pieces together, we have proved that

$$C_1 - \lambda\log n \le \sup_{\rho\in\mathcal{F}_{MF}}\Psi_n(\rho) \le C_2 - \lambda\log n,$$

for some constants C_1, C_2 free of n. This completes the proof.

4.1 Algorithmic achievability of the lower bound

In this section, we investigate the "algorithmic achievability" of the bounds from Theorem 4.1. In particular, the theorem guarantees that the ELBO corresponding to the optimal variational approximation in the mean-field family achieves the correct asymptotic order, i.e. that of $\log \mathcal{Z}_K(n)$. However, the theorem does not guarantee that this global optimum can be achieved by any algorithm used to numerically optimize the ELBO, such as Coordinate Ascent Variational Inference (CAVI). In practice, this algorithm may get stuck at a local optima of the mean-field class that does not properly capture the correct asymptotic behavior of the ELBO. We empirically study the behavior of CAVI for the optimization problem $\sup_{\rho \in \mathcal{F}_{\text{MF}}} \Psi_n(\rho)$ in the $d = 2$ case, which naturally constrains the coordinate updates to lie in the family of densities $\{f_{k,h,\beta}\}$. Our empirical results suggest that the CAVI algorithm is sub-optimal and is only capable of recovering the leading order term of the asymptotic expansion of $\log \mathcal{Z}_K(n)$.

The Coordinate Ascent Variational Inference (CAVI) algorithm is popular in statistics and machine learning for maximizing an evidence lower bound over a mean-field family; see Chapter 10 of Bishop for a book-level treatment. The CAVI can be interpreted as a cyclical coordinate ascent algorithm which at any iteration $t \geq 1$ cycles through maximizing $\Psi_n(\rho)$ as a function of ρ_j, keeping $\{\rho_\ell\}_{\ell \neq j}$ fixed at their current value $\{\rho_\ell^{(t)}\}_{\ell \neq j}$. For example, in the $d = 2$ case, the iterates $\rho^{(t)} = \rho_1^{(t)} \otimes \rho_2^{(t)}$ for $t \geq 1$ are given by

$$\rho_1^{(t)} = \arg\max_{\rho_1} \Psi_n\big(\rho_1 \otimes \rho_2^{(t-1)}\big),$$
$$\rho_2^{(t)} = \arg\max_{\rho_2} \Psi_n\big(\rho_1^{(t)} \otimes \rho_2\big),$$

with an arbitrary initialization $\rho^{(0)} = \rho_1^{(0)} \otimes \rho_2^{(0)} \ll \varphi$, and assuming the first component gets updated first. The objective function $\Psi_n(\rho_1 \otimes \rho_2)$ is concave in each argument[9] so that the maximization problems in the update step above have unique solutions. Moreover, these maximizers admit a convenient integral representation, which facilitates tractability of the updates in conditionally conjugate models. It is straightforward to see that the successive CAVI iterates increase the objective function value, since for any $t \geq 1$,

$$\Psi_n\big(\rho_1^{(t)} \otimes \rho_2^{(t)}\big) \geq \Psi_n\big(\rho_1^{(t)} \otimes \rho_2^{(t-1)}\big) \geq \Psi_n\big(\rho_1^{(t-1)} \otimes \rho_2^{(t-1)}\big), \tag{22}$$

although convergence to a global optimum is not guaranteed in general.

[9] although, it is rarely jointly concave

Returning to the present case, consider the standard form of a singular model with parameter dimension $d = 2$,

$$\gamma_K^{(n)}(\xi_1, \xi_2) \propto \xi_1^{h_1} \xi_2^{h_2} \exp(-n\xi_1^{2k_1}\xi_2^{2k_2}), \quad (\xi_1, \xi_2) \in [0,1]^2, \tag{23}$$

resulting from setting $b(\xi) \equiv 1$ in equation (18). Let $\lambda_i = (h_i + 1)/2k_i$ for $i = 1, 2$ as usual; we assume without loss of generality that $\lambda_1 \leq \lambda_2$, implying the real log-canonical threshold for this model is $\lambda = \lambda_1$. The mean field family $\mathcal{F}_{\text{MF}}$ in this case consists of product distributions $\rho_1 \otimes \rho_2$, where ρ_1 and ρ_2 are absolutely continuous densities on $[0, 1]$. We derive in Appendix C that the tth iteration of the CAVI algorithm (22) in this case is $\rho^{(t)}(\xi) = \rho_1^{(t)}(\xi_1) \cdot \rho_2^{(t)}(\xi_2)$, with

$$\rho_1^{(t)}(\xi_1) = f_{k_1,h_1,n\mu_1^{(t)}}(\xi_1),$$
$$\rho_2^{(t)}(\xi_2) = f_{k_2,h_2,n\mu_2^{(t)}}(\xi_2), \tag{24}$$

where recall the density $f_{k,h,\beta}$ is defined in equation (21) and for $t \geq 1$,

$$\mu_1^{(t)} = G\big(\lambda_2, n\mu_2^{(t-1)}\big),$$
$$\mu_2^{(t)} = G\big(\lambda_1, n\mu_1^{(t)}\big). \tag{25}$$

We also record the value of the ELBO at iteration t,

$$\begin{aligned}\Psi_n(\rho^{(t)}) = &- n\, G\big(\lambda_1, n\mu_1^{(t)}\big)\, G\big(\lambda_2, n\mu_2^{(t)}\big) \\ &+ n\,\mu_1^{(t)}\, G\big(\lambda_1, n\mu_1^{(t)}\big) + n\,\mu_2^{(t)}\, G\big(\lambda_2, n\mu_2^{(t)}\big) \\ &+ \log B\big(h_1, k_1, n\mu_1^{(t)}\big) + \log B\big(h_2, k_2, n\mu_2^{(t)}\big).\end{aligned} \tag{26}$$

This shows that the ELBO's behavior is fully determined by the convergence properties of the discrete time dynamical system in Eq. (25). Let $\rho^\star_{\text{CAVI}}(\xi) = \lim_{t\to\infty} \rho^{(t)}(\xi)$ denote the optimal mean-field distribution computed using the CAVI algorithm.

Numerical experiments for $d = 2$, random initialization of the dynamical system $(\mu_1^{(t)}, \mu_2^{(t)})$, and different combinations of (λ_1, λ_2) suggest that the asymptotic behavior of the ELBO of the converged CAVI algorithm, denoted by $\Psi_n(q^*)$, is given by $-\lambda \log n + C$. Table 4.1 contains the estimated coefficients and the corresponding p-values for the regression $\Psi_n(q^*) = \beta_0 + \beta_1 \log n + \beta_2 \log\log n$. We see that this regression fails to properly capture the multiplicity $m = 2$ in $\lambda_1 = 1$, $\lambda_2 = 1$ and $\lambda_1 = 2.3$, $\lambda_2 = 2.3$. Figure 2 visually summarizes these findings.

Our empirical results suggest that CAVI is capable of recovering the true global optima of ELBO($\mathcal{F}_{MF}$) as predicted by Theorem 4.1 for dimension $d = 2$.

5. DISCUSSION

In this article, we have taken a tour through the fascinating literature on evidence approximation in singular statistical models, and complimented some existing asymptotic results with non-asymptotic probabilistic bounds.

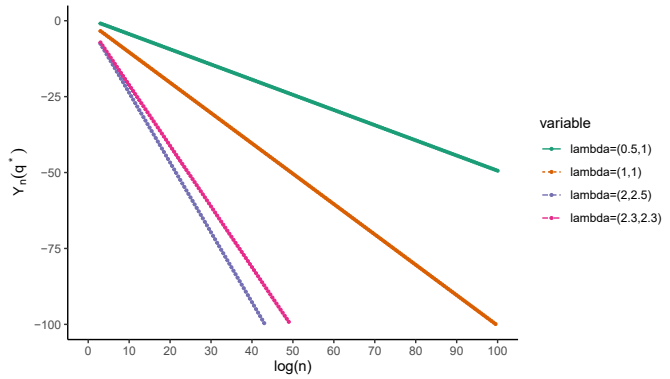

FIG 2. Optimized value of the ELBO for the example in Eq. (23) as a function of $\log n$ for different combinations of (λ_1, λ_2). Notice that the lines corresponding to $(2, 2.5)$ and $(2.3, 2.3)$ do not cross each other near $\log(n) \approx 30$. This indicates that the ELBO is missing the $\log\log(n)$ terms. Table 4.1 shows estimated coefficients and the corresponding p-values for the regressions $\Psi_n(q^*) = \beta_0 + \beta_1 \log n + \beta_2 \log\log n$.

(λ_1, λ_2)	Parameter	Estimate	p-value
(0.5, 1)	β_0	−-4.165e-01	2e-16
	β_1	−5.000e-01	2e-16
	β_2	5.336e-06	1.95e-06
(1,1)	β_0	−3.930e-01	2e-16
	β_1	−1.000e+00	2e-16
	β_2	2.538e-03	1.62e-09
(2,2.5)	β_0	−1.351e+00	2e-16
	β_1	−2.000e+00	2e-16
	β_2	4.834e-05	2.03e-06
(2.3,2.3)	β_0	−7.011e-01	2e-16
	β_1	−2.300e+00	2e-16
	β_2	2.842e-03	2.13e-08

TABLE 1
A table containing the estimated coefficients and P-values corresponding to the regressions $\Psi_n(q^*) = \beta_0 + \beta_1 \log n + \beta_2 \log\log n$ when $\Psi_n(q^*)$ is computed using mean-field VI. We see that the regression fails to correctly capture the multiplicity $m = 2$ in $\lambda_1 = 1$, $\lambda_2 = 1$ and $\lambda_1 = 2.3$, $\lambda_2 = 2.3$.

With the growing popularity of complex generative models in various applications, we hope the results and approaches discussed in this article aid future investigations into model selection guarantees involving singular models. We are also intrigued by the promise shown by mean-field variational inference in these problems. There are numerous open questions for variational inference in the singular setting.

The first line of open questions concerns the brief discussion in Sec. 4. Theorem 4.1 shows that $ELBO(\mathcal{F}_{MF})$, the optimal evidence lower bound over the mean-field class, recovers the leading order term of the asymptotic expansion of the log-marginal likelihood $\log \mathcal{Z}_K(n)$, but is not capable of recovering the lower order term $(m-1)\log\log n$. With the ELBO over the mean-field class provably unable to recover the correct asymptotic behavior of $\log \mathcal{Z}_K(n)$, a natural follow-up question would be to determine if this could be done with a more structured variational family?

The second line of open questions concerns the brief discussion in Sec. 4.1. Is it possible to achieve the theoretical bound in Theorem 4.1 with any of the numerical optimization algorithms such as CAVI or (stochastic) gradient descent? Our preliminary numerical results suggest that CAVI is able to recover the leading order term in the asymptotic expansion in dimension $d = 2$.

A third line of open questions stems from the fact that we have only studied the problem in the most simplified setting. Our first simplification arises from using the deterministic normalized posterior $\gamma_K^{(n)}$ instead of the true normalized posterior $\gamma_n \propto \exp\{-nK_n(\xi)\}\varphi(\xi)$. Additionally, we have not considered a model whose standard form is comprised of multiple sets of local coordinates. Theorem 4.1 will need to be re-verified in either of these more general contexts. New and interesting phenomena may arise as we relax these assumptions. It is entirely possible that the CAVI may fail to produce the correct asymptotic bounds for the ELBO when there are multiple standard form coordinate regions in the model.

Finally, we lack a full understanding of the role the standard form of the model plays in the algorithmic application of VI. The standard form of the model plays a central role in the derivations of both the asymptotic expansion in Sec. 2, and non-asymptotic expansion in Sec. 3, of the log-marginal likelihood as a change of variables for which we can easily study the behavior of the corresponding log-marginal likelihood. When studying the asymptotics of the ELBO using $\Psi_n(\rho)$ in Eq. (19), the standard form of the model also arises through a change of variables. However, when applying the CAVI algorithm to compute the optimal variational approximation, we can either compute the variational approximation in the original coordinate system, or first transform the model to standard form coordinate system and then compute the variational approximation in the standard coordinate system. It is not clear if these computations will result in optimal variational approximations that produce equivalent ELBOs. It may be the case that the ELBO is not numerically stable in the original coordinate system due to the singularity in the model, but is numerically stable in the standard form of the model, which resolves the singularities of the model. The answer to this question has important consequences for applications. If both coordinate systems produce equivalent behavior, then we could simply apply mean-field VI to singular models without having to determine the standard coordinates of the model.

Recall that the standard coordinate system for the model is found by computing the resolution of singularities of the model, which can be quite challenging[10] for even relatively simple examples such as the layered neural network example from Sec. 2.4.3. If the coordinate systems produce differing results, it may become necessary to determine the standard form of the model prior to apply the mean-field approximation. At this point, another possible approach to VI in singular model would be to use a more flexible transformation based variational family, such as the variational auto-encoder or normalizing flows, to learn a transformation to the standard form simultaneously while computing the optimal variational approximation. See [53] for recent work in this direction.

APPENDIX A: DEFINITION OF A SCHWARTZ DISTRIBUTION

Classically, functions are viewed in a point-wise manner. A function $f : \mathbb{R} \to \mathbb{R}$ is a maps which assigns to each point $x \in \mathbb{R}$ a numerical value $f(x) \in \mathbb{R}$. The Lebesgue spaces L^p, $1 \leq p \leq \infty$, hint at a different way to think about the idea of a function. Alternatively, a function $f \in L^p$ can be fully specified by studying the family of linear functionals $\int f\phi d\mu$, for $\phi \in L^q$, where $1/p + 1/q = 1$. An L^p function can be view as a linear map from $L^q \to \mathbb{R}$ instead of a point-wise map from $\mathbb{R}$ to $\mathbb{R}$. Schwartz distributions follow this same approach to defining generalized functions on the space of test functions. For $E \subset \mathbb{R}^d$, denote by $C_c^\infty(E)$ the space of all C^∞ functions whose support is compact and contained in E. The space of test functions over $E \subset \mathbb{R}^d$ is classically denoted by $\mathcal{D}(E)$ and represents the space $C_c^\infty(E)$ together with the topology defined by sequential convergence $\phi_j \to \phi$ if and only if $\partial^\alpha \phi_j \to \partial^\alpha \phi$ uniformly for all multi-indexes $\alpha \in \{0, 1, 2, 3, \ldots\}^d$, where $\partial^\alpha f$ denotes $\partial^{|\alpha|} f/(\partial^{\alpha_1} x_1 \partial^{\alpha_2} x_2 \cdots \partial^{\alpha_d} x_d)$ and $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_d$. The space of Schwartz distributions on E, denoted by $\mathcal{D}'(E)$, is the space of continuous linear functional on $C_c^\infty(E)$ with the weak* topology. For $F \in \mathcal{D}'(E)$ and $\phi \in \mathcal{D}(E)$ the value of F at ϕ will be denoted by $\langle F, \phi\rangle$. Two distributions $F, G \in \mathcal{D}'(E)$ are equal if $\langle F, \phi\rangle = \langle G, \phi\rangle$ for every $\phi \in \mathcal{D}(E)$. Our first example of a Schwartz distribution is the distribution defined by integration against a locally integrable function $f \in L^1_{\text{loc}}(E)$, $\langle f, \phi\rangle := \int_E f(x)\phi(x)dx$.[11] Similarly every Radon measure μ on E defines a distribution $\langle \mu, \phi\rangle := \int_E \phi d\mu$. Not all distributions arise from integrals. The point mass at the origin, denoted by δ, defines a distribution $\langle \delta, \phi\rangle := \phi(0)$. New distributions can be defined as linear transformations of existing distributions. Two important examples of this idea are differentiation and translation. For $F \in \mathcal{D}'(E)$, the derivative $\partial^\alpha F \in \mathcal{D}'(E)$, is given by $\langle \partial^\alpha F, \phi\rangle = (-1)^{|\alpha|}\langle F, \partial^\alpha \phi\rangle$ and the translation $\tau_y F \in \mathcal{D}'(E + y)$ is given by $\langle \tau_y F, \phi\rangle = \langle F, \tau_{-y}\phi\rangle$, where τ_y denotes translation by y, $\tau_y f(x) = f(x - y)$ and $E + y = \{x + y \,|\, x \in E\}$. In this manner the point mass at a point t is given by the distribution $\tau_t \delta$ and the derivative ∂H of distribution $H \in \mathcal{D}'(\mathbb{R})$ defined by the Heavyside step function $H(x) = \mathbb{1}_{\{x \geq 0\}}(x)$ is $\partial H = \delta$. Indeed for $\phi \in \mathcal{D}(\mathbb{R})$, $\langle \partial H, \phi\rangle = -\langle H, \partial\phi\rangle = -\int_0^\infty \phi'(t)dt = \phi(0) = \langle \delta, \phi\rangle$.

[10] The only practical model for which the full resolution map is currently known is Reduced Rank Regression [5].

[11] A function $f : \mathbb{R}^d \to \mathbb{R}$ is called locally integrable if $\int_K |f|dx < \infty$, for every bounded measurable set $K \subset \mathbb{R}^d$. The space of locally integrable functions is denoted by $L^1_{\text{loc}}(\mathbb{R}^d)$.

APPENDIX B: PROOFS FROM SECTION 3

B.1 Proof of Theorem 3.1

PROOF. The main idea behind our proof is to exploit the natural representation of $\mathcal{Z}_K(n)$ as the expectation of a non-negative random variable with respect to the prior measure. Specifically, let $T = K(\xi)$, where $\xi \sim \varphi$ is a random variable distributed according to the prior measure. Then, the non-negative random variable T takes values in the unit interval $[0, 1]$ and $\mathcal{Z}_K(n) = \mathbb{E}[e^{-nT}] = \int_0^1 e^{-nt}\varphi_K(t)dt$, where φ_K is the density function of T to be derived below. Before proceeding to simplify this expectation, we note some conventions and notation. Let $\bar{d} = \sum_{j=1}^d \mathbb{1}(k_j \neq 0)$, and without loss of generality assume that $k_j > 0$ for $j = 1, \ldots, \bar{d}$ and $k_j = 0$ for $j > \bar{d}$. Define $\lambda_j := (h_j + 1)/(2k_j)$ for $j = 1, \ldots, \bar{d}$, and without loss of generality, further assume that these are sorted in non-decreasing order $\lambda_1 \leq \lambda_2 \ldots \leq \lambda_{\bar{d}}$. By definition, $\bar{d} \geq m$, and the first m of the λ_js all equal λ. Throughout, we use the convention that an Expo(β) distribution has density $\beta e^{-\beta x}\mathbb{1}_{(0,\infty)}(x)$, that is, β denotes the rate parameter of the distribution.

The random variable $Z := -\log T$ can be expressed as $Z = \sum_{j=1}^{\bar{d}} Z_j$ with $Z_j = -\log(\xi_j^{2k_j})$ for $j = 1, \ldots, \bar{d}$. An application of the change of measure formula yields that $Z_j \sim \text{Expo}(\lambda_j)$ with $\lambda_j = (h_j + 1)/(2k_j)$ as defined above; interestingly, observe the quantities $(h_j + 1)/(2k_j)$s appear as the exponential rate parameters. Moreover, since the prior measure φ has a product form, the Z_js are independent across j. Letting $\Phi_K(\cdot)$ denote the cumulative distribution function of T, we then have, for any $t \in (0, 1)$,

$$\Phi_K(t) = P(T \leq t) = P(-\log T \geq -\log t) \tag{27}$$
$$= P\left(\sum_{j=1}^{\bar{d}} Z_j \geq \log(1/t)\right).$$

It follows from the above display that $\lim_{t\downarrow 0} \Phi_K(t) = 0$, $\lim_{t\uparrow 1} \Phi_K(t) = 1$, and Φ_K is an absolutely continuous cdf

that admits a density $\varphi_K(\cdot)$ with respect to the Lebesgue measure, given by,

$$\varphi_K(t) = \frac{1}{t}\, g_Z\big(\log(1/t)\big)\, \mathbb{1}_{(0,1)}(t), \tag{28}$$

where g_Z is the density of Z with respect to the Lebesgue measure. Our object of interest,

$$\begin{aligned}\mathcal{Z}_K(n) &= \int_0^1 e^{-nt}\varphi_K(t)\, dt \\ &= \int_0^n e^{-t}\, \frac{1}{t}\, g_Z\big(\log(n/t)\big)\, dt.\end{aligned} \tag{29}$$

Before proceeding to prove the theorem in its entire generality, we consider two special cases which are instructive in themselves and also help build towards the general proof.

First, consider the special case where $\lambda_j = \lambda$ for all $j = 1, \ldots, \bar{d}$. Then, $m = \bar{d}$ and $Z \sim \text{Gamma}(m, \lambda)$, where a $\text{Gamma}(\alpha, \beta)$ distribution has density

$$\big(\beta^\alpha/\Gamma(\alpha)\big)\, e^{-\beta x} x^{\alpha-1}\, \mathbb{1}_{(0,\infty)}(x).$$

It follows that for any $t \in (0, n)$,

$$g_Z\big(\log(n/t)\big) = \frac{\lambda^m}{\Gamma(m)}\, n^{-\lambda}\, t^{\lambda}\big(\log(n/t)\big)^{m-1}.$$

Substituting in equation (29), we obtain that

$$\begin{aligned}\mathcal{Z}_K(n) &= C n^{-\lambda} \underbrace{\int_0^n t^{\lambda-1}\, e^{-t}\, \big(\log(n/t)\big)^{m-1}\, dt}_{\mathcal{I}_m(n)} \\ &\asymp n^{-\lambda}\, (\log n)^{m-1}.\end{aligned} \tag{30}$$

The proof of the assertion that $\mathcal{I}_m(n) \asymp (\log n)^{m-1}$ for any $m \geq 1$ is straightforward and hence omitted. This completes the proof for this particular case.

As a second special case, suppose $\lambda_1 < \ldots < \lambda_{\bar{d}}$, which implies that $\lambda = \lambda_1$ and $m = 1$. This is known as the Hypoexponential distribution [35, 23] and an analytic expression for its density is available in the literature as quoted below.

THEOREM B.1 ([26, 8]). *Let $Z_k \overset{ind.}{\sim} Expo(\lambda_k)$ for $k = 1, \ldots, K$, with $\lambda_1 < \ldots < \lambda_K$. Then, the density g_Z of $Z = \sum_{k=1}^K Z_k$ is*

$$g_Z(z) = \sum_{k=1}^K \underbrace{\left(\prod_{r \neq k} \frac{\lambda_r}{\lambda_r - \lambda_k}\right)}_{b_k} g_k(z),$$

where $g_k(z) = \lambda_k e^{-\lambda_k z}$ is the density of Z_k.

The coefficients $\{b_k\}$ can be both positive and negative, and thus the above is not a mixture of exponential densities. However, the coefficient b_1 corresponding to the smallest rate parameter λ_1 is positive. We have, for any $t \in (0, n)$,

$$g_Z\big(\log(n/t)\big) = \sum_{j=1}^{\bar{d}} b_j\, \lambda_j\, n^{-\lambda_j}\, t^{\lambda_j}.$$

Substituting this expression in equation (B), we get

$$\begin{aligned}\mathcal{Z}_K(n) &= \sum_{j=1}^{\bar{d}} b_j\, \lambda_j\, n^{-\lambda_j} \int_0^n e^{-t} t^{\lambda_j - 1} dt \\ &\asymp \sum_{j=1}^{\bar{d}} b_j n^{-\lambda_j} \asymp n^{-\lambda_1}.\end{aligned} \tag{31}$$

This proves the theorem for this special case. The fact that $b_1 > 0$ has been crucially used to arrive at the last conclusion in the above display, along with the fact that $n^{-\alpha_1} > n^{-\alpha_2}$ for $\alpha_2 > \alpha_1 > 0$. This example carries the takeaway message that the exact form of the density g_Z is of secondary importance, and the focus should be on extracting the most significant contribution in terms of n. This is our strategy for the most general case.

In the general case, assume that there are $d^* \leq \bar{d}$ unique λ-values $\lambda_1^* < \lambda_2^* \ldots < \lambda_{d^*}^*$ among $\{\lambda_j\}_{j=1}^{\bar{d}}$ with corresponding multiplicities $m_1, \ldots, m_{d^*}$. It is then immediate that $\sum_{s=1}^{d^*} m_s = \bar{d}$. Also, $(\lambda_1^*, m_1) = (\lambda, m)$ from the theorem statement. Exploiting the independence of the Z_js, we write $Z = \sum_{s=1}^{d^*} W_s$, with $W_s \overset{ind.}{\sim} \text{Gamma}(m_s, \lambda_s^*)$ for $s = 1, \ldots, d^*$. While there exist expressions for the density of sum of independent Gamma random variables [26], they are much more cumbersome than the simpler case of exponentials in Theorem 3.2. Hence, we do not attempt to work with the density g_Z and instead aim to bound $\mathcal{Z}_K(n)$ from both sides. To that end, we crucially use the idea of stochastic ordering of random variables.

Recall that for real random variables X_1, X_2, X_1 is said to be stochastically smaller than X_2 if for every $x \in \mathbb{R}$, $P(X_2 > x) \geq P(X_1 > x)$. We use the notation $X_1 <_{\text{st}} X_2$ to denote this stochastic ordering. We now record a useful result.

LEMMA B.2. *Consider the random variable $Z = \sum_{s=1}^{d^*} W_s$, with $W_s \overset{ind.}{\sim} \text{Gamma}(m_s, \lambda_s^*)$. Assume $\lambda_1^* < \ldots < \lambda_{d^*}^*$ and let $\bar{d} = \sum_{s=1}^{d^*} m_s$. Define $Z_\ell = W_1$ and $Z_c = \sum_{s=2}^{d^*} \widetilde{W}_s$, where $\widetilde{W}_s \overset{ind.}{\sim} \text{Gamma}(m_s, \lambda_2^*)$ are also independent of W_1. Then, $Z_\ell \sim \text{Gamma}(m_1, \lambda_1^*)$, $Z_c \sim \text{Gamma}(\bar{d} - m_2, \lambda_2^*)$, Z_ℓ and Z_c are independent, and with $Z_u := Z_\ell + Z_c$,*

$$Z_\ell <_{\text{st}} Z <_{\text{st}} Z_u.$$

With this result in place, we now aim to bound $\mathcal{Z}_K(n) = Ee^{-nT}$. Since e^{-nT} is a non-negative random variable taking values in $(0,1)$, we have

$$Ee^{-nT} = \int_{u=0}^{1} P\big(e^{-nT} > u\big)du \tag{32}$$
$$= \int_{u=0}^{1} P\big(T < \log(1/u)/n\big)du$$
$$= \int_{u=0}^{1} P\big(Z > \log n - \log(\log 1/u)\big)du.$$

For any $z > 0$, we have the following two-sided inequality from Lemma B.2,

$$P(Z_\ell > z) < P(Z > z) < P(Z_u > z) < P(Z_\ell > z) + P(Z_c > z).$$

Here, the last inequality follows from an application of the union bound. Substituting this inequality at the end of equation (32) for every u and working backwards, we obtain

$$Ee^{-nT_\ell} < Ee^{-nT} < Ee^{-nT_\ell} + Ee^{-nT_c},$$

where $T_\ell = e^{-Z_\ell}$ and $T_c = e^{-Z_c}$. Since Z_ℓ and Z_c are both gamma random variables, it follows from equation (13) that $Ee^{-nT_\ell} > Cn^{-\lambda}(\log n)^{m-1}$ and $Ee^{-nT_\ell} + Ee^{-nT_c} < C_1 n^{-\lambda}(\log n)^{m-1} + C_2 n^{-\lambda_2}(\log n)^{m_2-1} < C_3 n^{-\lambda}(\log n)^{m-1}$. This delivers the desired bound. □

B.2 Proof of Theorem 3.4

First part: By abuse of notation we shall assume that φ corresponds to a product Beta density $\prod_{j=1}^{d}\text{Beta}(\xi_j \mid h_j+1, 1)$. Multiplying $Z(n)$ by $\prod_{j=1}^{d}(h_j+1)$ we have

$$\mathcal{Z}(n)\prod_{j=1}^{d}(h_j+1) = \int e^{-nK(\xi)-\sqrt{n}\xi^k W_n(\xi)}\varphi(\xi)d\xi$$
$$= \int_0^1 e^{-nt}\Big[\int e^{-\sqrt{nt}W_n(\xi)}\varphi_{\xi|T=t}(\xi_I)\varphi(\xi_J)d\xi\Big]\varphi_K(t)dt$$

where $\varphi_K(t)$ is the density of $T = K(\xi)$ as in §3.1. Substituting $\varphi_K(t) = (1/t)g_Z\big\{\log(1/t)\big\}$, where $g_Z(\cdot)$ is the pdf of a Gamma(λ, m) random variable, we have by another change of variable, $nt \mapsto t$ that,

$$\mathcal{Z}(n)\prod_{j=1}^{d}(h_j+1) =$$
$$\int_0^n e^{-t}\frac{1}{t}g_Z\Big(\log\frac{n}{t}\Big)\Big[\int e^{-\sqrt{t}W_n(\xi)}\varphi_{\xi|T=t/n}(\xi_I)\varphi(\xi_J)d\xi\Big]dt.$$

Noting,

$$g_Z\big(\log(n/t)\big) = \frac{\lambda^m}{\Gamma(m)}n^{-\lambda}t^{\lambda}\big(\log(n/t)\big)^{m-1}$$
$$= \frac{\lambda^m}{\Gamma(m)}n^{-\lambda}t^{\lambda}\big(\log n\big)^{m-1} + r_n(t)t^{\lambda}$$

it follows

$$\mathcal{Z}(n)\prod_{j=1}^{d}(h_j+1)$$
$$= \frac{\lambda^m}{\Gamma(m)}n^{-\lambda}\big(\log n\big)^{m-1}\int_0^n\int D_n(t,\xi)d\xi dt + R_n.$$

Second part: Since the maximum exponent of all logarithmic terms inside the summation is $(m-2)$ we have

$$\frac{|r_n(t)|}{n^{-\lambda}(\log n)^{m-1}} \le C(m)(\log n)^{-1}\sum_{j=1}^{m-2}|\log t|^{m-1-j}$$

for some constant $C(m)$ depending on m. Also $D_n(t) \le t^{\lambda-1}e^{-t+\sqrt{t}\|W_n\|_\infty}$ and hence

$$\frac{|R_n|}{n^{-\lambda}(\log n)^{m-1}} \le$$
$$\frac{C(m)}{\log n}\sum_{j=1}^{m-2}\int_{t=0}^{\infty}e^{-t+\sqrt{t}\|W_n\|_\infty}t^{\lambda-1}|\log t|^{m-1-j}dt$$

Since the function $e^{-t+\sqrt{t}\|W_n\|_\infty}t^{\lambda-1}|\log t|^{m-1-j}$ is integrable and $\|W_n\|_\infty = O_p(1)$, the result follows immediately.

B.3 Proof of Lemma B.2

For $G \sim \text{Gamma}(\alpha,\lambda)$ and any $t > 0$,

$$P(G \le t) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^t e^{-\lambda x}x^{\alpha-1}dx = \frac{1}{\Gamma(\alpha)}\int_0^{\lambda t}e^{-x}x^{\alpha-1}dx.$$

is an increasing function of λ (for fixed α and t). Thus, if $Z_i \sim \text{Gamma}(\alpha,\lambda_i)$ for $i = 1,2$ with $\lambda_1 > \lambda_2$, then $Z_1 <_{\text{st}} Z_2$.

The proof then follows from the fact that if Z_1, Z_2, Z_3 are non-negative random variables with $Z_1 <_{\text{st}} Z_2$, then $Z_1 + Z_3 <_{\text{st}} Z_2 + Z_3$.

B.4 Proof of Corollary 2.1

Since $b(\cdot)$ is analytic on U containing Ω, we have, for any $\xi \in \Omega$ that

$$b(\xi) = b(\mathbf{0}) + \sum_{|\alpha|\ge 1}\frac{\partial^\alpha b(\mathbf{0})}{\alpha!}\xi^\alpha,$$

where $\alpha = (\alpha_1,\ldots,\alpha_d)$ is a multi-index with $|\alpha| = \sum_{j-1}^{d}\alpha_j$, $\partial^\alpha b = \partial^{\alpha_1}\ldots\partial^{\alpha_d}b$, and $\alpha! = \alpha_1!\ldots\alpha_d!$. Now use the dominated convergence theorem to interchange the integral and sum, and observe that the constant term provides the dominating order.

B.5 Proof of Proposition B.3

We show that $W_n(\xi)$ converges weakly to the Gaussian process W^*. By Theorem 1.5.7 in [37] it suffices to show the marginal weak convergence and asymptotic

tightness of $W_n(\xi)$. We begin with the convergence of the marginals. For $\xi_1, \ldots, \xi_L \in [0,1]^d$ with integer $L > 0$. Applying the multivariate central limit theorem, as $n \to \infty$,

$$(W_n(\xi_1), \ldots, W_n(\xi_L)) \to N(0, C)$$

where $C = (c_w(\xi_i, \xi_j))_{1 \le i,j \le L}$. Next we show the asymptotic tightness of $W_n(\xi)$ by proving the following three sufficient conditions. First $[0,1]^d$ is totally bounded. The second condition is the tightness of $W_n(\xi_0)$ for a fixed ξ_0. Fix $\xi_0 \in [0,1]^d$, for $\epsilon > 0$. We need to show that there exists a compact set K, such that $\mathbb{P}\{W_n(\xi_0) \in K\} > 1 - \epsilon$. We construct $K = \{|W_n(\xi_0)| \le t\}$ with t chosen as follows. By **Assumption A1**, $\tilde{Z}_i(\xi_0)$ are independent centered sub-Gaussian and by Theorem 2.6.2 of [38]

$$\mathbb{P}(|W_n(\xi_0)| \ge t) \le \mathbb{P}\left(\left| \sum_{i=1}^n \tilde{Z}_i(\xi_0) \right| \ge \sqrt{n} t \right) \le 2\exp(-ct^2)$$

for some constant $c > 0$. Choosing $t = \sqrt{2\log(1/\epsilon)}$ completes the proof of tightness.

The third condition is that $W_n(\xi)$ is asymptotically uniformly d-equicontinuous, where $d(\xi, \zeta) = \|\xi - \zeta\|$ is the metric generated by the norm in **Assumption A1**. $W_n(\xi)$ is said to be asymptotically uniformly d-equicontinuous if for any $\eta, \epsilon > 0$, there exists a $\delta > 0$ such that

$$\mathbb{P}\left\{ \sup_{d(\xi,\zeta)<\delta} |W_n(\xi) - W_n(\zeta)| > \epsilon \right\} < \eta.$$

To that end,

$$\sup_{d(\xi,\zeta)<\delta} |W_n(\xi) - W_n(\zeta)|$$

$$\le \frac{1}{\sqrt{n}} \sum_{i=1}^n |\tilde{Z}_i(\xi) - \tilde{Z}_i(\zeta)| \le \frac{\delta}{\sqrt{n}} \sum_{i=1}^n L(X_i)$$

and

$$\mathbb{P}\Big\{ \sup_{d(\xi,\zeta)<\delta} |W_n(\xi) - W_n(\zeta)| > \epsilon \Big\}$$

$$\le \mathbb{P}\Big\{ \sum_{i=1}^n L(X_i) > \frac{\sqrt{n}\epsilon}{\delta} \Big\}$$

$$= \mathbb{P}\Big[\exp\Big\{ t \sum_{i=1}^n L(X_i) \Big\} > \exp\Big(t\frac{\sqrt{n}\epsilon}{\delta} \Big) \Big]$$

$$\le \exp\big\{ -t\sqrt{n}\epsilon/\delta + ntc_L^2/2 \big\}$$

for any $t > 0$, where the final inequality follows from Markov's and **Assumption A1**. Setting $t = \epsilon/(\delta\sqrt{n}c_L^2)$, we obtain

$$\mathbb{P}\Big\{ \sup_{d(\xi,\zeta)<\delta} |W_n(\xi) - W_n(\zeta)| > \epsilon \Big\} \le e^{-\epsilon^2/(2c_L^2\delta^2)}.$$

Choosing $\delta = \epsilon/(c_L\sqrt{2\log(1/\eta)})$ completes the proof of asymptotically uniformly d-equicontinuous. Therefore the conditions of Theorem 1.5.7 in [37] are met and $W_n(\xi)$ converges weakly to a Gaussian process.

B.6 Proof of Proposition B.4

We shall prove only the second part; the proof of the first part is very similar and is omitted. We use the notation

$$G(t, W) := t^{\lambda-1} e^{-t-\sqrt{t}W}.$$

Observe that $\int_0^n |D_n(t) - D(t)|dt$ is bounded above by

$$\int_0^n \Big\{ \int |G(t, W^*(\mathbf{0}, \xi_J)) - G(t, W^*(\xi_I, \xi_J))| \tag{33}$$

$$\varphi_{\xi|Z=-\log(t/n)}(\xi_I)\varphi(\xi_J) \Big\} d\xi dt +$$

$$\int_0^n \Big\{ \int |G(t, W_n(\xi_I, \xi_J)) - G(t, W^*(\xi_I, \xi_J))|$$

$$\varphi_{\xi|Z=-\log(t/n)}(\xi_I)\varphi(\xi_J) d\xi \Big\} dt$$

To control the first term, for given any $\epsilon > 0$, there exists $\delta > 0$, such that $\sup_{\{\|\xi_I\|<\delta\}} |e^{-\sqrt{t}W^*(\mathbf{0},\xi_J)} - e^{-\sqrt{t}W^*(\xi_I,\xi_J)}| < \epsilon$. Then the first term is less than

$$\epsilon + 2\int_0^n t^{\lambda-1} e^{-t+\sqrt{t}\|W^*\|_\infty} \Big[\tag{34}$$

$$\int_{\{\|\xi_I\| \ge \delta\}} \varphi_{\xi|Z=-\log(t/n)}(\xi_I) d\xi_I \Big] dt$$

Observe that the second term in the r.h.s of (34) is bounded above by

$$(m-1)\mathbb{P}\{\xi_1 > \delta/\sqrt{m-1} \mid Z = -\log(t/n)\}.$$

The one dimensional marginal $\xi_1 \mid Z$ of the conditional density (17) is given by

$$\varphi_{\xi_1|Z}(\xi_1) = \frac{2k_1}{\xi_1 Z}, \quad e^{-Z} \le \xi_1^{2k_1} \le 1.$$

Note that the sequence of random variables

$$f_n(\xi_1) = \mathbb{1}(\xi_1 > \delta/\sqrt{m-1})\varphi_{\xi_1|Z=-\log(t/n)}(\xi_1)$$

converges to zero and is bounded above by the integrable function $\varphi_{\xi_1|Z=-\log(t/n)}(\xi_1)$, hence an application of the dominated convergence theorem shows $\int f_n(\xi_1)d\xi_1$ converges to 0. Another application of DCT shows that the second term in (34) converges to 0.

The second term of (34) can be bounded above by

$$\int_0^\infty t^{\lambda-1} e^{-t} \Big\{ \int |e^{\sqrt{t}W_n(\xi_I,\xi_J)}$$

$$- e^{\sqrt{t}W^*(\xi_I,\xi_J)}| \varphi_{\xi|Z=-\log(t/n)}(\xi_I)\varphi(\xi_J) d\xi \Big\} dt$$

Since $W_n \xrightarrow{w} W^*$, by continuous mapping, the above converges weakly to 0.

Finally, we can reach

$$\int_n^\infty D(t)dt =$$

$$\int_{t=n}^{\infty} t^{\lambda-1} \int e^{-t-\sqrt{t}W^*(\mathbf{0},\xi_J)} \varphi_{\xi|Z=-\log(t/n)}(\xi_I)\varphi(\xi_J)d\xi_I dt$$
$$\leq \int_{t=n}^{\infty} t^{\lambda-1} e^{-t+\sqrt{t}\|W^*\|_\infty} dt$$

which converges to 0, concluding the proof.

B.7 Proof of asymptotic equivalence

An important ingredient of making this connection is to show a weak convergence of the sequence of stochastic processes W_n. The expected value and the covariance of $Z_i(\xi)$ are $\mathbb{E}Z_i(\xi) = 0, \text{cov}[Z_i(\xi), Z_i(\zeta)] := c_z(\xi, \zeta)$, respectively. For $W_n(\xi)$, the same quantities are given by

$$\mathbb{E}W_n(\xi) = \frac{\xi^{-\mathrm{k}}}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{E}Z_i(\xi) = 0,$$

$$\text{cov}[W_n(\xi), W_n(\zeta)] = \frac{c_z(\xi,\zeta)}{\xi^{\mathrm{k}}\zeta^{\mathrm{k}}} := c_w(\xi,\zeta).$$

Let W^* denote a mean zero Gaussian process on $\Omega = [0,1]^d$ with covariance kernel c_w. Under appropriate conditions on the stochastic processes $\{\tilde{Z}_i(\xi) = \xi^{-\mathrm{k}} Z_i(\xi) : \xi \in [0,1]^d\}$, we show in Proposition B.3 that W_n weakly converges to W^*. The proof requires sub-Gaussianity [38] of $\tilde{Z}_i(\xi)$. [12]

Assumption A1: Suppose that $\tilde{Z}_i(\xi)$ are iid sub-Gaussian. Furthermore suppose that there exists a positive function $L : \mathbb{R} \to \mathbb{R}^+$ with $\mathbb{E}e^{tL(X_1)} \leq e^{t^2/(2c_L)}$ for every $t > 0$ and for some constant $c_L > 0$, such that

$$|\tilde{Z}_i(\xi) - \tilde{Z}_i(\xi')| \leq L(X_i)\|\xi - \xi'\|.$$

PROPOSITION B.3. *Under Assumption* **A1**, $W_n \xrightarrow{w} W^*$.

Sections 10.4 and 10.5 of [52] provide heuristic arguments to study weak convergence of W_n; the assumptions require $\xi^{-\mathrm{k}} Z_i(\xi)$ to be at least $d/2+1$ times differentiable. On the other hand, our **Assumption A1** requires $\tilde{Z}_i(\xi)$ to be Lipschitz and sub-Gaussian.

The next ingredient in establishing the connection is to establish a weak limit of $D_n(t) = \int D_n(t,\xi)d\xi$. In Proposition B.4, we show that for each $t > 0$, $D_n(t)$ converges weakly to the following fixed stochastic process

$$D(t) = \int t^{\lambda-1} e^{-t-\sqrt{t}W^*(\mathbf{0},\xi_J)} \varphi(\xi_J) d\xi_J.$$

In addition, we also show in Proposition B.4 that $\int_0^n D_n(t)dt$ converges in distribution to $\int_0^\infty D(t)dt$.

[12] A real valued random variable X is called sub-Gaussian if there exists a constant $c_X > 0$ such that $\mathbb{P}(|X| > t) \leq 2e^{-t^2/(2c_X)}, \quad t \geq 0$.

PROPOSITION B.4. *If $W_n \xrightarrow{w} W^*$, $\|W^*\|_\infty = O_p(1)$ and $\xi_I \mapsto W^*(\xi_I, \xi_J)$ is almost surely continuous, then for each $t > 0$, $D_n(t) \xrightarrow{w} D(t)$ and $\int_0^n D_n(t)dt \xrightarrow{w} \int_0^\infty D(t)dt$.*

Using Theorem 3.4 and Proposition B.4,

(35)
$$\mathcal{Z}(n) \prod_{j=1}^{d} (h_j + 1) \sim n^{-\lambda} (\log n)^{m-1} \times$$
$$\frac{\lambda^m}{\Gamma(m)} \int_0^\infty t^{\lambda-1} e^{-t+\sqrt{t}W^*(\mathbf{0},\xi_J)} \varphi(\xi_J)\varphi(\xi_J) dt$$

(36)
$$\mathcal{Z}(n) \sim \frac{n^{-\lambda}(\log n)^{m-1}}{2^m (m-1)! \prod_{j=1}^m k_j} \int_0^\infty t^{\lambda-1} e^{-t+\sqrt{t}W^*(\mathbf{0},\xi_J)} \xi_J^{h_J} dt,$$

where $h_J = (h_{m+1}, \ldots, h_d)$. Using the properties of the Dirac delta function, we can write

$$D(t) = \int t^{\lambda-1} e^{-t-\sqrt{t}W^*(\xi)} \delta_{\mathbf{0}}(\xi_I)\varphi(\xi_J) d\xi.$$

where $\delta_{\mathbf{0}}(\xi_I)$ is a Dirac delta measure at $\mathbf{0}$, which also appears in Theorem 11 in [50] in the expansion of $\mathcal{Z}(n)$. Observing that that $k_j = 0$ for $j = m+1, \ldots d$, (35) exactly matches with the equation (5.32) of Theorem 10 or the expression under Theorem 11 in [50].

APPENDIX C: REMAINING PROOFS FROM SECTION 4

C.1 Proof of Lemma 4.3

Part (i) follows from a change of variable $v = \beta u^{2k}$. For part (ii), we have, using the definition of $B(k,h,\beta)$,

$$G(\lambda,\beta) = \frac{B(k,2k+h,\beta)}{B(k,h,\beta)}$$
$$= \frac{(2k)^{-1}\beta^{-(\lambda+1)}\Gamma(\lambda+1)\gamma(\lambda+1,\beta)}{(2k)^{-1}\beta^{-(\lambda)}\Gamma(\lambda)\gamma(\lambda,\beta)}$$
$$= \frac{\lambda}{\beta}\frac{\gamma(\lambda+1,\beta)}{\gamma(\lambda,\beta)}.$$

For the first part of part (iii), we have $\log B(k,h,\beta) = -\lambda \log\beta + \log\gamma(\lambda,\beta) +$ terms free of β. The conclusion follows since $\lim_{\beta\to\infty}\gamma(\lambda,\beta) = 1$.

For the second part of part (iii), use Remark 4.2 to write

$$G(\lambda,\beta) = \frac{\lambda}{\beta}\left(1 - \frac{\beta^\lambda e^{-\beta}}{\Gamma(\lambda+1)\gamma(\lambda,\beta)}\right).$$

From the above, the conclusion follows since $\lim_{\beta\to\infty}\beta^\lambda e^{-\beta} = 0$ and $\lim_{\beta\to\infty}\gamma(\lambda,\beta) = 1$.

For the first part of (iv) use (i) and Remark 4.2 to compute $\lim_{\beta\to 0}\log B(k,h,\beta)$.

For the second part (iv), use (ii) and Remark 4.2 to compute $\lim_{\beta\to 0} G(\lambda,\beta)$.

C.2 Derivation of Eq. (24)

It follows from variational calculus that the optimal marginals of the mean-field approximation to $\gamma_K(\xi)$ are of the form

$$\rho_j(\xi_j) \propto \exp\{\mathbb{E}_{-j}[\log \gamma_K(\xi)]\},\ 1 \le j \le d,$$

where $\mathbb{E}_{-j}$ denotes the expectation with respect to

$$\rho_{-j}(\xi_{-j}) := \prod_{k \neq j} \rho_k(\xi_k).$$

A straightforward computation shows that for $d = 2$ the optimal marginals are of the form, $\rho_1(\xi_1) \propto \exp\{-n\mu_1 \xi_1^{2k_1}\}$ and $\rho_2(\xi_2) \propto \exp\{-n\mu_2 \xi_2^{2k_2}\}$, where $\mu_1 = \int_0^1 \xi_2^{2k_2} \rho_2(\xi_2) d\xi_2$ and $\mu_2 = \int_0^1 \xi_1^{2k_1} \rho_1(\xi_1) d\xi_1$. The identities $\int_0^1 \xi_2^{2k_2} \rho_2(\xi_2) d\xi_2 = G(\lambda_2, n\mu_2)$ and $\int_0^1 \xi_2^{2k_1} \rho_2(\xi_1) d\xi_1 = G(\lambda_1, n\mu_1)$ follow from a simple change of variables. Finally the CAVI algorithm updates at time step t are given by

$$\rho_1^{(t)}(\xi_1) = \frac{\exp\{-n\mu_1^{(t-1)} \xi_1^{2k_1}\}}{\int_0^1 \exp\{-n\mu_1 \xi_1^{2k_1}\} d\xi_1} = f_{k_1, h_1, n\mu_1^{(t-1)}}(\xi_1),$$

$$\rho_2^{(t)}(\xi_2) = \frac{\exp\{-n\mu_2^{(t)} \xi_2^{2k_2}\}}{\int_0^1 \exp\{-n\mu_2 \xi_2^{2k_2}\} d\xi_2} = f_{k_2, h_2, n\mu_2^{(t)}}(\xi_2).$$

ACKNOWLEDGMENTS

The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

FUNDING

AB acknowledges NSF DMS 2210689, NSF DMS 1916371, and NSF CAREER 1653404 for partially supporting this project. DP acknowledges NSF DMS 2210689 and NSF DMS 1916371 for partially supporting this project. YY acknowledges NSF DMS 2210717 for partially supporting this project.

REFERENCES

[1] ABRAMOWITZ, M. and STEGUN, I. A. (1964). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables* **55**. Courier Corporation.

[2] ACQUISTAPACE, F., BROGLIA, F. and FERNANDO, J. F. (2022). *Topics in global real analytic geometry*. Springer.

[3] AOYAGI, M. (2010). Stochastic complexity and generalization error of a restricted Boltzmann machine in Bayesian estimation. *Journal of Machine Learning Research* **11** 1243–1272.

[4] AOYAGI, M. (2019). Learning Coefficient of Vandermonde Matrix-Type Singularities in Model Selection. *Entropy* **21** 561.

[5] AOYAGI, M. and WATANABE, S. (2005). Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Networks* **18** 924 - 933.

[6] AOYAGI, M. and WATANABE, S. (2005). Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Networks* **18** 924–933.

[7] ARNOLD, V. I., VARCHENKO, A. N. and GUSEIN-ZADE, S. M. (2012). *Singularities of differentiable maps: Volume II Monodromy and asymptotic integrals* **83**. Springer Science & Business Media.

[8] BIBINGER, M. (2013). Notes on the sum and maximum of independent exponentially distributed random variables with different scale parameters. *arXiv preprint arXiv:1307.3945*.

[9] BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning*. *Information Science and Statistics*. Springer.

[10] BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* **just-accepted**.

[11] COOPER, J. L. B. (1966). Laplace Transformations of Distributions. *Canadian Journal of Mathematics* **18** 1325–1332.

[12] COX, D., LITTLE, J. and OSHEA, D. (2013). *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra*. Springer Science & Business Media.

[13] DRTON, M., LIN, S., WEIHS, L. and ZWIERNIK, P. (2017). Marginal likelihood and model selection for Gaussian latent tree and forest models. *Bernoulli* **23** 1202–1232. https://doi.org/10.3150/15-BEJ775

[14] DRTON, M. and PLUMMER, M. (2017). A Bayesian information criterion for singular models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79** 323–380.

[15] FOLLAND, G. B. (1999). *Real analysis: modern techniques and their applications* **40**. John Wiley & Sons.

[16] FRIEDLANDER, F. G., JOSHI, M. S., JOSHI, M. and JOSHI, M. C. (1998). *Introduction to the Theory of Distributions*. Cambridge University Press.

[17] FULTON, W. (1989). *Algebraic Curves: An Introduction to Algebraic Geometry*. Addison-Wesley Publishing Company, Advanced Book Program.

[18] GHORBANI, B., JAVADI, H. and MONTANARI, A. (2018). An instability in variational inference for topic models. In *International Conference on Machine Learning*.

[19] HAYASHI, N. and WATANABE, S. (2017). Tighter upper bound of real log canonical threshold of non-negative matrix factorization and its application to Bayesian inference. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)* 1–8. IEEE.

[20] HOSINO, T., WATANABE, K. and WATANABE, S. (2005). Stochastic complexity of variational Bayesian hidden Markov models. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* **2** 1114–1119. IEEE.

[21] HOSINO, T., WATANABE, K. and WATANABE, S. (2006). Free Energy of Stochastic Context Free Grammar on Variational Bayes. In *ICONIP*.

[22] KASS, R. E., TIERNEY, L. and KADANE, J. B. (1990). The validity of posterior expansions based on Laplace's method. In *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard* (S. Geisser, J. Hodges, S. Press and A. Zellner, eds.) Elsevier Science Publishers B. V. (North Holland).

[23] LEVY, E. (2022). On the density for sums of independent exponential, Erlang and gamma variates. *Statistical Papers* **63** 693–721.

[24] LIN, S. (2011). Algebraic methods for evaluating integrals in Bayesian statistics, PhD thesis, UC Berkeley.

[25] MACKAY, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

[26] MATHAI, A. (1982). Storage capacity of a dam with gamma type inputs. *Annals of the Institute of Statistical Mathematics* **34** 591–597.

[27] MUKHERJEE, S. S., SARKAR, P., WANG, Y. R. and YAN, B. (2018). Mean field for the stochastic blockmodel: optimization landscape and convergence issues. In *Advances in Neural Information Processing Systems* 10694–10704.

[28] NAKAJIMA, S., SATO, I., SUGIYAMA, M., WATANABE, K. and KOBAYASHI, H. (2014). Analysis of variational bayesian latent dirichlet allocation: Weaker sparsity than MAP. *Advances in neural information processing systems* **27**.

[29] NAKAJIMA, S., WATANABE, K. and SUGIYAMA, M. (2019). *Variational Bayesian Learning Theory*. Cambridge University Press.

[30] NAKAJIMA, S. and WATANABE, S. (2007). Variational Bayes solution of linear neural networks and its generalization performance. *Neural Computation* **19** 1112–1153.

[31] PATI, D., BHATTACHARYA, A. and YANG, Y. (2018). On statistical optimality of variational Bayes. In *International Conference on Artificial Intelligence and Statistics* 1579–1588.

[32] ROBERT, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

[33] RUSAKOV, D. and GEIGER, D. (2005). Asymptotic model selection for naive Bayesian networks. *Journal of Machine Learning Research* **6** 1–35.

[34] SCHWARZ, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6** 461–464.

[35] SMAILI, K., KADRI, T. and KADRY, S. (2013). Hypoexponential distribution with different parameters.

[36] TIERNEY, L. and KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81** 82–86.

[37] VAN DER VAART, A. W. and WELLNER, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes* 16–28. Springer.

[38] VERSHYNIN, R. (2018). High-Dimensional Probability: An Introduction with Applications in Data Science. 2018. *URL https://www. math. uci. edu/˜ rvershyn/papers/HDP-book/HDP-book. pdf*.

[39] WAINWRIGHT, M. J. and JORDAN, M. I. (2008). *Graphical models, exponential families, and variational inference*. Now Publishers Inc.

[40] WANG, B., TITTERINGTON, D. et al. (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis* **1** 625–650.

[41] WATANABE, K., SHIGA, M. and WATANABE, S. (2009). Upper bound for variational free energy of Bayesian networks. *Machine Learning* **75** 199–215.

[42] WATANABE, K. and WATANABE, S. (2004). Lower bounds of stochastic complexities in variational Bayes learning of Gaussian mixture models. In *IEEE Conference on Cybernetics and Intelligent Systems, 2004.* **1** 99-104 vol.1.

[43] WATANABE, K. and WATANABE, S. (2006). Stochastic complexities of Gaussian mixtures in variational Bayesian approximation. *Journal of Machine Learning Research* **7** 625–644.

[44] WATANABE, K. and WATANABE, S. (2007). Stochastic complexity for mixture of exponential families in generalized variational Bayes. *Theoretical computer science* **387** 4–17.

[45] WATANABE, K. and WATANABE, S. (2007). Stochastic complexities of general mixture models in variational Bayesian learning. *Neural Networks* **20** 210 - 219.

[46] WATANABE, S. (1999). Algebraic analysis for singular statistical estimation. In *International Conference on Algorithmic Learning Theory* 39–50. Springer.

[47] WATANABE, S. (2001). Algebraic analysis for nonidentifiable learning machines. *Neural Computation* **13** 899–933.

[48] WATANABE, S. (2001). Algebraic geometrical methods for hierarchical learning machines. *Neural Networks* **14** 1049–1060.

[49] WATANABE, S. (2007). Almost All Learning Machines are Singular. In *2007 IEEE Symposium on Foundations of Computational Intelligence* 383-388.

[50] WATANABE, S. (2009). *Algebraic geometry and statistical learning theory* **25**. Cambridge University Press.

[51] WATANABE, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research* **14** 867–897.

[52] WATANABE, S. (2018). *Mathematical theory of Bayesian statistics*. CRC Press.

[53] WEI, S. and LAU, E. (2023). Variational Bayesian Neural Networks via Resolution of Singularities. *arXiv preprint arXiv:2302.06035*.

[54] YAMAZAKI, K. and WATANABE, S. (2003). Singularities in mixture models and upper bounds of stochastic complexity. *Neural networks* **16** 1029–1038.

[55] YANG, Y., PATI, D., BHATTACHARYA, A. et al. (2020). α-variational inference with statistical guarantees. *Annals of Statistics* **48** 886–905.

[56] ZEMANIAN, A. H. (1987). *Distribution Theory and Transform Analysis: An Introduction to Generalized Functions, with Applications*. *Dover Books on Advanced Mathematics*. Dover Publications.

[57] ZHANG, A. Y. and ZHOU, H. H. (2020). Theoretical and Computational Guarantees of Mean Field Variational Inference for Community Detection. *The Annals of Statistics (to appear)*.