Algebraic Methods for Evaluating Integrals in Bayesian Statistics

by

Shaowei Lin

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Mathematics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Bernd Sturmfels, Chair
Professor Lior Pachter
Associate Professor Yun S. Song

Spring 2011

Algebraic Methods for Evaluating Integrals in Bayesian Statistics

Abstract

Algebraic Methods for Evaluating Integrals
in Bayesian Statistics

by

Shaowei Lin

Doctor of Philosophy in Mathematics

University of California, Berkeley

Professor Bernd Sturmfels, Chair

The accurate evaluation of marginal likelihood integrals is a difficult fundamental problem in Bayesian inference that has important applications in machine learning and computational biology. Following the recent success of algebraic statistics [16,41,43] in frequentist inference and inspired by Watanabe's foundational approach to singular learning theory [58], the goal of this dissertation is to study algebraic, geometric and combinatorial methods for computing Bayesian integrals effectively, and to explore the rich mathematical theories that arise in this connection between statistics and algebraic geometry. For these integrals, we investigate their exact evaluation for small samples and their asymptotics for large samples.

According to Watanabe, the key to understanding singular models lies in desingularizing the Kullback-Leibler function $K(\omega)$ of the model at the true distribution. This step puts the model in a standard form so that various central limit theorems can be applied. While general algorithms exist for desingularizing any analytic function, applying them to non-polynomial functions such as $K(\omega)$ can be computationally expensive. Many singular models are however represented as regular models whose parameters are polynomial functions of new parameters. Discrete models and multivariate Gaussian models are all examples. We call them *regularly parametrized* models. One of our main contributions is showing how this polynomiality can be exploited by defining *fiber ideals* for singular models and relating the properties of these algebraic objects to the statistics. In particular, we prove that a model is put in standard form if we monomialize the corresponding fiber ideal. As a corollary, the learning coefficient of a model is equal to the *real log canonical threshold* (RLCT) of the fiber ideal.

While *complex* log canonical thresholds are well-studied in algebraic geometry, little is known about their real analogs. In Chapter 4, we prove their fundamental properties and simple rules of computation. We also extend Varchenko's notion [54] of Newton polyhedra and nondegeneracy for functions to ideals. Using these methods, we discover a formula for the RLCT of a monomial ideal *with respect to* a monomial amplitude. For all other ideals, this formula is an upper bound for their RLCT. Our tools are then applied to a difficult statistical example involving a naïve Bayesian network with two ternary random variables.

Because our statistical models are defined over compact semianalytic parameter spaces Ω, we need to extend standard asymptotic theory [3] of real analytic functions over neighborhoods of the origin to functions over domains like Ω. Chapter 3 summarizes these results which are critical for other proofs in this dissertation. We also give explicit formulas for the full asymptotic expansion of a Laplace integral over Ω in terms of the Laurent coefficients of the associated zeta function. In Chapter 5, we apply these formulas to Laplace integrals $Z(n)$ with nondegenerate phase functions, and describe algorithms for computing the coefficient C in the first term asymptotics $Z(n) \approx Cn^{-\lambda}(\log n)^{\theta-1}$. Procedures for calculating all higher order coefficients are also developed and explained.

Watanabe's treatment of singular models assumes knowledge of the true distribution. In this dissertation, we also explore marginal likelihood integrals of exponential families given data where the true distribution is unknown. This is the context in which Schwarz [48], Haughton [27] and Geiger and Rusakov [45] studied the Bayesian Information Criterion (BIC). We find here that the log likelihood ratio of the data is equal to the Kullback-Leibler function of the model at the maximum likelihood distribution. Therefore, all the methods we developed for Kullback-Leibler functions apply, so we describe how to compute the full asymptotics of the marginal likelihood integral by monomializing the associated fiber ideal.

Lastly, to complement developments in asymptotic estimation as well as in Markov Chain Monte Carlo (MCMC) estimation, we present, in Chapter 2, symbolic algorithms for computing marginal likelihood integrals *exactly* for discrete data of small samples. The underlying statistical models are mixtures of independent distributions, or, in geometric language, secant varieties of Segre-Veronese varieties. For these models, the numerical value of the integral is a rational number, and exact evaluation means computing that rational number rather than a floating point approximation. These exact results provide a gold standard with which approximation methods can be compared.

To my Brother

Contents

Acknowledgments

First and foremost, I want to thank my advisor, Bernd Sturmfels, for inspiring me to pursue this research topic. Through the years, he taught by example what it means to be a successful researcher. I learned to always ask good questions and to always ask for help. I also want to thank his family for being like a family to me in Berlin and in Berkeley.

Second, I am grateful to two great statisticians, Sumio Watanabe and Mathias Drton, who were very patient in imparting their insights on this subject area. Watanabe taught me to approximate Kullback-Leibler functions with sums of squares, while Drton inspired me to extend my results for discrete models to regularly parametrized models. Thank you for being so generous with your ideas, advice and encouragement.

Third, I am thankful for Zhiqiang Xu, who was a co-author on one of my paper. I enjoyed the many hours we spent coding and discussing problems in Berlin and in Beijing.

I want to thank my committee members Lior Pachter and Yun S. Song for organizing memorable classes, attending my qualifying exam, and taking time to read this thesis.

Special thanks go to Christine Berkesch, Nero Budur, Anne Frühbis-Krüger, Luis García-Puente, Anton Leykin, Robin Pemantle, Josef Schicho, Zach Teitler and Piotr Zwiernik for teaching me so much. I am also grateful to Morgan Brown, Dustin Cartwright, María Cueto, Jesús De Loera, Daniel Erman, Alex Fink, Christopher Hillar, Edward Kim, Matthias Köppe, Jiawang Nie, Luke Oeding, Philipp Rostalski, Raman Sanyal, Caroline Uhler and Cynthia Vinzant for all the enlightening discussions.

Of course, my PhD would not be possible without funding and guidance from the Agency for Science, Technology and Research (A*STAR), Singapore. I am also indebted to the UC Berkeley Mathematics Department and to MSRI for the many opportunities to grow.

In my personal life, few people have given me as much encouragement and support as Diogo Oliveira e Silva, Dewen Soh, Jeremy Chan, my sister Qianwen and my loving parents. Thank you for believing in me!

To my wonderful fiancée and best friend Cynthia Yu: nobody knows what goes on behind the scenes besides you. Thank you for holding my hands in this journey.

To Father God, Jesus and Holy Spirit: thank you for never letting go. You have opened my eyes to see. Thank you for setting me free to play, to dream and to fly.

Chapter 1

Integrals in Bayesian Statistics

Bayesian statistics is foundational in applications such as machine learning and computational biology. A fundamental problem in Bayesian statistics is the accurate evaluation of integrals. This chapter is an introduction to some of the theory that has developed around this problem, and a summary of the major contributions in this dissertation. As a start, in Section 1.1, we survey some of the integrals which arise in this field, while in Section 1.4 we review some important classes of models used in this dissertation.

We will primarily be interested in two kinds of integrals. The first has the form

$$\int_{\Omega} p_1(\omega)^{u_1} \cdots p_k(\omega)^{u_k} d\omega$$

where $\Omega \subset \mathbb{R}^d$ is a polytope, the $p_i(\omega)$ are polynomials in $\omega = (\omega_1, \ldots, \omega_d)$ and the u_i are integers. In Chapter 2, we study efficient algorithms for computing this integral *exactly* for a special class of discrete statistical models.

The second kind of integrals has the form

$$Z(n) = \int_{\Omega} e^{-nf(\omega)} \varphi(\omega) d\omega$$

where $\Omega \subset \mathbb{R}^d$ is a compact semianalytic subset, and $f(\omega)$ and $\varphi(\omega)$ are real analytic functions. We will be interested in estimating this integral for large n, where n usually refers to the sample size. The asymptotics of such integrals is well understood for regular statistical models, but little was known for singular models until a breakthrough in 2001 due to Sumio Watanabe [57]. His insight was to put the model in a suitable standard form by employing the technique of resolution of singularities from algebraic geometry. To this standard form, various central limit theorems can be applied. We briefly describe the basic ideas behind his results in Sections 1.2 and 1.3.

Watanabe's work provided the theoretical foundation for understanding singular models, but applying this theory proved to be challenging computationally. The main difficulty lies

in finding resolutions of singularities for Kullback-Leibler functions which are real analytic log integrals. Our largest contribution to these developments is providing effective algebraic tools for this computation. We show that for a very general class of models known as *regularly parametrized models*, instead of desingularizing the Kullback-Leibler function, we only need to monomialize an associated ideal of polynomials called the *fiber ideal*. Parametrized discrete models, multivariate Gaussian models and graphical models are all examples of regularly parametrized models, so our methods are widely applicable. We summarize our main results in Section 1.5. We show that the *learning coefficient* of a model equals the *real log canonical threshold* of the fiber ideal. Through this exploration, we uncover many algebraic, geometric and combinatorial results which are interesting mathematically in their own right.

In studying singular models, Watanabe was primarily interested in their behavior for large sample sizes while assuming knowledge of the true distribution. For instance, in one of his main theorems, he computes the asymptotics of the *expected* log marginal likelihood integral. Meanwhile, in many practical situations, the true distribution is unknown but we are given large-sample data and we want to estimate the corresponding marginal likelihood integral. In Section 1.6, we study this scenario for exponential families and show that their marginal likelihood integrals have a connection to Kullback-Leibler functions of models over some true distribution. This allows us to apply our results from Section 1.5. In particular, we prove that for regularly parametrized models, under some conditions, the asymptotics of their likelihood integrals can be computed by monomializing the associated fiber ideals. Using this approach, we will be able to compute higher order asymptotics of the integrals, using formulas from Chapter 5.

1.1 Model Selection

The fundamental problem in statistical learning theory is choosing a statistical model that best describes the given data. More precisely, let X be a random variable with state space $\mathcal{X}$, and let $x_1, x_2, \ldots, x_N$ be N independent random samples of X. A *statistical model* $\mathcal{M}$ on $\mathcal{X}$ is a family of probability distributions on $\mathcal{X}$ parametrized by a space Ω. The distribution corresponding to $\omega \in \Omega$ is denoted by $p(x|\omega, \mathcal{M})dx$. In this dissertation, $\mathcal{X}$ will either be a discrete space $[k] := \{1, 2, \ldots, k\}$ or a real vector space $\mathbb{R}^k$, while Ω will be compact subset of $\mathbb{R}^d$. Now, given the data $x_1, \ldots, x_N$, we select the "best" model by computing a *criterion* or score for each model, and picking the model that maximizes this criterion. Different criteria exist for different frameworks and purposes. We examine some important ones below.

1.1.1 Maximum Likelihood

The frequentist approach is to compute the *maximum likelihood*

$$\max_{\omega\in\Omega} p(x_1,\ldots,x_N|\omega,\mathcal{M}) = \max_{\omega\in\Omega} \prod_{i=1}^{N} p(x_i|\omega,\mathcal{M}) \tag{1.1}$$

as a criterion for each model $\mathcal{M}$, and a parameter $\omega^* \in \Omega$ which achieves this optimal value is known as a *maximum likelihood estimate*. The idea behind this approach is to find among all the models a distribution that is mostly likely to produce the given data.

Techniques for attacking such optimization problems are studied intensively in statistics. A common *numerical* technique used, especially for graphical models (see Section 1.4.4), is the Expectation-Maximization (EM) algorithm [14,56]. In recent years, *algebraic* techniques for solving this problem are also being explored in the fast-growing field of algebraic statistics. For example, Gröbner bases methods are being used to solve the Lagrange equations for the maximum likelihood estimates [16, 30, 41].

1.1.2 Marginal Likelihood

The Bayesian approach is to pick the model which maximizes the posteriori probability

$$p(\mathcal{M}|x_1,\ldots,x_N) \;=\; \frac{p(x_1,\ldots,x_N,\mathcal{M})}{p(x_1,\ldots,x_N)} \;\propto\; p(x_1,\ldots,x_N,\mathcal{M}).$$

In this approach, we often assume that there is prior distribution $p(\omega|\mathcal{M})d\omega$ on Ω, and that each model $\mathcal{M}$ is assigned a prior probability $p(\mathcal{M})$. Then,

$$p(x_1,\ldots,x_N,\mathcal{M}) = p(\mathcal{M})\, p(x_1,\ldots,x_N|\mathcal{M})$$

where $p(x_1,\ldots,x_N|\mathcal{M})$ is the *marginal likelihood integral*

$$\int_\Omega p(x_1,\ldots,x_N|\omega,\mathcal{M})\, p(\omega|\mathcal{M})d\omega = \int_\Omega \prod_{i=1}^{N} p(x_i|\omega,\mathcal{M})\, p(\omega|\mathcal{M})d\omega. \tag{1.2}$$

Here, we integrate with respect to the measure prescribed by the prior $p(\omega|\mathcal{M})d\omega$. In this dissertation, $d\omega$ is often the standard Lebesgue measure on a subset $\Omega \subset \mathbb{R}^d$. Computationally, the greatest difficulty lies in evaluating this integral. To approximate this integral *numerically*, statisticians often use Markov chain Monte Carlo (MCMC) methods. One major goal of this dissertation is to study how *algebraic* methods can be employed to evaluate or approximate this integral efficiently. In so doing, we hope to expand the scope of algebraic statistics from the study of derivatives in the frequentist approach to that of integrals in the Bayesian approach. Some of these algebraic methods include lattice point enumeration in

polytopes, resolution of singularities, and toric geometry. We will describe them in detail in the coming chapters.

There is a strong relationship between the maximum likelihood and the marginal likelihood integral. Fixing the sample size N and data $x_1, \ldots, x_N$, we define the log likelihood

$$f(\omega) = -\frac{1}{N} \sum_{i=1}^{N} \log p(x_i|\omega, \mathcal{M})$$

and consider the function

$$L(n) = \int_{\Omega} e^{-nf(\omega)}\, p(\omega|\mathcal{M}) d\omega \tag{1.3}$$

where n is a positive integer. Then, the marginal likelihood integral (1.2) equals $L(N)$. Now, with some mild assumptions on f, $p(\omega|\mathcal{M})$ and Ω, we can show that asymptotically

$$L(n) \;\approx\; C \cdot (e^{-f_0})^n \cdot n^{-\lambda} (\log n)^{\theta-1}$$

as $n \to \infty$ (see Chapter 3). Here, $C \in \mathbb{R}, \lambda \in \mathbb{Q}, \theta \in \mathbb{Z}$ are positive constants, and

$$f_0 = \min_{\omega \in \Omega} f(\omega).$$

Thus, $(e^{-f_0})^N$ is the maximum likelihood (1.1) from the frequentist approach, so this number is a good first approximation of the marginal likelihood integral from the Bayesian approach. We discuss how this approximation can be improved in Section 1.3.

Generally, computing the maximum likelihood involves understanding the zero set

$$\{\omega \in \Omega : \frac{df}{d\omega}(\omega) = 0\}$$

while computing the marginal likelihood integral involves understanding the behavior of the function f in a sufficiently small neighborhood of this zero set via resolution of singularities. As a result, the algebraic methods for studying this neighborhood is fundamentally different from that of earlier investigations in algebraic statistics [16, 30, 41].

1.1.3 Cross Validation

A Bayesian technique for estimating the distribution of X uses the *predictive distribution*

$$p(x|x_1, \ldots, x_N, \mathcal{M}) = \frac{\displaystyle\int_{\Omega} p(x|\omega, \mathcal{M}) \prod_{i=1}^{N} p(x_i|\omega, \mathcal{M})\, p(\omega|\mathcal{M}) d\omega}{\displaystyle\int_{\Omega} \prod_{i=1}^{N} p(x_i|\omega, \mathcal{M})\, p(\omega|\mathcal{M}) d\omega}.$$

Assuming that we have new data $x_{N+1}, \ldots, x_{N+M}$, we can then test our predictive distribution against this data by computing the likelihood

$$p(x_{N+1}, \ldots, x_{N+M} | x_1, \ldots, x_N, \mathcal{M}) = \prod_{i=1}^{M} p(x_{i+N} | x_1, \ldots, x_N, \mathcal{M}) \tag{1.4}$$

and picking the model that minimizes this likelihood.

Cross validation is a model selection method that capitalizes on this principle. The given data is randomly partitioned into two sets. The first set, called the *training set*, allows us to produce a predictive distribution for each model $\mathcal{M}$. Meanwhile, the second set, called the *validation set*, is used for computing the likelihood (1.4) as a selection criterion.

Given two models $\mathcal{M}_1$ and $\mathcal{M}_2$ such that $\mathcal{M}_1$ is a subset of $\mathcal{M}_2$, the maximum likelihood criterion always selects the more complex model $\mathcal{M}_2$. This is a problem known as *overfitting*, because ideally, we want the simplest model that describes the data well. Cross validation overcomes this problem, penalizing overfitted models by requiring a good fit with the validation set. The marginal likelihood approach also overcomes this problem, since integrals over higher dimensional parameter spaces suffer larger penalties (see Section 1.3).

1.2 Regular and Singular Models

In statistical learning theory, it is important to have a measure of how far apart two probability distributions are. Given two distributions $q(x)dx$ and $p(x)dx$ on a space $\mathcal{X}$, we define the *Kullback-Leibler divergence* $K(q\|p)$ from q to p to be the integral

$$K(q\|p) = \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x)} dx.$$

When $\mathcal{X}$ is a finite discrete space, the distributions $q(x)dx$ and $p(x)dx$ are discrete measures and this integral becomes the finite sum

$$K(q\|p) = \sum_{x \in \mathcal{X}} q(x) \log \frac{q(x)}{p(x)}.$$

This function plays an important role in many applications, as we shall see in this chapter. As p varies over all distributions on $\mathcal{X}$, the function $K(q\|p)$ is minimized when $p = q$, so

$$K(q\|p) \geq 0 \quad \text{and} \quad K(q\|p) = 0 \Leftrightarrow q = p$$

for all distributions q and p. Because the formula for K is not symmetric in q and p,

$$K(q\|p) \neq K(p\|q)$$

in general. Nonetheless, $K(q\|p)$ is sometimes referred to the Kullback-Leibler *distance*.

We now define what it means for a model to be regular. Let $\mathcal{M}$ be a statistical model on $\mathcal{X}$ with parameter space $\Omega \subset \mathbb{R}^d$. We say that $\mathcal{M}$ is *identifiable* if

$$p(x|\omega_1, \mathcal{M}) = p(x|\omega_2, \mathcal{M}) \ \forall x \quad \Rightarrow \quad \omega_1 = \omega_2.$$

Given $\tilde{\omega} \in \Omega$, consider the function $K_{\tilde{\omega}} : \Omega \to \mathbb{R}$,

$$K_{\tilde{\omega}}(\omega) = \int_{\mathcal{X}} p(x|\tilde{\omega}, \mathcal{M}) \log \frac{p(x|\tilde{\omega}, \mathcal{M})}{p(x|\omega, \mathcal{M})} dx.$$

This integral is the Kullback-Leibler divergence from $p(x|\tilde{\omega}, \mathcal{M})dx$ to $p(x|\omega, \mathcal{M})dx$. Define the *Fisher information matrix* $I(\tilde{\omega})$ to be the $d \times d$ Hessian matrix of $K_{\tilde{\omega}}(\omega)$ at $\omega = \tilde{\omega}$, i.e.

$$I_{jk}(\tilde{\omega}) = \frac{\partial^2 K_{\tilde{\omega}}}{\partial \omega_j \partial \omega_k}(\tilde{\omega}), \quad 1 \leq j, k \leq d.$$

Because $K_{\tilde{\omega}}(\omega)$ attains its minimum at $\omega = \tilde{\omega}$, the symmetric matrix $I(\tilde{\omega})$ is positive semidefinite. The model $\mathcal{M}$ is *regular* if $\mathcal{M}$ is identifiable and its Fisher information matrix $I(\omega)$ is positive definite for all $\omega \in \Omega$. Otherwise, we say that $\mathcal{M}$ is *singular*.

Regular models have many desirable properties. For instance, if the data is drawn from a fixed distribution in the model, then the distribution of the maximum likelihood estimator approaches a Gaussian distribution as the sample size grows large [58, §1.2.1]. The marginal likelihood integral also behaves well asymptotically, as we shall see in Section 1.3. Unfortunately, many models studied in statistical learning theory are singular, often because of the existence of hidden variables. Some such examples will be described in Section 1.4. Recently, using advanced techniques from algebraic geometry, Watanabe made significant progress in understanding the asymptotic behavior of singular models [58]. His singular learning theory will form the foundation of this chapter as we study fiber ideals of statistical models.

Example 1.1 (Coin Toss)**.** Let us illustrate the above concepts with a simple example. Suppose we have a random variable X with state space $\mathcal{X} := \{H, T\}$ representing the outcomes of a coin toss. Let the data $x_1, \ldots, x_{100}$ be a sequence of 100 independent observations of X with 53 heads and 47 tails in summary. We propose three models to explain this data.

The first model $\mathcal{M}_1$ assumes that the data comes from tossing a fair coin, i.e.

$$p(H|\mathcal{M}_1) = p(T|\mathcal{M}_1) = \frac{1}{2}.$$

Here, the maximum likelihood and marginal likelihood are both trivially equal to

$$\left(\frac{1}{2}\right)^{100} \approx 7.88860905 \times 10^{-31}.$$

The second model $\mathcal{M}_2$ assumes that the coin is biased with heads occurring with probability $\omega \in \Omega := [0,1]$. This model is then parametrized by

$$p(H|\omega, \mathcal{M}_2) = \omega, \quad p(T|\omega, \mathcal{M}_2) = 1 - \omega.$$

The maximum likelihood estimate will then be the relative frequency $\omega^* = 53/100$ of heads in the data, and the maximum likelihood is

$$\left(\frac{53}{100}\right)^{53} \left(\frac{47}{100}\right)^{47} \approx 9.44540125 \times 10^{-31}.$$

Meanwhile, assuming the uniform prior distribution on the parameter space $[0,1]$, the marginal likelihood is

$$\begin{aligned}\int_0^1 \omega^{53}(1-\omega)^{47} d\omega &= \frac{1}{852576221558946798965230169760 0} \\ &\approx 1.17291566 \times 10^{-31}.\end{aligned}$$

Thus, using maximum likelihood as a criterion for model selection, we would have chosen the more complex $\mathcal{M}_2$, while comparison of the marginal likelihood integrals would have given us $\mathcal{M}_1$. This demonstrates how the maximum likelihood approach suffers from overfitting.

Now, consider a third model $\mathcal{M}_3$ which involves a hidden variable Y. A coin with sides colored blue and red is first flipped. If the outcome is blue, we then toss a fair coin. If the outcome is red, we toss a biased coin where heads occur with probability $\omega \in [0,1]$. Suppose the colored coin comes up blue with probability $t \in [0,1]$. Then, the model is parametrized by the polynomials

$$\begin{aligned}p(H|t, \omega, \mathcal{M}_3) &= t\left(\frac{1}{2}\right) + (1-t)w \\ p(T|t, \omega, \mathcal{M}_3) &= t\left(\frac{1}{2}\right) + (1-t)(1-w)\end{aligned}$$

where $(t, \omega) \in [0,1]^2$. One can check that the Fisher information matrix for $\mathcal{M}_2$ is $I(\omega) = 1$ for all $\omega \in [0,1]$, so $\mathcal{M}_2$ is regular. Meanwhile, for $\mathcal{M}_3$, when $\omega = 1/2$,

$$I(t, \frac{1}{2}) = \begin{pmatrix} 0 & 0 \\ 0 & 8(1-t)^2 \end{pmatrix} \quad \text{for all } t \in [0,1]$$

so $\mathcal{M}_3$ is an example of a singular model. Observe that $\mathcal{M}_3$ is also not identifiable, because the parameters $\omega = 1/2, t \in [0,1]$ all give the same distribution $p(H) = p(T) = 1/2$. □

1.3 Model Asymptotics

In Section 1.1.2, we observed that the maximum likelihood is a good first approximation of the marginal likelihood integral. We now study how this approximation can be improved. Under certain regularity conditions, we may apply a Laplace asymptotic approximation to the integral $L(n)$ in equation (1.3). This gives us the *Bayesian Information Criterion* (BIC) proposed by Schwarz [48]. Not much was known about extending the BIC to singular models, until the recent work of Watanabe [58]. We discuss his results in this section.

1.3.1 Bayesian Information Criterion

Let Ω be a compact subset of $\mathbb{R}^d$. Let $f : \Omega \to \mathbb{R}$ and $\varphi : \Omega \to \mathbb{R}$ be functions which are real analytic over Ω. We will consider *Laplace integrals* of the form

$$L(n) = \int_\Omega e^{-nf(\omega)}\, \varphi(\omega) d\omega$$

where we will be interested in the asymptotics of $L(n)$ as n tends to ∞. Here, the functions f and φ are known as the *phase* and *amplitude* functions respectively. An example of such an integral is the marginal likelihood (1.3).

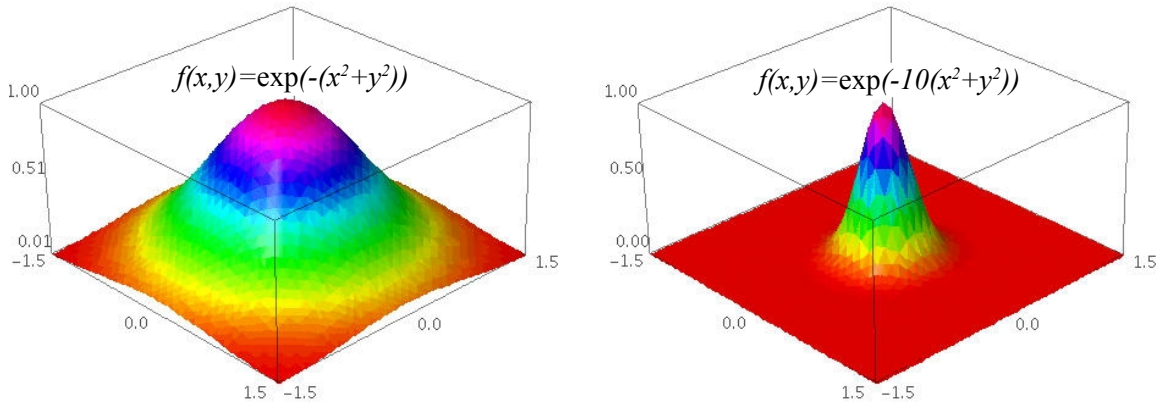


Figure 1.1: Integral asymptotics depends on maximum of integrand.

Now, as n grows large, the value of $L(n)$ becomes increasingly dominated by the behavior of the integrand near its maximum points, which correspond to minimum points of the phase function $f(\omega)$. If $f(\omega)$ has a unique minimum point and satisfies some regularity condition near this point, then the asymptotics can be given explicitly.

Proposition 1.2 (Laplace approximation)**.** *Let Ω be a compact subset of $\mathbb{R}^d$ and f, φ be functions $\Omega \to \mathbb{R}$ which are real analytic over Ω. Suppose f attains its minimum uniquely*

at $\omega^* \in \Omega$ and is defined in a small $\mathbb{R}^d$-neighborhood of ω^*. If the Hessian $H(\omega^*)$ of f at ω^* is positive definite and $\varphi(\omega^*)$ is positive, then as $n \to \infty$,

$$L(n) = \int_\Omega e^{-nf(\omega)}\, \varphi(\omega) d\omega \quad \to \quad (e^{-f(\omega^*)})^n \sqrt{\frac{(2\pi)^d}{\det H(\omega^*)}}\, \varphi(\omega^*)\, n^{-d/2}.$$

Proof. Consider the Taylor expansion of f around ω^*, and apply the formula

$$\int_{\mathbb{R}^d} e^{-\frac{1}{2}\omega^{\mathsf{T}} H \omega} d\omega = \sqrt{\frac{(2\pi)^d}{\det H}}$$

for a Gaussian integral with positive definite matrix H. □

If the above regularity conditions hold for the marginal likelihood integral of a statistical model, then by taking logarithms we get the *Bayesian information criterion*

$$\log L(n) = -nf(\omega^*) - \frac{d}{2} \log n + O(1)$$

where $-nf(\omega^*)$ is the log maximum likelihood. Applying this approximation to linear exponential models [48], Schwarz showed that the BIC selects the correct model with probability 1 as the sample size $n \to \infty$. Since then, this model selection criterion has also been applied to other regular models [27, 35, 45], such as curved exponential models and graphical models without hidden variables. In the next section, we study how this analysis could be extended to singular models, such as those with hidden variables.

1.3.2 Singular Learning Theory

Let X be a random variable with state space $\mathcal{X}$ subject to a true distribution $q(x)dx$, and let $x_1, \ldots, x_N$ be N independent samples of X. In statistical learning, one may think of $q(x)$ as the probability density with which a *teacher machine* is generating data x_i. The goal of the *student machine* is to estimate this density from the x_i, via some learning algorithm. In this section, we summarize the basics of Watanabe's singular learning theory [58, §1.4]. For convenience, given a model $\mathcal{M}$ and a true distribution $q(x)dx$ which lies in $\mathcal{M}$, we consider $q(x)dx$ as part of the information defining $\mathcal{M}$.

Up to this point in the chapter, we have been looking at marginal likelihood integrals corresponding to some given data. For singular learning theory, we shift gears and assume that the data is subject to some true distribution, and we will be interested in the *expected behavior* of the marginal likelihood integral. It is important to understand this statistical distinction before we continue. We discuss this distinction in greater detail in Remark 1.4. We return to marginal likelihood integrals of large-sample data without assuming knowledge of the true distribution in Section 1.6.

A critical step in analyzing the asymptotic behavior of a singular model $\mathcal{M}$ is understanding the log likelihood ratio

$$K_N(\omega) = \frac{1}{N} \log \frac{\prod_{i=1}^N q(x_i)}{\prod_{i=1}^N p(x_i|\omega, \mathcal{M})} = \frac{1}{N} \sum_{i=1}^N \log \frac{q(x_i)}{p(x_i|\omega, \mathcal{M})}.$$

The expected value of $K_N(\omega)$ over all data is the Kullback-Leibler divergence

$$K(\omega) = \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x|\omega, \mathcal{M})} dx.$$

For $\omega \in \Omega$ such that $K(\omega) \neq 0$, let us define a random process

$$\psi_N(\omega) = \frac{NK_N(\omega) - NK(\omega)}{\sqrt{NK(\omega)}}$$

so that we have the relation

$$K_N(\omega) = K(\omega) + \sqrt{\frac{K(\omega)}{N}}\, \psi_N(\omega).$$

One can show that under some mild assumptions, this process converges in law to a Gaussian process as $N \to \infty$. Unfortunately, this process is not defined for points where $K(\omega) = 0$.

Watanabe's insight was to use a technique in algebraic geometry known as resolution of singularities (see Section 3.1) to analyze the log likelihood ratio $K_N(\omega)$. In particular, the theory says that if $K(\omega)$ is real analytic over Ω, there exists a real d-dimensional manifold $\mathscr{M}$ and a real analytic map $\rho : \mathscr{M} \to \Omega$ with the following property: for each $y \in \mathscr{M}$, there exist local coordinates $\mu = (\mu_1, \ldots, \mu_d)$ such that y is the origin and

$$\begin{aligned} K(\rho(\mu)) &= \mu_1^{2\kappa_1} \cdots \mu_d^{2\kappa_d} = \mu^{2\kappa} \\ \det \rho'(\mu) &= h(\mu) \mu_1^{\tau_1} \cdots \mu_d^{\tau_d} = h(\mu) \mu^\tau \end{aligned} \tag{1.5}$$

for some non-negative integers $\kappa_1, \ldots, \kappa_d, \tau_1, \ldots, \tau_d$ and non-vanishing real analytic function $h(\mu)$. Using the *desingularization* map ρ, one can show that

$$\log \frac{q(x)}{p(x|g(\mu), \mathcal{M})} = a(x, \mu) \mu^\kappa$$

for some real analytic function $a(x, \mu)$. We may now use $a(x, u)$ to define a random process

$$\xi_N(\mu) = \frac{1}{\sqrt{N}} \sum_{i=1}^N (\mu^\kappa - a(x_i, \mu))$$

which is related to the log likelihood ratio $K_N(g(\mu))$ via the relation

$$\begin{aligned} K_N(g(\mu)) &= K(g(\mu)) + \sqrt{\frac{K(g(\mu))}{N}}\,\xi_N(\mu) \\ &= \mu^{2\kappa} + \frac{1}{\sqrt{N}}\mu^{\kappa}\xi_N(\mu). \end{aligned} \tag{1.6}$$

Watanabe showed that as $N \to \infty$, $\xi_N(\mu)$ tends to a Gaussian process over the manifold $\mathscr{M}$. He termed the formula (1.6) as the *standard form* of the log likelihood ratio.

Desingularizing the Kullback-Leibler divergence $K(\omega)$ is crucial for putting the log likelihood ratio in standard form. Watanabe discovered that this desingularization is also critical in the asymptotic expansion of the log marginal likelihood integral. Recall that given samples $x_1, \ldots, x_N$, the marginal likelihood integral for our model $\mathcal{M}$ is

$$L(N) = \int_\Omega \prod_{i=1}^{N} p(x_i|\omega, \mathcal{M})\, p(\omega|\mathcal{M}) d\omega.$$

Because the x_i are random variables subject to a true distribution $q(x)dx$, the integral $L(N)$ is also a random variable. To analyze the asymptotic behavior of $L(N)$ as the sample size N grows large, our first step is to define the *zeta function*

$$\zeta(z) = \int_\Omega K(\omega)^{-z}\varphi(\omega)d\omega, \quad z \in \mathbb{C}$$

where $\varphi(\omega)$ is the prior $p(\omega|\mathcal{M})$ on the parameter space Ω. Standard asymptotic theory (see Chapter 3) tells us that the full Laurent expansion of this meromorphic function gives the full asymptotic expansion of the Laplace integral

$$Z(n) = \int_\Omega e^{-nK(\omega)}\varphi(\omega)d\omega, \quad n \in \mathbb{Z}_+.$$

In fact, if (λ, θ) is the smallest pole and its multiplicity of the zeta function $\zeta(z)$, then

$$Z(n) \;\approx\; Cn^{-\lambda}(\log n)^{\theta-1}$$

asymptotically. In algebraic geometry, when $\varphi(\omega)$ is a positive constant function, the smallest pole λ is known as the *real log canonical threshold* of $K(\omega)$.

Unlike the random variable $L(N)$, the integral $Z(n)$ is deterministic. Nonetheless, Watanabe discovered a close connection between their asymptotics. Using some technical stochastic arguments, he generalizes the Bayesian information criterion [58, §6.2].

Theorem 1.3 (Watanabe)**.** *Suppose $q(x)$, $\varphi(\omega)$, $p(x|\omega, \mathcal{M})$ and Ω satisfy some mild analyticity and integrality conditions (see [58, §6] for more details).*

Let (λ, θ) be coefficients appearing in the asymptotic expansion

$$\log Z(n) = -\lambda \log n + (\theta - 1) \log \log n + O(1).$$

Then as $N \to \infty$, the log marginal likelihood integral has the asymptotic expansion

$$\log L(N) = \sum_{i=1}^{N} \log q(x_i) - \lambda \log N + (\theta - 1) \log \log N + F(\xi_N)$$

where the random variable $F(\xi_N)$ is a function of ξ_N. Consequently, after taking expectations,

$$\mathbb{E}[\log L(N)] = N \int_{\mathcal{X}} q(x) \log q(x) dx - \lambda \log N + (\theta - 1) \log \log N + \mathbb{E}[F(\xi_N)]$$

where $\mathbb{E}[F(\xi_N)]$ converges to a constant.

The constant λ appearing in the asymptotics of $\log L(N)$ is called the *learning coefficient* of the model $\mathcal{M}$ subject to the true distribution $q(x)dx$. Surprisingly, to prove Watanabe's theorem, resolution of singularities is not necessary [58, §4.5]. The desingularization map ρ comes in only when we want to compute the pair (λ, θ). Indeed, by applying ρ as a change of variables to the zeta function, we can show that

$$(\lambda, \theta) = \min_{y \in \mathscr{M}} (\lambda_y, \theta_y)$$

where we define using the integers κ_i, τ_j described in (1.5) for each $y \in \mathscr{M}$,

$$\lambda_y = \min_{1 \le i \le d} \frac{\tau_i + 1}{2\kappa_i}, \quad \theta_y = \#\min_{1 \le i \le d} \frac{\tau_i + 1}{2\kappa_i}.$$

Here, #min denotes the number of times the minimum is attained, and the pairs (λ_y, θ_y) are ordered such that $(\lambda_1, \theta_1) < (\lambda_2, \theta_2)$ if for sufficiently large N,

$$\lambda_1 \log N - \theta_1 \log \log N \ < \ \lambda_2 \log N - \theta_2 \log \log N.$$

This proves a deep result that the learning coefficient is a positive rational number.

In summary, many practical applications of statistical learning theory depend on resolving the singularities of the Kullback-Leibler divergence $K(\omega)$ of a statistical model $\mathcal{M}$ and a true distribution $q(x)$. A major goal of this dissertation is to make this computation feasible for a large class of models, in particular *regularly parametrized* models (see Section 1.5). In physics and information theory, the Kullback-Leibler divergence is also known as relative entropy, so the results of this dissertation may be applied to problems in these fields as well.

Remark 1.4. In this section, we considered the asymptotics of many kinds of integrals from Bayesian statistics. The difference between these integrals may be confusing for the reader, so we list them side by side below and explain how they differ.

For this discussion, let us fix the sample size N. The integral that we are most interested in is the marginal likelihood integral $L(N)$ which we rewrite as

$$L(N) = \prod_{i=1}^{N} q(x_i) \cdot \int_{\Omega} e^{-NK_N(\omega)} \varphi(\omega) d\omega = \prod_{i=1}^{N} q(x_i) \cdot \mathscr{L}(K_N)$$

where $K_N(\omega)$ is the log likelihood ratio and

$$\mathscr{L}(f) = \int_{\Omega} e^{-Nf(\omega)} \varphi(\omega) d\omega$$

for any function $f : \Omega \to \mathbb{R}$. On the other hand, the Laplace integral

$$Z(n) = \mathscr{L}(K) = \mathscr{L}(\mathbb{E}(K_N))$$

comes from replacing the log likelihood ratio $K_N(\omega)$ with its expectation $K(\omega)$, the Kullback-Leibler function. Thus, we have three expressions of interest:

$$\mathbb{E}\left[\log \mathscr{L}(K_N)\right], \quad \log \mathbb{E}\left[\mathscr{L}(K_N)\right], \quad \log \mathscr{L}(\mathbb{E}\left[K_N\right]).$$

Their numerical values are not necessarily equal to one another. In Theorem 1.3, Watanabe relates the asymptotics of the first expression to that of the third expression.

Watanabe's analysis assumes knowledge of the true distribution. In many practical situations, we do not have knowledge of the true distribution, but we are given large-sample data. We want to estimate the marginal likelihood integral for this data using some asymptotic methods. This question will be discussed for exponential families in Section 1.6. □

1.4 Important Classes of Models

In this section, we define and discuss several classes of statistical models which will be used throughout this dissertation.

1.4.1 Discrete Models

A *discrete* model is a statistical model whose state space $\mathcal{X}$ is a finite set. Writing $\mathcal{X} = [k] := \{1, 2, \ldots, k\}$, we define p_i to be the probability of the i-th outcome. The set of probability distributions on $[k]$ is the $(k-1)$-dimensional simplex

$$\Delta_{k-1} := \{p \in \mathbb{R}^k : p_i \geq 0 \ \forall i, \sum_{i=1}^{k} p_i = 1\}.$$

Thus, we may represent a discrete model with parameter space Ω by a map $p : \Omega \to \Delta_{k-1}$. For instance, the model $\mathcal{M}_3$ in Example 1.1 is parametrized over $\Omega = [0,1]^2$ by polynomials

$$p_1(t,\omega) = t/2 + \omega - t\omega,$$
$$p_2(t,\omega) = 1 - t/2 - \omega + t\omega.$$

Note that $p_1(t,\omega) + p_2(t,\omega) = 1$ and $p_1, p_2 \geq 0$ for all $(t,\omega) \in [0,1]^2$.

We say that a discrete model is *free* if p is the identity map $\Delta_{k-1} \to \Delta_{k-1}$. It is easy to see that free discrete models are regular.

1.4.2 Multivariate Gaussian Models

A *multivariate Gaussian* model $\mathcal{N}(\mu, \Sigma)$ with mean $\mu \in \mathbb{R}^k$ and covariance matrix $\Sigma \in \mathbb{R}^{k\times k}_{\succ 0}$ is a statistical model with state space $\mathcal{X} = \mathbb{R}^k$ and probability density function given by

$$p(x) = \frac{1}{\sqrt{(2\pi)^k \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu)\right).$$

Here, $\mathbb{R}^{k\times k}_{\succ 0}$ is the cone of all $k\times k$ positive definite real matrices. It is also possible to define multivariate Gaussian models for positive semidefinite matrices Σ with zero determinant, but we will not study them in this chapter. In Sections 1.4.4 and 1.5, we will look at examples of multivariate Gaussian models where the mean and covariance matrix are parametrized over some space Ω by a polynomial map $(\mu, \Sigma) : \Omega \to \mathbb{R}^k \times \mathbb{R}^{k\times k}_{\succ 0}$.

If (μ, Σ) is the identity map $\mathbb{R}^k \times \mathbb{R}^{k\times k}_{\succ 0} \to \mathbb{R}^k \times \mathbb{R}^{k\times k}_{\succ 0}$, we say that the associated multivariate Gaussian model is *free*. Such models are regular. Indeed, it is not hard to see that the model is identifiable, and a little bit of work shows that the Fisher information matrix is Σ^{-1} which is positive definite if $\Sigma \in \mathbb{R}^{k\times k}_{\succ 0}$.

1.4.3 Exponential Families

An *exponential family* is a statistical model whose probability density function corresponding to a parameter $\omega \in \Omega$ can be written in the form

$$p(x|\omega) = h(x) \exp\big(\eta(\omega)^\mathsf{T} T(x) - A(\omega)\big)$$

for some functions $h : \mathcal{X} \to \mathbb{R}$, $T : \mathcal{X} \to \mathbb{R}^k$, $A : \Omega \to \mathbb{R}$ and $\eta : \Omega \to \mathbb{R}^k$, and where

$$\eta(\omega)^\mathsf{T} T(x) := \sum_{i=1}^{k} \eta_i(\omega) T_i(x)$$

is the dot product. Now, because $\int_\mathcal{X} p(x|\omega)dx = 1$, it follows that

$$A(\omega) = \log \int_\mathcal{X} h(x) \exp\big(\eta(\omega)^\mathsf{T} T(x)\big)\, dx \tag{1.7}$$

so $A(\omega)$ acts as a normalization factor and is called the *log-partition function*. Observe that the value of $A(\omega)$ only depends on that of $\eta(\omega)$, so we will sometimes write $A(\eta)$ to emphasize this dependence. In fact, the density $p(x|\omega)$ also only depends on the value of $\eta(\omega)$. For this reason, η is called the *natural parameter*.

The log-partition function plays an important role in maximum likelihood estimation. Let X be a random variable with true distribution $p(x|\omega)dx$ for some unknown $\omega \in \Omega$, and suppose $x_1, x_2, \ldots, x_N$ are N independent samples of X. Then, the likelihood of the data is

$$\prod_{i=1}^{N} p(x_i|\omega) = \prod_{i=1}^{N} h(x_i) \cdot \exp\left(\eta(\omega)^{\mathsf{T}} \sum_{i=1}^{N} T(x_i) - NA(\omega)\right). \tag{1.8}$$

Maximizing this likelihood is equivalent to maximizing

$$A(\omega) - \eta(\omega)^{\mathsf{T}}\hat{\mu}$$

where $\hat{\mu}$ is the sample mean

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} T(x_i).$$

Because the sample mean depends only on the $T(x_i)$, we call $T(x)$ the *sufficient statistic*. A maximum likelihood estimate $\hat{\eta}$ for the natural parameter η is then a solution to

$$\nabla A(\eta) = \hat{\mu}. \tag{1.9}$$

Consequently, if $\hat{\eta}$ is in the image of the map $\eta : \Omega \to \mathbb{R}^k$, the maximum likelihood estimates for the parameter ω will be the set

$$\eta^{-1}(\hat{\eta}) = \{\omega \in \Omega : \eta(\omega) = \hat{\eta}\}.$$

The log-partition function also has an interesting connection with the marginal likelihood integral, via the Kullback-Leibler divergence. We study this connection in Section 1.6.

Discrete models and multivariate Gaussian models are examples of exponential families. For discrete models with state probabilities $p_i(\omega)$ for $i = 1, \ldots, k$, we may write

$$p(i|\omega) = \exp(\eta(\omega)^{\mathsf{T}} T(i)), \quad i \in \{1, \ldots, k\}$$

where $\eta(\omega) \in \mathbb{R}^k$ with $\eta_i(\omega) = \log p_i(\omega)$, and $T(i) \in \mathbb{R}^k$ is the i-th standard basis vector. For multivariate Gaussian models $\mathcal{N}(\mu, \Sigma)$, where μ and Σ are functions of $\omega \in \Omega$, we set

$$\begin{aligned} h(x) &= \frac{1}{(2\pi)^{k/2}} \\ \eta(\Sigma, \mu) &= (\Sigma^{-1}, \Sigma^{-1}\mu) \\ T(x) &= (-\frac{1}{2}xx^T, x) \\ A(\Sigma, \mu) &= \frac{1}{2}\left(\log|\Sigma| + \langle \mu\mu^T, \Sigma^{-1}\rangle\right). \end{aligned}$$

Here, $\langle A, B\rangle$ is the matrix inner product, i.e. the trace of the matrix product AB.

1.4.4 Graphical Models

A graphical model is a statistical model that describes the conditional independence relationships between some random variables $X_1, X_2, \ldots, X_s$ by means of a graph $G = (V, E)$. Here, $V = \{1, 2, \ldots, s\}$ is the set of vertices which are in one-to-one correspondence with the random variables, and E is the set of edges which can be directed or undirected. In practice, the random variables X_i are usually discrete variables or Gaussian variables.

In this section, we will not go into the various ways conditional independence relationships are represented by the edges of the graph, but we refer the reader to books [16, 32, 35] which offer good introductions to the rich theory of graphical models. Because we are interested in Bayesian integrals which arise from parametrized models, we will describe parametrizations of three classes of graphical models, namely:

- Discrete models
  - Directed graphical models
  - Undirected graphical models
- Gaussian models
  - Mixed graph models

Now, let X be the collection $(X_1, \ldots, X_s)$ of random variables where each X_i is a discrete variable with k_i states. The state space of X is then the cartesian product $\prod_{i=1}^{s}\{1, \ldots, k_i\}$. Observations of X will be denoted by the lower case x. Given a subset $S \subset \{1, \ldots, s\}$, let X_S denote the random variable $(X_i)_{i \in S}$. Let $G = (V, E)$ be a *directed acyclic graph*, that is, a directed graph with no directed cycles. For each $i \in V = \{1, \ldots, s\}$, we define the *parents* $\mathrm{pa}(i)$ of i to be the set of all vertices j such that the directed edge $j \to i$ is in E. We say that a probability distribution on X *factors* according to G if

$$p(X = x) = \prod_{i=1}^{s} p(X_i = x_i | X_{\mathrm{pa}(i)} = x_{\mathrm{pa}(i)}). \tag{1.10}$$

A discrete *directed graphical model* is the statistical model $\mathcal{M}$ of all probability distributions on X which factor according to G. The model parameters are the conditional probabilities $p(X_i = x_i | X_{\mathrm{pa}(i)} = x_{\mathrm{pa}(i)})$ and the root probabilities $p(X_i = x_i)$ where $\mathrm{pa}(i) = \emptyset$. Therefore, in this model, a directed edge between two vertices indicates a direct *causal* relationship between the associated random variables.

We can also define discrete graphical models for *undirected* graphs. Suppose $G = (V, E)$ is an undirected *simple* graph, that is, a graph without self-loops or multiple edges between two vertices. A subset of the vertices is called a *clique* if every two vertices are connected by an edge. Let $\mathscr{C}$ be the set of maximal cliques in G. A discrete *undirected* graphical model is

a statistical model on X parametrized by

$$p(X = x) = \frac{1}{Z(\omega)} \prod_{C \in \mathscr{C}} \omega_{x_C}^{(C)} \tag{1.11}$$

where the $\omega_{x_C}^{(C)} \in \mathbb{R}_{\geq 0}$ are model parameters and $Z(\omega)$ is the required normalization factor so that all the probabilities sum to one. In the model, the cliques indicate subsets of random variables which are correlated with one another.

Finally, we describe Gaussian models for mixed graphs [51]. A mixed graph $G = (V, E)$ is a graph with three types of edges: undirected edges $i - j$, directed edges $i \to j$ and bidirected edges $i \leftrightarrow j$. We will assume that we have a partition $U \cup B$ of the vertices V such that all undirected edges have their vertices in U and all bidirected edges have their vertices in B. As for the directed edges, they are allowed to point $U \to U$, $B \to B$ or $U \to B$, but not $B \to U$, and we assume that the subgraph formed by the directed edges is acyclic. Between two vertices, we allow multiple edges of different types, but not more than one of each type. A *Gaussian mixed graph model* on G is a multivariate Gaussian model $\mathcal{N}(0, \Sigma)$ with zero mean and covariance matrix $\Sigma \in \mathbb{R}^{k \times k}, k = |V|$, parametrized as follows. First, we assume that the vertices of G are labeled $1, \ldots, k$ such that $u < b$ for all $u \in U$ and $b \in B$, and that $i < j$ for all directed edges $i \to j$. Let Λ be a $k \times k$ matrix with $\Lambda_{ij} = 0$ if $i \to j \notin E$ and $\Lambda_{ii} = 0$ for all i. Let K and Φ be symmetric positive definite matrices, with rows and columns indexed by U and by B respectively, such that $K_{ij} = 0$ if $i - j \notin E$, $\Phi_{ij} = 0$ if $i \leftrightarrow j \notin E$ and $K_{ii}, \Phi_{ii} > 0$ for all i. Now, we parametrize Σ using

$$\Sigma = (I - \Lambda)^{-\intercal} \begin{pmatrix} K^{-1} & 0 \\ 0 & \Phi \end{pmatrix} (I - \Lambda)^{-1}. \tag{1.12}$$

Hence, the model is parametrized by the $\Lambda_{ij}, K_{ij}, \Phi_{ij}$ corresponding to directed, undirected and bidirected edges of G, as well as the K_{ii}, Φ_{ii} corresponding to vertices in U and B. These parameters are subject to the positive definite conditions on K and Φ. Ancestral graphs [44] and chain graphs [1] are special cases of mixed graphs, and mixed graph models are used frequently in structural equation modeling [23].

1.4.5 Mixtures of Independence Models

We consider a collection of discrete random variables

$$\begin{matrix} X_1^{(1)}, & X_2^{(1)}, & \ldots, & X_{s_1}^{(1)}, \\ X_1^{(2)}, & X_2^{(2)}, & \ldots, & X_{s_2}^{(2)}, \\ \vdots & \vdots & \ddots & \vdots \\ X_1^{(k)}, & X_2^{(k)}, & \ldots, & X_{s_k}^{(k)}, \end{matrix}$$

where $X_1^{(i)}, \ldots, X_{s_i}^{(i)}$ are identically distributed with state space $\{0, 1, \ldots, t_i\}$. Note that here the integer 0 is included in the state spaces, for notational reasons.

The *independence model* $\mathcal{M}$ for these variables is a toric model [41, §1.2] represented by an integer $d \times n$-matrix A with

$$d \,=\, t_1 + t_2 + \cdots + t_k + k \quad \text{and} \quad n \,=\, \prod_{i=1}^{k} (t_i + 1)^{s_i}. \tag{1.13}$$

The columns of the matrix A are indexed by elements v of the state space

$$\{0, 1, \ldots, t_1\}^{s_1} \times \{0, 1, \ldots, t_2\}^{s_2} \times \cdots \times \{0, 1, \ldots, t_k\}^{s_k}. \tag{1.14}$$

The rows of the matrix A are indexed by the model parameters, which are the d coordinates of the points $\theta = (\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(k)})$ in the polytope

$$P \;\;=\;\; \Delta_{t_1} \times \Delta_{t_2} \times \cdots \times \Delta_{t_k}, \tag{1.15}$$

and the model $\mathcal{M}$ is the subset of the simplex Δ_{n-1} given parametrically by

$$p_v \;\;=\;\; \text{Prob}\big(X_j^{(i)} = v_j^{(i)} \text{ for all } i, j \big) \;\;=\;\; \prod_{i=1}^{k} \prod_{j=1}^{s_i} \theta^{(i)}_{v_j^{(i)}}. \tag{1.16}$$

This is a monomial in d unknowns. The matrix A is defined by taking its column a_v to be the exponent vector of this monomial.

The independence model may also be thought of as a discrete graphical model represented by Figure 1.2. In this diagram, the boxes around the random variables are called *plates*. Each plate indicates that we have several independent and identically distributed variables, while the number of copies is shown in its corner.

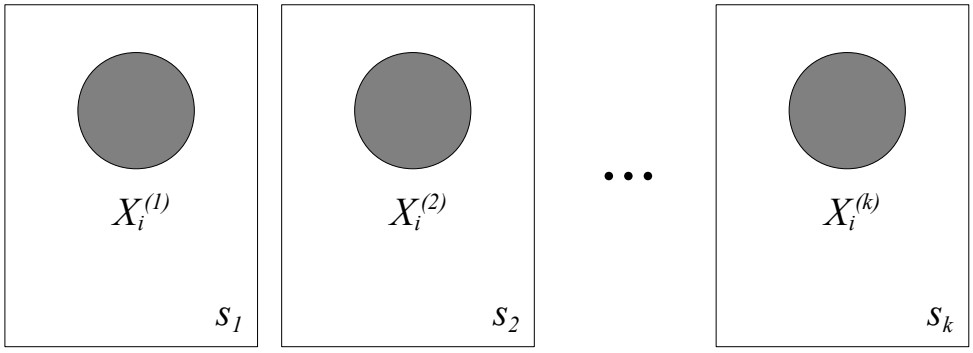


Figure 1.2: Graphical model representation of the independence model.

In algebraic geometry, the model $\mathcal{M}$ is known as *Segre-Veronese variety*

$$\mathbb{P}^{t_1} \times \mathbb{P}^{t_2} \times \cdots \times \mathbb{P}^{t_k} \quad \hookrightarrow \quad \mathbb{P}^{n-1}, \tag{1.17}$$

where the embedding is given by the line bundle $\mathcal{O}(s_1, s_2, \ldots, s_k)$. The manifold $\mathcal{M}$ is the toric variety of the polytope P. Both objects have dimension $d-k$, and they are identified with each other via the moment map [21, §4].

Example 1.5. Consider three binary random variables where the last two random variables are identically distributed. In our notation, this corresponds to $k = 2$, $s_1 = 1$, $s_2 = 2$ and $t_1 = t_2 = 1$. We find that $d = 4, n = 8$, and

$$A \;=\; \begin{array}{c} \\ \theta_0^{(1)} \\ \theta_1^{(1)} \\ \theta_0^{(2)} \\ \theta_1^{(2)} \end{array} \begin{array}{c} \begin{array}{cccccccc} p_{000} & p_{001} & p_{010} & p_{011} & p_{100} & p_{101} & p_{110} & p_{111} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 2 & 1 & 1 & 0 & 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 2 & 0 & 1 & 1 & 2 \end{pmatrix} \end{array}.$$

The columns of this matrix represent the monomials in the parametrization (1.16). The model $\mathcal{M}$ lies in the 5-dimensional subsimplex of Δ_7 given by $p_{001} = p_{010}$ and $p_{101} = p_{110}$, and it consists of all rank one matrices

$$\begin{pmatrix} p_{000} & p_{001} & p_{100} & p_{101} \\ p_{010} & p_{011} & p_{110} & p_{111} \end{pmatrix}.$$

In algebraic geometry, the surface $\mathcal{M}$ is called a *rational normal scroll*. □

The matrix A has repeated columns whenever $s_i \geq 2$ for some i. It is sometimes convenient to represent the model $\mathcal{M}$ by the matrix $\tilde{A}$ which is obtained from A by removing repeated columns. We label the columns of $\tilde{A}$ by elements $v = (v^{(1)}, \ldots, v^{(k)})$ of (1.14) whose components $v^{(i)} \in \{0, 1, \ldots, t_i\}^{s_i}$ are weakly increasing. Hence $\tilde{A}$ is a $d \times \tilde{n}$-matrix with

$$\tilde{n} \;=\; \prod_{i=1}^{k} \binom{s_i + t_i}{s_i}. \tag{1.18}$$

The model $\mathcal{M}$ and its mixtures are subsets of a subsimplex $\Delta_{\tilde{n}-1}$ of Δ_{n-1}.

The *mixture model* $\mathcal{M}^{(2)}$ is the set of distributions which are convex combinations of two distributions in $\mathcal{M}$. The natural parameter space of this model is the polytope

$$\Theta \;=\; \Delta_1 \,\times\, P \,\times\, P.$$

Let $a_v \in \mathbb{N}^d$ be the column vector of A indexed by the state v, which is either in (1.14) or in $\{1, 2, \ldots, n\}$. The parametrization (1.16) can be written simply as $p_v = \theta^{a_v}$. The mixture model $\mathcal{M}^{(2)}$ is defined to be the subset of Δ_{n-1} with the parametric representation

$$p_v \;=\; \sigma_0 \cdot \theta^{a_v} \,+\, \sigma_1 \cdot \rho^{a_v} \qquad \text{for } (\sigma, \theta, \rho) \,\in\, \Theta.$$

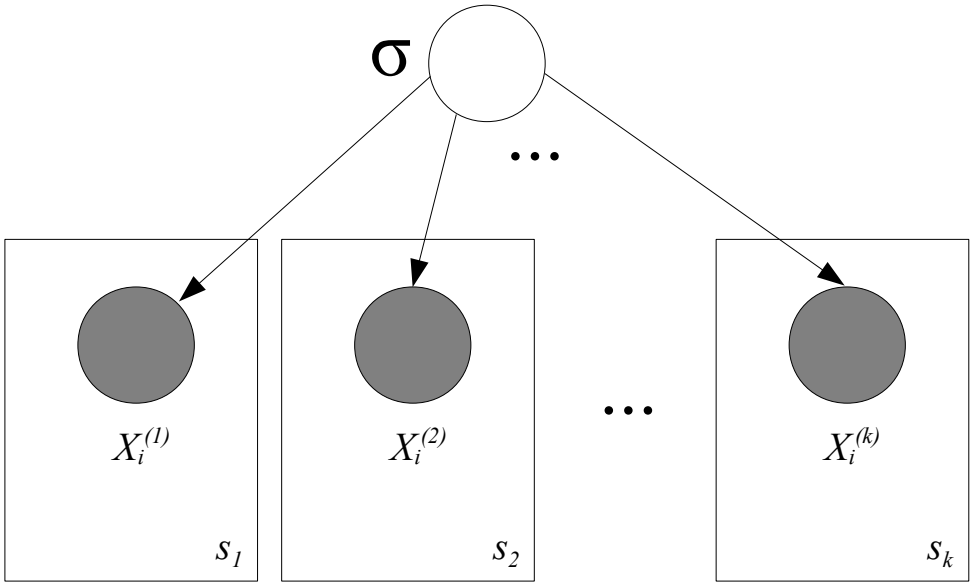


Figure 1.3: Graphical model representation of the mixture model.

This mixture model can also be represented by a discrete directed graphical model, as shown in Figure 1.3. In this diagram, the vertex labelled σ corresponds to a binary random variable. The shaded vertices are observed random variables, while the unshaded vertices are hidden.

In algebraic geometry, the model $\mathcal{M}^{(2)}$ is known as the first secant variety of the Segre-Veronese variety (1.17). We could also consider the higher secant varieties $\mathcal{M}^{(l)}$ which correspond to mixtures of l independent distributions, and much of our analysis can be extended to that case, but for simplicity we restrict ourselves to $l = 2$. The variety $\mathcal{M}^{(2)}$ is embedded in the projective space $\mathbb{P}^{\tilde{n}-1}$ with $\tilde{n}$ as in (1.18). Note that $\tilde{n}$ can be much smaller than n. If this is the case, it is convenient to aggregate states whose probabilities are identical and represent the data by a vector $\tilde{U} \in \mathbb{N}^{\tilde{n}}$. Here is an example.

Example 1.6. Let k=1, s_1=4 and t_1=1, so $\mathcal{M}$ is the independence model for four identically distributed binary random variables. Then $d = 2$ and $n = 16$. The corresponding integer matrix and its row and column labels are

$$A \;=\; \begin{matrix} \\ \theta_0 \\ \theta_1 \end{matrix} \begin{matrix} \begin{matrix} p_{0000} & p_{0001} & p_{0010} & p_{0100} & p_{1000} & p_{0011} & \cdots & p_{1110} & p_{1111} \end{matrix} \\ \begin{pmatrix} 4 & 3 & 3 & 3 & 3 & 2 & \cdots & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 2 & \cdots & 3 & 4 \end{pmatrix} \end{matrix}.$$

However, this matrix has only $\tilde{n} = 5$ distinct columns, and we instead use

$$\tilde{A} \;=\; \begin{matrix} \\ \theta_0 \\ \theta_1 \end{matrix} \begin{matrix} \begin{matrix} p_0 & p_1 & p_2 & p_3 & p_4 \end{matrix} \\ \begin{pmatrix} 4 & 3 & 2 & 1 & 0 \\ 0 & 1 & 2 & 3 & 4 \end{pmatrix} \end{matrix}.$$

The mixture model $\mathcal{M}^{(2)}$ is the subset of Δ_4 given by the parametrization

$$p_i \quad = \quad \binom{4}{i} \cdot \left(\sigma_0 \cdot \theta_0^{4-i} \cdot \theta_1^i \;+\; \sigma_1 \cdot \rho_0^{4-i} \cdot \rho_1^i\right) \qquad \text{for } i = 0, 1, 2, 3, 4.$$

In algebraic geometry, this threefold is the secant variety of the rational normal curve in $\mathbb{P}^4$. This is the cubic hypersurface with the implicit equation

$$\det \begin{bmatrix} 12p_0 & 3p_1 & 2p_2 \\ 3p_1 & 2p_2 & 3p_3 \\ 2p_2 & 3p_3 & 12p_4 \end{bmatrix} \quad = \quad 0.$$

In [30, Example 9], the likelihood function (2.4) was studied for the data

$$\tilde{U} \quad = \quad (\tilde{U}_0, \tilde{U}_1, \tilde{U}_2, \tilde{U}_3, \tilde{U}_4) \quad = \quad (51, 18, 73, 25, 75).$$

Using Gröbner bases techniques, we find that it has three local maxima (modulo swapping θ and ρ) whose coordinates are algebraic numbers of degree 12. □

In Chapter 2, we examine marginal likelihood integrals for this class of mixture models for discrete data. Our study augments developments in the asymptotic theory of these integrals by providing tools for exact symbolic integration when the sample size is small. These exact results can then serve as a gold standard against which the accuracy of approximation and asymptotic methods can be ascertained.

The numerical value of the integral we have in mind is a rational number, and exact evaluation means computing that rational number rather than a floating point approximation. For a first example consider the integral

$$\int_\Theta \prod_{i,j \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \left(\, \pi \lambda_i^{(1)} \lambda_j^{(2)} + \tau \rho_i^{(1)} \rho_j^{(2)} \,\right)^{U_{ij}} d\pi \, d\tau \, d\lambda \, d\rho, \tag{1.19}$$

where Θ is the 13-dimensional polytope $\Delta_1 \times \Delta_3 \times \Delta_3 \times \Delta_3 \times \Delta_3$. The factors in this product are the probability simplices

$$\begin{array}{rcll} \Delta_1 & = & \{(\pi, \tau) \in \mathbb{R}^2_{\geq 0} : \pi + \tau = 1\}, & \\ \Delta_3 & = & \{(\lambda_{\mathtt{A}}^{(k)}, \lambda_{\mathtt{C}}^{(k)}, \lambda_{\mathtt{G}}^{(k)}, \lambda_{\mathtt{T}}^{(k)}) \in \mathbb{R}^4_{\geq 0} & : \sum_i \lambda_i^{(k)} = 1\}, \quad k = 1, 2, \\ \Delta_3 & = & \{(\rho_{\mathtt{A}}^{(k)}, \rho_{\mathtt{C}}^{(k)}, \rho_{\mathtt{G}}^{(k)}, \rho_{\mathtt{T}}^{(k)}) \in \mathbb{R}^4_{\geq 0} & : \sum_i \rho_i^{(k)} = 1\}, \quad k = 1, 2. \end{array}$$

and we integrate with respect to Lebesgue probability measure on Θ. If we take the exponents U_{ij} to be the entries of the particular contingency table

$$U = \begin{pmatrix} 4 & 2 & 2 & 2 \\ 2 & 4 & 2 & 2 \\ 2 & 2 & 4 & 2 \\ 2 & 2 & 2 & 4 \end{pmatrix}, \tag{1.20}$$

then the exact value of the integral (1.19) is the rational number

$$\frac{571 \cdot 773426813 \cdot 17682039596993 \cdot 625015426432626533}{2^{31} \cdot 3^{20} \cdot 5^{12} \cdot 7^{11} \cdot 11^{8} \cdot 13^{7} \cdot 17^{5} \cdot 19^{5} \cdot 23^{5} \cdot 29^{3} \cdot 31^{3} \cdot 37^{3} \cdot 41^{3} \cdot 43^{2}}. \tag{1.21}$$

The table (1.20) is taken from Example 1.3 of [41], where the integrand

$$\prod_{i,j \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \big(\pi \lambda_i^{(1)} \lambda_j^{(2)} + \tau \rho_i^{(1)} \rho_j^{(2)} \big)^{U_{ij}} \tag{1.22}$$

was studied using the EM algorithm, and the problem of validating its global maximum over Θ was raised. See [19, §4.2] and [50, §3] for further discussions. That optimization problem, which was widely known as the 100 *Swiss Francs problem*, has since been solved [22].

The main difficulty in performing computations such as (1.19) = (1.21) lies in the fact that the expansion of the integrand has many terms. A first naive upper bound on the number of monomials in the expansion of (1.22) would be

$$\prod_{i,j \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} (U_{ij}+1) \quad = \quad 3^{12} \cdot 5^4 \quad = \quad 332,150,625.$$

However, the true number of monomials is only $3,892,097$, and we obtain the rational number (1.21) by summing the values of the corresponding integrals

$$\int_\Theta \pi^{a_1} \tau^{a_2} (\lambda^{(1)})^u (\lambda^{(2)})^v (\rho^{(1)})^w (\rho^{(2)})^x d\pi\, d\tau\, d\lambda\, d\rho \quad =$$
$$\frac{a_1!\, a_2!}{(a_1+a_2+1)!} \cdot \frac{3! \prod_i u_i!}{(\sum_i u_i + 3)!} \cdot \frac{3! \prod_i v_i!}{(\sum_i v_i + 3)!} \cdot \frac{3! \prod_i w_i!}{(\sum_i w_i + 3)!} \cdot \frac{3! \prod_i x_i!}{(\sum_i x_i + 3)!}.$$

The geometric idea behind our approach is that the Newton polytope of (1.22) is a *zonotope* and we are summing over its lattice points. Definitions for these geometric objects are given in Section 2.2, and algorithms implementing these ideas are described in Section 2.3. The `Maple` library for our algorithms is made available at

`http://math.berkeley.edu/~shaowei/integrals.html`.

1.5 Regularly Parametrized Models

Previously, we saw in Section 1.3.2 that the crux to understanding a singular model lies in desingularizing its Kullback-Leibler distance, which is an integral or sum of log functions. While general algorithms for desingularizing any analytic function exist [6,25], applying them to non-polynomial functions such as the Kullback-Leibler distance can be computationally prohibitive. Many singular models are however defined by polynomial maps. Our goal is to exploit this polynomiality in understanding such singular models.

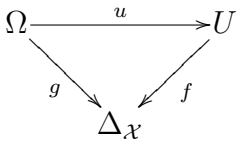
In this section, we accomplish our goal by introducing *fiber ideals* for a general class of statistical models known as *regularly parametrized models*. Parametrized discrete models and multivariate Gaussian models are all examples of such models. We show that monomializing the fiber ideal allows us to construct desingularizations of the Kullback-Leibler distance. In fact, many invariants of statistical models such as the learning coefficient can be computed directly from fiber ideals. Computationally, monomializing a polynomial ideal is often easier than monomializing a non-polynomial analytic function. In some cases, this monomialization can be achieved simply by inspection.

1.5.1 Fiber Ideals

Let us introduce what it means for a model to be regularly parametrized. Informally, we may think of such models as regular models whose parameters are functions of new parameters. Consider a regular model $\mathcal{M}_f$ on a state space $\mathcal{X}$ parametrized by a space U and whose probability density function at each $u \in U$ is $f_x(u) := p(x|u, \mathcal{M}_f)$. Suppose $\mathcal{M}_g$ is another model on $\mathcal{X}$ parametrized by a space Ω whose probability density function at each $\omega \in \Omega$ is $g_x(\omega) := p(x|\omega, \mathcal{M}_g)$. Let $u : \Omega \to U$ be a real analytic map.

Definition 1.7. We say that $\mathcal{M}_g$ is *regularly parametrized via u with base $\mathcal{M}_f$* if $g_x(\omega) = f_x(u(\omega))$ for each $\omega \in \Omega$.

We represent this relationship by the following commutative diagram, where $\Delta_{\mathcal{X}}$ denotes the set of probability distributions on $\mathcal{X}$.



In other words, $g = f \circ u$ and the model $\mathcal{M}_g$ factors through the regular model $\mathcal{M}_f$. It is computationally favorable for the map u to be polynomial, but we will not require this here. Free discrete models and free multivariate Gaussian models are regular, so parametrized versions of these models are examples of regularly parametrized models.

Now, let $\hat{u}$ be a point in the parameter space U of the regular base model $\mathcal{M}_f$. Suppose that $U \subset \mathbb{R}^k$ so the map u has coordinate functions $u_1, \ldots, u_k$.

Definition 1.8. The *fiber ideal* $I_{\hat{u}}$ of $\mathcal{M}_g$ at $\hat{u}$ is the ideal

$$\langle u(\omega) - \hat{u} \rangle := \langle u_1(\omega) - \hat{u}_1, \ldots, u_k(\omega) - \hat{u}_k \rangle$$

in the ring $\mathcal{A}_\Omega$ of real-valued analytic functions on Ω. The variety of this ideal is the fiber $\{\omega \in \Omega : u(\omega) = \hat{u}\}$ in Ω of the map u over the point $\hat{u}$.

Suppose also that $\Omega \subset \mathbb{R}^d$, so each point ω has coordinates $\omega_1, \ldots, \omega_d$. If the map u is polynomial in these coordinates, then the polynomial ideal in $\mathbb{R}[\omega_1, \ldots, \omega_d] \subset \mathcal{A}_\Omega$ generated by $u_1(\omega) - \hat{u}_1, \ldots, u_k(\omega) - \hat{u}_k$ will be contained as a set in the fiber ideal $I_{\hat{u}}$. The ring $\mathcal{A}_\Omega$ also has many polynomial functions such as $1 + \omega_1^2$ which are nonzero over Ω and are therefore units in the ring. Multiplication or division by these units leaves the fiber ideal unchanged.

Example 1.9. Discrete models are regularly parametrized, so we may define fiber ideals at each point $\hat{p}$ in the simplex Δ_{k-1} where k is the number of states. If the discrete model is described by a map $p : \Omega \to \Delta_{k-1}$, then the fiber ideal $I_{\hat{p}}$ at $\hat{p} \in \Delta_{k-1}$ is

$$\langle p_1(\omega) - \hat{p}_1, \ldots, p_k(\omega) - \hat{p}_k \rangle.$$

In fact, we may leave out the generator $p_k(\omega) - \hat{p}_k$ because

$$p_k(\omega) - \hat{p}_k = -\sum_{i=1}^{k-1} \big(p_i(\omega) - \hat{p}_i\big).$$

For discrete directed graphical models, each $p_i(\omega)$ is a polynomial (1.10) in the conditional and root probabilities, so the fiber ideal is finitely generated by polynomials. For undirected graphical models, each $p_i(\omega)$ is a rational function (1.11) but the normalization factor $Z(\omega)$ is a polynomial which does not vanish over the parameter space Ω. Hence, $Z(\omega)$ is a unit in $\mathcal{A}_\Omega$ and the fiber ideal is once again finitely generated by the polynomials

$$\prod_{C \in \mathscr{C}} \omega_{x_C}^{(C)} \;-\; \hat{p}_x Z(\omega), \quad \text{for } x \in \prod_{i=1}^{s} \{1, \ldots, k_i\}.$$

Note that the ambient ring $\mathcal{A}_\Omega$ of this ideal changes with the parameter space Ω. □

Example 1.10. Multivariate Gaussian models are regularly parametrized, and the fiber ideal at a point $(\hat{\mu}, \hat{\Sigma}) \in \mathbb{R} \times \mathbb{R}^{k \times k}_{\succ 0}$ is generated by

$$\begin{aligned} \mu_i(\omega) - \hat{\mu}_i, &\quad \text{for } 1 \le i \le k, \\ \Sigma_{ij}(\omega) - \hat{\Sigma}_{ij}, &\quad \text{for } 1 \le i, j \le k. \end{aligned}$$

For Gaussian mixed graph models, $\Sigma(\omega)$ is parametrized using a product (1.12) of inverse matrices, and each entry in an inverse matrix is a cofactor divided by the matrix determinant. The matrix $I - \Lambda$ is upper triangular with ones in the diagonal, so its determinant is 1. As for the matrix K, it is positive definite by definition, so its determinant is nonzero. Thus, the $\Sigma_{ij}(\omega)$ are rational functions whose denominator $\det K$ does not vanish over the parameter space Ω. Multiplying by this unit, we see that the fiber ideal is again finitely generated by the polynomials

$$\Sigma_{ij}(\omega) - \hat{\Sigma}_{ij} \det K, \quad \text{for } 1 \le i, j \le k.$$

Each $\Sigma_{ij}(\omega)$ can be expressed as a sum of path monomials over all *treks* between vertices i and j in the mixed graph [51]. □

1.5.2 Real Log Canonical Thresholds

Just as we defined real log canonical thresholds (RLCTs) for functions in Section 1.3.2, we can also define RLCTs for ideals. In fact, the learning coefficient of a regularly parametrized model can be computed from the real log canonical threshold of its fiber ideal.

Let $\Omega \subset \mathbb{R}^d$ be a compact semianalytic set where *semianalytic* means that Ω is defined by analytic inequalities

$$\Omega = \{\omega \in \mathbb{R}^d : g_1(\omega) \geq 0, \ldots, g_l(\omega) \geq 0\}.$$

We also require the interior of Ω to be nonempty; otherwise, integrals over Ω will be identically zero. Let $I = \langle f_1, \ldots, f_r \rangle$ be an ideal in the ring $\mathcal{A}_\Omega$ of real analytic functions over Ω. Let $\varphi : \Omega \to \mathbb{R}$ be *nearly analytic*, i.e. φ is a product $\varphi_a \varphi_s$ of functions where φ_a is real analytic and φ_s is positive and smooth. Define the *real log canonical threshold* $\mathrm{RLCT}_\Omega(I; \varphi)$ of I to be the pair (λ, θ) where λ is the smallest pole of the zeta function

$$\zeta(z) = \int_\Omega \left(f_1(\omega)^2 + \cdots + f_r(\omega)^2\right)^{-z/2} |\varphi(\omega)| d\omega, \quad z \in \mathbb{C}$$

and θ its multiplicity. This definition is independent of the choice of generators $f_1, \ldots, f_r$ of the ideal I. In Chapter 4, we prove fundamental properties of RLCTs of ideals and explore their relationship to Newton polyhedra in nondegenerate cases. In particular, we show how to compute the RLCT of a monomial ideal when Ω is a sufficiently small neighborhood of the origin and φ is a monomial function. A `Singular` library which checks the nondegeneracy of functions and ideals, and computes the RLCT of monomial ideals, is made available at

`http://math.berkeley.edu/~shaowei/rlct.html`.

For regularly parametrized models, we now state the relationship between their learning coefficients and real log canonical thresholds of their fiber ideals. This theorem is one of the main contributions of this dissertation.

Theorem 1.11. *Let $\mathcal{M}$ and $\mathcal{M}_R$ be models for X with parameter spaces U and Ω respectively such that $\mathcal{M}$ is regularly parametrized with base $\mathcal{M}_R$ via the map $u : \Omega \to U$. Assuming that the true distribution is $q(x)dx = p(x|\hat{u}, \mathcal{M}_R)dx$ for some $\hat{u} \in U$, let (λ, θ) denote the learning coefficient of $\mathcal{M}$ subject to $q(x)dx$. Let $I_{\hat{u}} = \langle u(\omega) - \hat{u} \rangle$ be the fiber ideal of $\mathcal{M}$ at $\hat{u}$.*

If the variety $\mathcal{V} = \{\omega \in \Omega : u(\omega) = \hat{u}\}$ is nonempty, then

$$(2\lambda, \theta) = \min_{\omega \in \mathcal{V}} \mathrm{RLCT}_{\Omega_\omega}(I_{\hat{u}}; \varphi)$$

where each Ω_ω is a sufficiently small neighborhood of ω in Ω.

Proof. Let $K : U \to \mathbb{R}$ be the Kullback-Leibler function

$$u \mapsto \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x|u, \mathcal{M}_R)} dx.$$

Since $q(x) = p(x|\hat{u}, \mathcal{M}_R)$, this function achieves its minimum at $u = \hat{u}$ so $K(\hat{u}) = 0$ and $\nabla K(\hat{u}) = 0$. The identifiability of $\mathcal{M}_R$ ensures that this minimum point is unique. According to Theorem 1.3, the learning coefficient (λ, θ) is given by

$$(\lambda, \theta) = \mathrm{RLCT}_{\Omega}(K \circ u(\omega); \varphi).$$

Note that $\{\omega \in \Omega : K \circ u(\omega) = 0\} = \{\omega \in \Omega : u(\omega) = \hat{u}\} = \mathcal{V}$. Using Proposition 4.2,

$$(\lambda, \theta) = \min_{\omega \in \mathcal{V}} \mathrm{RLCT}_{\Omega_\omega}(K \circ u(\omega); \varphi).$$

Finally, because $\mathcal{M}_R$ is regular, the Fisher information matrix $\nabla^2 K(\hat{u})$ is positive definite so we apply Proposition 4.4 to get

$$(2\lambda, \theta) = \min_{\omega \in \mathcal{V}} \mathrm{RLCT}_{\Omega_\omega}(\langle u(\omega) - \hat{u} \rangle; \varphi). \qquad \square$$

Remark 1.12. The inspiration to reduce the Kullback-Leibler function to a sum of squares of polynomial parametrizing the model came from a discussion with Sumio Watanabe about learning coefficients of discrete models in September 2008. This eventually led to the idea of defining fiber ideals and studying their real log canonical thresholds.

During a converseration in October 2010, Mathias Drton asked if fiber ideals can also be defined for Gaussian models. His question prompted the author to introduce the concept of regularly parametrized models and to extend fiber ideals to all such models. $\square$

1.5.3 Desingularizing the Kullback-Leibler Function

In the previous section, we saw that the learning coefficient can be computed by monomializing the fiber ideal I of our singular model. This monomialization can be described by a real analytic manifold M covered by coordinate charts M_i and a proper real analytic map $\rho : M \to \Omega$ such that the *pullback ideal* $\rho^* I = \{f \circ \rho \in \mathcal{A}_M : f \in I\}$ is monomial in each chart, where $\mathcal{A}_M$ represents the ring of real analytic functions on M. Furthermore, we will also require that the map ρ is an isomorphism between $\Omega \setminus \mathcal{V}(I)$ and $M \setminus \mathcal{V}(\rho^* I)$. Here, $\mathcal{V}(I)$ is the analytic variety $\{\omega \in \Omega : f(\omega) = 0 \ \forall f \in I\}$, and similarly for $\mathcal{V}(\rho^* I)$. If ρ satisfies this property, we say that it is a *change of variable away from* $\mathcal{V}(I)$. If in each chart with coordinates $x_1, \ldots, x_d$, the Jacobian determiant of ρ also equals

$$|\rho'(x)| = b(x) x_1^{t_1} \cdots x_d^{t_d}$$

where the t_1 are non-negative integers and $b(x)$ does not vanish for all x, we say that ρ is a monomialization map for I.

In fact, this monomialization step provides much more information than just the learning coefficient. It is the key to desingularizing the Kullback-Leibler function of the model, thus allowing us to express the log likelihood ratio in standard form (see Section 1.3.2). The last ingredient we need for desingularizing the Kullback-Leibler function is *principalizing* our monomial fiber ideal. The principalization of ideals is a topic of great interest in algebraic geometry, see [11, 52, 59] and [34, §3.3]. We now explain one approach to this process.

Let M be a real analytic manifold, and let I be an ideal in the ring $\mathcal{A}_M$ of real analytic functions on M. Suppose we can cover M with coordinate charts M_i so that the ideal I is monomial in each chart. Then, according to Goward [24], there is a real analytic manifold $\mathscr{M}$ covered by coordinate charts $\mathscr{M}_i$ and a proper real analytic map $\rho_G : \mathscr{M} \to M$ such that the pullback ideal $\rho_G^* I$ is monomial and principal in each chart. Here, *principal* means that the ideal is generated by exactly one function. Furthermore, ρ_G is the composition of a sequence of blowups whose centers are the intersection of coordinate hyperplanes in each chart M_i or charts generated by previous blowups. Goward showed that there is a simple combinatorial algorithm for calculating this sequence of blowups. We call ρ_G the *Goward principalization map* for the ideal I.

Let us describe briefly what these blowups look like. Suppose we have a chart $V \subset \mathbb{R}^d$ with coordinates $\omega_1, \ldots, \omega_d$, and we want to blow up this chart with respect to the center $\{\omega \in \mathbb{R}^d : \omega_1 = \cdots = \omega_r = 0\}, 2 \leq r \leq d$. We define the *blowup space*

$$\tilde{V} = \{((\omega_1, \ldots, \omega_d), (\xi_1 : \ldots : \xi_r)) \in V \times \mathbb{P}^{r-1} : \omega_i \xi_j = \omega_j \xi_i, \quad i, j = 1, \ldots, r\}$$

where $(\xi_1 : \ldots : \xi_r)$ are homogeneous coordinates of the $(r-1)$-dimensional projective space $\mathbb{P}^{r-1}$. The *blowup map* $\pi : \tilde{V} \to V$ is the projection $(\omega, \xi) \mapsto \omega$. The coordinate charts

$$\tilde{V}_i = \{(\omega, \xi) \in \tilde{V} : \xi_i \neq 0\}, \quad i = 1, \ldots, r.$$

with natural local coordinates $(\omega_1^{(i)}, \ldots, \omega_d^{(i)})$ satisfying

$$\omega_j^{(i)} = \begin{cases} \omega_j, & j = i \text{ or } j > r, \\ \xi_j / \xi_i, & \text{otherwise.} \end{cases}$$

cover the blowup space $\tilde{V}$, and the blowup map can be expressed as

$$\omega_j = \begin{cases} \omega_j^{(i)}, & j = i \text{ or } j > r, \\ \omega_i^{(i)} \omega_j^{(i)}, & \text{otherwise} \end{cases}$$

in terms of the local coordinates of the chart $\tilde{V}_i$.

If $M \subset \mathbb{R}^d$ is a subset with coordinates $\omega_1, \ldots, \omega_d$ and the ideal I is monomial in these coordinates, then we may also use the Newton polyhedra methods in Section 4.2 to principalize the ideal. More specifically, if $\mathcal{F}$ is a smooth refinement of the normal fan of the Newton polyhedron $\mathcal{P}(I)$, then the associated toric blowup $\rho_{\mathcal{F}} : \mathbb{P}(\mathcal{F}) \to M$ is a principalization map for I. This technique is commonly used for the resolution of singularities in toric varieties. In comparison, Goward's principalization method is more general, because it applies to locally monomial ideal sheaves over *any* real analytic manifold.

With these principalization maps in place, we may now desingularize the Kullback-Leibler function for regularly parametrized models. Our next result may be thought of as an extension of Proposition 4.4.

Theorem 1.13. *Let $U \subset \mathbb{R}^d$. Let the maps $u : \Omega \to U$ and $K : U \to \mathbb{R}$ be real analytic at $0 \in \Omega$ and $\hat{u} = u(0) \in U$ respectively. Suppose that $K(\hat{u}) = 0, \nabla K(\hat{u}) = 0$ and the Hessian $\nabla^2 K(\hat{u})$ is positive definite. Let $I \subset \mathcal{A}_\Omega$ be the ideal $\langle u(\omega) - \hat{u} \rangle$.*

Let $\rho : M \to \Omega$ be a monomialization map for I, and suppose $\rho_G : \mathscr{M} \to M$ is the Goward principalization map or a toric principalization map for $\rho^* I$. Then, $\rho \circ \rho_G$ desingularizes $K \circ u$ at the origin.

Proof. By applying a translation to the subset U of $\mathbb{R}^d$, we may assume without loss of generality that $\hat{u}$ is the origin $0 \in U$. Since we are only interested in desingularizing $K \circ u$ at the origin $0 \in \Omega$, we may assume that Ω and U are sufficiently small neighborhoods of their origins such that $K(u) = 0$ if and only if $u = 0$.

The proof of Proposition 4.4 tells us that there is a linear change of coordinates $T : \mathbb{R}^d \to \mathbb{R}^d$ such that the power series expansion of K is $v_1^2 + \cdots + v_d^2 + O(v^3)$ where $(v_1, \ldots, v_d) = T(u_1, \ldots, u_d)$. The Morse lemma [40, §2.2] says that in fact

$$K \circ T^{-1}(v) = (v_1 + g_1(v))^2 + \ldots + (v_d + g_d(v))^2$$

for some functions $g_i(v) = O(v^2)$ which are real analytic at the origin.

Now, if we consider $v(\omega) = T \circ u(\omega)$, note that the coordinate functions $v_1(\omega), \ldots, v_d(\omega)$ generate the same ideal as $u_1(\omega), \ldots, u_d(\omega)$. In each chart of $\mathscr{M}$, the functions

$$u_1 \circ \rho \circ \rho_G(\mu), \ldots, u_d \circ \rho \circ \rho_G(\mu)$$

generate a principal monomial ideal $\langle \mu_1^{\kappa_1} \cdots \mu_d^{\kappa_d} \rangle$ for some non-negative integers κ_i. Thus,

$$v_i \circ \rho \circ \rho_G(\mu) = \mu^\kappa q_i(\mu)$$

for some functions $q_i(\mu) \in \mathcal{A}_{\mathscr{M}}$. Because μ^κ is generated by the $v_i \circ \rho \circ \rho_G(\mu)$, we have

$$\mu^\kappa = (q_1(\mu) r_1(\mu) + \cdots + q_d(\mu) r_d(\mu))\, \mu^\kappa$$

for some functions $r_i(\mu) \in \mathcal{A}_{\mathscr{M}}$. Since the interior of $\mathscr{M}$ is nonempty, the ring $\mathcal{A}_{\mathscr{M}}$ does not have any zero divisors, so we get

$$1 = q_1(\mu) r_1(\mu) + \cdots + q_d(\mu) r_d(\mu). \tag{1.23}$$

Applying the change of variable $\rho \circ \rho_G$ to $K \circ u$, we have

$$\begin{aligned} K \circ u \circ \rho \circ \rho_G(\mu) &= K \circ T^{-1}(v \circ \rho \circ \rho_G(\mu)) \\ &= (\mu^\kappa q_1 + g_1(\mu^\kappa q))^2 + \ldots + (\mu^\kappa q_d + g_d(\mu^\kappa q))^2 \\ &= \mu^\kappa((q_1 + \mu^\kappa h_1)^2 + \ldots + (q_d + \mu^\kappa h_d)^2) \end{aligned}$$

for some functions $h_i(\mu)$. To prove that $\rho \circ \rho_G$ desingularizes $K \circ u$, we claim that the function

$$a(\mu) = (q_1 + \mu^\kappa h_1)^2 + \ldots + (q_d + \mu^\kappa h_d)^2$$

does not vanish in $(\rho \circ \rho_G)^{-1}(\Omega)$. Indeed, if $a(\mu) = 0$, then $K(u \circ \rho \circ \rho_G(\mu)) = 0$. Therefore, $u \circ \rho \circ \rho_G(\mu)$ is the origin, and so is $v \circ \rho \circ \rho_G(\mu)$. This implies that $\mu^\kappa q_i(\mu) = 0$ for all i. If $\mu^\kappa \neq 0$, then $q_i(\mu) = 0$ for all i, which contradicts (1.23). Now, suppose $\mu^\kappa = 0$. Because $a(\mu) = 0$ implies $q_i + \mu^\kappa h_i = 0$ for all i, we get as a consequence that $q_i(\mu) = 0$ for all i, which again contradicts (1.23).

To finish up the proof that $\rho \circ \rho_G$ is a resolution of singularities for $K \circ u$, it remains to show that in each chart of $\mathscr{M}$, the Jacobian determinant of $\rho \circ \rho_G$ equals

$$|(\rho \circ \rho_G)'| = b(\mu)\mu_1^{\tau_1} \cdots \mu_d^{\tau_d}$$

where the τ_i are non-negative integers and $b(\mu)$ does not vanish in $(\rho \circ \rho_G)^{-1}(\Omega)$. Indeed, the Jacobian determinant of ρ already has this form, and the Goward principalization and toric principalization only makes monomial substitutions and products to this determinant. □

Corollary 1.14. *Let $\mathcal{M}$ and $\mathcal{M}_R$ be models for X with parameter spaces U and Ω respectively such that $\mathcal{M}$ is regularly parametrized with base $\mathcal{M}_R$ via the map $u : \Omega \to U$. Suppose the true distribution is $q(x)dx = p(x|\hat{\omega}, \mathcal{M})dx$ for some $\hat{\omega} \in \Omega$. We translate Ω so that $\hat{w}$ is the origin $0 \in \Omega$. Let $\hat{u} = u(0)$ and $I_{\hat{u}} = \langle u(\omega) - \hat{u} \rangle$ be the fiber ideal of $\mathcal{M}$ at $\hat{u}$. Let*

$$K(\omega) = \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x|\omega, \mathcal{M})} dx$$

be the Kullback-Leibler function for $\mathcal{M}$ at the true distribution.

Let $\rho : M \to \Omega$ be a monomialization map for I, and suppose $\rho_G : \mathscr{M} \to M$ is the Goward principalization map or a toric principalization map for $\rho^* I$. Then, $\rho \circ \rho_G$ desingularizes K at the origin.

This corollary gives us a new perspective on the difference between regular and singular models. For regular models, monomializing the fiber ideal is equivalent to finding a linear change of variables $v = T(u)$ so that the ideal is generated by the coordinate functions $v_1, \ldots, v_d$. This coordinate change allows us to apply the Laplace approximation, the Bayesian information criterion and the central limit theorem. For singular models, this change of variables may no longer be linear. Because monomializing the fiber ideal allows us to compute learning coefficients and formulate central limit theorems, we may think of this monomial representation of the fiber ideal as a standard form of the singular model.

1.6 Marginal Likelihood of Exponential Families

In Section 1.1.2, we suggested a method of estimating the marginal likelihood integral that is different from Watanabe's approach in Theorem 1.3 and Remark 1.4. Recall that he assumes the data come from a true distribution, and he computes the first term asymptotics of the expected log marginal likelihood. On the other hand, our suggested method does not assume that the true distribution is known. Instead, we consider the function

$$L(n) = \Big(\prod_{i=1}^{N} p(x_i|\hat{\omega}, \mathcal{M})^{1/N} \Big)^n \cdot \int_{\Omega} e^{-nK_N(\omega)} \varphi(\omega) d\omega \tag{1.24}$$

where $\varphi(\omega) = p(\omega|\mathcal{M})$ and $K_N(\omega)$ is the log likelihood ratio

$$K_N(\omega) = \frac{1}{N} \sum_{i=1}^{N} \log \frac{p(x_i|\hat{\omega}, \mathcal{M})}{p(x_i|\omega, \mathcal{M})} \tag{1.25}$$

for some maximum likelihood estimate $\hat{\omega}$. Note that the integer variable n is different from the sample size N in the formula for $L(n)$, and that $L(N)$ is precisely the marginal likelihood integral for the given data $x_1, \ldots, x_N$.

Our goal is to describe the asymptotics of $L(n)$ as $n \to \infty$. To do that, we need to find a resolution of singularities for $K_N(\omega)$ but this is a difficult problem. We want to find a way of relating the desingularization of $K_N(\omega)$ to that of some fiber ideal. This idea works out well for exponential families. Indeed, if $\mathcal{M}$ is an exponential family with probability density $p(x|\omega, \mathcal{M}) = h(x) \exp(\eta(\omega)^{\mathsf{T}} T(x) - A(\omega))$, then

$$K_N(\omega) = \big(A(\omega) - \eta(\omega)^{\mathsf{T}} \hat{\mu} \big) \; - \; \big(A(\hat{\omega}) - \eta(\hat{\omega})^{\mathsf{T}} \hat{\mu} \big), \quad \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} T(x_i)$$

where $\hat{\mu}$ is the mean of the sufficient statistics $T(x_i)$. Recall that $A(\omega)$ depends only on the natural parameter η. Thus, we may define another exponential family $\mathcal{M}_N$ whose probability density is $p(x|\eta, \mathcal{M}_N) = h(x) \exp(\eta^{\mathsf{T}} T(x) - A(\eta))$, and so the model $\mathcal{M}$ factors through $\mathcal{M}_N$. If $\eta(\hat{\omega})$ is also a maximum likelihood estimate for $\mathcal{M}_N$, we say that the estimate $\hat{\omega}$ is *natural*.

Remark 1.15. For a discrete model $\mathcal{M}$ parametrized by state probabilities $p_1(\omega), \ldots, p_k(\omega)$, there exist a natural MLE if and only if the set $S = \{\omega \in \Omega : p(\omega) = \hat{q}\}$ is nonempty where $\hat{q} = (\hat{q}_1, \ldots, \hat{q}_k)$ is the vector of relative frequencies coming from the data. We have a similar statement for a multivariate Gaussian model parametrized by mean $\mu(\omega)$ and covariance $\Sigma(\omega)$. If $\hat{\mu}$ and $\hat{\Sigma}$ are the sample mean and sample covariance matrix, then the model has a natural MLE if and only if the set $S = \{\omega \in \Omega : \mu(\omega) = \hat{\mu}, \, \Sigma(\omega) = \hat{\Sigma}\}$ is nonempty. In fact, for both discrete and Gaussian models, if a natural MLE exists, then all MLEs are natural and the set of MLEs is precisely the set S.

Let us study this condition more closely for a Gaussian mixed graph model $\mathcal{M}$. Suppose d is the number of parameters in the model. The parameter space Ω is an open subset of $\mathbb{R}^d$, because it is subject to the positive definite conditions on the matrices K and Φ in (1.12). Thus, the MLE may not exist, because it is possible that the likelihood function attains its maximum only on the boundary of Ω. Geometrically, to find the set U of all MLEs, we first consider, in the image $\Sigma(\Omega) \subset \mathbb{R}^{k\times k}$, the set T of covariance matrices where the likelihood function is maximized. Then, U is the preimage $\Sigma^{-1}(T)$. Therefore, MLEs exist if and only if T is nonempty, and *natural* MLEs exist if and only if T contains only the sample covariance $\hat{\Sigma}$. When the underlying graph is a directed graph, explicit conditions for the existence and the uniqueness of the MLE are investigated in [13, 53]. □

The next result shows that if natural MLEs exist, then $K_N(\omega)$ is precisely equal to the Kullback-Leibler function of $\mathcal{M}$ at the maximum likelihood distribution.

Proposition 1.16. *Let $\mathcal{M}$ and $\mathcal{M}_N$ be exponential families as described above. Given some data, suppose $\hat{\omega}$ is a natural maximum likelihood estimate for $\mathcal{M}$. Then, the Kullback-Leibler divergence $K(\omega)$ of $\mathcal{M}$ from the maximum likelihood distribution $p(x|\hat{\omega})dx$ to the distribution $p(x|\omega)dx, \omega \in \Omega$, depends only on $\eta(\omega)$. It equals*

$$K(\eta) \;=\; \big(A(\eta) - \eta^{\mathsf{T}}\hat{\mu}\big) \;-\; \big(A(\hat{\eta}) - \hat{\eta}^{\mathsf{T}}\hat{\mu}\big), \quad \hat{\eta} = \eta(\hat{\omega}).$$

Proof. The first statement follows from the fact that the distributions in the exponential family depend only on the natural parameter η. Now, by definition, $K(\eta)$ equals

$$\int_{\mathcal{X}} h(x)\, e^{\hat{\eta}^{\mathsf{T}}T(x) - A(\hat{\eta})} \,\big(A(\eta) - \eta^{\mathsf{T}}T(x) - A(\hat{\eta}) + \hat{\eta}^{\mathsf{T}}T(x)\big)\, dx.$$

The proposition follows if we have

$$\begin{aligned} 1 &= \int_{\mathcal{X}} h(x)\, e^{\hat{\eta}^{\mathsf{T}}T(x) - A(\hat{\eta})} dx, \\ \hat{\mu} &= \int_{\mathcal{X}} h(x)\, e^{\hat{\eta}^{\mathsf{T}}T(x) - A(\hat{\eta})}\, T(x) dx. \end{aligned}$$

The first equation comes from $\int p(x|\hat{\eta})dx = 1$ while the second is the result of differentiating (1.7) with respect to the natural parameter η and applying (1.9). □

Now, because $K_N(\omega)$ is equal to the Kullback-Leibler function of the singular model at some true distribution, we may use the results of Section 1.5 to desingularize $K_N(\omega)$ and obtain the full asymptotics of $L(n)$ for regularly parametrized exponential families.

Theorem 1.17. *Let $\mathcal{M}$ and $\mathcal{M}_R$ be models for X with parameter spaces U and Ω respectively such that $\mathcal{M}$ is regularly parametrized with base $\mathcal{M}_R$ via the map $u : \Omega \to U$. Let $x_1, \ldots, x_N$ be independent random samples of X. Suppose that $\mathcal{M}$ is an exponential family and that $\hat{\omega}$*

is a natural maximum likelihood estimate for the data. We translate Ω *so that* $\hat{\omega}$ *is the origin* $0 \in \Omega$*. Let* $\hat{u} = u(0)$ *and* $I_{\hat{u}} = \langle u(\omega) - \hat{u} \rangle$ *be the fiber ideal of* $\mathcal{M}$ *at* $\hat{u}$*. Let* $L(n)$ *and* $K_N(\omega)$ *be functions defined by* (1.24) *and* (1.25)*.*

Then, asymptotically as $n \to \infty$*,*

$$L(n) \;\approx\; \Big(\prod_{i=1}^{N} p(x_i|\hat{\omega}, \mathcal{M})^{1/N} \Big)^n \cdot C n^{-\lambda} (\log n)^{\theta - 1}$$

where C *is a positive constant and*

$$(2\lambda, \theta) = \min_{\omega:\, u(\omega) = \hat{u}} \mathrm{RLCT}_{\Omega_\omega}(I_{\hat{u}}; \varphi).$$

Moreover, if $\rho : M \to \Omega$ *is a monomialization map for* $I_{\hat{u}}$*, and* $\rho_G : \mathscr{M} \to M$ *is the Goward principalization map or a toric principalization map for* $\rho^* I$*, then* $\rho \circ \rho_G$ *desingularizes* $K_N(\omega)$ *at the origin.*

Proof. Because the maximum likelihood estimate $\hat{\omega}$ is natural, it follows from Proposition 1.16 that $K_N(\omega)$ equals the Kullback-Leibler function $K(\omega)$ of $\mathcal{M}$ at the true distribution $p(x|\hat{\omega}, \mathcal{M})dx$. Therefore, the asymptotics of $L(n)$ is given by the real log canonical threshold (λ, θ) of $K(\omega)$. By Theorem 1.11, this RLCT is given by the formula stated above. Moreover, Theorem 1.13 tells us how to desingularize $K(\omega)$ after monomializing the fiber ideal. □

The resolution of singularities $\rho \circ \rho_G$ described in the theorem allows us to compute, not just the first term, but the full asymptotics of $L(n)$. This is accomplished by applying the desingularization map to the zeta function associated to $K_N(\omega)$, in order to compute its poles and Laurent coefficients. We can then employ Theorem 3.16 to compute the desired asymptotic expansion.

Alternatively, we can also use the methods of Chapter 5, without computing the Goward principalization map ρ_G. Indeed, suppose we want to compute the asymptotic expansion up to $O(n^{-\lambda_0}(\log n)^{\theta_0 - 1})$ for some (λ_0, θ_0). We first find the set S of points $\mu \in M$ where the RLCT of the pullback $\rho^* I_{\hat{u}}$ is at most (λ_0, θ_0). We then try to cover S with hypercubic patches $[0, 1]^d$ and apply either Theorem 5.11 or Theorem 5.13 to compute the desired asymptotic coefficients. These two theorems pertain to nondegenerate functions only, but we are able to apply them, because the pullback $\rho^* I_{\hat{u}}$ is monomial, and so by Proposition 4.22, $K_N \circ \rho(\mu)$ is nondegenerate at every point $\mu \in M$.

Example 1.18. To demonstrate how fiber ideals can be used to compute asymptotics of marginal likelihood integrals given large-sample data, we revisit the coin toss model $\mathcal{M}_3$ in Example 1.1. Recall that the model is parametrized by polynomials

$$p_1(\omega, t) = \frac{1}{2} t + (1 - t)\omega$$
$$p_2(\omega, t) = \frac{1}{2} t + (1 - t)(1 - \omega)$$

where the parameters (ω, t) lie in $[0,1]^2$. Suppose we have N independent random samples with relative frequencies q_1 and q_2. The marginal likelihood integral of the data is

$$\int_{[0,1]^2} p_1(\omega,t)^{Nq_1} p_2(\omega,t)^{Nq_2}\, d\omega dt = (q_1^{q_1} q_2^{q_2})^N \int_{[0,1]^2} e^{-NK(\omega,t)} d\omega dt$$

where $K(\omega, t)$ is the Kullback-Leibler function

$$K(\omega,t) = q_1 \log \frac{q_1}{p_1(\omega,t)} + q_2 \log \frac{q_2}{p_2(\omega,t)}.$$

Fixing q_1 and q_2 so that the function $K(\omega, t)$ is independent of N, we are interested in the first term asymptotics of the Laplace integral

$$L(N) = \int_{[0,1]^2} e^{-NK(\omega,t)} d\omega dt$$

as $N \to \infty$. For this discrete model, the fiber ideal is

$$I = \langle p_1(\omega,t) - q_1, p_2(\omega,t) - q_2 \rangle = \langle p_1(\omega,t) - q_1 \rangle.$$

Case 1: $(q_1, q_2) \neq (1/2, 1/2)$**.**

Without loss of generality, let us assume $q_1 > 1/2$. The variety $\mathcal{V}$ of the fiber ideal is

$$\mathcal{V} = \left\{ (\omega,t) \in [0,1]^2 : q_1 \leq \omega \leq 1,\ (\omega - \frac{1}{2})(1-t) = q_1 - \frac{1}{2} \right\}.$$

A graph of this variety is shown in Figure 1.4. This variety is nonempty, so by Remark 1.15, there exist natural maximum likelihood estimates for the model. We may use Theorem 1.17 in deriving the asymptotics of $L(N)$. According to Theorem 3.16, the asymptotics does not change if we limit the domain of integration to a small neighborhood of the variety $\mathcal{V}$. Let

$$\Omega = \{(\omega,t) : \delta \leq \omega \leq 1,\ 0 \leq t \leq 1\}$$

be the new domain, where $1/2 < \delta < q_1$. Because our goal is to monomialize the fiber ideal, let us make the substitution

$$t = \frac{-s + (\omega - q_1)}{\omega - 1/2}$$

so that the fiber ideal becomes $\langle s \rangle$. This substitution is real analytic over our new domain Ω. If we also substitute $\omega = q_2 u + q_1$, the domain Ω becomes the triangle

$$\Delta = \{(u,s) : -(q_1 - 1/2) \leq s \leq q_2 u,\ -\frac{q_1 - \delta}{q_2} \leq u \leq 1\}$$

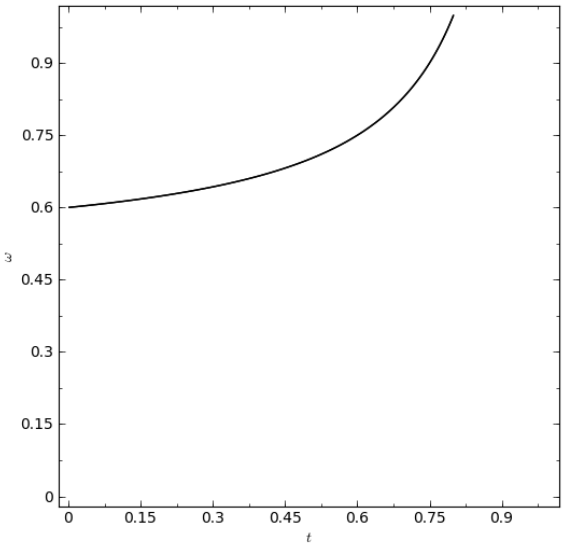


Figure 1.4: The variety of the fiber ideal for $q_1 = 0.6$.

while the variety of the fiber ideal becomes

$$\{(u, s) : 0 \leq u \leq 1,\ s = 0\}.$$

At this point, we could easily compute the learning coefficient (λ, θ) of the model by finding the RLCT of the fiber ideal. However, because we are ultimately interested in the first term asymptotics $CN^{-\lambda}(\log N)^{\theta-1}$, we will do a little bit more work so that we can use Theorem 5.11 to compute the constant C. To apply this theorem directly, we would like to integrate over the rectangular domain

$$\Xi = \{(u, s) : 0 \leq u \leq 1,\ -1 \leq s \leq 1\}.$$

instead of over the triangular domain Δ. This is possible if we adjusted for the asymptotics of the integral over the regions

$$\Delta_1 = \{(u, s) : 0 \leq q_2 u \leq s \leq \varepsilon\}$$
$$\Delta_2 = \{(u, s) : -\varepsilon \leq s \leq q_2 u \leq 0\}$$

for some small $\varepsilon > 0$. We show later that the integrals over these regions do not affect the first term asymptotics of the integral over Ξ. The Jacobian determinant of our substitutions is $q_2/(q_2 u + q_1 - 1/2)$, while the integrals over the domains $\{0 \leq u \leq 1,\ -1 \leq s \leq 0\}$ and $\{0 \leq u \leq 1,\ 0 \leq s \leq 1\}$ are equal. Therefore, the integral over Ξ equals

$$2 \int_{[0,1]^2} e^{-NK(u,s)} \frac{q_2}{q_2 u + q_1 - 1/2}\, du ds \tag{1.26}$$

where

$$K(u,s) = q_1 \log \frac{q_1}{q_1+s} + q_2 \log \frac{q_2}{q_2-s}.$$

Now, the fiber ideal $\langle s \rangle$ is monomial, so by Proposition 4.22, $K(u,s)$ is nondegenerate. The Newton polyhedron of $K(u,s)$ is twice that of the fiber ideal. In particular, it is the positive orthant with the origin translated to the point $(0,2)$. Therefore, the RLCT of $K(u,s)$ is $(\lambda,\theta) = (1/2,1)$. The normal fan $\mathcal{F}$ of this polyhedron has only one cone, namely the positive orthant. Thus, the fan is already smooth. By Theorem 5.11, the leading coefficient C is

$$\Gamma(1/2) \int_{[0,1]} g(u,0)^{-1/2} \frac{q_2}{q_2 u + q_1 - 1/2}\, du$$

where $g(u,s)$ is the strict transform $K(u,s)/s^2$. A simple calculation shows that

$$g(u,0) = \frac{1}{2}\left(\frac{1}{q_1} + \frac{1}{q_2}\right) = \frac{1}{2q_1 q_2}.$$

Consequently, the first term asymptotics of $L(N)$ is

$$L(N) \;\approx\; \sqrt{2\pi q_1 q_2}\, \log \frac{1}{2q_1 - 1}\; N^{-1/2}. \tag{1.27}$$

Finally, to prove that the integrals over the regions Δ_1 and Δ_2 do not affect the first term asymptotics of $L(N)$, we blow up the origin so that both regions become rectangular. In the region Δ_1, after substituting $u = xy$ and $s = y$, the integral becomes

$$\int_0^{\varepsilon} \int_0^{1/q_2} e^{-NK(y)} \frac{q_2 y}{q_2 xy + q_1 - 1/2}\, dx dy$$

where $K(y)$ is $K(u,s)$ with $s = y$. Now, by Proposition 3.9,

$$\mathrm{RLCT}_{[0,\frac{1}{q_2}]\times[0,\varepsilon]} \left(K(y); \frac{q_2 y}{q_2 xy + q_1 - 1/2} \right) \;=\; \mathrm{RLCT}_{[0,\frac{1}{q_2}]\times[0,\varepsilon]} \left(K(y); y \right)$$

because $q_2/(q_2 xy + q_1 - 1/2)$ is positive in the above domain. Applying the Newton polyhedra method to the nondegenerate function $K(y)$, we learn that the latter RLCT equals $(1,1)$. Hence, the integral over Δ_1 is grows like N^{-1} asymptotically, and it does not affect the first term asymptotics of $L(N)$. The same is true for Δ_2.

Case 2: $(q_1, q_2) = (1/2, 1/2)$**.**

This is the more interesting scenario in our coin toss example because of the singularities involved. The variety $\mathcal{V}$ of the fiber ideal is the union of lines

$$\{\omega = 1/2\} \;\cup\; \{t = 1\} \quad \subset [0,1]^2.$$

After substituting $\omega = (1+u)/2$ and $t = 1-s$, the fiber ideal becomes the monomial ideal $\langle us \rangle$ while the domain of integration becomes $\{(u,s) : -1 \leq u \leq 1,\ 0 \leq s \leq 1\}$. Meanwhile, the Jacobian determinant of the substitutions is $1/2$, but the integrals over the domains $\{-1 \leq u \leq 0, 0 \leq s \leq 1\}$ and $\{0 \leq u \leq 1, 0 \leq s \leq 1\}$ are equal. Therefore, we can write the marginal likelihood integral $L(N)$ as

$$\int_{[0,1]^2} e^{-NK(u,s)} du ds \tag{1.28}$$

where

$$K(u,s) = \frac{1}{2}\log\frac{1}{1+us} + \frac{1}{2}\log\frac{1}{1-us}.$$

By applying Theorems 1.17 and 5.11, we find that the RLCT of $K(u,s)$ is $(\lambda, \theta) = (1/2, 2)$ and that the first term asymptotics of $L(N)$ is

$$L(N) \ \approx \ \sqrt{\frac{\pi}{8}}\, N^{-1/2} \log N.$$

The computations are not difficult in this case, so we leave them to the reader. In Example 5.15, we derive higher order asymptotics of this integral. □

Remark 1.19. For regularly parametrized exponential families, the asymptotics of marginal likelihood integrals given data depends very much on the sample mean $\hat{u}$ appearing in Theorem 1.17. One may argue that in practice, the data almost always gives a sample mean $\hat{u}$ at which the likelihood integral can be estimated using the BIC or the Laplace approximation. Consequently, it seems unnecessary to study these integrals at singular points $\hat{u}$.

Example 1.18 demonstrates how this argument can be misleading. In this example, the sample mean $\hat{u}$ is the vector $q = (q_1, q_2)$ of relative frequencies of the data. Firstly, for all values of q, the variety of the fiber ideal is not a collection of isolated points, so the Laplace approximation cannot be applied directly. Secondly, the BIC only gives the exponents (λ, θ) of the first term asymptotics $CN^{-\lambda}(\log N)^{\theta-1}$ but not the leading coefficient C. It is in this constant C that a lot of interesting asymptotic behavior occurs. For instance, let us consider relative frequencies $(q_1, q_2) = (1/2 + 1/N, 1/2 - 1/N)$ for Example 1.18. This point q is close to the singular point $(1/2, 1/2)$. If we approximate the marginal likelihood integral using the formula (1.27), then for large N,

$$L(N) \ \approx \ \sqrt{\frac{\pi}{2}}\, N^{-1/2} \log N.$$

This $N^{-1/2}\log N$ behavior agrees up to a scalar with the asymptotics of the integral at the point $(1/2, 1/2)$. It suggests that singular learning theory is useful for understanding marginal likelihood integrals at regular data points which are close to singular ones. □

To summarize, in Sections 1.1–1.4, we introduced basic concepts from model selection and Watanabe's singular learning theory which are required for this dissertation. We discussed the distinction between maximum likelihood and marginal likelihood integrals, and between regular and singular models. We described some important classes of statistical models used in this dissertation, such as discrete models, Gaussian models, and exponential families. We saw that the key to deciphering a singular model is resolving the singularities of the Kullback-Leibler function of the model at the true distribution.

In the last two sections, we described our main statistical contributions. In Section 1.5, we defined a new general class of models, known as regularly parametrized models. We also introduced fiber ideals for these models at given true distributions. We saw that for regularly parametrized models, to desingularize the Kullback-Leibler function at the true distribution, we only need to monomialize the corresponding fiber ideal. In Theorem 1.13 and Corollary 1.14, we gave details of the construction of this desingularization, and showed that the last ingredient needed is principalization of the monomialized fiber ideal. We also proved in Theorem 1.11 that the learning coefficient of the model is the real log canonical threshold of the fiber ideal.

In Section 1.6, we studied the classic problem of estimating marginal likelihood integrals for large-sample data without knowledge of the true distribution. Our aim was to use our new techniques involving fiber ideals in approximating such integrals for regularly parametrized exponential families. Theorem 1.17 stated how this can be accomplished when the maximum likelihood estimate is natural. We finished this chapter with an example which uses many of the new tools from this dissertation in computing the complete first term asymptotics $Cn^{-\lambda}(\log n)^{\theta-1}$ of the marginal likelihood integral.

In recent years, there have been increasing interest in computing learning coefficients of statistical models and applying them to the approximation of marginal likelihood integrals. For example, Aoyagi, Watanabe and Yamazaki [2, 60, 61] has clarified this for certain classes of mixture models and reduced-rank regression models, while Geiger and Rusakov [45] has accomplished this for naïve Bayesian networks with binary states. Their computations involve finding real log canonical thresholds of functions rather than that of ideals. In Section 4.3, we use the ideal-theoretic approach developed in this dissertation to simplify computations of the learning coefficient of a mixture model with ternary features. This algebraic approach was also applied by Zwiernik [63] in his analysis of learning coefficients of binary tree models.

Chapter 2

Exact Evaluation

Evaluation of marginal likelihood integrals is central to Bayesian statistics. It is generally assumed that these integrals cannot be evaluated exactly, except in trivial cases, and a wide range of numerical techniques (e.g. MCMC) have been developed to obtain asymptotics and numerical approximations [39]. The aim of this chapter is to show that exact integration is more feasible than is surmised in the literature.

We examine marginal likelihood integrals for mixtures of discrete independence models defined in Section 1.4.5. Bayesian inference for these models arises in many contexts, including machine learning and computational biology. In Chapter 1, we described how recent work in these fields has made a connection to singularities in algebraic geometry [15, 45, 57, 60, 61]. Our methods in this chapter complement those developments by providing tools for symbolic integration when the sample size is small. The values of the integrals we study are rational numbers, and exact evaluation means computing those rational numbers rather than floating point approximations.

This chapter is organized as follows. In Section 2.1 we look at marginal likelihood integrals of mixture models and investigate their basic properties. In Section 2.2 we examine the Newton zonotopes of mixture models, and we derive formulas for marginal likelihood evaluation using tools from geometric combinatorics. Our algorithms and their implementations are described in detail in Section 2.3. Section 2.4 is concerned with applications in Bayesian statistics. We show how *Dirichlet priors* can be incorporated into our approach, we discuss the evaluation of *Bayes factors*, we compare our setup with that of [39], and we illustrate the scope of our methods by computing an integral arising from a data set of [18].

This chapter is joint work with Bernd Sturmfels and Zhiqiang Xu. A preliminary draft version was first published in Section 5.2 of the Oberwolfach lecture notes [16]. We refer to that volume for further information on the use of computational algebra in Bayesian statistics. A full version of this chapter was later published in [38].

2.1 Independence Models and their Mixtures

Let $\mathcal{M}$ be the independence model defined in (1.16). We now consider marginal likelihood integrals arising from this model. All our domains of integration in this chapter will be polytopes that are products of standard probability simplices. On each such polytope we fix the standard Lebesgue probability measure. In other words, our discussion of Bayesian inference refers to the uniform prior on each parameter space. Naturally, other prior distributions, such as Dirichlet priors, are of interest, and our methods are extended to these in Section 2.4. In what follows, we simply work with uniform priors.

We identify the state space (1.14) with the set $\{1, \ldots, n\}$. A *data vector* $U = (U_1, \ldots, U_n)$ is thus an element of $\mathbb{N}^n$. The *sample size* of these data is $U_1 + U_2 + \cdots + U_n = N$. If the sample size N is fixed then the probability of observing these data is

$$\mathbf{L}_U(\theta) \quad = \quad \frac{N!}{U_1! U_2! \cdots U_n!} \cdot p_1(\theta)^{U_1} \cdot p_2(\theta)^{U_2} \cdot \cdots \cdot p_n(\theta)^{U_n}.$$

This expression is a function on the polytope P which is known as the *likelihood function* of the data U with respect to the independence model $\mathcal{M}$. The marginal likelihood of the data U with respect to the model $\mathcal{M}$ equals

$$\int_P \mathbf{L}_U(\theta)\, d\theta.$$

The value of this integral is a rational number which we now compute explicitly. The data U will enter this calculation by way of the *sufficient statistic* $b = A \cdot U$, which is a vector in $\mathbb{N}^d$. The coordinates of this vector are denoted $b_j^{(i)}$ for $i = 1, \ldots, k$ and $j = 0, \ldots, t_k$. Thus $b_j^{(i)}$ is the total number of times the value j is attained by one of the random variables $X_1^{(i)}, \ldots, X_{s_i}^{(i)}$ in the i-th group. Clearly, the sufficient statistics satisfy

$$b_0^{(i)} + b_1^{(i)} + \cdots + b_{t_i}^{(i)} \quad = \quad s_i \cdot N \qquad \text{for all } \; i = 1, 2, \ldots, k. \tag{2.1}$$

The likelihood function $\mathbf{L}_U(\theta)$ is the constant $\frac{N!}{U_1! \cdots U_n!}$ times the monomial

$$\theta^b \quad = \quad \prod_{i=1}^{k} \prod_{j=0}^{t_i} (\theta_j^{(i)})^{b_j^{(i)}}.$$

The logarithm of this function is concave on the polytope P, and its maximum value is attained at the point $\hat{\theta}$ with coordinates $\hat{\theta}_j^{(i)} = b_j^{(i)}/(s_i \cdot N)$.

Lemma 2.1. The integral of the monomial θ^b over the polytope P equals

$$\int_P \theta^b d\theta \quad = \quad \prod_{i=1}^{k} \frac{t_i!\, b_0^{(i)}!\, b_1^{(i)}! \cdots b_{t_i}^{(i)}!}{(s_i N + t_i)!}.$$

The product of this number with the multinomial coefficient $N!/(U_1! \cdots U_n!)$ equals the marginal likelihood of the data U for the independence model $\mathcal{M}$.

Proof. Since P is the product of simplices (1.15), this follows from the formula

$$\int_{\Delta_t} \theta_0^{b_0}\theta_1^{b_1}\cdots\theta_t^{b_t} d\theta \quad = \quad \frac{t! \cdot b_0! \cdot b_1! \cdots b_t!}{(b_0+b_1+\cdots+b_t+t)!} \tag{2.2}$$

for the integral of a monomial over the standard probability simplex Δ_t. □

Now, we shift our focus to the mixture model $\mathcal{M}^{(2)}$. Our objective is to compute marginal likelihood integrals for this model. Recall that its parameter space is the polytope

$$\Theta \;=\; \Delta_1 \,\times\, P \,\times\, P$$

and that the model is parametrized by

$$p_v \;=\; \sigma_0 \cdot \theta^{a_v} \,+\, \sigma_1 \cdot \rho^{a_v} \qquad \text{for } (\sigma,\theta,\rho) \in \Theta. \tag{2.3}$$

The likelihood function of a data vector $U \in \mathbb{N}^n$ for $\mathcal{M}^{(2)}$ equals

$$\mathbf{L}_U(\sigma,\theta,\rho) \quad = \quad \frac{N!}{U_1!U_2!\cdots U_n!}\, p_1(\sigma,\theta,\rho)^{U_1} \,\cdots\, p_n(\sigma,\theta,\rho)^{U_n}. \tag{2.4}$$

The marginal likelihood of the data U with respect to the model $\mathcal{M}^{(2)}$ equals

$$\int_\Theta \mathbf{L}_U(\sigma,\theta,\rho)\, d\sigma d\theta d\rho \;=\; \frac{N!}{U_1!\cdots U_n!} \int_\Theta \prod_v (\sigma_0\theta^{a_v} + \sigma_1\rho^{a_v})^{U_v} d\sigma\, d\theta\, d\rho. \tag{2.5}$$

The following proposition shows that we can evaluate this integral *exactly*.

Proposition 2.2. *The marginal likelihood (2.5) is a rational number.*

Proof. The likelihood function $\mathbf{L}_U$ is a $\mathbb{Q}_{\geq 0}$-linear combination of monomials $\sigma^a\theta^b\rho^c$. The integral (2.5) is the same $\mathbb{Q}_{\geq 0}$-linear combination of the numbers

$$\int_\Theta \sigma^a\theta^b\rho^c d\sigma\, d\theta\, d\rho \quad = \quad \Big(\int_{\Delta_1} \sigma^a d\sigma\Big)\cdot\Big(\int_P \theta^b d\theta\Big)\cdot\Big(\int_P \rho^c d\rho\Big).$$

Each of the three factors is an easy-to-evaluate rational number, by (2.2). □

Example 2.3. The integral (1.19) expresses the marginal likelihood of a 4×4-table of counts $U = (U_{ij})$ with respect to the mixture model $\mathcal{M}^{(2)}$. Specifically, the marginal likelihood of the data (1.20) equals the normalizing constant $40!\cdot(2!)^{-12}\cdot(4!)^{-4}$ times the number (1.21). The model $\mathcal{M}^{(2)}$ consists of all non-negative 4×4-matrices of rank ≤ 2 whose entries sum to one. Note that here the parametrization (2.3) is not identifiable, because $\dim(\mathcal{M}^{(2)}) = 11$ but $\dim(\Theta) = 13$. In this example, $k = 2$, $s_1{=}s_2{=}1$, $t_1{=}t_2{=}3$, $d = 8$, $n = 16$. □

Example 2.4. Consider again the setup of Example 1.6. Using the methods to be described in the next two sections, we computed the exact value of the marginal likelihood for the data $\tilde{U} = (51, 18, 73, 25, 75)$ with respect to $\mathcal{M}^{(2)}$. The rational number (2.5) is found to be the ratio of two relatively prime integers having 530 digits and 552 digits, and its numerical value is approximately $0.778871633883867861133574286090 \cdot 10^{-22}$. □

2.2 Summation over a Zonotope

Our starting point is the observation that the Newton polytope of the likelihood function (2.4) is a zonotope. Recall that the *Newton polytope* of a polynomial is the convex hull of all exponent vectors appearing in the expansion of that polynomial, and a polytope is a *zonotope* if it is the image of a standard cube under a linear map. See [12, §7] and [62, §7] for further discussions. We are here considering the zonotope

$$Z_A(U) \;\; = \;\; \sum_{v=1}^{n} U_v \cdot [0, a_v],$$

where $[0, a_v]$ represents the line segment between the origin and the point $a_v \in \mathbb{R}^d$, and the sum is a Minkowski sum of line segments. We write $Z_A = Z_A(1, 1, \ldots, 1)$ for the basic zonotope spanned by the vectors a_v. Hence $Z_A(U)$ is obtained by stretching Z_A along those vectors by factors U_v respectively. Assuming that the counts U_v are all positive, we have

$$\dim(Z_A(U)) \;\; = \;\; \dim(Z_A) \;\; = \;\; \mathrm{rank}(A) \;\; = \;\; d - k + 1. \tag{2.6}$$

The zonotope Z_A is related to the polytope $P = \mathrm{conv}(A)$ in (1.15) as follows. The dimension $d - k = t_1 + \cdots + t_k$ of P is one less than $\dim(Z_A)$, and P appears as the *vertex figure* of the zonotope Z_A at the distinguished vertex 0.

Remark 2.5. For higher mixtures $\mathcal{M}^{(l)}$, the Newton polytope of the likelihood function is isomorphic to the Minkowski sum of $(l-1)$-dimensional simplices in $\mathbb{R}^{(l-1)d}$. Only when $l = 2$, this Minkowski sum is a zonotope. □

The marginal likelihood (2.5) we wish to compute is the integral

$$\int_{\Theta} \prod_{v=1}^{n} (\sigma_0 \theta^{a_v} + \sigma_1 \rho^{a_v})^{U_v} d\sigma d\theta d\rho \tag{2.7}$$

times the constant $N!/(U_1! \cdots U_n!)$. Our approach to this computation is to sum over the lattice points in the zonotope $Z_A(U)$. If the matrix A has repeated columns, we may replace A with the reduced matrix $\tilde{A}$ and U with the corresponding reduced data vector $\tilde{U}$. If one desires the marginal likelihood for the reduced data vector $\tilde{U}$ instead of the original data vector U, the integral remains the same while the normalizing constant becomes

$$\frac{N!}{\tilde{U}_1! \cdots \tilde{U}_{\tilde{n}}!} \cdot \alpha_1^{\tilde{U}_1} \cdots \alpha_{\tilde{n}}^{\tilde{U}_{\tilde{n}}},$$

where α_i is the number of columns in A equal to the i-th column of $\tilde{A}$. In what follows we ignore the normalizing constant and focus on computing the integral (2.7) with respect to the original matrix A.

For a vector $b \in \mathbb{R}^d_{\geq 0}$ we let $|b|$ denote its L^1-norm $\sum_{t=1}^d b_t$. Recall from (1.16) that all columns of the $d \times n$-matrix A have the same coordinate sum

$$a \;:=\; |a_v| \;=\; s_1 + s_2 + \cdots + s_k, \quad \text{for all } v = 1, 2, \ldots, n,$$

and from (2.1) that we may denote the entries of a vector $b \in \mathbb{R}^d$ by $b_j^{(i)}$ for $i = 1, \ldots, k$ and $j = 0, \ldots, t_k$. Also, let $\mathbb{L}$ denote the image of the linear map $A : \mathbb{Z}^n \to \mathbb{Z}^d$. Thus $\mathbb{L}$ is a sublattice of rank $d - k + 1$ in $\mathbb{Z}^d$. We abbreviate $Z_A^{\mathbb{L}}(U) := Z_A(U) \cap \mathbb{L}$. Now, using the binomial theorem, we have

$$(\sigma_0 \theta^{a_v} + \sigma_1 \rho^{a_v})^{U_v} \;=\; \sum_{x_v=0}^{U_v} \binom{U_v}{x_v} \sigma_0^{x_v} \sigma_1^{U_v - x_v} \theta^{x_v \cdot a_v} \rho^{(U_v - x_v) \cdot a_v}.$$

Therefore, in the expansion of the integrand in (2.7), the exponents of θ are of the form of $b = \sum_v x_v a_v \in Z_A^{\mathbb{L}}(U),\ 0 \leq x_v \leq U_v$. The other exponents may be expressed in terms of b. This gives us

$$\prod_{v=1}^n (\sigma_0 \theta^{a_v} + \sigma_1 \rho^{a_v})^{U_v} \;=\; \sum_{\substack{b \in Z_A^{\mathbb{L}}(U) \\ c = AU - b}} \phi_A(b, U) \cdot \sigma_0^{|b|/a} \cdot \sigma_1^{|c|/a} \cdot \theta^b \cdot \rho^c. \tag{2.8}$$

Writing $\mathbf{D}(U) = \{(x_1, \ldots, x_n) \in \mathbb{Z}^n : 0 \leq x_v \leq U_v, v = 1, \ldots, n\}$, the coefficient in (2.8) is

$$\phi_A(b, U) \;=\; \sum_{\substack{Ax = b \\ x \in \mathbf{D}(U)}} \prod_{v=1}^n \binom{U_v}{x_v}. \tag{2.9}$$

Thus, by formulas (2.2) and (2.8), the integral (2.7) evaluates to

$$\sum_{\substack{b \in Z_A^{\mathbb{L}}(U) \\ c = AU - b}} \phi_A(b, U) \cdot \frac{(|b|/a)!\,(|c|/a)!}{(|U|+1)!} \cdot \prod_{i=1}^k \left(\frac{t_i!\, b_0^{(i)}! \cdots b_{t_i}^{(i)}!}{(|b^{(i)}| + t_i)!} \; \frac{t_i!\, c_0^{(i)}! \cdots c_{t_i}^{(i)}!}{(|c^{(i)}| + t_i)!} \right). \tag{2.10}$$

We summarize the result of this derivation in the following theorem.

Theorem 2.6. *The marginal likelihood of the data U in the mixture model $\mathcal{M}^{(2)}$ is equal to the sum (2.10) times the normalizing constant $N!/(U_1! \cdots U_n!)$.*

Each individual summand in the formula (2.10) is a ratio of factorials and hence can be evaluated symbolically. The challenge in turning Theorem 2.6 into a practical algorithm lies in the fact that both of the sums (2.9) and (2.10) are over very large sets. We shall discuss these challenges and present techniques from both computer science and mathematics for addressing them.

We first turn our attention to the coefficients $\phi_A(b, U)$ of the expansion (2.8). These quantities are written as an explicit sum in (2.9). The first useful observation is that these coefficients are also the coefficients of the expansion

$$\prod_v (\theta^{a_v} + 1)^{U_v} \quad = \quad \sum_{b \in Z_A^{\mathbb{L}}(U)} \phi_A(b, U) \cdot \theta^b, \tag{2.11}$$

which comes from substituting $\sigma_i = 1$ and $\rho_j = 1$ in (2.8). When the cardinality of $Z_A^{\mathbb{L}}(U)$ is sufficiently small, the quantity $\phi_A(b, U)$ can be computed quickly by expanding (2.11) using a computer algebra system. We used MAPLE for this and all other symbolic computations in this chapter.

If the expansion (2.11) is not feasible, then it is tempting to compute the individual $\phi_A(b, U)$ via the sum-product formula (2.9). This method requires summation over the set $\{x \in \mathbf{D}(U) \,:\, Ax = b\}$, which is the set of lattice points in an $(n - d + k - 1)$-dimensional polytope. Even if this loop can be implemented, performing the sum in (2.9) symbolically requires the evaluation of many large binomials, causing the process to be rather inefficient.

An alternative is offered by the following recurrence formula:

$$\phi_A(b, U) \quad = \quad \sum_{x_n=0}^{U_n} \binom{U_n}{x_n} \phi_{A \setminus a_n}(b - x_n a_n, U \setminus U_n). \tag{2.12}$$

This is equivalent to writing the integrand in (2.7) as

$$\left(\prod_{v=1}^{n-1} (\sigma_0 \theta^{a_v} + \sigma_1 \rho^{a_v})^{U_v} \right) (\sigma_0 \theta^{a_n} + \sigma_1 \rho^{a_n})^{U_n}.$$

More generally, for each $0 < i < n$, we have the recurrence

$$\phi_A(b, U) = \sum_{b' \in Z_{A'}^{\mathbb{L}}(U')} \phi_{A'}(b', U') \cdot \phi_{A \setminus A'}(b - b', U \setminus U'),$$

where A' and U' consist of the first i columns and entries of A and U respectively. This corresponds to the factorization

$$\left(\prod_{v=1}^{i} (\sigma_0 \theta^{a_v} + \sigma_1 \rho^{a_v})^{U_v} \right) \left(\prod_{v=i+1}^{n} (\sigma_0 \theta^{a_v} + \sigma_1 \rho^{a_v})^{U_v} \right).$$

This formula gives flexibility in designing algorithms with different payoffs in time and space complexity, to be discussed in Section 2.3.

The next result records useful facts about the quantities $\phi_A(b, U)$.

Proposition 2.7. *Suppose $b \in \mathbb{Z}_A^{\mathbb{L}}(U)$ and $c = AU - b$. Then, the following quantities are all equal to $\phi_A(b, U)$:*
(1) $\#\{z \in \{0,1\}^N : A^U z = b\}$, where A^U is the extended matrix

$$A^U \;:=\; (\underbrace{a_1, \ldots, a_1}_{U_1}, \underbrace{a_2, \ldots, a_2}_{U_2}, \ldots, \underbrace{a_n, \ldots, a_n}_{U_n}),$$

(2) $\phi_A(c, U)$,
(3)

$$\sum_{\substack{Ax=b \\ l_j \le x_j \le u_j}} \prod_{v=1}^{n} \binom{U_v}{x_v},$$

where $u_j = \min\{U_j\} \cup \{b_m/a_{jm}\}_{m=1}^n$ and $l_j = U_j - \min\{U_j\} \cup \{c_m/a_{jm}\}_{m=1}^n$.

Proof. (1) This follows directly from (2.11).
(2) For each $z \in \{0,1\}^N$ satisfying $A^U z = b$, note that $\bar{z} = (1, 1, \ldots, 1) - z$ satisfies $A^U \bar{z} = c$, and vice versa. The conclusion thus follows from (1).
(3) We require $Ax = b$ and $x \in \mathbf{D}(U)$. If $x_j > u_j = b_m/a_{jm}$ then $a_{jm}x_j > b_m$, which implies $Ax \neq b$. The lower bound is derived by a similar argument. □

One aspect of our approach is the decision, for any given model A and data set U, whether or not to attempt the expansion (2.11) using computer algebra. This decision depends on the cardinality of the set $Z_A^{\mathbb{L}}(U)$. In what follows, we compute the number exactly when A is unimodular. When A is not unimodular, we obtain useful lower and upper bounds.

Let S be any subset of the columns of A. We call S *independent* if its elements are linearly independent in $\mathbb{R}^d$. With S we associate the integer

$$\text{index}(S) \;:=\; [\mathbb{R}S \cap \mathbb{L} : \mathbb{Z}S].$$

This is the index of the abelian group generated by S inside the possibly larger abelian group of all lattice points in $\mathbb{L} = \mathbb{Z}A$ that lie in the span of S. The following formula is due to R. Stanley and appears in [49, Theorem 2.2]:

Proposition 2.8. *The number of lattice points in the zonotope $Z_A(U)$ equals*

$$\#Z_A^{\mathbb{L}}(U) \;=\; \sum_{S \subseteq A \text{ indep.}} \text{index}(S) \cdot \prod_{a_v \in S} U_v. \tag{2.13}$$

In fact, the number of monomials in (2.8) equals $\#M_A(U)$, where $M_A(U)$ is the set $\{b \in Z_A^{\mathbb{L}}(U) : \phi_A(b, U) \neq 0\}$, and this set can be different from $Z_A^{\mathbb{L}}(U)$. For that number we have the following bounds. The proof, which uses methods in [49, §2], will be omitted here.

Theorem 2.9. *The number $\#M_A(U)$ of monomials in the expansion (2.8) of the likelihood function to be integrated satisfies the two inequalities*

$$\sum_{S \subseteq A \text{ indep.}} \prod_{v \in S} U_v \;\leq\; \#M_A(U) \;\leq \sum_{S \subseteq A \text{ indep.}} \text{index}(S) \cdot \prod_{v \in S} U_v. \tag{2.14}$$

By definition, the matrix A is *unimodular* if $\text{index}(S) = 1$ for all independent subsets S of the columns of A. In this case, the upper bound coincides with the lower bound, and so $M_A(U) = Z_A^{\mathbb{L}}(U)$. This happens in the classical case of two-dimensional contingency tables ($k = 2$ and $s_1 = s_2 = 1$). In general, $\#Z_A^{\mathbb{L}}(U)/\#M_A(U)$ tends to 1 when all coordinates of U tend to infinity. This is why we believe that for computational purposes, $\#Z_A^{\mathbb{L}}(U)$ is a good approximation of $\#M_A(U)$.

Remark 2.10. There exist integer matrices A for which $\#M_A(U)$ does not agree with the upper bound in Theorem 2.9. However, we conjecture that $\#M_A(U) = \#Z_A^{\mathbb{L}}(U)$ holds for matrices A of Segre-Veronese type as in (1.16) and strictly positive data vectors U. □

Example 2.11. Consider the 100 *Swiss Francs* example in Section 1.4.5. Here A is unimodular and it has 16145 independent subsets S. The corresponding sum of 16145 squarefree monomials in (2.13) gives the number of terms in the expansion of (1.22). For the data U in (1.20) this sum evaluates to $3,892,097$. □

Example 2.12. We consider the matrix and data from Example 1.6.

$$\begin{aligned} \tilde{A} &= \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 & 0 \end{pmatrix} \\ \tilde{U} &= \begin{pmatrix} 51, 18, 73, 25, 75 \end{pmatrix} \end{aligned}$$

By Theorem 2.9, the lower bound is 22,273 and the upper bound is 48,646. Here the number $\#M_{\tilde{A}}(\tilde{U})$ of monomials agrees with the latter. □

We next present a formula for $\text{index}(S)$ when S is any linearly independent subset of the columns of the matrix A. After relabeling we may assume that $S = \{a_1, \ldots, a_k\}$ consists of the first k columns of A. Let $H = VA$ denote the row Hermite normal form of A. Here $V \in SL_d(\mathbb{Z})$ and H satisfies

$$H_{ij} = 0 \text{ for } i > j \;\text{ and }\; 0 \leq H_{ij} < H_{jj} \text{ for } i < j.$$

Hermite normal form is a built-in function in computer algebra systems. For instance, in MAPLE the command is `ihermite`. Using the invertible matrix V, we may replace A with H, so that $\mathbb{R}S$ becomes $\mathbb{R}^k$ and $\mathbb{Z}S$ is the image over $\mathbb{Z}$ of the upper left $k \times k$-submatrix of H. We seek the index of that lattice in the possibly larger lattice $\mathbb{Z}A \cap \mathbb{Z}^k$. To this end we compute the column Hermite normal form $H' = HV'$. Here $V' \in SL_n(\mathbb{Z})$ and H' satisfies

$$H'_{ij} = 0 \text{ if } i > j \text{ or } j > d \;\text{ and }\; 0 \leq H_{ij} < H_{ii} \text{ for } i < j.$$

The lattice $\mathbb{Z}A \cap \mathbb{Z}^k$ is spanned by the first k columns of H', and this implies

$$\text{index}(S) \quad = \quad \frac{H_{11}H_{22}\cdots H_{kk}}{H'_{11}H'_{22}\cdots H'_{kk}}.$$

2.3 Algorithms

In this section we discuss algorithms for computing the integral (2.7) exactly, and we discuss their advantages and limitations. In particular, we examine four main techniques which represent the formulas (2.10), (2.11), (2.6) and (2.12) respectively. The practical performance of the various algorithms is compared by computing the integral in Example 1.6.

A MAPLE library which implements our algorithms is made available at

`http://math.berkeley.edu/~shaowei/integrals.html`.

The input for our MAPLE code consists of parameter vectors $s = (s_1, \ldots, s_k)$ and $t = (t_1, \ldots, t_k)$ as well as a data vector $U \in \mathbb{N}^n$. This input uniquely specifies the $d \times n$-matrix A. Here d and n are as in (1.13). The output features the matrices A and $\tilde{A}$, the marginal likelihood integrals for $\mathcal{M}$ and $\mathcal{M}^{(2)}$, as well as the bounds in (2.14).

We tacitly assume that A has been replaced with the reduced matrix $\tilde{A}$. Thus from now on we assume that A has no repeated columns. This requires some care concerning the normalizing constants. All columns of the matrix A have the same coordinate sum a, and the convex hull of the columns is the polytope $P = \Delta_{t_1} \times \Delta_{t_2} \times \cdots \times \Delta_{t_k}$. Our domain of integration is the following polytope of dimension $2d - 2k + 1$:

$$\Theta \;=\; \Delta_1 \times P \times P.$$

We seek to compute the rational number

$$\int_\Theta \prod_{v=1}^{n} (\sigma_0 \theta^{a_v} + \sigma_1 \rho^{a_v})^{U_v} d\sigma d\theta d\rho, \tag{2.15}$$

where integration is with respect to Lebesgue probability measure. Our MAPLE code outputs this integral multiplied with the statistically correct normalizing constant. That constant will be ignored in what follows. In our complexity analysis, we fix A while allowing the data U to vary. The complexities will be given in terms of the sample size $N = U_1 + \cdots + U_n$.

2.3.1 Ignorance is Costly

Given an integration problem such as (2.15), a first attempt is to use the symbolic integration capabilities of a computer algebra package such as MAPLE. We refer to this method as *ignorant integration*:

```
U  := [51, 18, 73, 25, 75]:
f  := (s*t^4        +(1-s)*p^4        )^U[1] *
      (s*t^3*(1-t)  +(1-s)*p^3*(1-p)  )^U[2] *
      (s*t^2*(1-t)^2+(1-s)*p^2*(1-p)^2)^U[3] *
      (s*t  *(1-t)^3+(1-s)*p  *(1-p)^3)^U[4] *
      (s    *(1-t)^4+(1-s)    *(1-p)^4)^U[5]:
II := int(int(int(f,p=0..1),t=0..1),s=0..1);
```

In the case of mixture models, recognizing the integral as the sum of integrals of monomials over a polytope allows us to avoid the expensive integration step above by using (2.10). To demonstrate the power of using (2.10), we implemented a simple algorithm that computes each $\phi_A(b, U)$ using the naive expansion in (2.9). We computed the integral in Example 1.6 with a small data vector $U = (2, 2, 2, 2, 2)$, which is the rational number

$$\frac{66364720654753}{590573839872170153399400000},$$

and summarize the run-times and memory usages of the two algorithms in the table below. All experiments reported in this section are done in MAPLE.

	Time(seconds)	Memory(bytes)
Ignorant Integration	16.331	155,947,120
Naive Expansion	0.007	458,668

For the remaining comparisons in this section, we no longer consider the ignorant integration algorithm because it is computationally too expensive.

2.3.2 Symbolic Expansion of the Integrand

While ignorant use of a computer algebra system is unsuitable for computing our integrals, we can still exploit its powerful polynomial expansion capabilities to find the coefficients of (2.11). A major advantage is that it is very easy to write code for this method. We compare the performance of this symbolic expansion algorithm against that of the naive expansion algorithm. The table below concerns computing the coefficients $\phi_A(b, U)$ for the original data $U = (51, 18, 73, 25, 75)$. The column "Extract" refers to the time taken to extract the coefficients $\phi_A(b, U)$ from the expansion of the polynomial, while the column "Sum" shows the time taken to evaluate (2.10) after all the needed values of $\phi_A(b, U)$ had been computed and extracted.

	Time(seconds)				Memory
	$\phi_A(b, U)$	Extract	Sum	Total	(bytes)
Naive Expansion	2764.35	-	31.19	2795.54	10,287,268
Symbolic Expansion	28.73	962.86	29.44	1021.03	66,965,528

2.3.3 Storage and Evaluation of $\phi_A(b, U)$

Symbolic expansion is fast for computing $\phi_A(b, U)$, but it has two drawbacks: high memory usage and the long time it takes to extract the values of $\phi_A(b, U)$. One solution is to create specialized data structures and algorithms for expanding (2.11), rather using than those offered by MAPLE.

First, we tackle the problem of storing the coefficients $\phi_A(b, U)$ for $b \in Z_A^{\mathbb{L}}(U) \subset \mathbb{R}^d$ as they are being computed. One naive method is to use a d-dimensional array $\phi[\cdot]$. However, noting that A is not row rank full, we can use a d_0-dimensional array to store $\phi_A(b, U)$, where $d_0 = \text{rank}(A) = d - k + 1$. Furthermore, by Proposition 2.7(2), the expanded integrand is a symmetric polynomial, so only half the coefficients need to be stored. We will leave out the implementation details so as not to complicate our discussions. In our algorithms, we will assume that the coefficients are stored in a d_0-dimensional array $\phi[\cdot]$, and the entry that represents $\phi_A(b, U)$ will be referred to as $\phi[b]$.

Next, we discuss how $\phi_A(b, U)$ can be computed. One could use the naive expansion (2.9), but this involves evaluating many binomials coefficients and products, so the algorithm is inefficient for data vectors with large coordinates. A more efficient solution uses the recurrence formula (2.12):

Algorithm 2.13 (RECURRENCE(A, U))**.**
Input: The matrix A and the vector U.
Output: The coefficients $\phi_A(b, U)$.
Step 1: Create a d_0-dimensional array ϕ of zeros.
Step 2: For each $x \in \{0, 1, \ldots, U_1\}$ set

$$\phi[xa_1] \;:=\; \binom{U_1}{x}.$$

Step 3: Create a new d_0-dimensional array ϕ'.
Step 4: For each $2 \le j \le n$ do
1. Set all the entries of ϕ' to 0.
2. For each $x \in \{0, 1, \ldots, U_j\}$ do
   For each non-zero entry $\phi[b]$ in ϕ do
   Increment $\phi'[b + xa_j]$ by $\binom{U_j}{x}\phi[b]$.
3. Replace ϕ with ϕ'.

Step 5: Output the array ϕ.

The space complexity of this algorithm is $O(N^{d_0})$ and its time complexity is $O(N^{d_0+1})$. By comparison, the naive expansion algorithm takes $O(N^d)$ space and $O(N^{n+1})$ time.

We now turn our attention to computing the integral (2.15). One major issue is the lack of memory to store all the terms of the expansion of the integrand. We overcome this problem by writing the integrand as a product of smaller factors which can be expanded

separately. In particular, we partition the columns of A into submatrices $A^{[1]}, \ldots, A^{[m]}$ and let $U^{[1]}, \ldots, U^{[m]}$ be the corresponding partition of U. Thus the integrand becomes

$$\prod_{j=1}^{m} \prod_{v} (\sigma_0 \theta^{a_v^{[j]}} + \sigma_1 \rho^{a_v^{[j]}})^{U_v^{[j]}},$$

where $a_v^{[j]}$ is the v-th column in $A^{[j]}$. The resulting algorithm for evaluating the integral is:

Algorithm 2.14 (Fast Integral)**.**
Input: The matrices $A^{[1]}, \ldots, A^{[m]}$, vectors $U^{[1]}, \ldots, U^{[m]}$ and the vector t.
Output: The value of the integral (2.15) in exact rational arithmetic.
Step 1: For $1 \leq j \leq m$, compute $\phi^{[j]} :=$ RECURRENCE$(A^{[j]}, U^{[j]})$.
Step 2: Set $I := 0$.
Step 3: For each non-zero entry $\phi^{[1]}[b^{[1]}]$ in $\phi^{[1]}$ do

$\vdots$

For each non-zero entry $\phi^{[m]}[b^{[m]}]$ in $\phi^{[m]}$ do
Set $b := b^{[1]} + \cdots + b^{[m]}$, $c := AU - b$, $\phi := \prod_{j=1}^{m} \phi^{[j]}[b^{[j]}]$.
Increment I by

$$\phi \cdot \frac{(|b|/a)!(|c|/a)!}{(|U|+1)!} \cdot \prod_{i=1}^{k} \frac{t_i!\, b_0^{(i)}! \cdots b_{t_i}^{(i)}!}{(|b^{(i)}|+t_i)!} \; \frac{t_i!\, c_0^{(i)}! \cdots c_{t_i}^{(i)}!}{(|c^{(i)}|+t_i)!}.$$

Step 4: Output the sum I.

The algorithm can be sped up by precomputing the factorials used in the product in Step 3. The space and time complexity of this algorithm is $O(N^S)$ and $O(N^T)$ respectively, where $S = \max_i \operatorname{rank} A^{[i]}$ and $T = \sum_i \operatorname{rank} A^{[i]}$. From this, we see that the splitting of the integrand should be chosen wisely to achieve a good pay-off between the two complexities.

In the table below, we compare the naive expansion algorithm and the fast integral algorithm for the data $U = (51, 18, 73, 25, 75)$. We also compare the effect of splitting the integrand into two factors, as denoted by $m = 1$ and $m = 2$. For $m = 1$, the fast integral algorithm takes significantly less time than naive expansion, and requires only about 1.5 times more memory.

	Time(minutes)	Memory(bytes)
Naive Expansion	43.67	9,173,360
Fast Integral (m=1)	1.76	13,497,944
Fast Integral (m=2)	139.47	6,355,828

2.3.4 Limitations and Applications

While our algorithms are optimized for exact evaluation of integrals for mixtures of independence models, they may not be practical for applications involving large sample sizes.

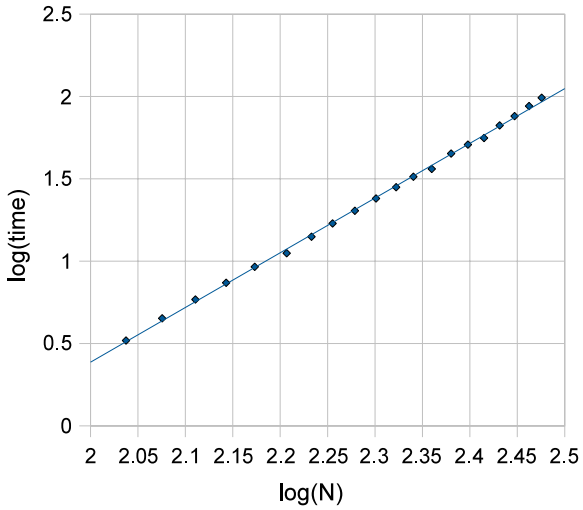


Figure 2.1: Comparison of computation time against sample size.

To demonstrate their limitations, we vary the sample sizes in Example 1.6 and compare the computation times. The data vectors U are generated by scaling $U = (51, 18, 73, 25, 75)$ according to the sample size N and rounding off the entries. Here, N is varied from 110 to 300 by increments of 10. Figure 2.1 shows a logarithmic plot of the results. The times taken for $N = 110$ and $N = 300$ are 3.3 and 98.2 seconds respectively. Computation times for larger samples may be extrapolated from the graph. Indeed, a sample size of 5000 could take more than 13 days.

For other models, such as the 100 *Swiss Francs* model in Section 1.4.5 and that of the schizophrenic patients in Example 2.20, the limitations are even more apparent. In the table below, for each example we list the sample size, computation time, rank of the corresponding A-matrix and the number of terms in the expansion of the integrand. Despite having smaller sample sizes, the computations for the latter two examples take a lot more time. This may be attributed to the higher ranks of the A-matrices and the larger number of terms that need to be summed up in our algorithm.

	Size	Time	Rank	#Terms
Coin Toss	242	45 sec	2	48,646
100 Swiss Francs	40	15 hrs	7	3,892,097
Schizophrenic Patients	132	16 days	5	34,177,836

Despite their high complexities, we believe that our algorithms are important because they provide a gold standard with which approximation methods such as those studied in [39] can be compared. Below, we use our exact methods to ascertain the accuracy of asymptotic formula derived in [57] and [60,61] using desingularization methods from algebraic geometry.

Example 2.15. We revisit the model from Example 1.6. Let us consider data vectors $U = (U_0, U_1, U_2, U_3, U_4)$ with $U_i = Nq_i$ where N is a multiple of 16 and

$$q_i = \frac{1}{16}\binom{4}{i}, \quad i = 0, 1, \ldots, 4.$$

Let $I_N(U)$ be the integral (2.15). Define

$$F_N(U) = N \sum_{i=0}^{4} q_i \log q_i - \log I_N(U).$$

According to [61], for large N we have the asymptotics

$$\mathrm{E}_U[F_N(U)] \;=\; \frac{3}{4}\log N + O(1) \tag{2.16}$$

where the expectation E_U is taken over all U with sample size N under the distribution defined by $q = (q_0, q_1, q_2, q_3, q_4)$. Thus, we should expect

$$F_{16+N} - F_N \;\approx\; \frac{3}{4}\log(16+N) - \frac{3}{4}\log N =: g(N).$$

We compute $F_{16+N} - F_N$ using our exact methods and list the results below.

N	$F_{16+N} - F_N$	$g(N)$
16	0.21027043	0.225772497
32	0.12553837	0.132068444
48	0.08977938	0.093704053
64	0.06993586	0.072682510
80	0.05729553	0.059385934
96	0.04853292	0.050210092
112	0.04209916	0.043493960

Clearly, the table supports our conclusion. The coefficient $3/4$ of $\log N$ in the formula (2.16) is the *learning coefficient* of the statistical model, which was discussed in Section 1.3.2. By Theorem 1.11, this coefficient equals the *real log canonical threshold* of the polynomial ideal

$$\begin{aligned}\langle \sigma_0\theta_0^3\theta_1 + \sigma_1\rho_0^3\rho_1 - 1/16,\ \sigma_0\theta_0^2\theta_1^2 + \sigma_1\rho_0^2\rho_1^2 - 1/16,\\ \sigma_0\theta_0\theta_1^3 + \sigma_1\rho_0\rho_1^3 - 1/16,\ \sigma_0\theta_1^4 + \sigma_1\rho_1^4 - 1/16\rangle.\end{aligned}$$

The example suggests that Proposition 4.2a could be developed into a numerical technique for computing the real log-canonical thresholds of polynomial ideals. □

2.4 Back to Bayesian Statistics

In this section we discuss how the exact integration approach presented here interfaces with issues in Bayesian statistics. The first concerns the rather restrictive assumption that our marginal likelihood integral be evaluated with respect to the uniform distribution (Lebesgue measure) on the parameter space Θ. It is standard practice to compute such integrals with respect to *Dirichlet priors*, and we shall now explain how our algorithms can be extended to Dirichlet priors. That extension is also available as a feature in our MAPLE implementation.

Recall that the *Dirichlet distribution* $\mathrm{Dir}(\alpha)$ is a continuous probability distribution parametrized by a vector $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_m)$ of positive reals. It is the multivariate generalization of the beta distribution and is conjugate prior (in the Bayesian sense) to the multinomial distribution. This means that the probability distribution function of $\mathrm{Dir}(\alpha)$ specifies the belief that the probability of the i-th among $m+1$ events equals θ_i given that it has been observed $\alpha_i - 1$ times. More precisely, the probability density function $f(\theta;\alpha)$ of $\mathrm{Dir}(\alpha)$ is supported on the m-dimensional simplex

$$\Delta_m \;=\; \bigl\{(\theta_0,\ldots,\theta_m) \in \mathbb{R}^m_{\geq 0} \,:\, \theta_0 + \cdots + \theta_m = 1\bigr\},$$

and it equals

$$f(\theta_0,\ldots,\theta_m;\alpha_0,\ldots,\alpha_m) \;\;=\;\; \frac{1}{\mathbb{B}(\alpha)} \cdot \theta_0^{\alpha_0-1}\theta_1^{\alpha_1-1}\cdots\theta_m^{\alpha_m-1} \;\;=:\;\; \frac{\theta^{\alpha-1}}{\mathbb{B}(\alpha)}.$$

Here the normalizing constant is the multinomial beta function

$$\mathbb{B}(\alpha) \;\;=\;\; \frac{m!\Gamma(\alpha_0)\Gamma(\alpha_1)\cdots\Gamma(\alpha_m)}{\Gamma(\alpha_0+\alpha_1+\cdots+\alpha_m)}.$$

Note that, if the α_i are all integers, then this is the rational number

$$\mathbb{B}(\alpha) \;=\; \frac{m!(\alpha_0-1)!(\alpha_1-1)!\cdots(\alpha_m-1)!}{(\alpha_0+\cdots+\alpha_m-1)!}.$$

Thus the identity (2.2) is the special case of the identity $\int_{\Delta_m} f(\theta;\alpha)d\theta = 1$ for the density of the Dirichlet distribution when all $\alpha_i = b_i + 1$ are integers.

We now return to the marginal likelihood for mixtures of independence models. To compute this quantity with respect to Dirichlet priors means the following. We fix positive real numbers α_0, α_1, and $\beta_j^{(i)}$ and $\gamma_j^{(i)}$ for $i = 1,\ldots,k$ and $j = 0,\ldots,t_i$. These specify Dirichlet distributions on Δ_1, P and P. Namely, the Dirichlet distribution on P given by the $\beta_j^{(i)}$ is the product probability measure given by taking the Dirichlet distribution with parameters $(\beta_0^{(i)}, \beta_1^{(i)}, \ldots, \beta_{t_i}^{(i)})$ on the i-th factor Δ_{t_i} in the product (1.15) and similarly for the $\gamma_j^{(i)}$. The resulting product probability distribution on the polytope $\Theta = \Delta_1 \times P \times P$ is

called the *Dirichlet distribution* with parameters (α, β, γ). Its probability density function is the product of the respective densities:

$$f(\sigma,\theta,\rho;\alpha,\beta,\gamma) \;=\; \frac{\sigma^{\alpha-1}}{\mathbb{B}(\alpha)} \cdot \prod_{i=1}^{k} \frac{(\theta^{(i)})^{\beta^{(i)}-1}}{\mathbb{B}(\beta^{(i)})} \cdot \prod_{i=1}^{k} \frac{(\rho^{(i)})^{\gamma^{(i)}-1}}{\mathbb{B}(\gamma^{(i)})}. \tag{2.17}$$

By the marginal likelihood with Dirichlet priors we mean the integral

$$\int_{\Theta} \mathbf{L}_U(\sigma,\theta,\rho)\, f(\sigma,\theta,\rho;\alpha,\beta,\gamma) d\sigma d\theta d\rho. \tag{2.18}$$

This is a modification of (2.5) and it depends not just on the data U and the model $\mathcal{M}^{(2)}$ but also on the choice of Dirichlet parameters (α, β, γ). When the coordinates of these parameters are arbitrary positive reals but not integers, then the value of the integral (2.18) is no longer a rational number. Nonetheless, it can be computed exactly as follows. We abbreviate the product of Gamma functions in the denominator of the density (2.17) as

$$\mathbb{B}(\alpha,\beta,\gamma) \;:=\; \mathbb{B}(\alpha) \cdot \prod_{i=1}^{k} \mathbb{B}(\beta^{(i)}) \cdot \prod_{i=1}^{k} \mathbb{B}(\gamma^{(i)}).$$

Instead of the integrand (2.8) we now need to integrate

$$\sum_{\substack{b \in Z_A^{\mathbb{I}}(U) \\ c = AU-b}} \frac{\phi_A(b,U)}{\mathbb{B}(\alpha,\beta,\gamma)} \cdot \sigma_0^{|b|/a+\alpha_0-1} \cdot \sigma_1^{|c|/a+\alpha_1-1} \cdot \theta^{b+\beta-1} \cdot \rho^{c+\gamma-1}$$

with respect to Lebesgue probability measure on Θ. Doing this term by term, as before, we obtain the following modification of Theorem 2.6.

Corollary 2.16. *The marginal likelihood of the data U in the mixture model $\mathcal{M}^{(2)}$ with respect to Dirichlet priors with parameters (α, β, γ) equals*

$$\frac{N!}{U_1!\cdots U_n! \cdot \mathbb{B}(\alpha,\beta,\gamma)} \cdot \sum_{\substack{b \in Z_A^{\mathbb{I}}(U) \\ c = AU-b}} \phi_A(b,U)\, \frac{\Gamma(|b|/a+\alpha_0)\Gamma(|c|/a+\alpha_1)}{\Gamma(|U|+|\alpha|)} \cdot \prod_{i=1}^{k} \Big(\frac{t_i!\Gamma(b_0^{(i)}+\beta_0^{(i)})\cdots\Gamma(b_{t_i}^{(i)}+\beta_{t_i}^{(i)})}{\Gamma(|b^{(i)}|+|\beta^{(i)}|)} \;\; \frac{t_i!\Gamma(c_0^{(i)}+\gamma_0^{(i)})\cdots\Gamma(c_{t_i}^{(i)}+\gamma_{t_i}^{(i)})}{\Gamma(|c^{(i)}|+|\gamma^{(i)}|)} \Big).$$

A well-known experimental study [39] compares different methods for computing numerical approximations of marginal likelihood integrals. The model considered in the study is the *naive-Bayes model*, which, in the language of algebraic geometry, corresponds to arbitrary secant varieties of Segre varieties. In this chapter we considered the first secant variety of arbitrary Segre-Veronese varieties. In what follows we restrict our discussion to the intersection of both classes of models, namely, to the first secant variety of Segre varieties. For the remainder of this section we fix

$$s_1 = s_2 = \cdots = s_k = 1$$

but we allow $t_1, t_2, \ldots, t_k$ to be arbitrary positive integers. Thus in the model of [39, Equation 1], we fix $r_C = 2$, and the n there corresponds to our k.

To keep things as simple as possible, we shall fix the uniform distribution as in Sections 2.1–2.3 above. Thus, in the notation of [39, §2], all Dirichlet hyperparameters α_{ijk} are set to 1. This implies that, for any data $U \in \mathbb{N}^n$ and any of our models, the problem of finding the maximum a posteriori (MAP) configuration is equivalent to finding the maximum likelihood (ML) configuration. To be precise, the *MAP configuration* is the point $(\hat{\sigma}, \hat{\theta}, \hat{\rho})$ in Θ which maximizes the likelihood function $\mathbf{L}_U(\sigma, \theta, \rho)$ in (2.4). This maximum may not be unique, and there will typically be many local maxima. Chickering and Heckerman [39, §3.2] used the EM algorithm [41, §1.3] to approximate the MAP configuration numerically.

The Laplace approximation and the BIC score (Section 1.3.1) depend on the idea that the MAP configuration can be found with high accuracy and that the data U were actually drawn from the corresponding distribution $p(\hat{\sigma}, \hat{\theta}, \hat{\rho})$. Let $\mathbf{H}(\sigma, \theta, \rho)$ denote the Hessian matrix of the log-likelihood function $\log \mathbf{L}(\sigma, \theta, \rho)$. Then the Laplace approximation [39, Equation 15] states that the logarithm of the marginal likelihood can be approximated by

$$\log \mathbf{L}(\hat{\sigma}, \hat{\theta}, \hat{\rho}) \;-\; \frac{1}{2}\log|\det \mathbf{H}(\hat{\sigma}, \hat{\theta}, \hat{\rho})| \;+\; \frac{2d - 2k + 1}{2}\log(2\pi). \tag{2.19}$$

The Bayesian information criterion (BIC) suggests the coarser approximation

$$\log \mathbf{L}(\hat{\sigma}, \hat{\theta}, \hat{\rho}) \;-\; \frac{2d - 2k + 1}{2}\log(N), \tag{2.20}$$

where $N = U_1 + \cdots + U_n$ is the sample size.

In algebraic statistics, we do not content ourselves with the output of the EM algorithm but, to the extent possible, we seek to actually solve the likelihood equations [30] and compute all local maxima of the likelihood function. We consider it a difficult problem to reliably find $(\hat{\sigma}, \hat{\theta}, \hat{\rho})$, and we are concerned about the accuracy of approximations like (2.19) or (2.20).

Example 2.17. Consider the 100 *Swiss Francs* table (1.20) discussed in Section 1.4.5. Here $k = 2$, $s_1 = s_2 = 1$, $t_1 = t_2 = 3$, the matrix A is unimodular, and (1.17) is the Segre embedding $\mathbb{P}^3 \times \mathbb{P}^3 \hookrightarrow \mathbb{P}^{15}$. The parameter space Θ is 13-dimensional, but the model $\mathcal{M}^{(2)}$ is 11-dimensional, so the given parametrization is not identifiable [19]. This means that the Hessian matrix $\mathbf{H}$ is singular, and hence the Laplace approximation (2.19) is not defined. □

Example 2.18. We compute (2.19) and (2.20) for the model and data in Example 1.6. According to [30, Example 9], the likelihood function $p_0^{51}p_1^{18}p_2^{73}p_3^{25}p_4^{75}$ has three local maxima $(\hat{p}_0, \hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4)$ in the model $\mathcal{M}^{(2)}$, and these translate into six local maxima $(\hat{\sigma}, \hat{\theta}, \hat{\rho})$ in the parameter space Θ, which is the 3-cube. The two global maxima are

$$(0.3367691969, 0.0287713237, 0.6536073424),$$
$$(0.6632308031, 0.6536073424, 0.0287713237).$$

Both of these points in Θ give the same point in the model:

$$(\hat{p}_0, \hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4) \;=\; (0.12104, 0.25662, 0.20556, 0.10758, 0.30920).$$

The likelihood function evaluates to $0.1395471101 \times 10^{-18}$ at this point. The following table compares the various approximations. Here, "Actual" refers to the base-10 logarithm of the marginal likelihood in Example 1.6.

BIC	-22.43100220
Laplace	-22.39666281
Actual	-22.10853411

□

The method for computing the marginal likelihood which was found to be most accurate in the experimental study is the *candidate method* [39, §3.4]. This is a Monte-Carlo method which involves running a Gibbs sampler. The basic idea is that one wishes to compute a large sum, such as (2.10) by sampling among the terms rather than listing all terms. In the candidate method one uses not the sum (2.10) over the lattice points in the zonotope but the more naive sum over all 2^N hidden data that would result in the observed data represented by U. The value of the sum is the number of terms, 2^N, times the average of the summands, each of which is easy to compute. A comparison of the results of the candidate method with our exact computations, as well as a more accurate version of Gibbs sampling which is adapted for (2.10), will be the subject of a future study.

One of the applications of marginal likelihood integrals lies in model selection. An important concept in that field is that of *Bayes factors*. Given data and two competing models, the Bayes factor is the ratio of the marginal likelihood integral of the first model over the marginal likelihood integral of the second model. In our context it makes sense to form that ratio for the independence model $\mathcal{M}$ and its mixture $\mathcal{M}^{(2)}$. To be precise, given any independence model, specified by positive integers $s_1, \ldots, s_k, t_1, \ldots, t_k$ and a corresponding data vector $U \in \mathbb{N}^n$, the Bayes factor is the ratio of the marginal likelihood in Lemma 2.1 and the marginal likelihood in Theorem 2.6. Both quantities are rational numbers and hence so is their ratio.

Corollary 2.19. *The Bayes factor which discriminates between the independence model $\mathcal{M}$ and the mixture model $\mathcal{M}^{(2)}$ is a rational number. It can be computed exactly using Algorithm 2.14 (and our* MAPLE*-implementation).*

Example 2.20. We conclude by applying our method to a data set taken from the Bayesian statistics literature. The study in [18, §3] analyzed the association between length of hospital stay (in years Y) of 132 schizophrenic patients and the frequency with which they are visited

by relatives. Their data set is the following 3×3 contingency table:

$$U \;=\; \begin{array}{lcccc} & 2{\leq}Y{<}10 & 10{\leq}Y{<}20 & 20{\leq}Y & \textit{Totals} \\ \text{Visited regularly} & 43 & 16 & 3 & \textit{62} \\ \text{Visited rarely} & 6 & 11 & 10 & \textit{27} \\ \text{Visited never} & 9 & 18 & 16 & \textit{43} \\ \textit{Totals} & \textit{58} & \textit{45} & \textit{29} & \mathbf{132} \end{array} \tag{2.21}$$

They present estimated posterior means and variances for these data, where *"each estimate requires a* 9*-dimensional integration"* [18, p. 561]. Computing their integrals is essentially equivalent to ours, for $k = 2, s_1 = s_2 = 1, t_1 = t_2 = 2$ and $N = 132$. The authors emphasize that *"the dimensionality of the integral does present a problem"* [18, p. 562], and they point out that *"all posterior moments can be calculated in closed form .... however, even for modest* N *these expressions are far to complicated to be useful"* [18, p. 559].

We differ on that conclusion. In our view, the closed form expressions in Section 2.2 are quite useful for modest sample size N. Using Algorithm 2.14, we computed the integral (2.15). It is the rational number with numerator

$$\begin{array}{l} 27801948853106338912064360032498932910387614080 5 \\ 28524283958209256935726588667532284587409752803 3 \\ 99493069713103633199906939405711180837568853737 \end{array}$$

and denominator

$$\begin{array}{l} 12288402873591935400678094796599848745442833177572204 \\ 50448819979286456995185542195946815073112429169997801 \\ 33503900169921912167352239204153786645029153951176422 \\ 43298328046163472261962028461650432024356339706541132 \\ 34375318471880274818667657423749120000000000000000. \end{array}$$

To obtain the marginal likelihood for the data U above, that rational number (of moderate size) still needs to be multiplied with the normalizing constant

$$\frac{132!}{43! \cdot 16! \cdot 3! \cdot 6! \cdot 11! \cdot 10! \cdot 9! \cdot 18! \cdot 16!}.$$

□

In this chapter, we studied marginal likelihood integrals for mixtures of independence models, and our main contribution is a formula for this integral as a sum over lattice points of a zonotope. To evaluate this formula efficiently and exactly, we prescribed various tricks, including a recurrence relation for the coefficients $\phi_A(b, U)$. We counted the number of lattice points in the zonotope, as a measure of the computational complexity of this evaluation. Last but not least, we extended our results to compute integrals with respect to *Dirichlet priors*. These exact results complement recent developments in the approximation of such integrals using asymptotic theory, including the approach described in the chapters that follow.

Chapter 3

Asymptotic Theory

In this chapter, we study the full asymptotic expansion of Laplace integrals of the form

$$Z(n) = \int_\Omega e^{-nf(\omega)}\varphi(\omega)d\omega, \quad n \to \infty \tag{3.1}$$

given some analyticity conditions on Ω, f and φ. The case where $\Omega = \mathbb{R}^d$ and φ is smooth with compact support was studied intensively in relation to oscillatory integrals by Arnol'd, Guseĭn-Zade and Varchenko [3, §6-7]. They showed that $Z(n)$ has the asymptotic expansion

$$Z(n) \approx \sum_\alpha \sum_{i=1}^{d} c_{\alpha,i} n^{-\alpha} (\log n)^{i-1} \tag{3.2}$$

where the α are positive rational numbers and the $c_{\alpha,i}$ are real constants. Unfortunately, in Bayesian statistics we often encounter integrals where Ω is a *semianalytic* set, i.e.

$$\Omega = \{\omega \in \mathbb{R}^d : g_1(\omega) \geq 0, \ldots, g_l(\omega) \geq 0\}$$

is defined by real analytic inequalities. Furthermore, critical points of the phase function f may lie on the boundary of Ω. Therefore, we are interested primarily in two questions:

1. Do our Bayesian integrals also have asymptotic expansions of the form (3.2)?
2. If so, how do we compute the exponents and coefficients of this expansion?

It turns out that the key idea is *simultaneous resolution* of the singularities of the phase function and the boundary inequalities. Together with a standard treatment of zeta functions and their transforms, we will be able to answer these questions. In Section 3.1, we explore different formulations of Hironaka's theorem on the resolution of singularities which will be used in this dissertation. In Section 3.2, we study local and global properties of zeta functions. Mellin and Laplace transforms will be discussed in Section 3.3. Finally, in Section 3.4, we apply these transforms to our zeta functions to get the asymptotic expansion results. Some of these results have been published in a preprint [37].

Remark 3.1. The idea of applying simultaneous resolutions to semianalytic sets is due to Watanabe [58], but many of the results in this chapter were derived independently of his work. In [58, Remark 4.5], Watanabe comments that integrals over semianalytic subsets Ω do indeed have asymptotic expansions of the form (3.2). His expression for the coefficients of this expansion takes a different form from our expression in Theorem 3.16. □

Remark 3.2. When we say that $Z(n)$ has an asymptotic expansion (3.2), we mean that for each $\alpha_0 > 0$ and $i_0 > 0$, as $n \to \infty$, we have

$$Z(n) - \sum_{(\alpha,i)<(\alpha_0,i_0)} c_{\alpha,i} n^{-\alpha} (\log n)^{i-1} = O(n^{-\alpha_0} (\log n)^{i_0-1}).$$

Here, the pairs (α, i) in the sum are reverse-ordered by the value of $n^{-\alpha_0}(\log n)^{i_0-1}$ for large n, and $O(\,\cdot\,)$ is the big-O notation.

However, for a fixed integer $n > 0$, it is not necessarily true that the infinite series (3.2) converges to $Z(n)$. For instance, by Corollary 5.8, we have the equality

$$Z_1(n) := \int_{\mathbb{R}} e^{-nx^2} dx \;=\; \sqrt{\pi} n^{-1/2}.$$

Changing the domain of integration from $\mathbb{R}$ to the interval $[-1, 1]$ and using Theorem 3.16, we can show that, as $n \to \infty$,

$$Z_2(n) := \int_{-1}^{1} e^{-nx^2} dx \;\approx\; \sqrt{\pi} n^{-1/2}.$$

This is the full asymptotic expansion of $Z_2(n)$, but $Z_2(n) < Z_1(n) = \sqrt{\pi} n^{-1/2}$. □

Let us introduce some notation for the rest of this chapter. Given $x \in \mathbb{R}^d$, let $\mathcal{A}_x(\mathbb{R}^d)$ be the ring of real-valued functions $f : \mathbb{R}^d \to \mathbb{R}$ that are analytic at x. We sometimes shorten the notation to $\mathcal{A}_x$ when it is clear we are working with $\mathbb{R}^d$. When $x = 0$, it is convenient of think of $\mathcal{A}_0$ as a subring of the formal power series ring $\mathbb{R}[[\omega_1, \ldots, \omega_d]] = \mathbb{R}[[\omega]]$. It consists of power series which are convergent in some neighborhood of the origin. For all x, $\mathcal{A}_x \simeq \mathcal{A}_0$ by translation. Now, given a subset $\Omega \subset \mathbb{R}^d$, let $\mathcal{A}_\Omega$ be the ring of real functions analytic at each point $x \in \Omega$. Locally, each function can be represented as a power series centered at x. Given $f \in \mathcal{A}_\Omega$, define the analytic variety $\mathcal{V}_\Omega(f) = \{\omega \in \Omega : f(\omega) = 0\}$ while for an ideal $I \subset \mathcal{A}_\Omega$, we set $\mathcal{V}_\Omega(I) = \cap_{f \in I} \mathcal{V}_\Omega(f)$. Let Γ represent the Gamma function and $\Gamma^{(i)}$ its i-th derivative. Let the vector 1 represent the all-ones vector. Lastly, given a finite set $S \subset \mathbb{R}$, let $\#\min S$ denote the number of times the minimum is attained in S.

3.1 Resolution of Singularities

Hironaka's celebrated theorem [29] on the resolution of singularities is a deep result in algebraic geometry. It was first proved in 1964 and since then, many different variations of this

theorem has been developed. The version we state below is due to Atiyah [4] and was used by Watanabe [58, Theorem 2.3] in constructing singular learning theory.

Theorem 3.3 (Resolution of Singularities)**.** *Let f be a non-constant real analytic function in some neighborhood $\Omega \subset \mathbb{R}^d$ of the origin with $f(0) = 0$. There exists a triple (M, W, ρ) where*

a. $W \subset \Omega$ is a neighborhood of the origin,

b. M is a d-dimensional real analytic manifold,

c. $\rho : M \to W$ is a real analytic map,

satisfying the following properties:

i. ρ is proper, i.e. the inverse image of any compact set is compact,

ii. ρ is a real analytic isomorphism between $M \setminus \mathcal{V}_M(f \circ \rho)$ and $W \setminus \mathcal{V}_W(f)$,

iii. for any $y \in \mathcal{V}_M(f \circ \rho)$, there exists a local chart M_y with coordinates $\mu = (\mu_1, \mu_2, \dots \mu_d)$ such that y is the origin and

$$f \circ \rho(\mu) = a(\mu)\mu_1^{\kappa_1}\mu_2^{\kappa_2} \cdots \mu_d^{\kappa_d} = a(\mu)\mu^{\kappa}$$

where $\kappa_1, \kappa_2, \dots, \kappa_d$ are non-negative integers and a is a real analytic function with $a(\mu) \neq 0$ for all μ. Furthermore, the Jacobian determinant equals

$$|\rho'(\mu)| = h(\mu)\mu_1^{\tau_1}\mu_2^{\tau_2} \cdots \mu_d^{\tau_d} = h(\mu)\mu^{\tau}$$

where $\tau_1, \tau_2, \dots, \tau_d$ are non-negative integers and h is a real analytic function with $h(\mu) \neq 0$ for all μ.

We say that (M, W, ρ) is a *resolution of singularities* or a *desingularization* of f at the origin. In fact, we can desingularize a list of functions simultaneously.

Corollary 3.4 (Simultaneous Resolutions)**.** *Let $f_1, \dots, f_l$ be non-constant real analytic functions in some neighborhood $\Omega \subset \mathbb{R}^d$ of the origin with all $f_i(0) = 0$. Then there exists a triple (M, W, ρ) that desingularizes each f_i at the origin.*

Proof. The idea is to desingularize the product $f_1(\omega) \cdots f_l(\omega)$ and to show that this resolution is also a resolution for each f_i. See [57, Thm 11] and [25, Lemma 2.3] for details. □

Resolution of singularities for a function is synonymous with its monomialization. Given a finitely generated ideal I in the ring $\mathcal{A}_W$ of real analytic functions over W, we can also find a map $\rho : M \to W$ that monomializes the ideal, i.e. the *pullback ideal* $\rho^* I = \{f \circ \rho \in \mathcal{A}_M : f \in I\}$ in each chart of M is generated by monomial functions. One naïve way to prove this result is to simultaneously resolve a system of generators $f_1, \dots, f_r$ for I, which is equivalent to

resolving their product $f_1 \cdots f_r$. So it seems that we only need to know how to monomialize functions. However, as pointed out by Bierstone and Milman [5], Hironaka's inductive proof eventually involves passing to higher codimensional varieties defined by multiple functions, and taking the product of these functions breaks the induction.

In this dissertation, we have chosen to take a coordinate-dependent perspective on the resolution of singularities. Indeed, we speak of coordinate charts on manifolds such that the function of interest is monomial. We have also chosen to study a local version of the resolution theorem by focusing on desingularizing a function at the origin. There is a natural geometric extension of these concepts to schemes, ideal sheafs and divisors. For instance, the ideal sheaf of a normal crossing divisor is one which can be represented locally as a principal monomial ideal. Below, we state Kollár's version of the theorem on the strong monomialization of ideal sheafs. We will not define the terms from algebraic geometry appearing here, but refer the reader to Kollár's book [34] on this subject. Other good expositions on the technical scope and algorithmic aspects of resolutions of singularities have been written by Hauser [28], Bierstone and Milman [6] and Bravo, Encinas and Villamayor [11].

Theorem 3.5 ([34], Thm 3.35)**.** *There is a blowup sequence functor $\mathcal{BP}$ defined on all triples (X, I, E), where X is a smooth scheme of finite type over a field of characteristic zero, $I \subset \mathcal{O}_X$ is an ideal sheaf that is not zero on any irreducible component of X and E is a simple normal crossing divisor on X. The functor $\mathcal{BP}$ satisfies the following conditions.*

1. *In the blowup sequence $\mathcal{BP}(X, I, E) =$*

$$\begin{array}{lccccccccc} \Pi : X_r & \xrightarrow{\pi_{r-1}} & X_{r-1} & \xrightarrow{\pi_{r-2}} & \cdots & \xrightarrow{\pi_1} & X_1 & \xrightarrow{\pi_0} & X_0 & = X, \\ & & \cup & & & & \cup & & \cup & \\ & & Z_{r-1} & & \cdots & & Z_1 & & Z_0 & \end{array}$$

   all centers of blowups are smooth and have simple normal crossing with E.

2. *The pullback $\Pi^* I \subset \mathcal{O}_{X_r}$ is the ideal sheaf of a simple normal crossing divisor.*

3. *$\Pi : X_r \to X$ is an isomorphism over $X \setminus \operatorname{cosupp} I$.*

4. *$\mathcal{BP}$ commutes with smooth morphisms and with change of fields.*

5. *$\mathcal{BP}$ commutes with closed embeddings whenever $E = \emptyset$.*

Because resolution of singularities is an algorithmically delicate process, there are only a few software libraries available for computing desingularizations of functions or ideals. The first such library was written in 2000 by Bodnár and Schicho [9, 10] in `Maple`. Since then, Frühbis-Krüger and Pfister [20] has written another implementation in `Singular` that is faster and uses fewer charts. Typically, computing a resolution of singularities using these libraries for our statistical examples, such as the one in Section 4.3, take an extremely long time unless we apply some clever tricks to simplify the problem. A large number of charts are also produced in the process. Our dissertation hopes to ease both problems by emphasizing the use of fiber ideals in desingularizing statistical models (see Section 1.5).

3.2 Zeta Functions

In this section, we derive real log canonical thresholds of monomial functions, and demonstrate how resolution of singularities allows us to find the thresholds of non-monomial functions. We show that the global threshold of a function over a compact set is the minimum of local thresholds, and present an example where the threshold at a boundary point depend on the boundary inequalities. We end this section with a conjecture about the location of singularities with the smallest threshold.

Recall that in Section 1.3.2, we provided an informal definition of the real log canonical threshold (RLCT) of a function. To make this definition formal, we need to discuss why the zeta function is meromorphic over the whole complex plane and why all its poles are real. This is one of the goals of this section.

Definition 3.6. Given a compact subset Ω of $\mathbb{R}^d$, a function $f \in \mathcal{A}_\Omega$ which is real analytic over Ω, and a smooth function $\varphi : \Omega \to \mathbb{R}$, consider the zeta function

$$\zeta(z) = \int_\Omega \big|f(\omega)\big|^{-z} \, |\varphi(\omega)| \, d\omega, \quad z \in \mathbb{C}. \tag{3.3}$$

This function is well-defined for $z \in \mathbb{R}_{\leq 0}$. If $\zeta(z)$ can be continued analytically to the whole complex plane $\mathbb{C}$, then all its poles are isolated points in $\mathbb{C}$. Moreover, if all its poles are real, then it has a smallest pole λ which is positive. Let θ be the multiplicity of this pole. The pole λ is the *real log canonical threshold* of f with respect to φ over Ω. If $\zeta(z)$ has no poles, we set $\lambda = \infty$ and leave θ undefined. Let $\mathrm{RLCT}_\Omega(f;\varphi)$ be the pair (λ, θ). By abuse of notation, we sometimes refer to this pair as the real log canonical threshold of f. We order these pairs such that $(\lambda_1, \theta_1) < (\lambda_2, \theta_2)$ if for sufficiently large n,

$$\lambda_1 \log n - \theta_1 \log\log n \;<\; \lambda_2 \log n - \theta_2 \log\log n.$$

In other words, $(\lambda_1, \theta_1) < (\lambda_2, \theta_2)$ if $\lambda_1 > \lambda_2$, or $\lambda_1 = \lambda_2$ and $\theta_1 < \theta_2$. Lastly, we let $\mathrm{RLCT}_\Omega f$ denote $\mathrm{RLCT}_\Omega(f; 1)$ where 1 is the constant unit function.

Necessary analyticity conditions on f, φ and Ω such that the real log canonical threshold is well-defined will be given in Corollary 3.10.

3.2.1 Monomialization

There is a simple class of functions, namely monomials $\omega_1^{\kappa_1} \cdots \omega_d^{\kappa_d} = \omega^\kappa$, for which it is easy to compute the real log canonical threshold.

Proposition 3.7. *Let Ω be the positive orthant $\mathbb{R}^d_{\geq 0}$ and $\phi : \Omega \to \mathbb{R}$ be a compactly supported smooth function with $\phi(0) > 0$. Suppose $\kappa = (\kappa_1, \ldots, \kappa_d)$ and $\tau = (\tau_1, \ldots, \tau_d)$ are vectors of non-negative integers. Then,* $\mathrm{RLCT}_\Omega(\omega^\kappa; \omega^\tau \phi) = (\lambda, \theta)$ *where*

$$\lambda = \min_{1 \leq j \leq d} \Big\{ \frac{\tau_j + 1}{\kappa_j} \Big\}, \qquad \theta = \#\min_{1 \leq j \leq d} \Big\{ \frac{\tau_j + 1}{\kappa_j} \Big\}.$$

Proof. See [3, Lemma 7.3]. The idea is to express $\phi(\omega)$ as $T_s(\omega) + R_s(\omega)$ where T_s is the s-th degree Taylor polynomial and R_s the difference. We then integrate the main term $|f|^{-z} T_s$ explicitly and show that the integral of the remaining term $|f|^{-z} R_s$ does not have larger poles. This process gives the analytic continuation of $\zeta(z)$ to the whole complex plane, so we have the Laurent expansion

$$\zeta(z) = \sum_{\alpha > 0} \sum_{i=1}^{d} \frac{d_{\alpha,i}}{(z-\alpha)^i} + P(z) \tag{3.4}$$

where the poles α are positive rational numbers and $P(z)$ is a polynomial. □

For non-monomial $f(\omega)$, Hironaka's theorem allows us to reduce it to the monomial case. Moreover, we can show that the presence of a positive smooth factor in the amplitude does not change the real log canonical threshold. For the rest of this section, let

$$\Omega = \{\omega \in \mathbb{R}^d : g_1(\omega) \geq 0, \ldots, g_l(\omega) \geq 0\}$$

be compact and semianalytic. We also assume that $f, \varphi \in \mathcal{A}_\Omega$.

Lemma 3.8. *For each $x \in \Omega$, there is a neighborhood Ω_x of x in Ω such that for all smooth functions ϕ on Ω_x with $\phi(x) > 0$,*

$$\mathrm{RLCT}_{\Omega_x}(f; \varphi\phi) = \mathrm{RLCT}_{\Omega_x}(f; \varphi).$$

This real log canonical threshold is a positive rational number.

Proof. Let $x \in \Omega$. If $f(x) \neq 0$, then by the continuity of f, there exists a small neighborhood Ω_x where $0 < c_1 < |f(\omega)| < c_2$ for some constants c_1, c_2. Hence, for all smooth functions ϕ, the zeta functions

$$\int_{\Omega_x} \big|f(\omega)\big|^{-z} |\varphi(\omega)\phi(\omega)|\, d\omega \quad \text{and} \quad \int_{\Omega_x} \big|f(\omega)\big|^{-z} |\varphi(\omega)|\, d\omega$$

do not have any poles, so the lemma follows in this case.

Suppose $f(x) = 0$. By Corollary 3.4, we have a simultaneous local resolution of singularities (M, W, ρ) for the functions $f, \varphi, g_1, \ldots, g_l$ vanishing at x. For each point y in the fiber $\rho^{-1}(x)$, we have a local chart satisfying property (iii) of Theorem 3.3. Since ρ is proper, the fiber $\rho^{-1}(x)$ is compact so there is a finite subcover $\{M_y\}$. We claim that the image $\rho(\bigcup M_y)$ contains a neighborhood W_x of x in $\mathbb{R}^d$. Indeed, otherwise, there exists a bounded sequence $\{x_1, x_2, \ldots\}$ of points in $W \setminus \rho(\bigcup M_y)$ whose limit is x. We pick a sequence $\{y_1, y_2, \ldots\}$ such that $\rho(y_i) = x_i$. Since the x_i are bounded, the y_i lie in a compact set so there is a convergent subsequence with limit y_*. The y_i are not in the open set $\bigcup M_y$ so nor is y_*. But $\rho(y_*) = \lim \rho(y_i) = x$ so $y_* \in \rho^{-1}(x) \subset M_y$, a contradiction.

Now, define $\Omega_x = W_x \cap \Omega$ and let $\{\mathcal{M}_y\}$ be the collection of all sets $\mathcal{M}_y = M_y \cap \rho^{-1}(\Omega_x)$ which have positive measure. Picking a partition of unity $\{\sigma_y(\mu)\}$ subordinate to $\{\mathcal{M}_y\}$ such that σ_y is positive at each y, we write the zeta function $\zeta(z) = \int_{\Omega_x} |f(\omega)|^{-z} |\varphi(\omega)\phi(\omega)|\, d\omega$ as

$$\sum_y \int_{\mathcal{M}_y} \left|f \circ \rho(\mu)\right|^{-z} |\varphi \circ \rho(\mu)||\phi \circ \rho(\mu)||\rho'(\mu)|\sigma_y(\mu)\, d\mu.$$

For each y, the boundary conditions $g_i \circ \rho(\mu) \geq 0$ become monomial inequalities, so $\mathcal{M}_y$ is the union of orthant neighborhoods of y. The integral over $\mathcal{M}_y$ can thus be expressed as a sum of integrals of the form

$$\zeta_y(z) = \int_{\mathbb{R}^d_{\geq 0}} \mu^{-\kappa z + \tau} \psi(\mu) d\mu$$

where κ and τ are non-negative integer vectors while ψ is a compactly supported smooth function with $\psi(0) > 0$. Note that κ and τ do not depend on ϕ nor on the choice of orthant at y. By Proposition 3.7, the smallest pole of $\zeta_y(z)$ is

$$\lambda_y = \min_{1 \leq j \leq d} \{\frac{\tau_j + 1}{\kappa_j}\}, \quad \theta_y = \#\min_{1 \leq j \leq d} \{\frac{\tau_j + 1}{\kappa_j}\}.$$

Now, $\mathrm{RLCT}_{\Omega_x}(f; \varphi\phi) = \min_y\{(\lambda_y, \theta_y)\}$ which is a positive rational number. This formula is independent of ϕ, so we set $\phi = 1$ and the lemma follows. □

Abusing notation, we now let $\mathrm{RLCT}_{\Omega_x}(f; \varphi)$ represent the real log canonical threshold for a sufficiently small neighborhood Ω_x of x in Ω. If x is an interior point of Ω, we denote the threshold at x by $\mathrm{RLCT}_x(f; \varphi)$. More generally, we say that Ω is *full* at x if x lies in the closure of the interior of Ω. Note that if Ω is not full at x, then the integral (3.3) over small Ω_x is zero so $\mathrm{RLCT}_{\Omega_x}(f; \varphi) = (\infty, -)$.

3.2.2 Localization

The global RLCT over a subset Ω is the minimum of local RLCTs at points in Ω.

Proposition 3.9. *The set* $\{\mathrm{RLCT}_{\Omega_x}(f; \varphi) : x \in \Omega\}$ *has a minimum and*

$$\mathrm{RLCT}_{\Omega}(f; \varphi\phi) = \min_{x \in \Omega} \mathrm{RLCT}_{\Omega_x}(f; \varphi)$$

for all positive and smooth functions $\phi : \Omega \to \mathbb{R}$. *In fact, it suffices to consider the minimum over all* x *in the variety* $\mathcal{V}_\Omega(f)$, *and the RLCT is* $(\infty, -)$ *if this variety is empty.*

Proof. Lemma 3.8 associates a small neighborhood to each point in the compact set Ω, so there exists a subcover $\{\Omega_x : x \in S\}$ where S is finite. Let $\{\sigma_x(\omega)\}$ be a partition of unity subordinate to this subcover. Then,

$$\int_\Omega \left|f(\Omega)\right|^{-z} |\varphi(\omega)\phi(\omega)|\, d\omega = \sum_{x\in S} \int_{\Omega_x} \left|f(\Omega)\right|^{-z} |\varphi(\omega)\phi(\omega)|\, \sigma_x(\omega)\, d\omega.$$

From this finite sum, we have

$$\mathrm{RLCT}_\Omega(f;\varphi\phi) = \min_{x\in S} \mathrm{RLCT}_{\Omega_x}(f;\varphi\phi\sigma_x) = \min_{x\in S} \mathrm{RLCT}_{\Omega_x}(f;\varphi).$$

Now, if $y \in \Omega \setminus S$, let Ω_y be a neighborhood of y prescribed by Lemma 3.8 and consider the cover $\{\Omega_x : x \in S\} \cup \{\Omega_y\}$ of Ω. After choosing a partition of unity subordinate to this cover and repeating the above argument, we get

$$\mathrm{RLCT}_\Omega(f;\varphi\phi) \le \mathrm{RLCT}_{\Omega_y}(f;\varphi) \quad \text{for all } y \in \Omega.$$

Combining the two previously displayed equations proves the proposition. The last statement follows from the fact that the RLCT is infinite for points $x \notin \mathcal{V}_\Omega(f)$. □

We say that $\varphi : \Omega \to \mathbb{R}$ is *nearly analytic* if φ is a product $\varphi_a\varphi_s$ of functions where φ_a is real analytic and φ_s is positive and smooth. Our results up to this point allow us to conclude that for f, φ and Ω satisfying the following analyticity conditions, the zeta function (3.3) is meromorphic so the real log canonical threshold is well-defined.

Corollary 3.10. *Given a compact semianalytic set $\Omega \subset \mathbb{R}^d$, a function $f \in \mathcal{A}_\Omega$ satisfying $f(x) = 0$ for some $x \in \Omega$, and a nearly analytic function $\varphi : \Omega \to \mathbb{R}$, the zeta function (3.3) can be continued analytically to $\mathbb{C}$. It has a Laurent expansion (3.4) whose poles are positive rational numbers with a smallest element.*

Proof. The proofs of Lemma 3.8 and Proposition 3.9 outline a way to compute the Laurent expansion of the zeta function (3.3). □

3.2.3 Comparability of Phase Functions

If the function whose RLCT we are finding is complicated, we may replace it with a simpler function that bounds it. More precisely, given $f, g \in \mathcal{A}_\Omega$, we say that f and g are *comparable* in Ω if $c_1 f \le g \le c_2 f$ in Ω for some $c_1, c_2 > 0$. The next two results are due to Watanabe.

Proposition 3.11. *Given $f, g \in \mathcal{A}_\Omega$, suppose $0 \le cf \le g$ in Ω for some constant $c > 0$. Then, $\mathrm{RLCT}_\Omega(f;\varphi) \le \mathrm{RLCT}_\Omega(g;\varphi)$.*

Proof. See [58, §7]. □

Corollary 3.12. *If f, g are comparable in Ω, then $\mathrm{RLCT}_\Omega(f;\varphi) = \mathrm{RLCT}_\Omega(g;\varphi)$.*

3.2.4 Boundary of Domain of Integration

We now show that the threshold at a boundary point depends on the shape of the boundary.

Example 3.13. Consider the following two small neighborhoods of the origin.

$$\Omega_1 = \{(x, y) \in \mathbb{R}^2 : 0 \leq x \leq y \leq \varepsilon\}$$
$$\Omega_2 = \{(x, y) \in \mathbb{R}^2 : 0 \leq y \leq x \leq \varepsilon\}$$

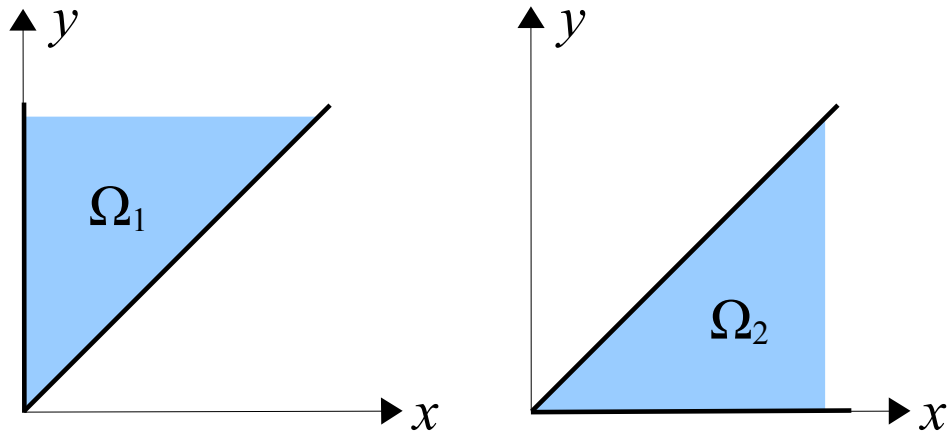


Figure 3.1: RLCT depends on the boundary inequalities.

To compute the real log canonical threshold of the function xy^2 over these sets, we have the corresponding zeta functions below.

$$\zeta_1(z) = \int_0^{\varepsilon}\int_0^{y} x^{-z}y^{-2z}\,dx\,dy = \frac{\varepsilon^{-3z+2}}{(-z+1)(-3z+2)}$$
$$\zeta_2(z) = \int_0^{\varepsilon}\int_0^{x} x^{-z}y^{-2z}\,dy\,dx = \frac{\varepsilon^{-3z+2}}{(-2z+1)(-3z+2)}$$

This shows that $\mathrm{RLCT}_{\Omega_1}(xy^2) = 2/3$ while $\mathrm{RLCT}_{\Omega_2}(xy^2) = 1/2$. □

Note that the RLCT does not depend as much on the analytic inequalities defining the boundary as it does on the set of points that the inequalities cut out. For instance, the above two neighborhoods of the origin may also be expressed as follows.

$$\Omega_1 = \{(x, y) \in \mathbb{R}^2 : 0 \leq x^2 \leq y^2 \leq \varepsilon\}$$
$$\Omega_2 = \{(x, y) \in \mathbb{R}^2 : 0 \leq y^2 \leq x^2 \leq \varepsilon\}$$

Changing the inequalities may change the resolution of singularities required to desingularize the phase f and the boundary of Ω, but the RLCT remains unchanged because the domain of integration for the corresponding zeta function $\zeta(z)$ is unchanged.

3.2.5 Deepest Singularities

Because the real log canonical threshold over a set $\Omega \subset \mathbb{R}^d$ is the minimum of thresholds at points $x \in \Omega$, we want to know where this minimum is achieved. Let us study this problem topologically. Consider a locally finite collection $\mathcal{S}$ of pairwise disjoint submanifolds $S \subset \Omega$ such that $\Omega = \cup_{S \in \mathcal{S}} S$ and each S is locally closed, i.e. the intersection of an open and a closed subset. Let $\overline{S}$ be the closure of S. We say $\mathcal{S}$ is a *stratification* of Ω if $S \cap \overline{T} \neq \emptyset$ implies $S \subset \overline{T}$ for all $S, T \in \mathcal{S}$. A stratification $\mathcal{S}$ of Ω is a *refinement* of another stratification $\mathcal{T}$ if $S \cap T \neq \emptyset$ implies $S \subset T$ for all $S \in \mathcal{S}$ and $T \in \mathcal{T}$.

Let the amplitude $\varphi : \Omega \to \mathbb{R}$ be nearly analytic. Define $S_{\lambda,1}, \ldots, S_{\lambda,r}$ to be the connected components of the set $\{x \in \Omega : \mathrm{RLCT}_{\Omega_x}(f; \varphi) = \lambda\}$ and let $\mathcal{S}$ be the collection $\{S_{\lambda,i}\}$. Now, define the order $\mathrm{ord}_x f$ of f at a point $x \in \Omega$ to be the smallest degree of a monomial appearing in a series expansion of f at x. This number is independent of the choice of local coordinates $\omega_1, \ldots, \omega_d$ because it is the largest k such that $f \in \mathfrak{m}_x^k$ where $\mathfrak{m}_x = \{g \in A_x : g(x) = 0\}$ is the vanishing ideal of x. Define $T_{l,1}, \ldots, T_{l,s}$ to be the connected components of the set $\{x \in \Omega : \mathrm{ord}_x f = l\}$ and let $\mathcal{T}$ be the collection $\{T_{l,j}\}$. We conjecture the following relationship between $\mathcal{S}$ and $\mathcal{T}$. It implies that the minimum real log canonical threshold over a set must occur at a point of highest order. An example of this stratification may be found in Proposition 4.25.

Conjecture 3.14. *The collections $\mathcal{S}$ and $\mathcal{T}$ are stratifications of Ω. Furthermore, if the amplitude φ is a positive smooth function, then $\mathcal{S}$ refines $\mathcal{T}$.*

3.3 Laplace and Mellin Transforms

Laplace integrals such as (3.1) occur frequently in physics, statistics and other applications. At first, the relationship between their asymptotic expansions and the Laurent expansion of the zeta function (3.3) seems strange. The key is to write these integrals as

$$Z(n) = \int_\Omega e^{-n|f(\omega)|} |\varphi(\omega)| \, d\omega = \int_0^\infty e^{-nt} v(t) \, dt$$
$$\zeta(z) = \int_\Omega \big|f(\omega)\big|^{-z} |\varphi(\omega)| \, d\omega = \int_0^\infty t^{-z} v(t) \, dt$$

where $v(t)$ is the state density function [57] or Gelfand-Leray function [3]

$$v(t) = \frac{d}{dt} \int_{0<|f(\omega)|<t} |\varphi(\omega)| \, d\omega.$$

Formally, $Z(n)$ is the *Laplace transform* of $v(t)$ while $\zeta(z)$ is its *Mellin transform*. Note that contrary to its name, $v(t)$ is generally not a function but a Schwartz distribution. Next, we

study the series expansions

$$Z(n) \approx \sum_{\alpha} \sum_{i=1}^{d} c_{\alpha,i} n^{-a} (\log n)^{i-1} \tag{3.5}$$

$$v(t)dt \approx \sum_{\alpha} \sum_{i=1}^{d} b_{\alpha,i}\, t^{\alpha} (\log t)^{i-1} dt \tag{3.6}$$

$$\zeta(z) \sim \sum_{\alpha} \sum_{i=1}^{d} d_{\alpha,i} (z-\alpha)^{-i} \tag{3.7}$$

where the series (3.5) and (3.6) are asymptotic expansions while (3.7) is the principal part of the Laurent series expansion. Formulas relating their coefficients are then deduced from the Laplace and Mellin transforms of $t^{\alpha}(\log t)^{i}$. Detailed expositions on this subject have been written by Arnol'd–Guseĭn-Zade–Varchenko [3, §6-7], Watanabe [57, §4] and Greenblatt [26].

Proposition 3.15. *The asymptotic expansion of the Laplace transform of $t^{\alpha-1}(\log t)^{i}$ is*

$$\int_0^{\infty} e^{-nt}\, t^{\alpha-1} (\log t)^{i}\, dt \approx \sum_{j=0}^{i} \binom{i}{j} (-1)^{j} \Gamma^{(i-j)}(\alpha)\, n^{-\alpha} (\log n)^{j}$$

while the Mellin transform of $t^{\alpha-1}(\log t)^{i}$ is

$$\int_0^{1} t^{-z}\, t^{\alpha-1} (\log t)^{i}\, dt = -\, i!\, (z-\alpha)^{-(i+1)}.$$

Proof. See [3, Thm 7.4] and [57, Ex 4.7] respectively. □

3.4 Asymptotic Expansion

In this section, we employ standard techniques to derive the asymptotic expansion of the Laplace integral (3.1) from the Laurent expansion of the zeta function (3.3). Recall that Γ is the Gamma function and that $\Gamma^{(i)}$ is its i-th derivative.

Theorem 3.16. *Let $\Omega \subset \mathbb{R}^d$ be a compact semianalytic subset and $\varphi : \Omega \to \mathbb{R}$ be nearly analytic. If $f \in \mathcal{A}_{\Omega}$ with $f(x) = 0$ for some $x \in \Omega$, then the Laplace integral*

$$Z(n) = \int_{\Omega} e^{-n|f(\omega)|} |\varphi(\omega)|\, d\omega$$

has the asymptotic expansion

$$\sum_{\alpha} \sum_{i=1}^{d} c_{\alpha,i}\, n^{-\alpha} (\log n)^{i-1}. \tag{3.8}$$

The α in this expansion range over positive rational numbers which are poles of

$$\zeta(z) = \int_{\Omega_\delta} \big|f(\omega)\big|^{-z} |\varphi(\omega)|\, d\omega \tag{3.9}$$

where $\delta \in \mathbb{R}$ is any $\delta > 0$ and $\Omega_\delta = \{\omega \in \Omega : |f(\omega)| < \delta\}$. The coefficients $c_{\alpha,i}$ satisfy

$$c_{\alpha,i} = \frac{(-1)^i}{(i-1)!} \sum_{j=i}^{d} \frac{\Gamma^{(j-i)}(\alpha)}{(j-i)!}\, d_{\alpha,j} \tag{3.10}$$

where $d_{\alpha,j}$ is the coefficient of $(z-\alpha)^{-j}$ in the Laurent expansion of $\zeta(z)$.

Proof. First, set $\delta = 1$. We split the integral $Z(n)$ into two parts:

$$Z(n) = \int_{|f(\omega)|<1} e^{-n|f(\omega)|} |\varphi(\omega)|\, d\omega + \int_{|f(\omega)|\geq 1} e^{-n|f(\omega)|} |\varphi(\omega)|\, d\omega.$$

The second integral is bounded above by Ce^{-n} for some positive constant C, so asymptotically it goes to zero more quickly than any $n^{-\alpha}$. For the first integral, we write $\zeta(z)$ as the Mellin transform of the state density function $v(t)$.

$$\zeta(z) = \int_{|f(\omega)|<1} \big|f(\omega)\big|^{-z} |\varphi(\omega)|\, d\omega = \int_0^1 t^{-z} v(t)\, dt.$$

By Corollary 3.10, $\zeta(z)$ has a Laurent expansion (3.4). Moreover, since $|f(\omega)| < 1$, $\zeta(n) \to 0$ as $n \to -\infty$ so the polynomial part $P(z)$ is identically zero. Applying the inverse Mellin transform to $\zeta(z)$, we get a series expansion (3.6) of the state density function $v(t)$. Applying the Laplace transform to $v(t)$ in turn gives the asymptotic expansion (3.5) of $Z(n)$. Therefore, from Proposition 3.15, we get the relations

$$c_{\alpha,i} = (-1)^{i-1} \sum_{j=i}^{d} \binom{j-1}{i-1} \Gamma^{(j-i)}(\alpha)\, b_{\alpha-1,j}, \quad d_{\alpha,j} = -(j-1)!\, b_{\alpha-1,j}.$$

Equation (3.10) follows immediately. Finally, for all other values of δ, we write

$$\int_\Omega \big|f(\omega)\big|^{-z} |\varphi(\omega)| d\omega = \int_{\Omega_\delta} \big|f(\omega)\big|^{-z} |\varphi(\omega)| d\omega + \int_{|f(\omega)|\geq\delta} \big|f(\omega)\big|^{-z} |\varphi(\omega)| d\omega.$$

The last integral does not have any poles, so the principal parts of the Laurent expansions of the first two integrals are the same for all δ. □

In this chapter, we studied the asymptotic theory of Laplace integrals over semianalytic subsets. Among our main contributions are Lemma 3.8 and Proposition 3.9 which describe local properties of the RLCT, and Theorem 3.16 which gives an explicit formula for the asymptotic expansion of the Laplace integral. These results will be important for proofs in Chapter 4. We also discussed delicate issues such as the behavior of the RLCT at boundary points and the location of deepest singularities. In Chapter 5, we exploit Theorem 3.16 and develop algorithms which allows us to compute the asymptotics in Example 1.18.

Chapter 4

Real Log Canonical Thresholds

In Chapter 3, we defined real log canonical thresholds (RLCTs) for *functions*, and studied their connection to asymptotic expansions of Laplace integrals. In this chapter, we investigate RLCTs for *ideals*, and show that the informal definition given in Section 1.5.2 is independent of the choice of generators. Computing RLCTs of ideals is important for many statistical applications, as was described in Section 1.5 where we studied *fiber ideals* of models.

In Section 4.1, we investigate their fundamental properties which provides us with symbolic tools for computing the RLCT more efficiently. In nondegenerate cases, we can calculate the RLCT using a combinatorial geometric method involving Newton polyhedra which originates from toric geometry. The method has been applied to RLCTs of functions [3, §8]. Our contribution in Section 4.2 is to extend it to RLCTs of ideals and to give a formula for the RLCT of a monomial ideal with respect to a monomial amplitude. We finish with a difficult statistical example in Section 4.3 which employs the tools discussed in this chapter. In our applications, we will only study RLCTs of polynomial ideals, but the proofs in this chapter extend easily to analytic functions so we will state them in their full generality. Some of our results have been publish in a preprint [37].

In this section, let $\Omega \subset \mathbb{R}^d$ be a compact semianalytic subset and let $\varphi : \Omega \to \mathbb{R}$ be nearly analytic. Before we give a definition for the RLCT of an ideal, let us define the RLCT for a finite set of functions $\{f_1, \ldots, f_r\} \subset \mathcal{A}_\Omega$. Later, this set of functions will represent a choice of generators for the ideal, and we will show that the RLCT is independent of this choice. Indeed, let $\mathrm{RLCT}_\Omega(f_1, \ldots, f_r; \varphi)$ be the smallest pole and multiplicity of the zeta function

$$\zeta(z) = \int_\Omega \Big(f_1(\omega)^2 + \cdots + f_r(\omega)^2 \Big)^{-z/2} |\varphi(\omega)|\, d\omega. \tag{4.1}$$

Recall that these pairs are ordered by the rule $(\lambda_1, \theta_1) > (\lambda_2, \theta_2)$ if $\lambda_1 > \lambda_2$, or $\lambda_1 = \lambda_2$ and $\theta_1 < \theta_2$. For $x \in \Omega$, we define $\mathrm{RLCT}_{\Omega_x}(f_1, \ldots, f_r; \varphi)$ to be the threshold for a sufficiently small neighborhood Ω_x of x in Ω.

Remark 4.1. The (complex) log canonical threshold may be defined in a similar fashion. It is the smallest pole of the zeta function

$$\zeta(z) = \int_\Omega \Big(|f_1(\omega)|^2 + \cdots + |f_r(\omega)|^2 \Big)^{-z} d\omega.$$

Observe that the f_i^2 have been replaced by $|f_i|^2$ and the exponent $-z/2$ is changed to $-z$. Crudely, this factor of 2 comes from the fact that $\mathbb{C}^d$ is a real vector space of dimension $2d$. The complex threshold is often different from the RLCT [47]. In algebraic geometry, more is known about complex log canonical thresholds than about real log canonical thresholds. Many results in this chapter were motivated by their complex analogs [8, 31, 33, 36]. □

4.1 Fundamental Formulas

4.1.1 Equivalent definitions

We give several equivalent definitions of $\mathrm{RLCT}_\Omega(f_1, \ldots, f_r; \varphi)$ which are helpful in proving its fundamental properties.

Proposition 4.2. *Given real analytic functions $f_1, \ldots, f_r \in \mathcal{A}_\Omega$, the pairs (λ, θ) defined in the statements below are all equal.*

a. The logarithmic Laplace integral

$$\log Z(n) = \log \int_\Omega \exp\Big(-n \sum_{i=1}^r f_i(\omega)^2 \Big) |\varphi(\omega)|\, d\omega$$

is asymptotically $-\frac{\lambda}{2} \log n + (\theta - 1) \log\log n + O(1)$.

b. The zeta function

$$\zeta(z) = \int_\Omega \Big(\sum_{i=1}^r f_i(\omega)^2 \Big)^{-z/2} |\varphi(\omega)|\, d\omega$$

has a smallest pole λ of multiplicity θ.

c. The pair (λ, θ) is the minimum

$$\min_{x \in \Omega} \mathrm{RLCT}_{\Omega_x}(f_1, \ldots, f_r; \varphi).$$

In fact, it is enough to vary x over $\mathcal{V}_\Omega(\langle f_1, \ldots, f_r \rangle)$.

Proof. Item (b) is the original definition of the RLCT. The equivalence of (a) and (b) follows from Theorem 3.16, and that of (b) and (c) from Proposition 3.9. Some of the statements in this proposition are also proved in [57, Thm 7.1]. □

4.1.2 Choice of Generators

$\mathrm{RLCT}_\Omega(f_1^2+\cdots+f_r^2;\varphi)=(\lambda,\theta)$ implies $\mathrm{RLCT}_\Omega(f_1,\ldots,f_r;\varphi)=(2\lambda,\theta)$. From this, it seems that we should restrict ourselves to RLCTs of single and of not multiple functions. However, as the next proposition shows, multiple functions are important because they allow us to work with ideals for which different generating sets can be chosen. This gives us freedom to switch between single and multiple functions in powerful ways. For instance, special cases of this proposition such as Lemmas 3 and 4 of [2] have been used to simplify computations.

Proposition 4.3. *If two sets $\{f_1,\ldots,f_r\}$ and $\{g_1,\ldots,g_s\}$ of functions generate the same ideal $I\subset\mathcal{A}_\Omega$, then*

$$\mathrm{RLCT}_\Omega(f_1,\ldots,f_r;\varphi)=\mathrm{RLCT}_\Omega(g_1,\ldots,g_s;\varphi).$$

Define this pair (λ,θ) to be $\mathrm{RLCT}_\Omega(I;\varphi)$. Here, λ is a positive rational number.

Proof. Each g_j can be written as a combination $h_1f_1+\cdots+h_rf_r$ of the f_i where the h_i are real analytic over Ω. By the Cauchy-Schwarz inequality,

$$g_j^2\le\big(h_1^2+\cdots+h_r^2\big)\big(f_1^2+\cdots+f_r^2\big).$$

Because Ω is compact, the h_i are bounded. Thus, summing over all the g_j, there is some constant $c>0$ such that,

$$\sum_{j=1}^{s}g_j^2\le c\sum_{i=1}^{r}f_i^2.$$

By Proposition 3.11, $\mathrm{RLCT}_\Omega(g_1,\ldots,g_r;\varphi)\le\mathrm{RLCT}_\Omega(f_1,\ldots,f_r;\varphi)$ and by symmetry, the reverse is also true, so we have equality. The fact that the real log canonical threshold is a positive rational number follows from Corollary 3.10. □

Above, we defined the RLCT of an ideal $\langle f_1,\ldots,f_r\rangle$ in terms of the RLCT of a particular function $K(f_1,\ldots,f_r)=f_1^2+\ldots+f_r^2$. We now show that any function K can be used as long as $K(0)=0,\nabla K(0)=0$ and $\nabla^2K(0)\succ 0$, i.e. the Hessian is positive definite.

Proposition 4.4. *Let $U\subset\mathbb{R}^d$, and let the maps $f:\Omega\to U$ and $K:U\to\mathbb{R}$ be real analytic at $\hat\omega\in\Omega$ and $\hat f=f(\hat\omega)\in U$ respectively. Suppose $K(\hat f)=0,\nabla K(\hat f)=0$ and $\nabla^2K(\hat f)\succ 0$. Then, for all $\varphi(\omega)$ nearly analytic at $\hat\omega$,*

$$\mathrm{RLCT}_{\Omega_{\hat\omega}}(K\circ f(\omega);\varphi)=(\lambda,\theta)$$

where $(2\lambda,\theta)=\mathrm{RLCT}_{\Omega_{\hat\omega}}(\langle f(\omega)-\hat f\rangle;\varphi)$.

Proof. Without loss of generality, we assume that $\hat{f}$ is the origin. Because $\nabla^2 K(\hat{f})$ is positive definite, there exists a linear change of coordinates $T : \mathbb{R}^d \to \mathbb{R}^d$ such that the power series expansion of K is $u_1^2 + \cdots + u_d^2 + O(u^3)$ where $(u_1, \ldots, u_d) = T(f_1, \ldots, f_d)$. This implies that there is a sufficiently small neighborhood $\tilde{U} \subset U$ of the origin such that

$$c(u_1^2 + \cdots + u_d^2) \leq K \circ T^{-1}(u) \leq C(u_1^2 + \cdots + u_d^2), \quad \forall\, u \in \tilde{U}$$

for some positive constants c and C. Linear algebra tells us that

$$\lambda_d(f_1^2 + \cdots + f_d^2) \leq u_1^2 + \cdots + u_d^2 \leq \lambda_1(f_1^2 + \cdots + f_d^2), \quad \forall\, f \in T^{-1}\tilde{U}$$

where λ_1 is the largest eigenvalue of $T^{\mathsf{T}}T$ and λ_d the smallest. Hence,

$$c\lambda_d(f_1^2 + \cdots + f_d^2) \leq K(f) \leq C\lambda_1(f_1^2 + \cdots + f_d^2), \quad \forall\, f \in T^{-1}\tilde{U}.$$

Now, since $f : \Omega \to U$ is continuous at $\hat{\omega}$, there exists some neighborhood $\tilde{\Omega} \subset \Omega$ of $\hat{\omega}$ such that $f(\tilde{\Omega}) \subset T^{-1}\tilde{U}$. Thus,

$$c\lambda_d(f_1(\omega)^2 + \cdots + f_d(\omega)^2) \leq K \circ f(\omega) \leq C\lambda_1(f_1(\omega)^2 + \cdots + f_d(\omega)^2), \quad \forall\, \omega \in \tilde{\Omega}$$

and so by Proposition 3.11,

$$\mathrm{RLCT}_{\Omega_{\hat{\omega}}}(K \circ f(\omega); \varphi) = \mathrm{RLCT}_{\Omega_{\hat{\omega}}}(f_1(\omega)^2 + \cdots + f_d(\omega)^2; \varphi) = (\lambda, \theta).$$

Finally, by definition, $(2\lambda, \theta) = \mathrm{RLCT}_{\Omega_{\hat{\omega}}}(\langle f_1(\omega), \ldots, f_d(\omega)\rangle; \varphi)$. □

4.1.3 Sum, Product and Chain Rules

For the next result, let $f_1, \ldots, f_r \in \mathcal{A}_X$ and $g_1, \ldots, g_s \in \mathcal{A}_Y$ where $X \subset \mathbb{R}^m$ and $Y \subset \mathbb{R}^n$ are compact semianalytic subsets. This occurs, for instance, when the f_i and g_j are polynomials with disjoint sets of indeterminates $\{x_1, \ldots, x_m\}$ and $\{y_1, \ldots, y_n\}$. Let $\varphi_x : X \to \mathbb{R}$ and $\varphi_y : Y \to \mathbb{R}$ be nearly analytic. We define $(\lambda_x, \theta_x) = \mathrm{RLCT}_X(f_1, \ldots, f_r; \varphi_x)$ and $(\lambda_y, \theta_y) = \mathrm{RLCT}_Y(g_1, \ldots, g_s; \varphi_y)$.

By composing with projections $X \times Y \to X$ and $X \times Y \to Y$, we may regard the f_i and g_j as functions analytic over $X \times Y$. Let I_x and I_y be ideals in $\mathcal{A}_{X\times Y}$ generated by the f_i and g_j respectively. Recall that the sum $I_x + I_y$ is generated by all the f_i and g_j while the product $I_x I_y$ is generated by $f_i g_j$ for all i, j.

Proposition 4.5 (Sum and Product Rules)**.** *The RLCTs for the sum and product of ideals I_x and I_y with disjoint indeterminates are*

$$\begin{aligned}
\mathrm{RLCT}_{X\times Y}(I_x + I_y; \varphi_x\varphi_y) &= (\lambda_x + \lambda_y,\ \theta_x + \theta_y - 1), \\
\mathrm{RLCT}_{X\times Y}(I_x I_y; \varphi_x\varphi_y) &= \begin{cases} (\lambda_x,\ \theta_x) & \text{if } \lambda_x < \lambda_y, \\ (\lambda_y,\ \theta_y) & \text{if } \lambda_x > \lambda_y, \\ (\lambda_x,\ \theta_x + \theta_y) & \text{if } \lambda_x = \lambda_y. \end{cases}
\end{aligned}$$

Proof. Define $f(x) = f_1^2 + \cdots + f_r^2$ and $g(y) = g_1^2 + \cdots + g_s^2$, and let $Z_x(n)$ and $Z_y(n)$ be the corresponding Laplace integrals. By Proposition 4.2,

$$\begin{aligned} \log Z_x(n) &= -\tfrac{1}{2}\lambda_x \log n + (\theta_x - 1)\log\log n + O(1) \\ \log Z_y(n) &= -\tfrac{1}{2}\lambda_y \log n + (\theta_y - 1)\log\log n + O(1) \end{aligned}$$

asymptotically. If $(\lambda, \theta) = \mathrm{RLCT}_{X\times Y}(I_x + I_y; \varphi_x\varphi_y)$, then

$$\begin{aligned} &-\tfrac{1}{2}\lambda \log n + (\theta - 1)\log\log n + O(1) \\ &\quad = \log \textstyle\int_{X\times Y} e^{-nf(x)-ng(y)}|\varphi_x||\varphi_y|\,dx\,dy \\ &\quad = \log \big(\textstyle\int_X e^{-nf(x)}|\varphi_x|\,dx\big)\big(\textstyle\int_Y e^{-ng(y)}|\varphi_y|\,dy\big) \\ &\quad = \log Z_x(n) + \log Z_y(n) \\ &\quad = -\tfrac{1}{2}(\lambda_x + \lambda_y)\log n + (\theta_x + \theta_y - 2)\log\log n + O(1) \end{aligned}$$

and the first result follows. For the second result, note that

$$f(x)g(y) = f_1^2 g_1^2 + f_1^2 g_2^2 + \cdots + f_r^2 g_s^2.$$

Let $\zeta_x(z)$ and $\zeta_y(z)$ be the zeta functions corresponding to $f(x)$ and $g(y)$. By Proposition 4.2, (λ_x, θ_x) and (λ_y, θ_y) are the smallest poles of $\zeta_x(z)$ and $\zeta_y(z)$ while $\mathrm{RLCT}_{X\times Y}(I_xI_y; \varphi_x\varphi_y)$ is the smallest pole of

$$\begin{aligned} \zeta(z) &= \textstyle\int_{X\times Y} \big(f(x)g(y)\big)^{-z/2}|\varphi_x||\varphi_y|\,dx\,dy \\ &= \big(\textstyle\int_X f(x)^{-z/2}|\varphi_x|\,dx\big)\big(\textstyle\int_Y g(y)^{-z/2}|\varphi_y|\,dy\big) = \zeta_x(z)\zeta_y(z). \end{aligned}$$

The second result then follows from the relationship between the poles. □

Our last property tells us the behavior of RLCTs under a change of variables. Consider an ideal $I \subset \mathcal{A}_W$ where W is a neighborhood of the origin. Let M be a real analytic manifold and $\rho : M \to W$ a proper real analytic map. Then, the *pullback* $\rho^* I = \{f \circ \rho : f \in I\}$ is an ideal of real analytic functions on M. If ρ is an isomorphism between $M \setminus \mathcal{V}(\rho^* I)$ and $W \setminus \mathcal{V}(I)$, we say that ρ is a *change of variables away from* $\mathcal{V}(I)$. Let $|\rho'|$ denote the absolute value of the Jacobian determinant of ρ. We call $(\rho^* I; (\varphi \circ \rho)|\rho'|)$ the *pullback pair*.

Proposition 4.6 (Chain Rule)**.** *Let W be a neighborhood of the origin and $I \subset \mathcal{A}_W$ a finitely generated ideal. If M is a real analytic manifold, $\rho : M \to W$ is a change of variables away from $\mathcal{V}(I)$ and $\mathcal{M} = \rho^{-1}(\Omega \cap W)$, then*

$$\mathrm{RLCT}_{\Omega_0}(I; \varphi) = \min_{x \in \rho^{-1}(0)} \mathrm{RLCT}_{\mathcal{M}_x}(\rho^* I; (\varphi \circ \rho)|\rho'|).$$

Proof. Let $f_1, \ldots, f_r$ generate I and define $f = f_1^2 + \cdots + f_r^2$. Then, $\mathrm{RLCT}_{\Omega_0}(I; \varphi)$ is the smallest pole and multiplicity of the zeta function

$$\zeta(z) = \int_{\Omega_0} f(\omega)^{-z/2} |\varphi(\omega)| \, d\omega$$

where $\Omega_0 \subset W$ is a sufficiently small neighborhood of the origin in Ω. Applying the change of variables ρ, we have

$$\zeta(z) = \int_{\rho^{-1}(\Omega_0)} f \circ \rho(\mu)^{-z/2} |\varphi \circ \rho(\mu)| |\rho'(\mu)| \, d\mu.$$

The proof of Lemma 3.8 shows that if Ω_0 is sufficiently small, there are finitely many points $y \in \rho^{-1}(0)$ and a cover $\{\mathcal{M}_y\}$ of $\mathcal{M} = \rho^{-1}(\Omega_0)$ such that

$$\zeta(z) = \sum_y \int_{\mathcal{M}_y} f \circ \rho(\mu)^{-z/2} |\varphi \circ \rho(\mu)| |\rho'(\mu)| \sigma_y(\mu) \, d\mu$$

where $\{\sigma_y\}$ is a partition of unity subordinate to $\{\mathcal{M}_y\}$. Furthermore, the $f_i \circ \rho$ generate the pullback $\rho^* I$ and $f \circ \rho = (f_1 \circ \rho)^2 + \cdots + (f_r \circ \rho)^2$. Therefore,

$$\mathrm{RLCT}_{\mathcal{M}_y}(f \circ \rho; (\varphi \circ \rho)|\rho'|\sigma_y) = \mathrm{RLCT}_{\mathcal{M}_y}(\rho^* I; (\varphi \circ \rho)|\rho'|)$$

and the result follows from the two previously displayed equations. □

4.2 Newton Polyhedra

Newton polyhedra methods are useful for computing the RLCT of a function f at a point x which is in the interior of the parameter space Ω. By applying a translation, we may assume without loss of generality that x is the origin $0 \in \mathbb{R}^d$ and that f is analytic at this origin.

4.2.1 Nondegeneracy

Given an analytic function $f \in \mathcal{A}_0(\mathbb{R}^d)$, we pick local coordinates $\{w_1, \ldots, w_d\}$ in a neighborhood of the origin. This allows us to represent f as a convergent power series $\sum_\alpha c_\alpha \omega^\alpha$ where $\omega = (\omega_1, \ldots, \omega_d)$ and each $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$. Let $[\omega^\alpha] f$ denote the coefficient c_α of ω^α in this expansion. Define its *Newton polyhedron* $\mathcal{P}(f) \subset \mathbb{R}^d$ to be the convex hull

$$\mathcal{P}(f) = \mathrm{conv}\,\{\alpha + \alpha' : [\omega^\alpha] f \neq 0, \alpha' \in \mathbb{R}^d_{\geq 0}\}.$$

A subset $\gamma \subset \mathcal{P}(f)$ is a *face* if there exists $\beta \in \mathbb{R}^d$ such that

$$\gamma = \{\alpha \in \mathcal{P}(f) : \langle \alpha, \beta \rangle \leq \langle \alpha', \beta \rangle \text{ for all } \alpha' \in \mathcal{P}(f)\}.$$

where $\langle\ ,\ \rangle$ is the standard dot product. Dually, the *normal cone* at γ is the set of all vectors $\beta \in \mathbb{R}^d$ satisfying the above condition. Each β lies in the non-negative orthant $\mathbb{R}^d_{\geq 0}$ because otherwise, the linear function $\langle \cdot\, , \beta \rangle$ does not have a minimum over the unbounded set $\mathcal{P}(f)$. As a result, the union of all the normal cones gives a partition $\mathcal{F}(f)$ of the non-negative orthant called the *normal fan*. Given a *compact* subset $\gamma \subset \mathbb{R}^d$, define the *face polynomial*

$$f_\gamma = \sum_{\alpha \in \gamma} c_\alpha \omega^\alpha.$$

Recall that f_γ is singular at a point $x \in \mathbb{R}^d$ if $\operatorname{ord}_x f \geq 2$, i.e.

$$f_\gamma(x) = \frac{\partial f_\gamma}{\partial \omega_1}(x) = \cdots = \frac{\partial f_\gamma}{\partial \omega_d}(x) = 0.$$

We say that f is *nondegenerate* if f_γ is non-singular at all points in the torus $(\mathbb{R}^*)^d$ for all compact faces γ of $\mathcal{P}(f)$, otherwise we say f is *degenerate*. Now, we define the *distance* l of $\mathcal{P}(f)$ to be the smallest $t \geq 0$ such that $(t, t, \ldots, t) \in \mathcal{P}(f)$. The *multiplicity* θ of l is the codimension of the face of $\mathcal{P}(f)$ at this intersection of the diagonal with $\mathcal{P}(f)$. However, if $l = 0$, we leave θ undefined. These notions of nondegeneracy, distance and multiplicity were first coined and studied by Varchenko [54].

We now extend the above notions to ideals. For any ideal $I \subset \mathcal{A}_0$, define

$$\mathcal{P}(I) = \operatorname{conv}\{\alpha \in \mathbb{R}^d : [\omega^\alpha] f \neq 0 \text{ for some } f \in I\}.$$

Related to this geometric construction is the monomial ideal

$$\operatorname{mon}(I) = \langle \omega^\beta : \textstyle\sum_\alpha c_\alpha \omega^\alpha \in I, c_\beta \neq 0 \rangle.$$

Note that I and $\operatorname{mon}(I)$ have the same Newton polyhedron, and if I is generated by $f_1, \ldots, f_r$, then $\operatorname{mon}(I)$ is generated by the monomials ω^α appearing in the f_i. One consequence is that $\mathcal{P}(f_1^2 + \cdots + f_r^2)$ is the scaled polyhedron $2\mathcal{P}(I)$. More importantly, the threshold of I is bounded above by that of $\operatorname{mon}(I)$. To prove this result, we need the following lemma. Recall that by the Hilbert Basis Theorem or by Dickson's Lemma [17], $\operatorname{mon}(I)$ is finitely generated.

Lemma 4.7. *Given $f \in \mathcal{A}_0(\mathbb{R}^d)$, let S be a finite set of exponents α of monomials ω^α which generate* $\operatorname{mon}(\langle f \rangle)$. *Then, there is a positive constant c such that*

$$|f(\omega)| \leq c \sum_{\alpha \in S} |\omega|^\alpha$$

in a sufficiently small neighborhood of the origin.

Proof. Let $\sum_\alpha c_\alpha \omega^\alpha$ be the power series expansion of f. Because f is analytic at the origin, there exists $\varepsilon > 0$ such that

$$\sum_\alpha |c_\alpha|\, \varepsilon^{\alpha_1 + \cdots + \alpha_d} < \infty.$$

Now, let $S = \{\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(s)}\}$. Since the monomials $\omega^{\alpha^{(i)}}$ generate $\mathrm{mon}(I)$, we can write

$$f(\omega) = \omega^{\alpha^{(1)}} g_1(\omega) + \cdots + \omega^{\alpha^{(s)}} g_s(\omega)$$

for some power series $g_i(\omega)$. Each $g_i(\omega)$ is absolutely convergent in the ε-neighborhood of the origin because f is absolutely convergent in that neighborhood. Thus, the $g_i(\omega)$ are analytic. Their absolute values are bounded above by some constant c in a small neighborhood of the origin, and the lemma follows. □

Proposition 4.8. *Let $I \subset \mathcal{A}_0$ be a finitely generated ideal and $\varphi : \mathbb{R}^d \to \mathbb{R}$ be nearly analytic at the origin. Then,*

$$\mathrm{RLCT}_0(I; \varphi) \leq \mathrm{RLCT}_0(\mathrm{mon}(I); \varphi).$$

Proof. Suppose $f \in \mathcal{A}_0(\mathbb{R}^d)$ and S is a finite set of generating exponents α for $\mathrm{mon}(\langle f \rangle)$. By Lemma 4.7 and the Cauchy-Schwarz inequality, there exist constants $c, c' > 0$ such that

$$f^2 \leq \Big(c \sum_{\alpha \in S} |\omega|^{\alpha} \Big)^2 \leq c' \sum_{\alpha \in S} \omega^{2\alpha}$$

in a sufficiently small neighborhood of the origin. More generally, if $f_1, \ldots, f_r$ generate I, then $f_1^2 + \ldots + f_r^2$ is bounded by a constant multiple of the sum of squares of monomials generating $\mathrm{mon}(I)$. The result now follows from Propostion 3.11. □

Given a compact subset $\gamma \subset \mathbb{R}^d$, define the *face ideal*

$$I_\gamma = \langle\, f_\gamma : f \in I \,\rangle.$$

The next result shows that we can compute I_γ directly from generators $f_1, \ldots, f_r$ for I.

Proposition 4.9. *If $I = \langle f_1, \ldots, f_r \rangle$, then $I_\gamma = \langle f_{1\gamma}, \ldots, f_{r\gamma} \rangle$ for all compact faces $\gamma \in \mathcal{P}(I)$.*

Proof. By definition, $\langle f_{1\gamma}, \ldots, f_{r\gamma} \rangle \subset I_\gamma$. For the other inclusion, it is enough to show that $f_\gamma \in \langle f_{1\gamma}, \ldots, f_{r\gamma} \rangle$ for all $f \in I$. First, we claim that if $\omega^\alpha = \omega^{\alpha'} \omega^{\alpha''}$ with $\alpha \in \gamma$ and $\omega^{\alpha'} \in \mathrm{mon}(I)$, then $\omega^{\alpha''} = 1$. Indeed, for all $\beta \in \mathbb{R}^d_{\geq 0}$ normal to γ, we have $\langle \alpha, \beta \rangle = \langle \alpha', \beta \rangle + \langle \alpha'', \beta \rangle$, but $\langle \alpha, \beta \rangle \leq \langle \alpha', \beta \rangle$ so $\langle \alpha'', \beta \rangle = 0$. This implies that $\alpha' + k\alpha'' \in \gamma$ for all integers $k > 0$. Since γ is compact, α'' must be the zero vector so $\omega^{\alpha''} = 1$.

Now, if $f \in I$, then $f = h_1 f_1 + \cdots + h_r f_r$ for some analytic functions $h_1, \ldots, h_r$. Clearly, $f_\gamma = (h_1 f_1)_\gamma + \cdots + (h_r f_r)_\gamma$. By the above claim, $(h_i f_i)_\gamma = h_{i0} f_{i\gamma}$ where h_{i0} is the constant term in h_i. Hence, $f_\gamma = h_{10} f_{1\gamma} + \cdots + h_{r0} f_{r\gamma} \in \langle f_{1\gamma}, \ldots, f_{r\gamma} \rangle$ as required. □

Remark 4.10. We now explain why we do not run into Gröbner-basis issues in this proposition. Let β be a vector in the normal cone at the face γ of $\mathcal{P}(I)$. Now, consider the weight order associated to β, and let $\mathrm{in}_\beta f$ be the sum of all the terms of f that are maximal with respect to this order [17, §15]. Let $\mathrm{in}_\beta I$ be the initial ideal

$$\mathrm{in}_\beta I = \langle\, \mathrm{in}_\beta f : f \in I \,\rangle.$$

Then, by definition, a set of functions $f_1, \ldots, f_r \in I$ is a Gröbner basis for I if and only if the initial ideal $\text{in}_\beta I$ is generated by the $\text{in}_\beta f_i$. Not all generating sets are Gröbner bases. But in our case, the face ideal I_γ is not the initial ideal $\text{in}_\beta I$. In fact, the face polynomial f_γ is not the initial form $\text{in}_\beta f$. For instance, suppose $I = \langle x, y \rangle$, $\beta = (1,1) \in \mathbb{R}^2$, and γ is the face of $\mathcal{P}(I)$ normal to β. If $f = x^2 + y^2 \in I$, then $\text{in}_\beta f = x^2 + y^2$ but $f_\gamma = 0$. □

Lastly, we give several equivalent definitions of *sos-nondegeneracy* for ideals I, where *sos* stands for *sum-of-squares*.

Proposition 4.11. *Let $I \subset \mathcal{A}_0$ be an ideal. The following statements are equivalent:*

1. *For some generating set $\{f_1, \ldots, f_r\}$ for I, $f_1^2 + \cdots + f_r^2$ is nondegenerate.*
2. *For all generating sets $\{f_1, \ldots, f_r\}$ for I, $f_1^2 + \cdots + f_r^2$ is nondegenerate.*
3. *For all compact faces $\gamma \subset \mathcal{P}(I)$, the variety $\mathcal{V}(I_\gamma)$ does not intersect the torus $(\mathbb{R}^*)^d$.*

If the ideal I satisfies any of these conditions, then we say that I is sos-nondegenerate.

Proof. Let $f_1, \ldots, f_r$ generate I and let $f = f_1^2 + \cdots + f_r^2$. If γ is a compact face of $\mathcal{P}(I)$, then (2γ) is a compact face of $\mathcal{P}(f) = 2\mathcal{P}(I)$. Furthermore, $f_{(2\gamma)} = f_{1\gamma}^2 + \cdots + f_{r\gamma}^2$ and

$$\frac{\partial f_{(2\gamma)}}{\partial \omega_i} = 2f_{1\gamma}\frac{\partial f_{1\gamma}}{\partial \omega_i} + \cdots + 2f_{r\gamma}\frac{\partial f_{r\gamma}}{\partial \omega_i}.$$

Now, $f_{1\gamma}^2 + \cdots + f_{r\gamma}^2 = 0$ if and only if $f_{1\gamma} = \cdots = f_{r\gamma} = 0$. It follows that f is nondegenerate if and only if $\mathcal{V}(\langle f_{1\gamma}, \ldots, f_{r\gamma} \rangle) \cap (\mathbb{R}^*)^d = \mathcal{V}(I_\gamma) \cap (\mathbb{R}^*)^d = \emptyset$ for all compact faces $\gamma \subset \mathcal{P}(I)$. This proves the equivalences (1) ⇔ (3) and (2) ⇔ (3). □

A `Singular` library which implements some of the algorithms discussed in this section is made available at the following website:

`http://math.berkeley.edu/~shaowei/rlct.html`

The library provides functions for determining the nondegeneracy of functions and of ideals, and computes the RLCT of monomial ideals.

Remark 4.12. After finishing this chapter, the author discovered another notion of nondegeneracy for ideals of *complex* formal power series due to Saia [46]. An ideal I is said to be *Newton nondegenerate* if there exists a generating set $\{f_1, \ldots, f_r\}$ of I such that for every compact face γ of $\mathcal{P}(I)$, the ideal of $\mathcal{A}_\gamma$ generated by $f_{1\gamma}, \ldots, f_{r\gamma}$ has finite colength in $\mathcal{A}_\gamma$. Here, $\mathcal{A}_\gamma$ is the ring $\mathcal{A}_0/J_\gamma$ where J_γ is the monomial ideal generated by all monomials ω^α such that α is not in the cone over γ. As mentioned by Bivià-Ausina [7, §2], this notion of nondegeneracy is equivalent to saying that for this generating set and for every compact face γ, the common complex zeros of $f_{1\gamma}, \ldots, f_{r\gamma}$ is contained in the coordinate hyperplanes $\{\omega \in \mathbb{R}^n : \omega_1 \cdots \omega_d = 0\}$. In fact, if I is Newton nondegenerate, then this condition is true for all generating sets of I. Thus, their notion of Newton nondegeneracy is in some sense the complex version of our notion of sos-nondegeneracy. □

4.2.2 Toric Resolutions

We recall some basic facts about toric varieties. We say a polyhedral cone σ is generated by vectors $v_1, \ldots, v_k \in \mathbb{R}^d$ if $\sigma = \{\sum_i \lambda_i v_i : \lambda_i \geq 0\}$. If σ is generated by lattice vectors $v_i \in \mathbb{Z}^d$, then σ is *rational*. If the origin is a face of σ, then σ is *pointed*. A *ray* is a pointed one-dimensional cone. Every rational ray has a lattice generator of minimal length called the *minimal generator*. Similarly, every pointed rational polyhedral cone σ is generated by the minimal generators of its edges. If these minimal generators are linearly independent over $\mathbb{R}$, then σ is *simplicial*. A simplicial cone is *smooth* if its minimal generators also form part of a $\mathbb{Z}$-basis of $\mathbb{Z}^d$. A collection $\mathcal{F}$ of pointed rational polyhedral cones in $\mathbb{R}^d$ is a *fan* if the faces of every cone in $\mathcal{F}$ are in $\mathcal{F}$ and the intersection of any two cones in $\mathcal{F}$ are again in $\mathcal{F}$. The *support* of $\mathcal{F}$ is the union of its cones as subsets of $\mathbb{R}^d$. If the support of $\mathcal{F}$ is the non-negative orthant, then $\mathcal{F}$ is *locally complete*. If every cone of $\mathcal{F}$ is simplicial (resp. smooth), then $\mathcal{F}$ is *simplicial* (resp. *smooth*). A fan $\mathcal{F}_1$ is a *refinement* of another fan $\mathcal{F}_2$ if the cones of $\mathcal{F}_1$ come from partitioning the cones of $\mathcal{F}_2$. See [21, 52] for more details.

Given a smooth locally complete fan $\mathcal{F}$, we have a smooth toric variety $\mathbb{P}(\mathcal{F})$ covered by open affines $U_\sigma \simeq \mathbb{R}^d$, one for each maximal cone σ of $\mathcal{F}$. Furthermore, we have a *blowup* map $\rho_{\mathcal{F}} : \mathbb{P}(\mathcal{F}) \to \mathbb{R}^d$ defined as follows: for each maximal cone σ of $\mathcal{F}$ minimally generated by $v_1, \ldots, v_d$ with $v_i = (v_{i1}, \ldots, v_{id})$, we have a monomial map $\rho_\sigma : U_\sigma \to \mathbb{R}^d$,

$$\begin{aligned}(\mu_1, \ldots, \mu_d) &\mapsto (\omega_1, \ldots, \omega_d).\\ \omega_1 &= \mu_1^{v_{11}} \mu_2^{v_{21}} \cdots \mu_d^{v_{d1}}\\ \omega_2 &= \mu_1^{v_{12}} \mu_2^{v_{22}} \cdots \mu_d^{v_{d2}}\\ &\vdots\\ \omega_d &= \mu_1^{v_{1d}} \mu_2^{v_{2d}} \cdots \mu_d^{v_{dd}}\end{aligned}$$

Let $v = v_\sigma$ be the matrix (v_{ij}) where each vector v_i forms a row of v. We represent the above monomial map by $\omega = \mu^v$. If v_{i+} represents the i-th row sum of v, the Jacobian determinant of this map is given by

$$(\det v)\mu_1^{v_{1+}-1} \cdots \mu_d^{v_{d+}-1}.$$

We are now ready to connect these concepts. The next two theorems are from Varchenko, see [54] and [3, §8.3]. His notion of degeneracy is weaker than ours because it does not include the condition $f_\gamma = 0$, but his proof [3, Lemma 8.9] actually supports the stronger notion. The set up is as follows: suppose f is analytic in a neighborhood W of the origin. Let $\mathcal{F}$ be any smooth refinement of the normal fan $\mathcal{F}(f)$ and $\rho_{\mathcal{F}}$ be the blowup associated to $\mathcal{F}$. Set $M = \rho_{\mathcal{F}}^{-1}(W)$. Let l be the distance of $\mathcal{P}(f)$ and θ its multiplicity.

Theorem 4.13. *If f is nondegenerate, then $(M, W, \rho_{\mathcal{F}})$ desingularizes f at 0.*

Theorem 4.14. *Suppose $(M, W, \rho_{\mathcal{F}})$ desingularizes f at 0. If f has a maximum or minimum at 0, then* $\mathrm{RLCT}_0\, f = (1/l, \theta)$.

We extend Theorem 4.14 to compute $\mathrm{RLCT}_0(f;\omega^\tau)$ for monomials ω^τ. Given a polyhedron $\mathcal{P}(f) \subset \mathbb{R}^d$ and a vector $\tau = (\tau_1, \ldots, \tau_d)$ of non-negative integers, let the *τ-distance* l_τ be the smallest $t \geq 0$ such that $t(\tau_1+1, \ldots, \tau_d+1) \in \mathcal{P}(f)$ and let the multiplicity θ_τ be the codimension of the face at this intersection.

Theorem 4.15. *Suppose $(M, W, \rho_\mathcal{F})$ desingularizes f at 0. If f has a maximum or minimum at 0, then* $\mathrm{RLCT}_0(f;\omega^\tau) = (1/l_\tau, \theta_\tau)$.

Proof. We follow roughly the proof in [3, §8] of Theorem 4.14. Let σ be a maximal cone of $\mathcal{F}$. Because $\mathcal{F}$ refines $\mathcal{F}(f)$, σ is a subset of some maximal cone σ' of $\mathcal{F}(f)$. Let $\alpha \in \mathbb{R}^d$ be the vertex of $\mathcal{P}(f)$ dual to σ'. Let v be the matrix whose rows are minimal generators of σ and ρ the monomial map $\mu \mapsto \mu^v$. Then,

$$\begin{aligned} f(\omega)^{-z}|\omega^\tau|\,d\omega &= f(\rho(\mu))^{-z}|\rho(\mu)^\tau||\rho'(\mu)|\,d\mu \\ &= g(\mu)^{-z}|(\det v)\mu^{-v\alpha z}\mu^{v\tau}\mu_1^{v_{1+}-1}\cdots\mu_d^{v_{d+}-1}|d\mu \end{aligned}$$

for some function $g(\mu)$. Because f has a maximum or minimum at 0, this ensures $g(\mu) \neq 0$ on the affine chart U_σ. Thus, for the cone σ,

$$(\lambda_\sigma, \theta_\sigma) = (\min S, \#\min S), \quad S = \left\{ \frac{\langle v_i, \tau+1\rangle}{\langle v_i, \alpha\rangle} : 1 \leq i \leq d \right\}$$

where $\tau+1 = (\tau_1+1, \ldots, \tau_d+1)$. We now give an interpretation for the elements of S. Fixing i, let P be the affine hyperplane normal to v_i passing through α. Then, $\langle v_i, \alpha\rangle / \langle v_i, \tau+1\rangle$ is the distance of P from the origin along the ray $\{t(\tau+1) : t \geq 0\}$. Since $\mathrm{RLCT}_0(f;\omega^\tau) = \min_\sigma(\lambda_\sigma, \theta_\sigma)$, the result follows. □

Remark 4.16. After finishing this chapter, the author discovered that a result similar to the previous theorem was proved by Vasil'ev [55] for *complex* analytic functions. □

4.2.3 Monomial Ideals

Monomial ideals play in special role in the theory of real log canonical thresholds of ideals, just as monomial functions are important in the theory of RLCTs of functions. The statement and proof of this next result is due to Piotr Zwiernik.

Proposition 4.17. *Monomial ideals are sos-nondegenerate.*

Proof. Let $f = f_1^2 + \cdots + f_r^2$ where $f_1, \ldots, f_r$ are monomials generating I. For each face γ of $\mathcal{P}(I)$, f_γ is also a sum of squares of monomials, so f_γ does not have any zeros in $(\mathbb{R}^*)^d$ and the result now follows from Proposition 4.11(3). □

We now come to the main theorem of this chapter. As a special case, we have a formula for the RLCT of a monomial ideal with respect to a monomial amplitude function. The analogous formula for *complex* log canonical thresholds of monomial ideals was discovered and proved by Howald [31].

Theorem 4.18. *If $I \subset \mathcal{A}_0$ is a finitely generated ideal, then*

$$\mathrm{RLCT}_0(I;\omega^\tau) \leq (1/l_\tau, \theta_\tau)$$

where l_τ is the τ-distance of $\mathcal{P}(I)$ and θ_τ its multiplicity. Equality occurs when I is monomial or, more generally, sos-nondegenerate.

Proof. If I is sos-nondegenerate, the equality follows from Theorem 4.15. For all other ideals, the inequality follows from Proposition 4.8 and Corollary 4.17. □

Remark 4.19. Define the principal part $f_\mathcal{P}$ of f to be $\sum_\alpha c_\alpha \omega^\alpha$ where the sum is over all α lying in some compact face of $\mathcal{P}(f)$. The above theorems imply that if f is nondegenerate, then $\mathrm{RLCT}_0 f = \mathrm{RLCT}_0 f_\mathcal{P}$. However, this equality is not true in general. For instance, if $f = (x+y)^2 + y^4$, then $f_\mathcal{P} = (x+y)^2$ but $\mathrm{RLCT}_0 f = (3/4, 1)$ and $\mathrm{RLCT}_0 f_\mathcal{P} = (1/2, 1)$. □

Our first corollary shows that the asymptotic correctness of the BIC is a special case of Watanabe's Theorem 1.3. Recall that for regular models where the true distribution is given by the parameter $\omega = 0$, the Kullback-Leibler function $K(\omega)$ satisfies $K(0) = 0$, $\nabla K(0) = 0$ and $\nabla^2 K(0) \succ 0$. For these models, the BIC states that the learning coefficient (λ, θ) equals $(d/2, 1)$, while Watanabe's Theorem states that (λ, θ) equals $\mathrm{RLCT}_0 K$.

Corollary 4.20. *If $K \in \mathcal{A}_0(\mathbb{R}^d)$ is such that $K(0) = 0$, $\nabla K(0) = 0$ and $\nabla^2 K(0) \succ 0$, then* $\mathrm{RLCT}_0 K = (d/2, 1)$.

Proof. Because its Hessian is full rank, there is a linear change of variables such that $K = \omega_1^2 + \cdots + \omega_d^2 + O(\omega^3)$. Hence, K is nondegenerate and the Newton polyhedron $\mathcal{P}(K)$ has distance $l = 2/d$ with $\theta = 1$. □

Corollary 4.21. *Let I be generated by $f_1, \ldots, f_s$ and suppose the Jacobian matrix $(\partial f_i / \partial \omega_j)$ has rank r at 0. Then,* $\mathrm{RLCT}_0 I \leq (\frac{1}{2}(r+d), 1)$.

Proof. Because the rank of $(\partial f_i / \partial \omega_j)$ is r, there is a linear change of variables such that the only linear monomials appearing in I are $\omega_1, \ldots, \omega_r$. It follows that $\mathcal{P}(I)$ lies in the halfspace $\alpha_1 + \cdots + \alpha_r + \frac{1}{2}(\alpha_{r+1} + \cdots + \alpha_d) \geq 1$ and its distance is at least $1/(r + \frac{d-r}{2}) = 2/(r+d)$. □

We saw in Propositions 4.11 and 4.17 that if $f_1, \ldots, f_r$ generate a monomial ideal, then the sum of squares $f_1^2 + \cdots + f_r^2$ is nondegenerate. As an extension to Proposition 4.4, we now show that $K(f_1, \ldots, f_r)$ is also nondegenerate for any function K such that $K(0) = 0, \nabla K(0) = 0$ and $\nabla^2 K(0) \succ 0$.

Proposition 4.22. *Let Ω and U be neighborhoods of the origin $0 \in \mathbb{R}^d$, and let $u : \Omega \to U$ and $K : U \to \mathbb{R}$ be real analytic maps satisfying $u(0) = 0$, $K(0) = 0$, $\nabla K(0) = 0$ and $\nabla^2 K(0) \succ 0$. If the ideal $I = \langle u_1(\omega), \ldots, u_d(\omega) \rangle$ is monomial, then the function $K \circ u(\omega)$ is nondegenerate at the origin.*

Proof. Because $K(0) = 0, \nabla K(0) = 0$ and $\nabla^2 K(0) \succ 0$, there is a linear change of coordinates $T : \mathbb{R}^d \to \mathbb{R}^d$ such that the power series expansion of K is $v_1^2 + \cdots + v_d^2 + O(v^3)$ where $(v_1, \ldots, v_d) = T(u_1, \ldots, u_d)$. Moreover, $v_1(\omega), \ldots, v_d(\omega)$ generate the same monomial ideal I as $u_1(\omega), \ldots, u_d(\omega)$. By Propositions 4.17 and 4.11(2), we see that $v_1(\omega)^2 + \cdots + v_d(\omega)^2$ is nondegenerate. To show that

$$K \circ u(\omega) = v_1(\omega)^2 + \cdots + v_d(\omega)^2 + O(v(\omega)^3)$$

is nondegenerate, it suffices to show that $K \circ u(\omega)$ and $v_1(\omega)^2 + \cdots + v_d(\omega)^2$ have the same *principal part* (see Remark 4.19 for a definition). We claim that the monomials appearing in the term $O(v(\omega)^3)$ cannot lie on any compact face of the Newton polyhedron $\mathcal{P}(K \circ u(\omega)) = \mathcal{P}(I^2)$. Now, any monomial in $O(v(\omega)^3)$ can be written as a product $\omega^\alpha \omega^{\alpha'} \omega^{\alpha''}$ of monomials in I. Suppose $\alpha + \alpha' + \alpha''$ lie in some compact face γ of $\mathcal{P}(I^2)$. Then, by the claim in the proof of Proposition 4.9, since $\omega^{\alpha+\alpha'} \in I^2$, we must have $\alpha'' = 0$, a contradiction. □

4.3 Applications to Statistical Models

In this section, we use our tools to compute learning coefficients of the discrete model $\mathcal{M}$ in Example 2.20. Recall that $\mathcal{M}$ is a naïve Bayesian network with two ternary random variables and two hidden states. It was designed by Evans, Gilula and Guttman [18] for investigating connections between the recovery time of 132 schizophrenic patients and the frequency of visits by their relatives. Their data is summarized in the 3×3 contingency table (2.21), which gives us the matrix

$$\hat{q} \;=\; \frac{1}{132} \begin{pmatrix} 43 & 16 & 3 \\ 6 & 11 & 10 \\ 9 & 18 & 16 \end{pmatrix} \tag{4.2}$$

of relative frequencies. The model is parametrized by the map

$$\begin{aligned} p: \quad \Omega = \Delta_1 \times \Delta_2 \times \Delta_2 \times \Delta_2 \times \Delta_2 \quad &\to \quad \Delta_8 \\ \omega = (t, a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2) \quad &\mapsto \quad (p_{ij}) \\ p_{ij} = t a_i b_j + (1-t) c_i d_j, \quad i, j \in \{1, 2, 3\} & \end{aligned}$$

where $a_3 = 1 - a_1 - a_2$, $a = (a_1, a_2, a_3) \in \Delta_2$ and similarly for b, c and d. Hence, a 3×3 matrix in the model is a convex combination of two rank one matrices, so it has rank at most two. In Example 2.20, the marginal likelihood integral

$$\mathcal{I} = \int_\Omega p_{11}^{43} \, p_{12}^{16} \, p_{13}^{3} \, p_{21}^{6} \, p_{22}^{11} \, p_{23}^{10} \, p_{31}^{9} \, p_{32}^{18} \, p_{33}^{16} \; d\omega$$

of the above data set was computed exactly.

We now estimate this integral using the asymptotic method described in Theorem 1.17. More specifically, we approximate $\mathcal{I}$ by considering the asymptotics of the function

$$L(N) = \int_\Omega \prod_{i,j} p_{ij}(\omega)^{-Nq_{ij}}\, d\omega$$

where the 3×3 matrix $q = (q_{ij})$ is some distribution lying in the model. Ideally, we want q to be the matrix $\hat{q}$ of relative frequencies coming from the data, but this matrix rarely lies in the model. However, we should be able to find a matrix in the model that is close to $\hat{q}$. This is a reasonable assumption because in practice, we want to study models which describe the data well. A good candidate for q is the maximum likelihood distribution. For instance, the matrix (4.2) of relative frequencies is not in the model because it is full rank. Using the EM algorithm, we compute the maximum likelihood distribution

$$q = \frac{1}{132}\begin{pmatrix} 43.00153927 & 15.99813189 & 3.000328847 \\ 5.979732739 & 11.12298188 & 9.897285383 \\ 9.018728012 & 17.87888620 & 16.10238577 \end{pmatrix}$$

which comes from the maximum likelihood estimate

$$\begin{aligned} t &= 0.5129202328 \\ (a_1, a_2) &= (0.09139459898, 0.3457903589), \\ (b_1, b_2) &= (0.1397061214, 0.4386217768), \\ (c_1, c_2) &= (0.8680689680, 0.05580725171), \\ (d_1, d_2) &= (0.7549807403, 0.2380125694). \end{aligned}$$

Observe that q is indeed very close to $\hat{q}$. In Remark 1.19, using another discrete model, we discuss how approximation of the likelihood integral varies with the choice of q.

Our next result states how the asymptotics of $L(N)$ depends on q. Let S_i denote the set of rank i matrices in $p(\Omega)$. Let S_{21} be the set of matrices in S_2 where there are permutations of the rows and of the columns such that $q_{11} = 0$ and q_{12}, q_{21}, q_{22} are all non-zero. Let S_{22} be the subset of S_2 where, up to permutations, $q_{11} = q_{22} = 0$ and q_{12}, q_{21} are non-zero. Before we prove this theorem, let us apply it to our statistical problem. Using the exact value of $\mathcal{I}$ from Example 2.20, we have

$$\log \mathcal{I} \;=\; -273.1911759.$$

Meanwhile, applying Theorem 4.23 with $q \in S_2 \setminus (S_{12} \cup S_{22})$, we obtain the approximation

$$\log L(132) \;\approx\; -275.9144024. \tag{4.3}$$

This estimate could be improved if we also computed the constant C in the asymptotics of $\log L(N)$ using techniques from Chapter 5. On the other hand, if the BIC was erroneously applied with the model dimension $d = 8$, we would get

$$BIC \;=\; -278.3558034.$$

Clearly, the approximation (4.3) is closer than the BIC to the exact value of $\log \mathcal{I}$.

Theorem 4.23. *The learning coefficient (λ, θ) of the model at q is given by*

$$(\lambda, \theta) = \begin{cases} (5/2, 1) & \text{if } q \in S_1, \\ (7/2, 1) & \text{if } q \in S_2 \setminus (S_{21} \cup S_{22}), \\ (4, 1) & \text{if } q \in S_{21} \setminus S_{22}, \\ (9/2, 1) & \text{if } q \in S_{22}. \end{cases}$$

Therefore, asymptotically as $N \to \infty$,

$$\log L(N) \;\approx\; N \sum_{i,j} q_{ij} \log q_{ij} - \lambda \log N + (\theta - 1) \log \log N + C$$

for some constant C.

We postpone the proof of this theorem to the end of this section. Let us begin with a few remarks about our approach to this problem. Firstly, Theorem 1.11 states that the learning coefficient (λ, θ) of the statistical model is given by

$$(2\lambda, \theta) = \min_{\omega^* \in \mathcal{V}} \mathrm{RLCT}_{\Omega_{\omega^*}} \langle p(\omega) - q \rangle$$

where $\mathcal{V}$ is the fiber $\{\omega \in \Omega : p(\omega) = q\}$ of the map p over q. Instead of restricting ourselves to a fixed q and its fiber $\mathcal{V}$, let us vary ω^* over all of Ω. For each $\omega^* \in \Omega$, we translate Ω so that ω^* is the origin and compute the RLCT of the fiber ideal $\langle p(\omega + \omega^*) - p(\omega^*) \rangle$. This is the content of Proposition 4.25. The proof of Theorem 4.23 will then consist of minimizing these RLCTs over the fiber $\mathcal{V}$ for each q in the model.

Secondly, in our computations, we will often be choosing different generators for our fiber ideal and making suitable changes of variables. Generators with few terms and small total degree will be highly desired. One useful trick is to multiply or divide the generators by functions $f(\omega)$ satisfying $f(0) \neq 0$. Such functions are units in the ring $\mathcal{A}_0$ of real analytic functions so this multiplication or division will not change the ideal generated. This next lemma also comes in handy in dealing with boundary issues.

Lemma 4.24. *Let $\Omega \subset \{(x_1, \ldots, x_d) \in \mathbb{R}^d\}$ be semianalytic. Let I be a monomial ideal and φ a monomial function in $x_1, \ldots, x_r$. If there exists a vector $\xi \in \mathbb{R}^{d-r}$ such that $\Omega_1 \times \Omega_2 \subset \Omega$ for sufficiently small ε where*

$$\begin{aligned} \Omega_1 &= \{(x_1, \ldots, x_r) \in [0, \varepsilon]^r\} \\ \Omega_2 &= \{(x_{r+1}, \ldots, x_d) = t(\xi + \xi') \text{ for all } t \in [0, \varepsilon], \xi' \in [-\varepsilon, \varepsilon]^{d-r}\}, \end{aligned}$$

then $\mathrm{RLCT}_{\Omega_0}(I; \varphi) = \mathrm{RLCT}_0(I; \varphi)$.

Proof. Because I and $|\varphi|$ remain unchanged by the flipping of signs of $x_1, \ldots, x_r$, the threshold of $(I;\varphi)$ does not depend on the choice of orthant, so $\mathrm{RLCT}_{\Omega_1}(I;\varphi) = \mathrm{RLCT}_0(I;\varphi)$. The lemma now follows from Proposition 4.5 and the fact that the threshold of the zero ideal over the cone neighborhood Ω_2 is $(\infty, -)$. □

We now come to the most computationally intensive result in this section. Let us define subsets $\Omega_u = \{\omega^* \in \Omega : t^* \in \{0,1\}\}$, $\Omega_m = \{\omega^* \in \Omega : t^* \notin \{0,1\}\}$ and

$$\begin{aligned}
\Omega_{m0} &= \{\omega^* \in \Omega_m : a^* = c^*, b^* = d^*\} \\
\Omega_{m0kl} &= \{\omega^* \in \Omega_{m0} : \#\{i : a_i^* = 0\} = k, \#\{i : b_i^* = 0\} = l\} \\
\Omega_{m1} &= \{\omega^* \in \Omega_m : (b^* \neq d^*, a^* = c^*) \text{ or } (a^* \neq c^*, b^* = d^*)\} \\
\Omega_{m10} &= \{\omega^* \in \Omega_{m1} : (a^* = c^*, \exists\, i\ a_i^* = 0) \text{ or } (b^* = d^*, \exists\, i\ b_i^* = 0)\} \\
\Omega_{m2} &= \{\omega^* \in \Omega_m : a^* \neq c^*, b^* \neq d^*\} \\
\Omega_{m2ad} &= \{\omega^* \in \Omega_{m2} : \exists\, i, j\ a_i^* = d_j^* = 0, c_i^* \neq 0, b_j^* \neq 0\} \\
\Omega_{m2bc} &= \{\omega^* \in \Omega_{m2} : \exists\, i, j\ b_i^* = c_j^* = 0, d_i^* \neq 0, a_j^* \neq 0\} \\
\Omega_{m21} &= \Omega_{m2ad} \cup \Omega_{m2bc} \\
\Omega_{m22} &= \Omega_{m2ad} \cap \Omega_{m2bc}.
\end{aligned}$$

Proposition 4.25. *Given $\omega^* \in \Omega$, let I be the ideal $\langle p(\omega + \omega^*) - p(\omega^*)\rangle$. Then,*

$$\mathrm{RLCT}_0\, I = \begin{cases}
(5,1) & \text{if } \omega^* \in \Omega_u, \\
(6,2) & \text{if } \omega^* \in \Omega_{m000}, \\
(6,1) & \text{if } \omega^* \in \Omega_{m010} \cup \Omega_{m001} \cup \Omega_{m020} \cup \Omega_{m002}, \\
(7,2) & \text{if } \omega^* \in \Omega_{m011}, \\
(7,1) & \text{if } \omega^* \in \Omega_{m012} \cup \Omega_{m021}, \\
(8,1) & \text{if } \omega^* \in \Omega_{m022}, \\
(6,1) & \text{if } \omega^* \in \Omega_{m1} \setminus \Omega_{m10}, \\
(7,1) & \text{if } \omega^* \in \Omega_{m10}, \\
(7,1) & \text{if } \omega^* \in \Omega_{m2} \setminus \Omega_{m21}, \\
(8,1) & \text{if } \omega^* \in \Omega_{m21} \setminus \Omega_{m22}, \\
(9,1) & \text{if } \omega^* \in \Omega_{m22}.
\end{cases}$$

Proof. The ideal I is generated by $g_{ij} = f_{ij}(\omega + \omega^*) - f_{ij}(\omega^*)$ where

$$f_{ij} = t a_i b_j + (1-t) c_i d_j, \quad i, j, k \in \{0, 1, 2\}$$

and $a_0 = b_0 = c_0 = d_0 = 1$. One can check that I is also generated by g_{10}, g_{20}, g_{01}, g_{02}, and

$g_{ij} - (d_j + d_j^*)g_{i0} - (a_i + a_i^*)g_{0j}, i, j \in \{1, 2\}$ which expand to give

$$\begin{gathered}
c_1(t_1^* - t) + a_1(t_0^* + t) + tu_1^* \\
c_2(t_1^* - t) + a_2(t_0^* + t) + tu_2^* \\
d_1(t_1^* - t) + b_1(t_0^* + t) + tv_1^* \\
d_2(t_1^* - t) + b_2(t_0^* + t) + tv_2^* \\
a_1d_1 - a_1t_0^*v_1^* + d_1t_1^*u_1^* \\
a_1d_2 - a_1t_0^*v_2^* + d_2t_1^*u_1^* \\
a_2d_1 - a_2t_0^*v_1^* + d_1t_1^*u_2^* \\
a_2d_2 - a_2t_0^*v_2^* + d_2t_1^*u_2^*
\end{gathered}$$

where $t_0^* = t^*, t_1^* = 1 - t^*, u_i^* = a_i^* - c_i^*, v_i^* = b_i^* - d_i^*$. Note that $\sum(a_i + a_i^*) = 1$ and $\sum a_i^* = 1$ so $\sum a_i = 0$ and similarly for b, c, d. Also, $\sum u_i^* = \sum a_i^* - c_i^* = 0$. The same is true for v^*. We now do a case-by-case analysis.

Case 1: $\omega^* \in \Omega_m$**.**

This implies $t_0^* \neq 0$ and $t_1^* \neq 0$. Since the indeterminates b_1, b_2, c_1, c_2 appear only in the first four polynomials, this suggests the change of variables

$$\begin{aligned}
c_i = &\quad (c_i' - tu_i^* - a_i(t_0^* + t))/(t_1^* - t), \quad i = 1, 2 \\
b_i = &\quad (b_i' - tv_i^* - d_i(t_1^* - t))/(t_0^* + t), \quad i = 1, 2
\end{aligned}$$

with new indeterminates $t, a_1, a_2, b_1', b_2', c_1', c_2', d_1, d_2$. In view of Proposition 4.6, the Jacobian determinant of this substitution is a constant.

Case 1.1: $\omega^* \in \Omega_{m1}$**.**

This implies $u^* \neq 0, v^* = 0$ or $u^* = 0, v^* \neq 0$. Without loss of generality, we assume $v^* = 0, u_1^* > 0$ and substitute

$$d_i = (d_i' + a_1t_0^*v_i^*)/(t_1^*u_1^* + a_1), \quad i = 1, 2.$$

The resulting pullback ideal is $\langle b_1', b_2', c_1', c_2', d_1', d_2' \rangle$. If ω^* lies in the interior of Ω, we use either Newton polyhedra or Proposition 4.5 to show that the RLCT of this monomial ideal is $(6, 1)$. If ω^* lies on the boundary of Ω, the situation is more complicated. Since we are considering a subset of a neighborhood of ω^*, the corresponding Laplace integral from Proposition 4.2a is smaller so the threshold is at least $(6, 1)$. To compute it exactly, we need blowups to separate the coordinate hyperplanes and the hypersurfaces defining the boundary.

Because $-u_1^* = u_2^* + u_3^*$, we cannot have $u_2^* = u_3^* = 0$. Suppose $u_2^* \neq 0$ and $u_3^* \neq 0$. We consider a blowup where one of the charts is given by the monomial map $t = s, a_i = sa_i', c_1' = rs, c_2' = rsc_2'', b_i' = rsb_i'', d_i' = rsd_i''$. Here, the pullback pair is $(\langle rs \rangle; r^5s^8)$. Now, we study the inequalities which are *active* at ω^*. For instance, if $b_1^* = 0$, then ω^* lies on the boundary defined by $0 \leq b_1 + b_1^*$. After the various changes of variables, the inequalities are as shown

below, where $b_3'' = -b_1'' - b_2''$ and similarly for c_3'', d_3'' and a_3'. Note that the inequality for $a_1^* = 0$ is omitted because $a_1^* = 0$ implies $u_1^* = -c_1^* \leq 0$. Similar conditions on the u_i^*, v_i^* hold for the other inequalities.

$$\begin{array}{lll}
b_i^* = 0: & 0 \leq rs(b_i'' - d_i''(t_1^* - s)/(t_1^* u_1^* + sa_1'))/(t_0^* + s) & \\
d_i^* = 0: & 0 \leq rsd_i''/(t_1^* u_1^* + sa_1') & \\
c_1^* = 0: & 0 \leq s(-u_1^* + a_1'(t_0^* + s) + r)/(t_1^* - s) & \\
c_2^* = 0: & 0 \leq s(-u_2^* + a_2'(t_0^* + s) + rc_2'')/(t_1^* - s) & u_2^* > 0 \\
c_3^* = 0: & 0 \leq s(-u_3^* + a_3'(t_0^* + s) - r - rc_2'')/(t_1^* - s) & u_3^* > 0 \\
a_2^* = 0: & 0 \leq sa_2' & u_2^* < 0 \\
a_3^* = 0: & 0 \leq sa_3' & u_3^* < 0
\end{array}$$

In applying Lemma 4.24, the choice of coordinates is important. For instance, if $b_2^* = b_3^* = 0$, we choose coordinates b_2'' and b_3'' and set $b_1'' = -b_2'' - b_3''$. The same is done for the d_i''. The pullback pair is unchanged by these choices. Now, with coordinates (r, s) and $(b_{i_1}'', b_{i_2}'', d_{j_1}'', d_{j_2}'', c_2'', a_2', a_3')$, we apply the lemma with the vector $\xi = (2, 2, u_1^*, u_1^*, 1, 1, 1)$, so the threshold is $\mathrm{RLCT}_0(rs; r^5 s^8) = (6, 1)$.

Now, if only one of u_2^*, u_3^* is zero, suppose $u_2^* = 0, u_3^* \neq 0$ without loss of generality. If $a_2^* = c_2^* \neq 0$, then the arguments of the previous paragraph show that the RLCT is again $(6, 1)$. If $a_2^* = c_2^* = 0$, we blow up the origin in $\mathbb{R}^7$ and consider the chart where $a_2 = s, c_i' = sc_i'', b_i' = sb_i'', d_i' = sd_i''$. The pullback pair is $(\langle sb_1'', sb_2'', sc_1'', sc_2'', sd_1'', sd_2'' \rangle; s^6)$. The active inequalities for $a_2^* = c_2^* = 0$ are

$$\begin{array}{ll}
c_2^* = 0: & 0 \leq s(c_2'' - t_0^* + t)/(t_1^* - t) \\
a_2^* = 0: & 0 \leq s.
\end{array}$$

Near the origin in $(s, b_1'', b_2'', c_1'', c_2'', d_1'', d_2'') \in \mathbb{R}^7$, these inequalities imply $s = 0$ so the new region $\mathcal{M}$ defined by the active inequalities is not full at the origin. Thus, we can ignore the origin in computing the RLCT. All other points on the exceptional divisor of this blowup lie on some other chart of the blowup where the pullback pair is $(s; s^6)$, so the RLCT is at least $(7, 1)$. In the chart where $c_2 = s, c_1 = sc_1'', a_2 = sa_2', b_i' = sb_i'', d_i' = sd_i''$, we have the active inequalities below. Note that $c_3^* \neq 0$ because $u_3^* = -u_1^* < 0$.

$$\begin{array}{ll}
b_i^* = 0: & 0 \leq s(b_i'' - d_i''(t_1^* - t)/(t_1^* u_1^* - (sa_2' + a_3))/(t_0^* + t) \\
d_i^* = 0: & 0 \leq sd_i''/(t_1^* u_1^* - (sa_2' + a_3)) \\
c_1^* = 0: & 0 \leq (sc_1'' - tu_1^* + (sa_2' + a_3)(t_0^* + t))/(t_1^* - t) \\
c_2^* = 0: & 0 \leq s(1 - a_2'(t_0^* + t))/(t_1^* - t) \\
a_2^* = 0: & 0 \leq sa_2' \\
a_3^* = 0: & 0 \leq a_3
\end{array}$$

Again, choosing suitable coordinates in the b_i'' and d_i'', we find that the RLCT is $(7, 1)$ by using Lemma 4.24 with $\xi = (2, 2, u_1^*, u_1^*, 1, 1, 1, -1)$ in coordinates $(b_{i_1}'', b_{i_2}'', d_{j_1}'', d_{j_2}'', a_2', a_3, c_1'', t)$.

Case 1.2: $\omega^* \in \Omega_{m2}$**.**

This implies $u^* \neq 0, v^* \neq 0$. Without loss of generality, suppose that $u_1^* \neq 0$. If $\omega^* \in \Omega_{m21}$, we further assume that $a_1^* = d_j^* = 0, u_1^* \neq 0, v_j^* \neq 0$. Substituting

$$\begin{aligned} d_i = &\quad (d_i' + a_1 t_0^* v_i^*)/(a_1 + t_1^* u_1^*), \quad i = 1, 2 \\ a_2 = &\quad (a_2' + a_1 u_2^*)/u_1^*, \end{aligned}$$

the pullback ideal is $\langle a_2', b_1', b_2', c_1', c_2', d_1', d_2' \rangle$ so the RLCT is at least $(7, 1)$. Note that $a_i = (a_2' w_i^* + a_1 u_i^*)/u_1^*$ for $i = 1, 2, 3$ where $w_i^* = 0, 1, -1$ respectively. If ω^* is not in Ω_{m21}, we consider the blowup chart $a_2' = s, b_i' = sb_i'', c_i' = sc_i'', d_i' = sd_i''$. The active inequalities are as follows. The symbol $v-$ denotes $v_i^* \leq 0$.

$$\begin{aligned} b_i^* = 0: &\quad 0 \leq [sb_i'' - tv_i^* - (sd_i'' + a_1 t_0^* v_i^*)(t_1^* - t)/(t_1^* u_1^* + a_1)]/(t_0^* + t) \quad v- \\ c_i^* = 0: &\quad 0 \leq [sc_i'' - tu_i^* - (sw_i^* + a_1 u_i^*)(t_0^* + t)/u_1^*]/(t_1^* - t) \quad u+ \\ a_i^* = 0: &\quad 0 \leq (sw_i^* + a_1 u_i^*)/u_1^* \quad u- \\ d_i^* = 0: &\quad 0 \leq (sd_i'' + a_1 t_0^* v_i^*)/(t_1^* u_1^* + a_1) \quad v+ \end{aligned}$$

The crux to understanding the inequalities is this: if $a_i^* = d_j^* = 0, u_i^* \neq 0, v_j^* \neq 0$, the coefficient of a_1 appears with different signs in the inequalities for $a_i^* = 0$ and $d_j^* = 0$. This makes it difficult to choose a suitable vector ξ for Lemma 4.24. Similarly, if $b_i^* = c_j^* = 0, v_i^* \neq 0, u_j^* \neq 0$, the coefficient of $u_1^* t + t_0^* a_1$ appears with different signs. Fortunately, since $\omega^* \notin \Omega_{m21}$, we do not have such obstructions and it is an easy exercise to find the vector ξ. Thus, the RLCT is $(7, 1)$.

If $\omega^* \in \Omega_{m21} \setminus \Omega_{m22}$, we blow up $a_1 = s, a_2' = sa_2'', b_i' = sb_i'', c_i = sc_i'', d_i = sd_i''$. The active inequalities for $a_1^* = d_j^* = 0$ imply that the new region $\mathcal{M}$ is not full at the origin of this chart. Thus, we shift our focus to the other charts of the blowup where the pullback pair is $(s; s^7)$, so the RLCT is at least $(8, 1)$. In the chart where $a_2' = s, a_1 = sa_1', b_i' = sb_i'', c_i = sc_i'', d_i = sd_i''$, we do not have obstructions coming from any $b_i^* = c_j^* = 0, v_i^* \neq 0, u_j^* \neq 0$ so it is again easy to find the vector ξ for Lemma 4.24. The threshold is exactly $(8, 1)$.

If $\omega^* \in \Omega_{m22}$, consider the following two charts out of the nine charts in the blowup of the origin in $\mathbb{R}^9$.

$$\begin{aligned} &\text{Chart 1:} \quad a_1 = s, t = st', a_2' = sa_2'', b_i' = sb_i'', c_i = sc_i'', d_i = sd_i'' \\ &\text{Chart 2:} \quad t = s, a_1 = sa_1', a_2' = sa_2'', b_i' = sb_i'', c_i = sc_i'', d_i = sd_i'' \end{aligned}$$

The inequalities for $a_i^* = d_j^* = 0, u_i^* \neq 0, v_j^* \neq 0$ and $b_i^* = c_j^* = 0, v_i^* \neq 0, u_j^* \neq 0$ imply that the new region $\mathcal{M}$ is not full at points outside of the other seven charts, so we may ignore these two charts in computing the RLCT. Indeed, for Chart 1, the active inequalities

$$\begin{aligned} a_i^* = 0: &\quad 0 \leq s(a_2'' w_i^* + u_i^*)/u_1^* \quad u- \\ d_i^* = 0: &\quad 0 \leq s(d_i'' + t_0^* v_i^*)/(t_1^* u_1^* + s) \quad v+ \end{aligned}$$

tell us that a_2'' or d_2'' must be non-zero for $\mathcal{M}$ to be full. In Chart 2, suppose $\mathcal{M}$ is full at some point x where $a_2'' = b_1'' = b_2'' = c_1'' = c_2'' = d_1'' = d_2'' = 0$. Then,

$$\begin{array}{lll} a_i^* = 0: & 0 \leq s(a_2'' w_i^* + a_1' u_i^*)/u_1^* & u- \\ d_i^* = 0: & 0 \leq s(d_i'' + a_1' t_0^* v_i^*)/(t_1^* u_1^* + s a_1') & v+ \end{array}$$

imply that $a_1' = 0$ at x. However, if this is the case, the inequalities

$$\begin{array}{lll} b_i^* = 0: & 0 \leq s[b_i'' - v_i^* - (d_i'' + a_1' t_0^* v_i^*)(t_1^* - s)/(t_1^* u_1^* + s a_1')]/(t_0^* + s) & v- \\ c_i^* = 0: & 0 \leq s[c_i'' - u_i^* - (a_2'' w_i^* + a_1' u_i^*)(t_0^* + s)/u_1^*]/(t_1^* - s) & u+ \end{array}$$

forces b_i'' or c_i'' to be non-zero for some i, a contradiction. Thus, we shift our focus to the other seven charts where the pullback pair is $(s; s^8)$ and the RLCT is at least $(9, 1)$. In the chart for $a_2' = s, a_1 = s a_1', t = s t', b_i' = s b_i'', c_i' = s c_i'', d_i' = s d_i''$, note that we cannot have both $a_2^* = 0$ and $a_3^* = 0$ because we assumed $a_1^* = 0$. It is now easy to find the vector ξ for Lemma 4.24, so the threshold is $(9, 1)$.

Case 1.3: $\omega^* \in \Omega_{m0}$**.**

This implies $u_i^* = v_i^* = 0$ for all i. The pullback ideal can be written as

$$\langle b_1', b_2', c_1', c_2' \rangle + \langle a_1, a_2 \rangle \langle d_1, d_2 \rangle$$

whose RLCT over an interior point of Ω is $(6, 2)$ by Proposition 4.5. This occurs in Ω_{m000} where none of the inequalities are active. Now, suppose the only active inequalities come from $a_1^* = c_1^* = 0$. We blow up the origin in $\{(a_1, c_1') \in \mathbb{R}^2\}$. In the chart given by $a_1 = a_1', c_1' = a_1' c_1''$, the new region $\mathcal{M}$ is not full at the origin, so we only need to study the chart where $c_1' = c_1'', a_1 = c_1'' a_1'$. The pullback pair becomes $(\langle c_1'' \rangle + \langle b_1', b_2', c_2' \rangle + \langle a_2 \rangle \langle d_1, d_2 \rangle; c_1'')$, and a simple application of Lemma 4.24 and Proposition 4.5 shows that the threshold is $(6, 1)$.

In this fashion, we study the different scenarios and summarize the pullback pairs and thresholds in the table below.

Inequalities	Pullback pair		RLCT
$-$	$(\langle b_1', b_2', c_1', c_2' \rangle + \langle a_1, a_2 \rangle \langle d_1, d_2 \rangle;$	$1)$	$(6, 2)$
$a_1^* = 0$	$(\langle b_1', b_2', c_1'', c_2' \rangle + \langle a_2 \rangle \langle d_1, d_2 \rangle;$	$c_1'')$	$(6, 1)$
$a_1^* = 0, b_1^* = 0$	$(\langle b_1'', b_2', c_1'', c_2' \rangle + \langle a_2 \rangle \langle d_2 \rangle;$	$b_1'' c_1'')$	$(7, 2)$
$a_1^* = 1$	$(\langle b_1', b_2', c_1'', c_2'' \rangle;$	$c_1'' c_2'')$	$(6, 1)$
$a_1^* = 1, b_1^* = 0$	$(\langle b_1'', b_2', c_1'', c_2'' \rangle;$	$b_1'' c_1'' c_2'')$	$(7, 1)$
$a_1^* = 1, b_1^* = 1$	$(\langle b_1'', b_2'', c_1'', c_2'' \rangle;$	$b_1'' b_2'' c_1'' c_2'')$	$(8, 1)$

For example, the case $a_3^* = c_3^* = 1$ corresponds to $a_1^* = a_2^* = c_1^* = c_2^* = 0$. Here, we blow up the origins in $\{(a_1, c_1') \in \mathbb{R}^2\}$ and $\{(a_2, c_2') \in \mathbb{R}^2\}$. As before, we can ignore the other charts and just consider the one where $a_1 = c_1'' a_1', c_1' = c_1'', a_2 = c_2'' a_2', c_2' = c_2''$. The pullback

pair is $(\langle c_1'' \rangle + \langle c_2'' \rangle + \langle b_1', b_2' \rangle, c_1'' c_2'')$. If $b_i^* \neq 0$ for all i, the RLCT is $(6,1)$ by Lemma 4.24 and Proposition 4.5.

Case 2: $\omega^* \in \Omega_u$**.**

Without loss of generality, assume $t^* = 0$ and substitute

$$\begin{aligned} c_i &= (c_i' - t(a_i + u_i^*))/(1-t) \quad i = 1,2 \\ d_i &= (d_i' - t(b_i + v_i^*))/(1-t) \quad i = 1,2. \end{aligned}$$

The pullback ideal is the sum of $\langle c_1', c_2', d_1', d_2' \rangle$ and

$$\langle t \rangle \langle a_1 + u_1^*, a_2 + u_2^* \rangle \langle b_1 + v_1^*, b_2 + v_2^* \rangle.$$

Since $c_3' = -c_1' - c_2'$ and similarly for the d_i', a_i, b_i, u_i^* and v_i^*, it is useful to write this ideal more symmetrically as the sum of $\langle c_1', c_2', c_3' \rangle$, $\langle d_1', d_2', d_3' \rangle$ and

$$\langle t \rangle \langle a_1 + u_1^*, a_2 + u_2^*, a_3 + u_3^* \rangle \langle b_1 + v_1^*, b_2 + v_2^*, b_3 + v_3^* \rangle.$$

Meanwhile, the inequalities are

$$\begin{aligned} a_i^* = 0: &\quad 0 \leq a_i \\ c_i^* = 0: &\quad 0 \leq (c_i' - t(a_i + u_i^*))/(1-t) \quad u_i^* \geq 0 \\ b_j^* = 0: &\quad 0 \leq b_j \\ d_j^* = 0: &\quad 0 \leq (d_j' - t(b_j + v_j^*))/(1-t) \quad v_j^* \geq 0. \end{aligned}$$

We now relabel the indices of the a_i and c_i', without changing the b_j and d_j', so that the active inequalities are among those from $a_1^* = 0, a_2^* = 0, c_{i_1}^* = 0, c_{i_2}^* = 0$. The b_j and d_j' are thereafter also relabeled so that the inequalities come from $b_1^* = 0, b_2^* = 0, d_{j_1}^* = 0, d_{j_2}^* = 0$. We claim that the new region $\mathcal{M}$ contains, for small ε, the orthant neighborhood

$$\{(a_1, a_2, b_1, b_2, c_{i_1}, c_{i_2}, d_{j_1}, d_{j_2}, -t) \in [0, \varepsilon]^9\}.$$

Indeed, the only problematic inequalities are

$$\begin{aligned} c_3^* = 0: &\quad 0 \leq (c_3' - t(-a_1 - a_2 + u_i^*))/(1-t) \quad u_3^* = 0 \\ d_3^* = 0: &\quad 0 \leq (d_3' - t(-b_1 - b_2 + v_j^*))/(1-t) \quad v_3^* = 0. \end{aligned}$$

However, these inequalities cannot occur because for instance, $u_3^* = 0$ and $c_3^* = 0$ implies $a_3^* = 0$, a contradiction since the a_i were relabeled to avoid this. Finally, the threshold of $\langle t \rangle$ is $(1,1)$ while that of $\langle a_1 + u_1^*, a_2 + u_2^* \rangle$ and $\langle b_1 + v_1^*, b_2 + v_2^* \rangle$ are at least $(2,1)$ each. By Proposition 4.5, the RLCT of their product is $(1,1)$ and that of the pullback ideal we were originally interested in is $(5,1)$. $\square$

Proof of Theorem 4.23. Given a matrix $q = (q_{ij})$, the learning coefficient (λ, θ) of the model at q is the minimum of RLCTs at points $\omega^* \in \Omega$ where $p(\omega^*) = q$. The first statement then follows logically from the five claims below:

$$p(\Omega_u) = S_1, \quad p(\Omega_{m0}) \subset S_1, \quad p(\Omega_{m0}) \subset S_1, \quad p(\Omega_{m21}) = S_{21}, \quad p(\Omega_{m22}) = S_{22}.$$

The first three claims are trivial. The proofs of the last two claims are similar, so we only show $p(\Omega_{m22}) = S_{22}$. First, it is easy to check that $p(\Omega_{m22}) \subset S_{22}$. Now, if $p(\omega^*) = q \in S_{22}$, then $q_{11} = t^* a_1^* b_1^* + (1 - t^*) c_1^* d_1^* = 0$ implies that $a_1^* = 0$ or $b_1^* = 0$ because the parameters are positive. Without loss of generality, suppose $a_1^* = 0$. Because $q_{12} \neq 0$, we have $c_1^* \neq 0$ which leads us to $d_1^* = 0$ and $b_1^* \neq 0$. The condition $q_{22} = 0$ then shows that $b_2^* = c_2^* = 0, a_2^* \neq 0, d_2^* \neq 0$. Therefore, $\omega^* \in \Omega_{m22}$ and $p(\Omega_{m22}) \supset S_{22}$.

The last statement of the theorem is a consequence of Theorem 1.17. □

Remark 4.26. This is a difficult example because of the algebraic interactions between the boundary of Ω and the fiber ideal I of the model. If we were computing the RLCT only at points ω^* which lie in the interior of Ω, the calculation would have been much easier. In future work, we hope to find an algorithm involving Newton polyhedra for computing the RLCT in situations where the boundary has normal crossings. □

In this chapter, we investigated the theory of real log canonical thresholds of ideals. A treatment of this topic was necessary for the analysis of our statistical models, but not much is known in the literature except for their relationship to jumping numbers and complex log canonical thresholds [47]. Hence, many of the results in this chapter are new, and they were inspired by analogous results for complex log canonical thresholds. In Section 4.1, we explored some of the fundamental properties of RLCTs of ideals. We gave several equivalent definitions, showed that they are independent of the choice of generators, and derived sum, product and chain rules for calculating them. In Section 4.2, we extended Varchenko's concept of Newton nondegeneracy to ideals, and introduced the notion of *sos-nondegeneracy*. These toric techniques allowed us to give an upper bound for the RLCT of arbitrary ideals (Proposition 4.8) and to compute them exactly for monomial ideals (Theorem 4.18). We applied these tools in Section 4.3 to a difficult statistical example where boundary issues complicate the computation. We derived the learning coefficient of the model and used it to approximate the marginal likelihood integral of actual health data [18].

Chapter 5

Higher Order Asymptotics

Let Ω be a compact semianalytic subset of $\mathbb{R}^d$, $f : \Omega \to \mathbb{R}$ be real analytic over Ω, and $\varphi : \Omega \to \mathbb{R}$ be nearly analytic over Ω. We consider the Laplace integral

$$Z(n) = \int_\Omega e^{-n|f(x)|}|\varphi(x)|dx.$$

By Theorem 3.16, $Z(n)$ has an asymptotic expansion

$$Z(n) \approx \sum_\alpha \sum_{i=1}^{d} c_{\alpha,i}\, n^{-\alpha}(\log n)^{i-1}, \quad n \to \infty$$

In this chapter, we are interested in computing the coefficients $c_{\alpha,i}$ in this expansion.

Definition 5.1. The *leading coefficient* $\mathrm{coef}_\Omega(f;\varphi)$ is the coefficient $c_{\lambda,\theta}$ of the leading term $c_{\lambda,\theta}\, n^{-\lambda}(\ln n)^\theta$ in the asymptotic expansion of $Z(n)$. Note that (λ,θ) is the real log canonical threshold $\mathrm{RLCT}_\Omega(f;\varphi)$.

Theorem 3.16 gives us a way to compute this leading coefficient from the Laurent expansion of the zeta function

$$\zeta(z) = \int_\Omega |f(x)|^{-z}|\varphi(x)|dx, \quad z \in \mathbb{C}$$

associated to $Z(n)$. Recall that Γ represents the Gamma function.

Proposition 5.2. *The leading coefficient* $\mathrm{coef}_\Omega(f;\varphi)$ *is given by*

$$c_{\lambda,\theta} = \frac{(-1)^\theta\,\Gamma(\lambda)}{(\theta-1)!}\, d_{\lambda,\theta}$$

where $d_{\lambda,\theta}$ *is the coefficient of* $(z-\lambda)^{-\theta}$ *in the Laurent expansion of* $\zeta(z)$.

Remark 5.3. This project to investigate leading coefficients and higher order coefficients of the asymptotic expansions of Laplace integrals began with a discussion with Robin Pemantle in April 2009 about asymptotics of generating functions. □

5.1 Sum, Product and Chain Rules

In Section 4.1.3, we saw sum, product and chain rules for the RLCT of ideals with disjoint sets of indeterminates. Incidentally, their proofs give us similar rules for the leading coefficients.

As before, let $f_x \in \mathcal{A}_X$ and $f_y \in \mathcal{A}_Y$ where $X \subset \mathbb{R}^m$ and $Y \subset \mathbb{R}^n$ are compact semianalytic subsets. It is useful to think of f_x and f_y as polynomials or analytic functions with disjoint sets of indeterminates $\{x_1, \ldots, x_m\}$ and $\{y_1, \ldots, y_n\}$. Let $\varphi_x : X \to \mathbb{R}$ and $\varphi_y : Y \to \mathbb{R}$ be nearly analytic. By composing with projections $X \times Y \to X$ and $X \times Y \to Y$, we may regard f_x, f_y, φ_x and φ_y as functions which are real analytic over $X \times Y$.

For the proofs, let $Z_x(n)$ and $Z_y(n)$ be the Laplace integrals corresponding to the triples (X, f_x, φ_x) and (Y, f_y, φ_y) respectively. Let $\zeta_x(z)$ and $\zeta_y(z)$ be the associated zeta functions. We define $(\lambda_x, \theta_x) = \mathrm{RLCT}_X(f_x; \varphi_x), c_x = \mathrm{coef}_X(f_x; \varphi_x)$ and similarly for (λ_y, θ_y) and c_y.

Proposition 5.4 (Sum Rule)**.** *For functions f_x, φ_x and f_y, φ_y with disjoint indeterminates,*

$$\mathrm{coef}_{X\times Y}(f_x + f_y; \varphi_x\varphi_y) = \mathrm{coef}_X(f_x; \varphi_x) \cdot \mathrm{coef}_Y(f_y; \varphi_y)$$

Proof. Let $(\lambda, \theta) = \mathrm{RLCT}_{X\times Y}(f_x + f_y; \varphi_x\varphi_y)$ and $c' = \mathrm{coef}_{X\times Y}(f_x + f_y; \varphi_x\varphi_y)$. Then,

$$\begin{aligned} Z_x(n) &\approx c_x\, n^{-\lambda_x} (\log n)^{\theta_x - 1}, \\ Z_y(n) &\approx c_y\, n^{-\lambda_y} (\log n)^{\theta_y - 1}, \\ Z(n) &\approx c\, n^{-\lambda} (\log n)^{\theta - 1} \end{aligned}$$

asymptotically. These Laplace integrals are related by the equation

$$\begin{aligned} Z(n) &\approx \textstyle\int_{X\times Y} e^{-nf_x - nf_y} |\varphi_x||\varphi_y|\, dx\, dy \\ &= \textstyle\int_X e^{-nf_x} |\varphi_x|\, dx \cdot \int_Y e^{-nf_y} |\varphi_y|\, dy = Z_x(n)\, Z_y(n) \end{aligned}$$

so $c' = c_x c_y$ as required. □

Proposition 5.5 (Product Rule)**.** *For functions f_x, φ_x and f_y, φ_y with disjoint indeterminates, if $\lambda_x = \lambda_y$, then*

$$\mathrm{coef}_{\Omega_x\times\Omega_y}(f_x f_y; \varphi_x\varphi_y) = \frac{(\theta_x - 1)!(\theta_y - 1)!}{(\theta_x + \theta_y - 1)!\, \Gamma(\lambda_x)} \cdot \mathrm{coef}_{\Omega_x}(f_x; \varphi_x) \cdot \mathrm{coef}_{\Omega_y}(f_y; \varphi_y).$$

On the other hand, if $\lambda_x < \lambda_y$, then

$$\mathrm{coef}_{\Omega_x\times\Omega_y}(f_x f_y; \varphi_x\varphi_y) = \mathrm{coef}_{\Omega_x}(f_x; \varphi_x) \cdot \zeta_y(\lambda_x).$$

Proof. Let $(\lambda, \theta) = \mathrm{RLCT}_{X\times Y}(f_x f_y; \varphi_x\varphi_y)$ and $c' = \mathrm{coef}_{X\times Y}(f_x f_y; \varphi_x\varphi_y)$. Then,

$$\begin{aligned} \zeta_x(z) &= d_x (z - \lambda_x)^{-\theta_x} + \cdots \\ \zeta_y(z) &= d_y (z - \lambda_y)^{-\theta_y} + \cdots \\ \zeta(z) &= d' (z - \lambda)^{-\theta} + \cdots \end{aligned}$$

are the Laurent expansions of the corresponding zeta functions for some coefficients d_x, d_y and d'. These zeta functions are related by the equation

$$\begin{aligned}\zeta(z) &= \int_{X\times Y} \left(f_x f_y\right)^{-z} |\varphi_x||\varphi_y|\, dx\, dy \\ &= \int_X f_x^{-z}|\varphi_x|\, dx \cdot \int_Y f_y^{-z}|\varphi_y|\, dy \quad = \quad \zeta_x(z)\zeta_y(z).\end{aligned}$$

Thus, if $\lambda_x = \lambda_y$, then $d' = d_x d_y$. On the other hand, if $\lambda_x < \lambda_y$, then $d' = d_x\, \zeta_y(\lambda_x)$. The two required formulas now follow from Proposition 5.2. □

Proposition 5.6 (Chain Rule)**.** *Let $\Omega \subset \mathbb{R}^d$ be a compact semianalytic subset and $f : \Omega \to \mathbb{R}$ a real analytic function. If W is an open neighborhood of Ω, M is a real analytic manifold, $\rho : M \to W$ is a change of variables away from $\mathcal{V}(f)$ and $\mathcal{M} = \rho^{-1}(\Omega)$, then*

$$\operatorname{coef}_\Omega(f;\varphi) = \operatorname{coef}_\mathcal{M}(f\circ\rho; (\varphi\circ\rho)|\rho'|).$$

Proof. Direct consequence of applying a change of variable to the Laplace integral. □

There are a few simple base cases where the leading coefficient computes easily.

Proposition 5.7. *Let $\Omega \subset \mathbb{R}$ be a compact neighborhood of the origin. For $k > 0, m \geq 0$,*

$$\operatorname{coef}_\Omega(x^k; x^m) = \frac{2}{k}\,\Gamma\!\left(\frac{m+1}{k}\right).$$

Proof. We perform the computation that was implicit in Proposition 3.7. For $\varepsilon > 0$,

$$\begin{aligned}\int_{-\varepsilon}^{\varepsilon} |x^{-zk+m}|dx &= 2\int_0^\varepsilon x^{-zk+m}dx \\ &= \frac{2\,\varepsilon^{-zk+m+1}}{-zk+m+1} \\ &= \frac{2}{-zk+m+1} e^{(-zk+m+1)\log\varepsilon} \\ &= \frac{-2/k}{z-(m+1)/k}\big(1 + (-zk+m+1)\log\varepsilon + \cdots\big)\end{aligned}$$

More generally, the coefficient at $z = (m+1)/k$ in the Laurent expansion of the zeta function

$$\zeta(z) = \int_\Omega |x^{-zk+m}|dx$$

is $-2/k$. The formula now follows from Proposition 5.2. □

We now give some consequences of applying the sum, product and chain rules to Proposition 5.7. Our next corollary is not an asymptotic result but an exact result. It was also proved by Pemantle and Wilson [42, §3] using non-asymptotic methods.

Corollary 5.8. *For $m \geq 0$,*

$$\int_{\mathbb{R}} e^{-nx^2}|x^m|dx \;=\; \begin{cases} \dfrac{m!\,\sqrt{\pi}}{(m/2)!\,2^m}\, n^{-(m+1)/2} & \text{for } m \text{ even,} \\ \left(\dfrac{m-1}{2}\right)!\, n^{-(m+1)/2} & \text{for } m \text{ odd.} \end{cases}$$

Proof. By Proposition 5.7, we have the asymptotics

$$Z(n) = \int_{\mathbb{R}} e^{-nx^2}|x^m|dx \;\approx\; \Gamma\left(\frac{m+1}{2}\right) n^{-(m+1)/2}.$$

Making the substitution $x = ty$, we get

$$Z(n) = \int_{\mathbb{R}} e^{-nt^2y^2}|t^m y^m|tdy = t^{m+1}Z(nt^2)$$

Now, because of the asymptotics, as $t \to \infty$,

$$\frac{Z(nt^2)}{\Gamma\left(\frac{m+1}{2}\right)(nt^2)^{-(m+1)/2}} \to 1.$$

This ratio is independent of t because it evaluates to

$$\frac{Z(n)}{\Gamma\left(\frac{m+1}{2}\right) n^{-(m+1)/2}}.$$

Thus, the asymptotic result is exact and the desired formula now follows from

$$\Gamma\left(\frac{m+1}{2}\right) = \begin{cases} \dfrac{m!\,\sqrt{\pi}}{(m/2)!\,2^m} & \text{for } m \text{ even,} \\ \left(\dfrac{m-1}{2}\right)! & \text{for } m \text{ odd} \end{cases}$$

which are standard formulas for the Gamma function. □

Using this corollary, the asymptotics of the Laplace integral

$$\int_{\mathbb{R}^d} e^{-n(x_1^2+\cdots+x_d^2)}|x_1^{m_1}\cdots x_d^{m_d}|dx = \prod_{i=1}^{d}\int_{\mathbb{R}} e^{-nx_i^2}|x_i^{m_i}|dx$$

can be computed by multiplying the formulas for each m_i. Lastly, as an extension of Proposition 3.7, we compute the leading coefficient of the asymptotics for a Laplace integral with monomial phase and amplitude functions.

Corollary 5.9. *For vectors $\kappa, \tau \in \mathbb{Z}^d_{\geq 0}$, suppose that $\kappa \neq (0, \ldots, 0)$ and that*

$$\lambda = \frac{\tau_1 + 1}{\kappa_1} = \cdots = \frac{\tau_\theta + 1}{\kappa_\theta} < \frac{\tau_{\theta+1} + 1}{\kappa_{\theta+1}} \leq \cdots \leq \frac{\tau_d + 1}{\kappa_d}$$

for some λ and θ. Then, asymptotically as $n \to \infty$,

$$\int_{[0,1]^d} e^{-n\omega^\kappa} \omega^\tau d\omega \;\approx\; \frac{\Gamma(\lambda)}{(\theta-1)! \prod_{i=1}^d \kappa_i \prod_{i=\theta+1}^d (-\lambda\kappa_i + \tau_i + 1)} \, n^{-\lambda} (\log n)^{\theta-1}.$$

Proof. We apply the product rule to the fact that

$$\operatorname{coef}_{[0,1]}(\omega_i^{\kappa_i}; \omega_i^{\tau_i}) = \frac{1}{\kappa_i} \Gamma\left(\frac{\tau_i + 1}{\kappa_i}\right)$$

and that for $\tilde{\omega} = (\omega_{\theta+1}, \ldots, \omega_d)$ and similarly for $\tilde{\kappa}$ and $\tilde{\tau}$, we compute

$$\tilde{\zeta}(\lambda) = \int_{[0,1]^{d-\theta}} \tilde{\omega}^{-\lambda\tilde{\kappa} + \tilde{\tau}} d\tilde{\omega} = \prod_{i=\theta+1}^{d} \frac{1}{-\lambda\kappa_i + \tau_i + 1}. \qquad \square$$

5.2 Leading Coefficient

In this section, we compute the leading coefficient $\operatorname{coef}_{[0,1]^d}(f; \omega^\tau)$ of the asymptotic expansion of a Laplace integral $Z(n)$ with a nondegenerate and nonnegative phase f and a monomial amplitude ω^τ in a unit hypercube $[0,1]^d$ at the origin.

Let us recall facts about nondegenerate functions and toric resolutions from Chapter 3. Suppose $f : W \to \mathbb{R}$ is real analytic in some neighborhood $W \subset \mathbb{R}^d$ of the origin. Let $\mathcal{F}$ be a smooth refinement of the normal fan $\mathcal{F}(f)$ of the Newton polyhedron $\mathcal{P}(f)$. We can associated to $\mathcal{F}$ a smooth toric variety $\mathbb{P}(f)$ and a blowup map $\rho_{\mathcal{F}} : \mathbb{P}(\mathcal{F}) \to \mathbb{R}^d$ which is defined by monomial maps ρ_σ on affine charts $U_\sigma \simeq \mathbb{R}^d$ for each maximal cone σ of $\mathcal{F}$. More specifically, if σ is minimally generated by $v_1, \ldots, v_d \in \mathbb{Z}^d$ with $v_i = (v_{i1}, \ldots, v_{id})$, the map $\rho_\sigma : U_\sigma \to \mathbb{R}^d$ is defined by $\mu \mapsto \omega = \mu^v$, i.e.

$$\begin{aligned}
\omega_1 &= \mu_1^{v_{11}} \mu_2^{v_{21}} \cdots \mu_d^{v_{d1}} \\
\omega_2 &= \mu_1^{v_{12}} \mu_2^{v_{22}} \cdots \mu_d^{v_{d2}} \\
&\vdots \\
\omega_d &= \mu_1^{v_{1d}} \mu_2^{v_{2d}} \cdots \mu_d^{v_{dd}}.
\end{aligned}$$

Here, v is the matrix (v_{ij}) where each generator v_i forms a row of v. Because $\mathcal{F}$ refines $\mathcal{F}(f)$, σ is contained in some maximal cone σ' of $\mathcal{F}(f)$. Let $\alpha \in \mathbb{R}^d$ be the vertex of $\mathcal{P}(f)$ that is dual to σ'. This means that in the matrix product $v\alpha$, for each $1 \leq i \leq d$,

$$(v\alpha)_i = \langle v_i, \alpha \rangle \leq \langle v_i, \alpha' \rangle \quad \text{for all } \alpha' \in \mathcal{P}(f).$$

As seen in the proof of Theorem 4.15, after the monomial change of variables $p_\sigma : \mu \mapsto \omega = \mu^v$, we have $f(\mu^v) = g(\mu)\mu^{v\alpha}$ where $g(\mu)$ is the *strict transform* of $f(\omega)$.

Given a vector $\tau \in \mathbb{Z}^d$ of nonnegative integers, let l_τ be the τ-distance of $\mathcal{P}(f)$ and θ_τ its multiplicity. Theorem 4.13 says that if f is nondegenerate, then $\rho_\mathcal{F}$ restricted to W resolves the singularity of f at the origin. Furthermore, if f is nonnegative, then asymptotically

$$\int_\Omega e^{-nf(\omega)}|\omega^\tau|d\omega \;\approx\; Cn^{-\lambda}(\log n)^{\theta-1}, \quad n \to \infty \tag{5.1}$$

where $(\lambda, \theta) = (1/l_\tau, \theta_\tau)$, $C > 0$ is a constant and $\Omega \subset W$ is a sufficiently small neighborhood of the origin. By scaling the coordinates appropriately, we may assume that Ω contains the hypercube $[-1, 1]^d$. Also, to simplify the computations, we break up $[-1, 1]^d$ into its 2^d unit orthants and consider the asymptotics in each orthant separately. Without loss of generality, we let $\Omega = [0, 1]^d$. Now, on the other hand, given a nondegenerate function f and $\Omega = [0, 1]^d$, how do we know if Ω is small enough for the asymptotics (5.1) hold? One sufficient condition is that the strict transform g does not vanish in $\rho_\mathcal{F}^{-1}\Omega$. Indeed, if this occurs, it follows by definition that $\rho_\mathcal{F}$ is a resolution of singularities for f over Ω, and the asymptotics (5.1) can then be computed explicitly from this resolution.

In computing the τ-distance, we intersected the ray $\{t(\tau + 1) : t \geq 0\}$ with the Newton polyhedron $\mathcal{P}(f)$. Let $\sigma_\tau \in \mathcal{F}(f)$ denote the cone corresponding to the face of $\mathcal{P}(f)$ at this intersection. We call σ_τ the *τ-cone* and note that its dimension is exactly $\theta = \theta_\tau$. Now, in the refinement $\mathcal{F}$ of $\mathcal{F}(f)$, let $\mathcal{F}_\tau$ be the set of maximal cones which intersect σ_τ in dimension θ. For each cone σ in $\mathcal{F}_\tau$, if v is the matrix whose rows v_i are the minimal generators of σ, we require that the first θ rows of v lie in the τ-cone σ_τ. Because of the special role played by the first θ coordinates, we write a vector $\mu \in U_\sigma \simeq \mathbb{R}^d$ as $(\hat{\mu}, \bar{\mu}) \in \mathbb{R}^\theta \times \mathbb{R}^{d-\theta}$.

Finally, before we present the formula for the leading coefficient $\operatorname{coef}_{[0,1]^d}(f; \omega^\tau)$, we need to understand the geometry of the blowup of the unit hypercube $[0, 1]^d$.

Lemma 5.10. *Let $\mathcal{F}$ be a smooth locally complete fan in $\mathbb{R}^d$. The blowup $\rho_\mathcal{F}^{-1}[0, 1]^d$ in $\mathbb{P}(\mathcal{F})$ of the unit hypercube $[0, 1]^d$ is the union of hypercubes*

$$\mathcal{H}_\sigma := [0, 1]^d \subset U_\sigma \simeq \mathbb{R}^d$$

as σ varies over maximal cones of $\mathcal{F}$. The $\mathcal{H}_\sigma$ intersect only along their boundaries.

Proof. For each maximal cone σ in $\mathcal{F}$, the blowup map $\rho_\mathcal{F}$ restricted to the open affine U_σ is given by $\mu \mapsto \omega = \mu^v$ where the rows of the matrix v are minimal generators v_i of σ. Taking logarithms of $\omega = \mu^v$, we get $(-\log \omega) = (-\log \mu)v$ where $\log \omega$ and $\log \mu$ are row vectors $(\log \omega_i)$ and $(\log \mu_i)$. As μ varies over the unit hypercube $\mathcal{H}_\sigma$, the vector $-\log \mu$ varies over the positive orthant. This implies that $-\log \omega$ takes all values in the cone σ generated by the v_i. Because $\mathcal{F}$ is locally complete, all points in the positive orthant have a unique preimage in $\cup\mathcal{H}_\sigma$ under the map $-\log \rho_\mathcal{F}$, except at the boundaries of the maximal cones. □

Theorem 5.11. *Let $\tau \in \mathbb{Z}^d_{\geq 0}$ and let the real analytic function $f : [0,1]^d \to \mathbb{R}$ be nondegenerate and nonnegative. Let $\mathcal{F}$ be a smooth refinement of $\mathcal{F}(f)$ such that the strict transform g of f under the blowup map $\rho_{\mathcal{F}}$ is positive over $\rho_{\mathcal{F}}^{-1}[0,1]^d$. Then, asymptotically*

$$\int_{[0,1]^d} e^{-nf(\omega)}\omega^\tau d\omega \quad \approx \quad Cn^{-\lambda}(\log n)^{\theta-1}, \quad n \to \infty$$

where $(\lambda, \theta) = (1/l_\tau, \theta_\tau)$ and

$$C = \mathrm{coef}_{[0,1]^d}(f; \omega^\tau) = \frac{\Gamma(\lambda)}{(\theta-1)!} \sum_{\sigma \in \mathcal{F}_\tau} \prod_{i=1}^{\theta} (v\alpha)_i^{-1} \int_{[0,1]^{d-\theta}} g(0,\bar{\mu})^{-\lambda} \bar{\mu}^{\overline{m}-1} d\bar{\mu}.$$

Remark 5.12. For each $\sigma \in \mathcal{F}_\tau$ in the above sum, v is the matrix of minimal generators of σ, $\alpha \in \mathcal{P}(f)$ is the vertex dual to σ, $m = (\hat{m}, \overline{m}) = v(-\lambda\alpha + \tau + 1)$ and $g(\hat{\mu}, \bar{\mu}) = f(\mu^v)\mu^{-v\alpha}$ is the strict transform of $f(\omega)$ in the affine chart U_σ. □

Proof. Given Theorem 4.14 and Proposition 5.2, we only need to compute the coefficient of $(z-\lambda)^{-\theta}$ in the Laurent expansion of the zeta function

$$\zeta(z) = \int_{[0,1]^d} f(\omega)^{-z}\omega^\tau d\omega. \tag{5.2}$$

Let $\mathcal{F}$ be any smooth refinement of $\mathcal{F}(f)$. Applying the blowup $\rho_{\mathcal{F}}$ as a change of variable to the zeta function and by Lemma 5.10,

$$\zeta(z) = \sum_{\sigma \in \mathcal{F}} \int_{H_\sigma} f(\mu^v)^{-z} \mu^{v(\tau+1)-1} d\mu.$$

where the sum is over maximal cones σ of $\mathcal{F}$. In this sum, the only cones which contribute a $(z-\lambda)^{-\theta}$ term to the Laurent expansion are cones which intersect the τ-cone in dimension θ. Also, $f(\mu^v) = g(\mu)\mu^{v\alpha}$. Thus, we want to compute the coefficient of $(z-\lambda)^\theta$ in

$$\sum_{\sigma \in \mathcal{F}_\tau} \int_{[0,1]^d} f(\mu^v)^{-z} \mu^{v(\tau+1)-1} d\mu = \sum_{\sigma \in \mathcal{F}_\tau} \int_{[0,1]^d} g(\mu)^{-z} \mu^{-v\alpha z + v(\tau+1)-1} d\mu. \tag{5.3}$$

Now, let us write the strict transform $g(\mu)$ as

$$g(\hat{\mu}, \bar{\mu}) = g(0, \bar{\mu}) + \mu_1 g_1(\mu) + \cdots + \mu_d g_d(\mu)$$

where $g(0, \bar{\mu})$ is the sum of terms in the power series of $g(\mu)$ which involve only the variables $\bar{\mu} = (\mu_{\theta+1}, \ldots, \mu_d)$, and $g_1(\mu), \ldots, g_d(\mu)$ are some analytic functions. Because $g(\mu)$ is positive over $H_\sigma = [0,1]^d$, so is $g(0, \bar{\mu})$. Applying the generalized multinomial expansion, we get

$$g(\mu)^{-z} = g(0,\bar{\mu})^{-z} + \text{terms involving } \mu_1, \ldots, \mu_\theta.$$

The terms involving $\mu_1, \ldots, \mu_\theta$ do not give a $(z-\lambda)^{-\theta}$ term, so we restrict our attention to

$$\sum_{\sigma \in \mathcal{F}_\tau} \prod_{i=1}^{\theta} \frac{1}{(-v\alpha z + v(\tau+1))_i} \int_{[0,1]^{d-\theta}} g(0,\bar{\mu})^{-z} \prod_{i=\theta+1}^{d} \mu_i^{(-v\alpha z + v(\tau+1)-1)_i} d\mu. \tag{5.4}$$

Now, since

$$\lambda = \frac{(v(\tau+1))_1}{(v\alpha)_1} = \cdots = \frac{(v(\tau+1))_\theta}{(v\alpha)_\theta} < \frac{(v(\tau+1))_i}{(v\alpha)_i} \quad \text{for } i = \theta+1, \ldots, d,$$

it follows that $-1 < (-v\alpha\lambda + v(\tau+1) - 1)_i$ for $i = \theta+1, \ldots, d$, so each integral in (5.4) is well-defined at $z = \lambda$. Therefore, the coefficient of $(z-\lambda)^\theta$ is

$$(-1)^\theta \sum_{\sigma \in \mathcal{F}_\tau} \prod_{i=1}^{\theta} (v\alpha)_i^{-1} \int_{[0,1]^{d-\theta}} g(0,\bar{\mu})^{-\lambda} \bar{\mu}^{\overline{m}-1} d\bar{\mu}$$

where $m = (\hat{m}, \overline{m}) = v(-\lambda\alpha + \tau + 1)$ and the result follows. □

5.3 Higher Order Coefficients

We now give an algorithm for computing the higher order asymptotics of Laplace integrals with a nondegenerate phase function. As before, we assume that the amplitude function is monomial, and that the domain of integration is the unit hypercube $[0,1]^d$.

Our main tool will be equation (3.10) in Theorem 3.16, which expresses the higher order asymptotics in terms of coefficients in the Laurent expansion of the associated zeta function. We will also need to work with subseries of the power series expansion of the strict transform $g(\mu)$. For instance, in Theorem 5.11, we used the function $g(0,\bar{\mu})$ which is the sum of terms not involving the variables $\mu_1, \ldots, \mu_\theta$. Let us introduce notation for these subseries. Suppose we have variables $\mu = (\mu_1, \ldots, \mu_d)$, a power series $g(\mu) = \sum_\alpha c_\alpha \mu^\alpha$, an integer vector $\delta \in \mathbb{Z}^d_{\geq 0}$ and a vector δ^+ which comes from annotating some of the entries of δ with a + sign. Let S be the set of all vectors $\alpha \in \mathbb{Z}^d$ such that for each i, $\alpha_i \geq \delta_i$ if δ_i^+ is annotated with a + sign, and $\alpha_i = \delta_i$ otherwise. We define the subseries

$$g[\delta^+](\mu) = \sum_{\alpha \in S} c_\alpha \mu^\alpha.$$

For example, suppose $\mu = (\mu_1, \mu_2, \mu_3)$ and let $g(\mu) = \sum_\alpha c_\alpha \mu^\alpha$ be a formal power series. Let $\delta = (2,0,1)$ and $\delta^+ = (2,0,1^+)$. Then, $S = \{\alpha \in \mathbb{Z}^d : \alpha_1 = 2, \alpha_2 = 0, \alpha_3 \geq 1\}$ and

$$g[2,0,1^+](\mu) = \sum_{\alpha_1=2, \alpha_2=0, \alpha_3 \geq 1} c_\alpha \mu^\alpha.$$

Recursively, we can compute the power series $g[\delta^+](\mu)$ using the formula

$$g[\ldots,(i_j+1)^+,\ldots] = g[\ldots,i_j^+,\ldots] - g[\ldots,i_j,\ldots], \quad i_j \geq 0.$$

The base cases consist of all vectors δ^+ where only the zeros are annotated with the $+$ signs. It is easy to see that

$$g[i_1,\ldots,i_k,0^+,\ldots,0^+](\mu) = \frac{\mu_1^{i_1}\cdots\mu_k^{i_k}}{i_1!\cdots i_k!} \cdot \frac{\partial^{i_1+\cdots+i_k} g}{\partial\mu_1^{i_1}\cdots\partial\mu_k^{i_k}}(0,\ldots,0,\mu_{k+1},\ldots,\mu_d).$$

This formula comes from considering the variables $\mu_{k+1},\ldots,\mu_d$ as constants, and applying a Taylor series expansion to the remaining $\mu_1\ \ldots,\mu_k$. The other base cases can be derived from this formula by permuting the variables.

Now, let us assume the setup of Theorem 5.11. Suppose we want to find the coefficient of $(z-\lambda)^{-\theta}$ in the Laurent expansion of the zeta function (5.2), where $(\lambda,\theta) \in (\mathbb{Q}_{>0},\mathbb{Z}_{>0})$ is not necessarily the real log canonical threshold. Given a matrix $v \in \mathbb{Z}_{\geq 0}^{d\times d}$ and vector $\alpha \in \mathbb{Z}_{\geq 0}^d$, let $m = v(-\lambda\alpha+\tau+1) \in \mathbb{Q}^d$ and let $D(m,\theta)$ be the set of all vectors $\delta \in \mathbb{Z}_{\geq 0}^d$ such that

1. $\delta_i \leq \max(0,\lfloor 1-m_i\rfloor)$ for each i,
2. $m+\delta$ has at least θ entries which are zeros.

Note that D is a finite set. For each $\delta \in D$, let δ^+ be the vector whose i-th entry is annotated with a $+$ sign if $(m+\delta)_i > 0$. Because of the special role played by the annotated entries, we write a vector $\mu \in \mathbb{R}^d$ as $(\hat{\mu},\bar{\mu})$ where $\bar{\mu}$ consists of coordinates corresponding to annotated entries in δ^+. Similarly, we write $m = (\hat{m},\bar{m})$. Let d_0 and d_+ be the number of zero and positive entries in the vector $m+\delta$ respectively. Lastly, let $K(m,\theta,\delta)$ be the set of all vectors $k = (k_0,\ldots,k_d) \in \mathbb{Z}^{d+1}$ such that

1. $k_i = -1$ if $(m+\delta)_i = 0$,
2. $k_i \geq 0$ if $(m+\delta)_i \neq 0$,
3. $k_0+\cdots+k_d = -\theta$.

With these notation in place, we may now state our next theorem.

Theorem 5.13. *Let $\tau \in \mathbb{Z}_{\geq 0}^d$ and let the real analytic function $f:[0,1]^d \to \mathbb{R}$ be nondegenerate and nonnegative. Let $\mathcal{F}$ be a smooth refinement of $\mathcal{F}(f)$ such that the strict transform g of f under the blowup map $\rho_{\mathcal{F}}$ is positive over $\rho_{\mathcal{F}}^{-1}[0,1]^d$. Then, the coefficient $c_{\lambda,t}$ in the asymptotic expansion*

$$\int_{[0,1]^d} e^{-nf(\omega)}\omega^\tau d\omega \quad \approx \quad \sum_{\lambda}\sum_{t=1}^{d} c_{\lambda,t}\, n^{-\lambda}(\log n)^{t-1}, \quad n\to\infty$$

is given by the formula

$$c_{\lambda,t} = \frac{(-1)^t}{(t-1)!} \sum_{\theta=t}^{d} \frac{\Gamma^{(\theta-t)}(\lambda)}{(\theta-t)!}\, d_{\lambda,\theta}$$

where

$$d_{\lambda,\theta} = (-1)^\theta \sum_{\sigma\in\mathcal{F}} \sum_{\delta\in D(m,\theta)} \sum_{k\in K(m,\theta,\delta)} C(\sigma,\delta,k)\, I(\sigma,\delta,k)$$

and

$$C(\sigma,\delta,k) = \frac{1}{(d_0-\theta)!} \prod_{i=1}^{d} (v\alpha)_i^{k_i} \prod_{(m+\delta)_i<0} (-1)^{k_i} k_i!\, (m+\delta)_i^{-(k_i+1)}$$

$$I(\sigma,\delta,k) = \int_{[0,1]^{d_+}} ((\log g)^{k_0} g^{-\lambda})[\delta^+](1,\bar{\mu})\, (\log\bar{\mu})^{\bar{k}}\, \bar{\mu}^{\overline{m}-1} d\bar{\mu}.$$

Remark 5.14. For each $\sigma \in \mathcal{F}$, v is the matrix of minimal generators of σ, $\alpha \in \mathcal{P}(f)$ is the vertex dual to σ, $m = v(-\lambda\alpha+\tau+1)$ and $g(\mu) = f(\mu^v)\mu^{-v\alpha}$ is the strict transform of $f(\omega)$ in the affine chart U_σ. For each $\delta \in D(m,\theta)$, δ^+ is the vector annotated according to the positive entries of $m+\delta$, $\bar{\mu}, \overline{m}, \bar{k}$ are subvectors of μ, m, k selected by this annotation, and d_0, d_+ are the number of zero, positive entries in $m+\delta$. In $I(\sigma,\delta,k)$, $((\log g)^{k_0} g^{-\lambda})[\delta^+]$ is a subseries of the power series $(\log g)^{k_0} g^{-\lambda}$. If $d_+ = 0$, this integral equals $((\log g)^{k_0} g^{-\lambda})[\delta^+](1)$. □

Proof. The formula for $c_{\lambda,t}$ comes from Theorem 3.16. As for the formula for $d_{\lambda,\theta}$, we follow the proof of Theorem 5.11 and compute the coefficient of $(z-\lambda)^{-\theta}$ in

$$\sum_{\sigma\in\mathcal{F}} \int_{[0,1]^d} g(\mu)^{-z} \mu^{m(z)-1} d\mu$$

where $m(z) = v(-z\alpha+\tau+1)$. Because $g(\mu)$ is positive over $[0,1]^d$, the function $g(\mu)^{-z}$ is well-defined for all $z \in \mathbb{R}$ and has a power series expansion

$$g(\mu)^{-z} = \sum_{\delta\in\mathbb{Z}^d_{\geq 0}} g_\delta(z)\, \mu^\delta \tag{5.5}$$

where the coefficients $g_\delta(z)$ do not have poles in $\mathbb{R}$. The zeta function now becomes

$$\sum_{\sigma\in\mathcal{F}} \sum_{\delta\in\mathbb{Z}^d_{\geq 0}} g_\delta(z) \int_{[0,1]^d} \mu^{m(z)+\delta-1} d\mu$$

$$= \sum_{\sigma\in\mathcal{F}} \sum_{\delta\in\mathbb{Z}^d_{\geq 0}} g_\delta(z) \prod_{i=1}^{d} \frac{1}{(m(z)+\delta)_i}.$$

Let $m = m(\lambda)$. A vector $\delta \in \mathbb{Z}^d_{\geq 0}$ contributes a $(z-\lambda)^{-\theta}$ to the Laurent series if the vector $m+\delta$ has at least θ zero entries. Let D' be the set of such vectors. This set is infinite, but every element $\delta' \in D'$ can be written uniquely as a sum $\delta + \delta''$ where $\delta \in D(m,\theta)$ and δ'' is a nonnegative integer vector satisfying $\delta''_i > 0$ only if $(m+\delta)_i > 0$. With these considerations, we restrict our attention to

$$\begin{aligned}
&\sum_{\sigma\in\mathcal{F}} \sum_{\delta'\in D'} g_{\delta'}(z) \prod_{i=1}^{d} \frac{1}{(m(z)+\delta')_i} \\
&= \sum_{\sigma\in\mathcal{F}} \sum_{\delta\in D(m,\theta)} \frac{(-1)^{d_0}}{(z-\lambda)^{d_0}} \frac{\displaystyle\sum_{\delta''\in\mathbb{Z}^{d_+}_{\geq 0}} g_{\delta+\delta''}(z) \prod_{(m+\delta)_i>0} (m(z)+\delta+\delta'')_i^{-1}}{\displaystyle\prod_{(m+\delta)_i=0} (v\alpha)_i \prod_{(m+\delta)_i<0} (m(z)+\delta)_i} \\
&= \sum_{\sigma\in\mathcal{F}} \sum_{\delta\in D(m,\theta)} \frac{(-1)^{d_0}}{(z-\lambda)^{d_0}} \frac{\displaystyle\int_{[0,1]^{d_+}} g^{-z}[\delta^+](1,\bar{\mu})\, \bar{\mu}^{\overline{m(z)}-1} d\bar{\mu}}{\displaystyle\prod_{(m+\delta)_i=0} (v\alpha)_i \prod_{(m+\delta)_i<0} (m(z)+\delta)_i}.
\end{aligned} \tag{5.6}$$

The last equality is a consequence of

$$g^{-z}[\delta^+](1,\bar{\mu}) = \sum_{\delta''\in\mathbb{Z}^{d_+}_{\geq 0}} g_{\delta+\delta''}(z)\, \bar{\mu}^{\overline{\delta+\delta''}}$$

which we derive from the definition (5.5). To find the coefficient of $(z-\lambda)^{-\theta}$ in each summand of (5.6), we need to find the coefficient of $(z-\lambda)^{d_0-\theta}$ in the function

$$h(z) = \frac{\displaystyle\int_{[0,1]^{d_+}} g^{-z}[\delta^+](1,\bar{\mu})\, \bar{\mu}^{\overline{m(z)}-1} d\bar{\mu}}{\displaystyle\prod_{(m+\delta)_i<0} (m(z)+\delta)_i}.$$

The Taylor formula for this coefficient is

$$\frac{1}{(d_0-\theta)!} \frac{\partial^{d_0-\theta}}{\partial z^{d_0-\theta}} h(\lambda)$$

and we apply to $h(z)$ the product rule for derivatives. Observe that we can bring the partial differential operator under the integral operator so

$$\frac{\partial}{\partial z} \int_{[0,1]^{d_+}} g^{-z}[\delta^+](1,\bar{\mu})\, \bar{\mu}^{\overline{m(z)}-1} d\bar{\mu} = \int_{[0,1]^{d_+}} \frac{\partial}{\partial z} g^{-z}[\delta^+](1,\bar{\mu})\, \bar{\mu}^{\overline{m(z)}-1} d\bar{\mu}.$$

It is also not difficult to show that

$$\begin{aligned}\frac{\partial}{\partial z} g^{-z}[\delta^+](1,\bar{\mu}) &= (-(\log g)g^{-z})[\delta^+](1,\bar{\mu}),\\ \frac{\partial}{\partial z}\mu_i^{m(z)_i-1} &= (-v\alpha)_i(\log\mu_i)\,\mu_i^{m(z)_i-1},\\ \frac{\partial}{\partial z}\frac{1}{(m(z)+\delta)_i^l} &= (-v\alpha)_i(-l)\frac{1}{(m(z)+\delta)_i^{l+1}},\quad l\geq 1.\end{aligned}$$

Combining these ideas gives us the desired formula. □

Example 5.15. Let us revisit Example 1.18 and compute higher order asymptotics of the integral (1.28), namely the coefficients $c_{\lambda,t}$ in

$$\int_{[0,1]^2} e^{-NK(u,s)}\,duds \;\approx\; \sum_{\lambda,t} c_{\lambda,t}\,N^{-\lambda}(\log N)^{t-1},\quad N\to\infty$$

where

$$K(u,s)=\frac{1}{2}\log\frac{1}{1+us}+\frac{1}{2}\log\frac{1}{1-us}=-\frac{1}{2}\log(1-u^2s^2).$$

We could write computer software that implements the algorithm in Theorem 5.13. However, for this example, it would be more instructive if we computed the asymptotics by following the proof of the theorem instead.

First, we want to find the poles of the zeta function

$$\zeta(z)=\int_{[0,1]^2} K(u,s)^{-z}\,duds.$$

After observing that

$$K(u,s)=\frac{u^2s^2}{2}\left(1+\frac{u^2s^2}{2}+\frac{u^4s^4}{3}+\cdots\right),$$

let us compute the power series expansion of $K(u,s)^{-z}$. The first term is $(u^2s^2/2)^{-z}$. Using Newton's generalized binomial theorem, every other term in the expansion is of the form

$$\left(\frac{u^2s^2}{2}\right)^{-z}\binom{-z}{\alpha_1}\binom{\alpha_1}{\alpha_2}\cdots\binom{\alpha_{k-1}}{\alpha_k}\left(\frac{u^2s^2}{2}\right)^{\alpha_1-\alpha_2}\left(\frac{u^4s^4}{3}\right)^{\alpha_2-\alpha_3}\cdots\left(\frac{u^{2k}s^{2k}}{k+1}\right)^{\alpha_k}$$

for some $k>0$, some integer vector $\alpha\in\mathbb{Z}^k$ such that $\alpha_1\geq\cdots\geq\alpha_k>0$, and where

$$\binom{-z}{\alpha_1}=\frac{(-z)(-z-1)\cdots(-z-\alpha_1+1)}{\alpha_1!}.$$

This means that for $\delta > 0$, the coefficient of the term $(us)^{-2z+2\delta}$ in this expansion is

$$h_\delta(z) = 2^z \sum_{\alpha \vdash \delta} \frac{\binom{-z}{\alpha_1}\binom{\alpha_1}{\alpha_2}\cdots\binom{\alpha_{k-1}}{\alpha_k}}{2^{\alpha_1-\alpha_2}3^{\alpha_2-\alpha_3}\cdots(k+1)^{\alpha_k}}$$

where we sum over all decreasing partitions α of δ, i.e. $\alpha_1 \geq \cdots \geq \alpha_k > 0$ and $\alpha_1+\cdots+\alpha_k = \delta$ for some $k > 0$. Integrating this term over the domain $[0,1]^2$, we get the contribution

$$\frac{h_\delta(z)}{4}\left(z - \frac{\delta+1}{2}\right)^{-2}$$

to the zeta function $\zeta(z)$. Now, let $\lambda = (\delta+1)/2$. It follows that the Laurent coefficient of $(z-\lambda)^{-2}$ is $d_{\lambda,2} = h_\delta(\lambda)/4$ while that of $(z-\lambda)^{-1}$ is $d_{\lambda,1} = h'_\delta(\lambda)/4$. Explicitly,

$$h'_\delta(z) = 2^z \sum_{\alpha \vdash \delta} \left(\frac{1}{z} + \frac{1}{z+1} + \cdots + \frac{1}{z+\alpha_1-1} + \frac{1}{\log 2}\right) \frac{\binom{-z}{\alpha_1}\binom{\alpha_1}{\alpha_2}\cdots\binom{\alpha_{k-1}}{\alpha_k}}{2^{\alpha_1-\alpha_2}3^{\alpha_2-\alpha_3}\cdots(k+1)^{\alpha_k}}.$$

Finally, by Theorem 3.16, we have the asymptotic coefficients

$$\begin{aligned} c_{\lambda,2} &= \Gamma(\lambda)\, d_{\lambda,2}, \\ c_{\lambda,1} &= -\Gamma(\lambda)\, d_{\lambda,1} - \Gamma^{(1)}(\lambda)\, d_{\lambda,2}. \end{aligned}$$

This gives us the following closed-form expressions

$$\begin{aligned} c_{\lambda,2} &= 2^{\lambda-2}\,\Gamma(\lambda) \sum_{\alpha \vdash 2\lambda-1} \frac{\binom{-\lambda}{\alpha_1}\binom{\alpha_1}{\alpha_2}\cdots\binom{\alpha_{k-1}}{\alpha_k}}{2^{\alpha_1-\alpha_2}3^{\alpha_2-\alpha_3}\cdots(k+1)^{\alpha_k}} \\ c_{\lambda,1} &= -2^{\lambda-2}\,\Gamma(\lambda) \sum_{\alpha \vdash 2\lambda-1} \frac{\binom{-\lambda}{\alpha_1}\binom{\alpha_1}{\alpha_2}\cdots\binom{\alpha_{k-1}}{\alpha_k}}{2^{\alpha_1-\alpha_2}3^{\alpha_2-\alpha_3}\cdots(k+1)^{\alpha_k}} H(2\lambda+2\alpha_1-2) \end{aligned}$$

where

$$H(k) = \begin{cases} \dfrac{1}{\log 2} - \gamma + 2\left(\dfrac{1}{2} + \dfrac{1}{4} + \cdots + \dfrac{1}{k}\right) & \text{if } k \text{ even}, \\ \dfrac{1}{\log 2} - \gamma + 2\left(\dfrac{1}{1} + \dfrac{1}{3} + \cdots + \dfrac{1}{k} - \log 2\right) & \text{if } k \text{ odd}. \end{cases}$$

and γ is the Euler-Mascheroni constant

$$\gamma = \lim_{n\to\infty}\left(\sum_{k=1}^{n}\frac{1}{k} - \log n\right) \approx 0.5772156649.$$

The above formulas allow us to compute the first few values of $c_{\lambda,t}$ easily:

$$\begin{aligned}
c_{\frac{1}{2},2} &= \sqrt{\frac{\pi}{8}}, & c_{\frac{1}{2},1} &= -\sqrt{\frac{\pi}{8}}\left(\frac{1}{\log 2} - 2\log 2 - \gamma\right), \\
c_{1,2} &= -\frac{1}{4}, & c_{1,1} &= \frac{1}{4}\left(\frac{1}{\log 2} + 1 - \gamma\right), \\
c_{\frac{3}{2},2} &= -\frac{\sqrt{2\pi}}{128}, & c_{\frac{3}{2},1} &= \frac{\sqrt{2\pi}}{128}\left(\frac{1}{\log 2} - 2\log 2 - \frac{10}{3} - \gamma\right), \\
c_{2,2} &= 0, & c_{2,1} &= -\frac{1}{24}.
\end{aligned}$$

Using these coefficients, we can get better approximations of the singular marginal likelihood integral (1.28) when the sample size N is large. □

In this chapter, we investigated algebraic methods for computing higher order asymptotics of Laplace integrals. Our contributions come in two flavors. The first flavor consists of sum, product and chain rules satisfied by leading coefficients of the asymptotics. These new results parallel those described in Section 4.1.3 for computing RLCTs of ideals. The second flavor is concerned with the asymptotics of nondegenerate functions. We gave explicit formulas for leading coefficients in Theorem 5.11 and for higher order coefficients in Theorem 5.13. We ended the chapter with an example where we computed such coefficients for the marginal likelihood integral of a discrete statistical model.

Bibliography

[1] Steen A. Andersson, David Madigan, and Michael D. Perlman. Alternative Markov properties for chain graphs. *Scand. J. Statist.*, 28(1):33–85, 2001.

[2] Miki Aoyagi and Sumio Watanabe. Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Netw.*, 18:924–933, September 2005.

[3] V. I. Arnol′d, S. M. Guseĭn-Zade, and A. N. Varchenko. *Singularities of differentiable maps. Vol. II*, volume 83 of *Monographs in Mathematics*. Birkhäuser Boston Inc., Boston, MA, 1988. Monodromy and asymptotics of integrals, Translated from the Russian by Hugh Porteous, Translation revised by the authors and James Montaldi.

[4] M. F. Atiyah. Resolution of singularities and division of distributions. *Comm. Pure Appl. Math.*, 23:145–150, 1970.

[5] Edward Bierstone and Pierre D. Milman. Local resolution of singularities. In *Real analytic and algebraic geometry (Trento, 1988)*, volume 1420 of *Lecture Notes in Math.*, pages 42–64. Springer, Berlin, 1990.

[6] Edward Bierstone and Pierre D. Milman. Resolution of singularities. In *Several complex variables (Berkeley, CA, 1995–1996)*, volume 37 of *Math. Sci. Res. Inst. Publ.*, pages 43–78. Cambridge Univ. Press, Cambridge, 1999.

[7] Carles Bivià-Ausina. Nondegenerate ideals in formal power series rings. *Rocky Mountain J. Math.*, 34(2):495–511, 2004.

[8] Manuel Blickle and Robert Lazarsfeld. An informal introduction to multiplier ideals. In *Trends in commutative algebra*, volume 51 of *Math. Sci. Res. Inst. Publ.*, pages 87–114. Cambridge Univ. Press, Cambridge, 2004.

[9] Gábor Bodnár and Josef Schicho. Automated resolution of singularities for hypersurfaces. *J. Symbolic Comput.*, 30(4):401–428, 2000.

[10] Gábor Bodnár and Josef Schicho. A computer program for the resolution of singularities. In *Resolution of singularities (Obergurgl, 1997)*, volume 181 of *Progr. Math.*, pages 231–238. Birkhäuser, Basel, 2000.

[11] Ana María Bravo, Santiago Encinas, and Orlando Villamayor U. A simplified proof of desingularization and applications. *Rev. Mat. Iberoamericana*, 21(2):349–458, 2005.

[12] David A. Cox, John Little, and Donal O'Shea. *Using algebraic geometry*, volume 185 of *Graduate Texts in Mathematics*. Springer, New York, second edition, 2005.

[13] A. P. Dempster. Covariance selection. *Biometrics*, 28(1):pp. 157–175, 1972.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977. With discussion.

[15] Mathias Drton. Likelihood ratio tests and singularities. *Ann. Statist.*, 37(2):979–1012, 2009.

[16] Mathias Drton, Bernd Sturmfels, and Seth Sullivant. *Lectures on algebraic statistics*, volume 39 of *Oberwolfach Seminars*. Birkhäuser Verlag, Basel, 2009.

[17] David Eisenbud. *Commutative algebra*, volume 150 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995. With a view toward algebraic geometry.

[18] Michael J. Evans, Zvi Gilula, and Irwin Guttman. Latent class analysis of two-way contingency tables by Bayesian methods. *Biometrika*, 76(3):557–563, 1989.

[19] Stephen E. Fienberg, Patricia Hersh, Alessandro Rinaldo, and Yi Zhou. Maximum likelihood estimation in latent class models for contingency table data. In *Algebraic and geometric methods in statistics*, pages 27–62. Cambridge Univ. Press, Cambridge, 2010.

[20] Anne Frühbis-Krüger and Gerhard Pfister. Algorithmic resolution of singularities. In *Singularities and computer algebra*, volume 324 of *London Math. Soc. Lecture Note Ser.*, pages 157–183. Cambridge Univ. Press, Cambridge, 2006.

[21] William Fulton. *Introduction to toric varieties*, volume 131 of *Annals of Mathematics Studies*. Princeton University Press, Princeton, NJ, 1993. The William H. Roever Lectures in Geometry.

[22] Shuhong Gao, Guangran Jiang, and Mingfu Zhu. Solving the 100 swiss francs problem. arXiv:0809.4627, 2008.

[23] Luis D. Garcia-Puente, Sarah Spielvogel, and Seth Sullivant. Identifying causal effects with computer algebra. In *Proceedings of the 26th Conference of Uncertainty in Artificial Intelligence*. AUAI Press, 2010.

[24] Russell A. Goward, Jr. A simple algorithm for principalization of monomial ideals. *Trans. Amer. Math. Soc.*, 357(12):4805–4812 (electronic), 2005.

[25] Michael Greenblatt. An elementary coordinate-dependent local resolution of singularities and applications. *J. Funct. Anal.*, 255(8):1957–1994, 2008.

[26] Michael Greenblatt. Resolution of singularities, asymptotic expansions of integrals and related phenomena. *Journal d'Analyse Mathmatique*, 111:221–245, 2010.

[27] Dominique M. A. Haughton. On the choice of a model to fit data from an exponential family. *Ann. Statist.*, 16(1):342–355, 1988.

[28] Herwig Hauser. The Hironaka theorem on resolution of singularities (or: A proof we always wanted to understand). *Bull. Amer. Math. Soc. (N.S.)*, 40(3):323–403 (electronic), 2003.

[29] Heisuke Hironaka. Resolution of singularities of an algebraic variety over a field of characteristic zero. I, II. *Ann. of Math. (2) 79 (1964), 109–203; ibid. (2)*, 79:205–326, 1964.

[30] Serkan Hoşten, Amit Khetan, and Bernd Sturmfels. Solving the likelihood equations. *Found. Comput. Math.*, 5(4):389–407, 2005.

[31] J. A. Howald. Multiplier ideals of monomial ideals. *Trans. Amer. Math. Soc.*, 353(7):2665–2671 (electronic), 2001.

[32] Michael I. Jordan. Graphical models. *Statist. Sci.*, 19(1):140–155, 2004.

[33] János Kollár. Singularities of pairs. In *Algebraic geometry—Santa Cruz 1995*, volume 62 of *Proc. Sympos. Pure Math.*, pages 221–287. Amer. Math. Soc., Providence, RI, 1997.

[34] János Kollár. *Lectures on resolution of singularities*, volume 166 of *Annals of Mathematics Studies*. Princeton University Press, Princeton, NJ, 2007.

[35] Steffen L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York, 1996. Oxford Science Publications.

[36] Robert Lazarsfeld. *Positivity in algebraic geometry. I, II*, volume 48, 49 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer-Verlag, Berlin, 2004.

[37] Shaowei Lin. Asymptotic approximation of marginal likelihood integrals. arXiv:1003.5338, 2010.

[38] Shaowei Lin, Bernd Sturmfels, and Zhiqiang Xu. Marginal likelihood integrals for mixtures of independence models. *J. Mach. Learn. Res.*, 10:1611–1631, 2009.

[39] David Maxwell Chickering and David Heckerman. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Mach. Learn.*, 29:181–212, November 1997.

[40] J. Milnor. *Morse theory*. Based on lecture notes by M. Spivak and R. Wells. Annals of Mathematics Studies, No. 51. Princeton University Press, Princeton, N.J., 1963.

[41] Lior Pachter and Bernd Sturmfels, editors. *Algebraic statistics for computational biology*. Cambridge University Press, New York, 2005.

[42] Robin Pemantle and Mark C. Wilson. Asymptotic expansions of oscillatory integrals with complex phase. In *Algorithmic probability and combinatorics*, volume 520 of *Contemp. Math.*, pages 221–240. Amer. Math. Soc., Providence, RI, 2010.

[43] Giovanni Pistone, Eva Riccomagno, and Henry P. Wynn. *Algebraic statistics*, volume 89 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL, 2001. Computational commutative algebra in statistics.

[44] Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *Ann. Statist.*, 30(4):962–1030, 2002.

[45] Dmitry Rusakov and Dan Geiger. Asymptotic model selection for naive Bayesian networks. *J. Mach. Learn. Res.*, 6:1–35 (electronic), 2005.

[46] Marcelo J. Saia. The integral closure of ideals and the Newton filtration. *J. Algebraic Geom.*, 5(1):1–11, 1996.

[47] Morihiko Saito. On real log canonical thresholds. arXiv:0707.2308, 2007.

[48] Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

[49] Richard P. Stanley. A zonotope associated with graphical degree sequences. In *Applied geometry and discrete mathematics*, volume 4 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pages 555–570. Amer. Math. Soc., Providence, RI, 1991.

[50] Bernd Sturmfels. Open problems in algebraic statistics. In *Emerging applications of algebraic geometry*, volume 149 of *IMA Vol. Math. Appl.*, pages 351–363. Springer, New York, 2009.

[51] Seth Sullivant, Kelli Talaska, and Jan Draisma. Trek separation for Gaussian graphical models. *Ann. Statist.*, 38(3):1665–1685, 2010.

[52] Bernard Teissier. Monomial ideals, binomial ideals, polynomial ideals. In *Trends in commutative algebra*, volume 51 of *Math. Sci. Res. Inst. Publ.*, pages 211–246. Cambridge Univ. Press, Cambridge, 2004.

[53] Caroline Uhler. Geometry of maximum likelihood estimation in Gaussian graphical models. arXiv:1012.2643, 2010.

[54] Alexander N. Varchenko. Newton polyhedra and estimation of oscillating integrals. *Funct. Anal. Appl.*, 10:175–196, 1976.

[55] V. A. Vasil′ev. Asymptotic behavior of exponential integrals in the complex domain. *Funktsional. Anal. i Prilozhen.*, 13(4):1–12, 96, 1979.

[56] Martin J Wainwright and Michael I Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., Hanover, MA, USA, 2008.

[57] Sumio Watanabe. Algebraic analysis for nonidentifiable learning machines. *Neural Comput.*, 13:899–933, April 2001.

[58] Sumio Watanabe. *Algebraic geometry and statistical learning theory*, volume 25 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2009.

[59] Jarosław Włodarczyk. Simple Hironaka resolution in characteristic zero. *J. Amer. Math. Soc.*, 18(4):779–822 (electronic), 2005.

[60] Keisuke Yamazaki and Sumio Watanabe. Singularities in mixture models and upper bounds of stochastic complexity. *Neural Netw.*, 16:1029–1038, September 2003.

[61] Keisuke Yamazaki and Sumio Watanabe. Newton diagram and stochastic complexity in mixture of binomial distributions. In *Algorithmic learning theory*, volume 3244 of *Lecture Notes in Comput. Sci.*, pages 350–364. Springer, Berlin, 2004.

[62] Günter M. Ziegler. *Lectures on polytopes*, volume 152 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995.

[63] Piotr Zwiernik. An asymptotic approximation of the marginal likelihood for general markov models. arXiv:1012.0753, 2010.