ETH Zurich

D-ITET

Semester project

Nonidentifiability in Neural Networks

Student: Stefan Stojanović *

Supervisor: Erwin Riegler †

Professor: Helmut Bölcskei †

June, 2021

Abstract. The aim of this project is to give a concise overview of statistical learning with singular models. Many of the commonly used statistical models such as layered neural networks, mixture models, hidden Markov models etc. are singular, i.e. nonidentifiable or without positive definite metric. On the other hand, asymptotic results (with respect to the number of data points) based on central limit theorem do not necessarily hold for singular models. Hence, some of the results of regular statistical learning, including asymptotic expansion of free energy, generalization and training errors are not valid in the singular case. Moreover, some of the criteria for model selection (BIC, AIC, etc) are not suitable for singular models. Comparison of these two types of models in terms of asymptotic behavior gives one of possible explanations for dominance of singular over regular models. Furthermore, theoretical understanding of singular models can lead to better practical implementations of learning algorithms. In order to present results which are valid both for regular and singular statistical models we directly follow the work of Prof. Watanabe [1] with special emphasis on layered neural networks.

The main idea of this theory is to examine singular models through the lens of algebraic geometry. Instead of a single point solution as in the regular case, in singular models the optimal set of parameters forms an analytic set with singularities. One of the possible frameworks for studying these sets is provided by algebraic geometry. The key result used is Hironaka's theorem of resolution of singularities. Quite remarkably, this theorem will provide us with enough insight into the singularities of the parameter space in order to determine asymptotic behavior of variables significant for statistical learning theory.

This paper is organized as follows. In chapter 1 we give an introduction to singular statistical learning while comparing it with learning in regular models. In chapter 2 we introduce main concepts from algebraic geometry that are needed for reaching our goal. This discussion is continued in chapter 3 where our main topic is Hironaka's theorem of resolution of singularities. Chapter 4 connects results established in algebraic geometry with statistical learning theory, and introduces results that are valid both for regular and singular models. Finally, in chapter 5 we consider neural networks as prototypes of singular models, and examine asymptotic behavior of single hidden layer neural networks with exponential activation functions.

*ETH Zurich, Switzerland (email: sstojanovic@student.ethz.ch)

†Chair for Mathematical Information Science, ETH Zurich, Switzerland (email: eriegler@mins.ee.ethz.ch, hboelcskei@ethz.ch)

Contents

1 Introduction

In recent decades there has been an immense progress in utilizing statistical models to a vast number of applications. Meanwhile there have also been attempts to examine their properties from theoretical point of view. What has been observed is that some of the most common used statistical models such as layered neural networks, Bayes networks, normal mixtures, reduced rank regressions, hidden Markov models, etc. have some common properties that differentiate them from the rest of the models. Namely, all of these models have some type of hierarchical structure or hidden variables. More abstractly, all of these models belong to a group of statistical models called singular models.

Because of their unique properties, singular models cannot be analyzed using tools from standard theory of statistical learning. In the following sections we will give an overview of an "alternative" theory of statistical learning based on a book by Professor Sumio Watanabe [1] that practically founded this field. Although this theory covers nonsingular models as well, level of sophistication needed for obtaining analogous results for singular models is higher. Therefore we will first provide reader with the background knowledge needed for stating main results of singular learning theory. In the following subsections we start with general results from statistical learning theory and show main differences between singular and nonsingular models.

1.1 Statistical learning

We will consider a general system which produces output y from output space Y given some input x from input space X. Based on observations of input-output pairs $(x, y) \in X \times Y$ and our prior knowledge, we can model the system using some input-output relation $f : X \to Y$. The aim of learning theory is to find *the best* model using available data. In this overview we will consider only parametric statistical models i.e. by a statistical model we will mean a conditional probability distribution $p(x, y|w)$ ($x \in X$, $y \in Y$) parametrized by parameter $w \in W$, where $W \subset \mathbb{R}^d$ for some natural number d.

Let $Z = X \times Y$ and $z^{(n)} \in Z^n = \overbrace{Z \times Z \times \cdots \times Z}^{n}$ be a set of identically distributed and independently sampled input-output pairs $z^{(n)} = \{(x_1, y_1),\ (x_2, y_2), \ldots,\ (x_n, y_n)\}$. Instead of modeling the system deterministically as a relation between input and output data points, we aim to model input and output as random variables with some probability distribution functions. Let's denote by $q(x, y)$ the joint probability distribution of input $x \in X$ and output $y \in Y$. Furthermore, let $q(x)$ be true marginal distribution of input, and $q(y|x) = \frac{q(x,y)}{q(x)}$ (for $x :\ q(x) > 0$) conditional distribution of output.

We will assume that the input x does not depend on the parameters of the model w, and therefore have that $p(x, y|w) = p(y|x, w)p(x|w) = p(y|x, w)p(x)$. Since $p(x)$ does not depend on parameter w, defining statistical model by distribution $p(x, y|w)$ or $p(y|x, w)$ is equivalent. Hence, we will mainly use the latter. For given data set $z^{(n)}$, model $p(x, y|w)$ and prior distribution of model $\varphi(w)$,

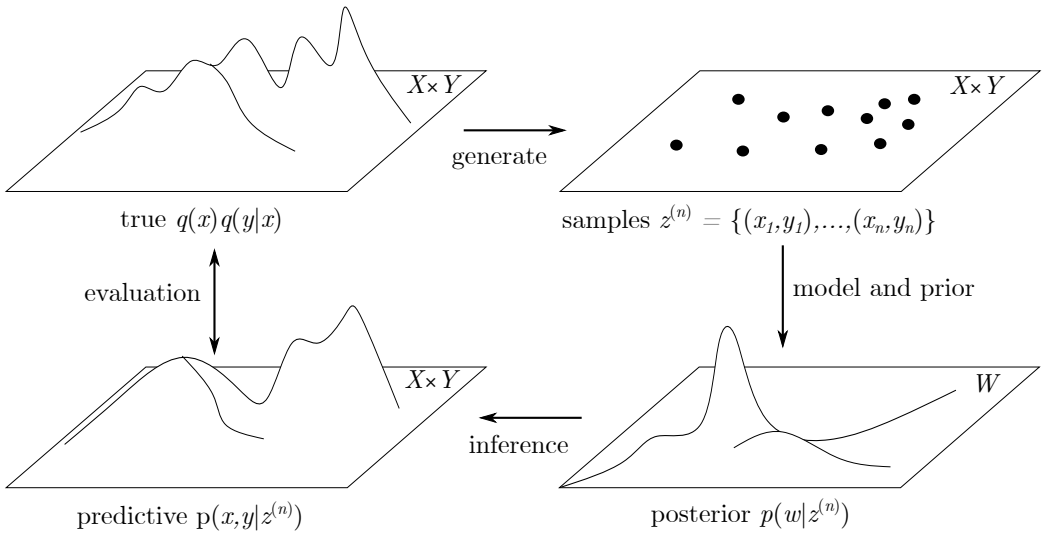


Figure 1: Learning process - obtaining predictive distribution based on data and prior knowledge. Figure inspired by Fig. 1.1. in [2].

posterior distribution is defined by:

$$p(w|z^{(n)}) = \frac{\varphi(w)\prod_{i=1}^{n} p(x_i, y_i|w)}{\int_W \varphi(w)\prod_{i=1}^{n} p(x_i, y_i|w)dw} = \frac{\varphi(w)\prod_{i=1}^{n} p(y_i|x_i, w)}{\int_W \varphi(w)\prod_{i=1}^{n} p(y_i|x_i, w)dw} = \frac{1}{Z_n}\varphi(w)\prod_{i=1}^{n} p(y_i|x_i, w) \tag{1}$$

where $Z_n \triangleq \int_W \varphi(w)\prod_{i=1}^{n} p(y_i|x_i, w)dw$ is the partition function. By examining posterior distribution (1) we could deduce which values of parameters are more likely to produce given data points and be in accordance with our prior knowledge given by $\varphi(w)$. The ultimate goal of statistical learning is to obtain a learning algorithm $z^{(n)} \mapsto p(y|x, z^{(n)})$, with $p(y|x, z^{(n)})$ being predictive probability density function which should approximate true distribution $q(y|x)$. Predictive (estimated, trained) distribution is computed from posterior distribution according to:

$$p(y|x, z^{(n)}) = \int_W p(y, w|x, z^{(n)})dw = \int_W p(y|x, w, z^{(n)})p(w|x, z^{(n)})dw = \int_W p(y|x, w)p(w|z^{(n)})dw \tag{2}$$

where we used that parameters w depend only on data points, and dependence of predictive distribution on data only through parameters of the model w. We will also define expectation with respect to $p(w|z^{(n)})$ as $\mathbb{E}_w[f(w)] := \int_W f(w)p(w|z^{(n)})dw$ and therefore write $p(y|x, z^{(n)}) = \mathbb{E}_w[p(y|x, w)]$. Once we determine the predictive distribution, we can use statistical criteria for evaluation of the obtained model. Based on achieved results, we can redefine model and/or prior in order to obtain more accurate predictions. This repeating cycle of statistical learning is shown in the Figure 1.

We are mainly interested in asymptotic behavior of statistical variables i.e. their expansion when the number of data points $n \to \infty$. In order to develop adequate theory, we will assume that true distribution $q(x, y) = q(x)q(y|x)$ is known. Even though this assumption makes results somewhat less applicable, they will provide intuition for behavior of statistical models. Moreover, this theory can be applied to form criteria for singular model selection which do not depend on the true distribution [3] and hence are useful for practical purposes.

Although we have introduced partition function Z_n as a normalization factor in equation (1), its significance for statistical learning is much higher. Namely, Z_n can be understood as a likelihood of the pair $p(y|x, w)$ and $\varphi(w)$, i.e. it shows how appropriate chosen model and prior are for given data set. We will see in section 4 that from asymptotic expansion of partition function, we can deduce asymptotic expansion of all variables that will be of our interest. This remark is in accordance with observations from statistical physics where, according to a course book[1], Z_n is the most important quantity and contains all information about the system.

In this paper we will focus our attention to statistical models called neural networks that are used mainly for supervised learning. Hence we will usually use models of the form $p(y|x, w)$. However, obtained results can be applied also to unsupervised learning with aim of predicting $p(x|w)$. In some sections (e.g. in section 1.4) we will prove the results for $p(x|w)$ in order to make notation less cumbersome. The same steps could be used for proving theorems in supervised case. Results we present are strongly motivated by [1, 2].

1.2 Neural networks

One of the commonly used model class in supervised learning are neural networks. Parameters of a neural network are weights and biases of the neurons, and for a network with d parameters the parameter space W is subset of $\mathbb{R}^d$. Let's now define a general neural network.

Definition 1 (Neural network, definition II.1 in [4])**.** *A neural network of length $L \in \mathbb{N}$, with layers of width $N_0, N_1, \ldots, N_L \in \mathbb{N}$ and a nonlinear activation function $\rho : \mathbb{R} \to \mathbb{R}$ is a map $\Phi : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ defined by:*

$$\Phi = \begin{cases} W_1, & L = 1 \\ W_2 \circ \rho \circ W_1, & L = 2 \\ W_L \circ \rho \circ W_{L-1} \circ \rho \circ \ldots \circ \rho \circ W_1, & L \geq 3 \end{cases} \tag{3}$$

where $W_l : \mathbb{R}^{N_{l-1}} \to \mathbb{R}^{N_l}$ with $W_l(x) = A_l x + b_l$, ($A_l \in \mathbb{R}^{N_l \times N_{l-1}}$, $b_l \in \mathbb{R}^{N_l}$) for $l = 1, 2, \ldots, L$ are the affine transformations, and nonlinear activation function acts component-wise i.e. $\rho(x) = [\rho(x_1), \rho(x_2), \ldots, \rho(x_k)]$ for $x = [x_1, x_2, \ldots, x_k] \in \mathbb{R}^k$.

[1] David Tong, *Statistical Physics*, University of Cambridge, link: https://www.damtp.cam.ac.uk/user/tong/statphys/sp.pdf

The choice of hyperparameters of the network such as length and widths of the layers, but also the choice of activation function, is governed by the application of interest. In the Figure 2 (right) are shown some of the commonly used activation functions.

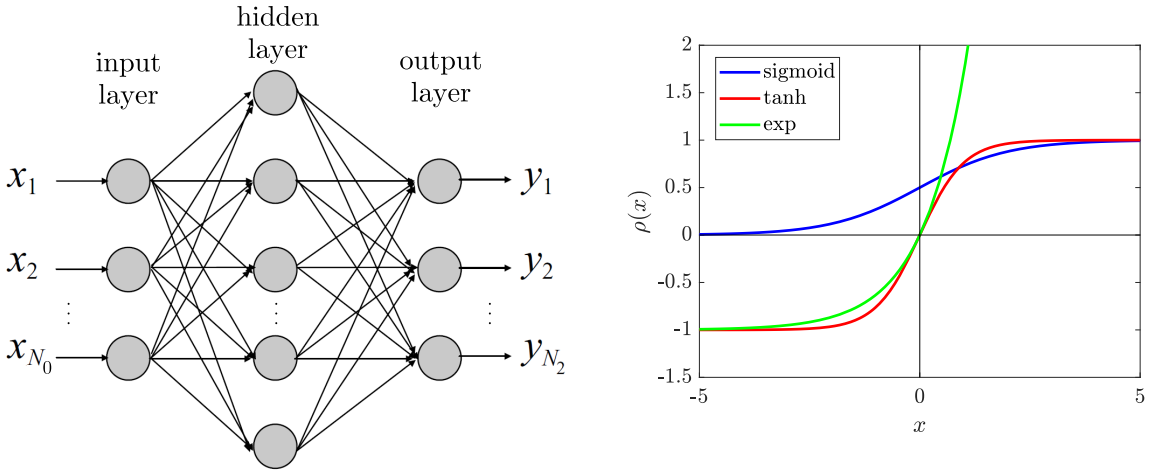


Figure 2: Left: architecture of single hidden layer neural network (adapted from Fig. 7.2 in [1]). Right: different nonlinearities used as activation functions in neural networks.

In order to gain more intuition of the neural networks, and make proofs and examples less cumbersome (and sometimes even only solvable), we will mainly use single hidden layer neural network, i.e. neural network of length $L = 2$. As shown in Figure 2 (left) these neural networks consist of input, hidden and output layer. Single hidden layer networks have the following form:

$$[\Phi(x)]_k = \sum_{i=1}^{N_1} a_{k,i}^{(2)} \rho\Big(\sum_{j=1}^{N_0} a_{i,j}^{(1)} x_j + b_i^{(1)} \Big) + b_k^{(2)}, \quad k = 1, 2, \ldots, N_2 \tag{4}$$

Example 1. *For a neural network Φ with scalar output ($N_L = 1$), we define a statistical model with parameters w (weights and biases) as follows:*

$$p(y|x, w) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}\big(y - \Phi(x)\big)^2 \Big) \tag{5}$$

By using this model, we assume that output is a random variable with normal distribution, whose expectation is given by $\Phi(x)$ and variance is unitary. Parameters w of the neural network determine the neural network Φ and we will use notation Φ_w to explicitly denote this relation.

1.3 Kullback-Leibler distance

In order to quantify how well a predictive distribution $p(y|x, z^{(n)})$ or model $p(y|x, w)$ approximate true distribution $q(y|x)$ we will measure distance between probability distributions using Kullback-Leibler distance.

Definition 2 (Kullback-Leibler distance)**.** *Let $q(x)$ and $p(x)$ be probability distributions of a continuous random variable defined on an open set $X \subseteq \mathbb{R}^N$. Then Kullback-Leibler (KL-) distance is given by:*

$$K(q||p) = \int_X q(x) \log \frac{q(x)}{p(x)} dx \tag{6}$$

where we use convention that $K(q||p) = \infty$ if $q(x) \neq 0$, $p(x) = 0$ for some $x \in X$ and $0 \log \frac{0}{0} = 0$.

Although it does not form a metric space, Kullback-Leibler distance is useful for quantifying similarity of two probability distributions. The fact that it is equal to zero if and only if two distributions are equal will be an important fact in the following discussion, and hence we give a short proof now.

Theorem 1. *Assume that $q(x)$ and $p(x)$ are as in definition 2. Then $K(q||p) \geq 0$ and equality holds if and only if $q = p$ almost everywhere on X.*

Proof. Note first that $\log y \leq y-1$, $\forall y \in \mathbb{R}, y>0$. Using this inequality and assumption that $p(x)$ and $q(x)$ are supported on X one has:

$$-K(q||p) = \int_X q(x) \log \frac{p(x)}{q(x)} dx \leq \int_X q(x) \Big(\frac{p(x)}{q(x)} - 1 \Big) dx = \int_X p(x)dx - \int_X q(x)dx = 1 - 1 = 0$$

Therefore $K(q||p) \geq 0$. If $q = p$ on X i.e. $q(x) = p(x)$, $x \in X$ then obviously $K(q||p) = 0$. If $K(q||p) = 0$ then the inequality above must hold with equality for all $x \in X$, i.e:

$$\int_X \underbrace{q(x)}_{\geq 0} \Big(\underbrace{\log \frac{p(x)}{q(x)} - \frac{p(x)}{q(x)} + 1}_{\leq 0} \Big) dx = 0$$

This implies that:

$$\log \frac{p(x)}{q(x)} = \frac{p(x)}{q(x)} - 1, \quad \implies p(x) = q(x), \quad \text{for a.a. } x \in X \text{ with } q(x) > 0$$

Now note that:

$$1 = \int_X p(x)dx = \int_{x \in X: q(x)>0} p(x)dx + \int_{x \in X: q(x)=0} p(x)dx = \underbrace{\int_{x \in X: q(x)>0} q(x)dx}_{=1} + \underbrace{\int_{x \in X; q(x)=0} p(x)dx}_{=0}$$

Hence $p(x) = 0$ for almost all $x \in X$ with $q(x) = 0$. We conclude that $p(x) = q(x)$ for almost all $x \in X$. □

In theory of statistical learning, given an estimated probability distribution p and true probability distribution q one define so-called generalization error of statistical model as $K(q||p)$. In the case of estimated probability distributions parametrized by some parameter $w \in W$ we write:

$$K(w) = K(q(x,y)||p(x,y|w)) = \int q(x,y) \log \frac{q(x,y)}{p(x,y|w)} dxdy = \int q(x,y) \log \frac{q(y|x)}{p(y|x,w)} dxdy \tag{7}$$

Since this quantity measures dissimilarity between true distribution $q(x,y)$ and modeled distribution $p(x,y|w)$, we seek to minimize $K(q||p)$ i.e. find density functions p (or only parameter w) such that they are as close to true distribution q as possible. Note that Kullback-Leibler distance in general is not symmetric: $K(q||p) \neq K(p||q)$. Hence it is important to determine distance $K(q||p)$ which measures expectation of dissimilarity of p and q but with respect to the true distribution q.

Let's now determine the Kullback-Leibler distance of a neural network Φ with conditional distribution as assumed in example 1.

Example 2. *Let $p(y|x,w)$ be given by equation (5) and let $w_0 \in W$ be true parameter i.e. assume $q(y|x) = p(y|x,w_0)$. Then Kullback-Leibler distance of true distribution $q(y|x)$ and model distribution $p(y|x,w)$ is given by:*

$$\begin{aligned}
K(w) &= \int q(x,y) \log \frac{q(y|x)}{p(y|x,w)} dxdy = \int q(x,y) \log \frac{\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(y - \Phi_{w_0}(x))^2)}{\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(y - \Phi_w(x))^2)} dxdy \\
&= \int q(x,y) \Big[\frac{1}{2}(\Phi_w^2(x) - \Phi_{w_0}^2(x)) + y(\Phi_{w_0}(x) - \Phi_w(x)) \Big] dxdy \\
&= \frac{1}{2} \int q(x)(\Phi_w^2(x) - \Phi_{w_0}^2(x))dx + \int q(x)(\Phi_{w_0}(x) - \Phi_w(x)) \underbrace{(\int q(y|x) y dy)}_{\mathbb{E}_{q(y|x)}(y) = \Phi_{w_0}(x)} dx \\
&= \frac{1}{2} \int q(x)(\Phi_w(x) - \Phi_{w_0}(x))^2 dx = \frac{1}{2} \mathbb{E}[(\Phi_w(x) - \Phi_{w_0}(x))^2]
\end{aligned} \tag{8}$$

Example 3. *Let's now consider a simple example with a line regression model of the following form:*

$$p(y|x,a,b) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2} \Big(y - (ax+b) \Big)^2 \Big) \tag{9}$$

with parameters $w = [a\ b] \in \mathbb{R}^2$. Assuming that $q(x)$ is standard normal distribution, it is straightforward to obtain expression for the KL-distance in the following form:

$$K(a,b) = \frac{1}{2}\int q(x)(ax+b-(a_0x+b_0))^2dx = \frac{1}{2}(a-a_0)^2 + \frac{1}{2}(b-b_0)^2 \tag{10}$$

It is evident that $K(a,b) = 0$ if and only if $a = a_0$ and $b = b_0$ implying that minimization of $K(a,b)$ leads to obtaining the true parameters (a_0, b_0).

Example 4. *Assume that neural network has the following form: $\Phi_w(x) = a\rho(bx) = a(e^{bx}-1)$ with $w = [a,b]$, $a, b \in \mathbb{R}$, and let true parameters be given by $w_0 = [0,0]$. Then according to the example 2, the KL-distance is given by:*

$$K(a,b) = \frac{1}{2}\int q(x)(a(e^{bx}-1))^2dx \tag{11}$$

and assuming that $q(x)$ is standard normal distribution we get:

$$K(a,b) = \frac{a^2}{2}\Big(e^{2b^2} - 2e^{b^2/2} + 1\Big) \tag{12}$$

Note that $K(a,b) = 0$ if and only if $a = 0$ or $b = 0$. This implies that we might have $K(a,b) = 0$ but $a \neq a_0 = 0$ or $b \neq b_0 = 0$. This is not in contradiction with theorem 1 since any choice of parameters $(a, 0)$ or $(0, b)$ gives the same probability distribution as (a_0, b_0). As we will show, the problem of not being able to uniquely determine true parameters comes from the fact that model proposed in this example is singular.

In the previous examples we have assumed that true distribution $q(x)$ is standard normal distribution. However, in the following discussion we are interested in results that hold for arbitrary input distribution $q(x)$ and hence will search for parameters which give $K(w) = 0$ independent of $q(x)$. Note that qualitative conclusions made in the previous two examples also hold for arbitrary $q(x)$.

1.4 Fisher information matrix

One of the fundamental results of statistics called Cramer-Rao theorem gives lower bound on variance of an unbiased estimator. This bound is inversely proportional to a matrix called Fisher information matrix. However, in the case of singular models Fisher information matrix is not invertible and therefore the Cramer-Rao theorem and the whole concept of bounding variance of an estimator is not applicable in singular models. In this section we provide reader with theorems that confirm claims made so far.

For results in this section we use probability distributions $p(x|w)$ instead of $p(x,y|w)$ in order to make notation less cumbersome. Conclusions we obtain are with slight modifications valid also for $p(x,y|w)$ since we can redefine x as $x \leftarrow \begin{bmatrix} x \\ y \end{bmatrix}$.

Definition 3 (Fisher information matrix)**.** *Let $p(x|w)$ be a statistical model parametrized by $w \in \mathbb{R}^d$ and $x \in X = \mathbb{R}^N$. The Fisher Information matrix $J(w) = \{J_{j,k}(w)\}$ for $1 \leq j,k \leq d$ is defined by:*

$$J_{j,k}(w) = \int_{\mathbb{R}^N} p(x|w)\frac{\partial}{\partial w_j}\log p(x|w)\frac{\partial}{\partial w_k}\log p(x|w)dx \tag{13}$$

if the integral is finite.

Note that if we define $A(x|w) \in \mathbb{R}^d$ as a row-vector with $\{\frac{\partial}{\partial w_j}\log p(x|w)\}_{j=1}^d$ as its elements, then we can write Fisher information matrix as follows:

$$J(w) = \mathbb{E}_{p(x|w)}[A^T(x|w)A(x|w)] \tag{14}$$

From this quadratic expression for $J(w)$ it is evident that $J(w)$ is symmetric and positive semi-definite matrix.

Lemma 1. *Fisher information matrix of a statistical model $p(x|w)$ is positive definite if and only if $\{\frac{\partial}{\partial w_j}\log p(x|w)\}_{j=1}^d$ is linearly independent as a function of x, for all x with $p(x|w) > 0$.*

Proof. Let Fisher information matrix J be positive definite. Then, for any nonzero $a \in \mathbf{R}^d$:

$$0 < a^T J(w) a = a^T \mathbb{E}_{p(x|w)}[A^T(x|w)A(x|w)]a = \mathbb{E}_{p(x|w)}[a^T A^T(x|w)A(x|w)a] = \mathbb{E}_{p(x|w)}[||A(x|w)a||^2]$$

Hence equality $A(x|w)a = 0$ holds only for $a = 0$ and vector $\{\frac{\partial}{\partial w_j} \log p(x|w)\}_{j=1}^d$ is linearly independent for all x with $p(x|w) > 0$.
Now assume that $\{\frac{\partial}{\partial w_j} \log p(x|w)\}_{j=1}^d$ is linearly independent on supp $p(\cdot|w)$. From linear independence of elements of $A(x|w)$ follows that $A^T(x|w)A(x|w)$ is positive definite matrix $\forall w \in \mathbb{R}^d$. Therefore, $\forall x \in$ supp $p(x|w)$ we have that J is positive definite. □

We will assume that statistical model $p(x|w)$ satisfies the following condition:

$$\frac{\partial}{\partial w_j} \int p(x|w)dx = \int \frac{\partial}{\partial w_j} p(x|w)dx, \quad j = 1, \ldots, d \tag{15}$$

Corollary 5.9 in [5] gives sufficient conditions for equation (15) to hold. In short, it is sufficient that integral on the left hand side exists for any w, that integrand on the right hand side exists and that there exist some integrable function on X that upper-bounds $\left|\frac{\partial}{\partial w_j} p(x|w)\right|$. Since models usually depend on parameters in a smooth way, all three conditions are satisfied for all models we will consider.

We now state an alternative form of the Fisher information matrix that will be useful in the further discussion.

Lemma 2. *Fisher information matrix of a statistical model $p(x|w)$ is equal to the negative expectation of Hessian matrix of $\log p(x|w)$, i.e:*

$$J_{j,k}(w) = -\mathbb{E}_{p(x|w)}\Big[H_{j,k}\Big(\log p(x|w)\Big)\Big] = -\int p(x|w) \frac{\partial^2}{\partial w_j w_k} \log p(x|w)dx, \quad 1 \le j, k \le d \tag{16}$$

Proof. Let $j, k \in \{1, \ldots, d\}$. From partial derivative of logarithmic function we obtain:

$$p(x|w) \frac{\partial}{\partial w_j} \log p(x|w) = \frac{\partial}{\partial w_j} p(x|w)$$

Using the fact that $\int p(x|w)dx = 1$, $\forall w \in \mathbb{R}^d$, the equation above and continuity condition (15), we get:

$$\int p(x|w) \frac{\partial}{\partial w_j} \log p(x|w)dx = \int \frac{\partial}{\partial w_j} p(x|w)dx = \frac{\partial}{\partial w_j} \int p(x|w)dx = 0 \tag{17}$$

Differentiating left and right hand side once more gives:

$$\int \frac{\partial}{\partial w_k} p(x|w) \frac{\partial}{\partial w_j} \log p(x|w)dx + \int p(x|w) \frac{\partial}{\partial w_k} \frac{\partial}{\partial w_j} \log p(x|w)dx = 0$$

Hence:

$$\int p(x|w) \frac{\partial}{\partial w_k} \log p(x|w) \frac{\partial}{\partial w_j} \log p(x|w) = -\int p(x|w) \frac{\partial^2}{\partial w_j w_k} \log p(x|w)dx$$

$$\implies J_{j,k}(w) = -\int p(x|w) \frac{\partial^2}{\partial w_j w_k} \log p(x|w)dx$$

□

Example 5. *In this example we derive general formula for Fisher information matrix of a neural network Φ_w (5). From statistical model of the form (5) we derive that:*

$$\frac{\partial \log p(y|x,w)}{\partial w_j} = (y - \Phi_w(x)) \frac{\partial \Phi_w(x)}{\partial w_j}$$

Substituting this expression in the definition of Fisher information matrix, we obtain:

$$J_{j,k}(w) = \int p(x,y|w) \frac{\partial \log p(y|x,w)}{\partial w_j} \frac{\partial \log p(y|x,w)}{\partial w_k} dxdy =$$

$$= \int p(x)p(y|x,w)\Big(y - \Phi_w(x)\Big)^2 \frac{\partial \Phi_w(x)}{\partial w_j} \frac{\partial \Phi_w(x)}{\partial w_k} dxdy =$$

$$= \int p(x) \frac{\partial \Phi_w(x)}{\partial w_j} \frac{\partial \Phi_w(x)}{\partial w_k} \underbrace{\left(\int p(y|x,w)(y - \Phi_w(x))^2 dy \right)}_{\mathrm{Var}_{p(y|x,w)}(y)=1} dx =$$

$$= \mathbb{E}\left[\frac{\partial \Phi_w(x)}{\partial w_j} \frac{\partial \Phi_w(x)}{\partial w_k} \right] \tag{18}$$

Now we give statement of Cramer-Rao lower bound theorem and prove it afterwards. We first prove the case for single dimensional parameters w in theorem 2, and then for multidimensional parameters in theorem 3.

Theorem 2 (Cramer-Rao lower bound, section 8.3 in [6]). *Let $p(x|w)$ be a statistical model with $x \in \mathbb{R}^N$, $w \in \mathbb{R}$ and let $T(x)$ be an unbiased estimator of parameter w. Assume that condition (15) holds for $d = 1$ and also:*

$$\frac{\partial}{\partial w} \int T(x) p(x|w) dx = \int T(x) \frac{\partial}{\partial w} p(x|w) dx \tag{19}$$

Then:

$$\mathrm{Var}_{p(x|w)}[T(x)] \geq \frac{1}{\mathbb{E}_{p(x|w)}\left[\left(\frac{\partial}{\partial w} \log p(x|w) \right)^2 \right]} \tag{20}$$

where $\mathrm{Var}_{p(x|w)}[\,\cdot\,]$ denotes variance of random variable with respect to $p(x|w)$.

Proof. $T(x)$ is an unbiased estimator of w, hence:

$$\mathbb{E}_{p(x|w)}[T(x) - w] = \int (T(x) - w) p(x|w) dx = 0$$

Differentiating with respect to parameter w yields:

$$\int (T(x) - w) \frac{\partial}{\partial w} p(x|w) dx - \int p(x|w) dx = 0$$

$$\implies \int (T(x) - w) p(x|w) \frac{\partial}{\partial w} \log p(x|w) dx = 1 \tag{21}$$

We can now rewrite the last equation and use Cauchy-Schwartz inequality to obtain:

$$\int \left((T(x) - w)\sqrt{p(x|w)} \right) \left(\sqrt{p(x|w)} \frac{\partial}{\partial w} \log p(x|w) \right) dx = 1$$

$$\implies \int \left((T(x) - w)\sqrt{p(x|w)} \right)^2 dx \int \left(\sqrt{p(x|w)} \frac{\partial}{\partial w} \log p(x|w) \right)^2 dx \geq 1$$

$$\implies \int (T(x) - w)^2 p(x|w) dx \int p(x|w) \left(\frac{\partial}{\partial w} \log p(x|w) \right)^2 dx \geq 1$$

Identifying the left term as variance of the estimator and the right one as expectation, we get equation we wanted to prove:

$$\implies \mathrm{Var}_{p(x|w)}[T(x)] \cdot \mathbb{E}_{p(x|w)}\left[\left(\frac{\partial}{\partial w} \log p(x|w) \right)^2 \right] \geq 1$$

□

Theorem 3 (Cramer-Rao lower bound - multiparameters, section 8.3 in [6]). *Let $p(x|w)$ be a statistical model, $x \in \mathbb{R}^N$ and $w \in \mathbb{R}^d$ with Fisher information matrix $J(w)$ and let $T(x) = [T_1(x),\ T_2(x), \ldots,\ T_d(x)]^T$ be an unbiased estimator of the parameters $w = [w_1,\ w_2, \ldots,\ w_d]^T$. Assume that condition (15) holds, and that condition (19) holds for every $T_i(x)$, $i = 1, 2, \ldots, d$. Then:*

$$\mathrm{Cov}_{p(x|w)}[T(x)] \geq J^{-1}(w) \tag{22}$$

where $\mathrm{Cov}_{p(x|w)}[T(x)]$ is the covariance matrix of the estimator $T(x)$ defined by:

$$\mathrm{Cov}_{p(x|w)}[T(x)] = \int (T(x) - w)(T(x) - w)^T p(x|w) dx \tag{23}$$

Proof. Define the following vectors:

$$V_1 = \begin{bmatrix} T_1(x) - w_1 & T_2(x) - w_2 & \dots & T_d(x) - w_d \end{bmatrix}^T$$

$$V_2 = \begin{bmatrix} \frac{\partial \log p(x|w)}{\partial w_1} & \frac{\partial \log p(x|w)}{\partial w_2} & \dots & \frac{\partial \log p(x|w)}{\partial w_d} \end{bmatrix}^T$$

and let $V = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$. Note that both vectors are dependent on x and w. Using equation (17) and assumption that $T(x)$ is unbiased we have $\mathbb{E}_{p(x|w)}[V] = 0$.
Employing equation (21) to unbiased estimators $T_i(x)$ we obtain:

$$\mathbb{E}_{p(x|w)}\left[(T_k(x) - w_k)\frac{\partial}{\partial w_k}\log p(x|w)\right] = 1, \quad k = 1, \dots, d$$

Similarly, for $k \neq j, k, j = 1 \dots, d$ we obtain:

$$\mathbb{E}_{p(x|w)}\left[(T_k(x) - w_k)\frac{\partial}{\partial w_j}\log p(x|w)\right] = \int (T_k(x) - w_k)\frac{\partial}{\partial w_j}p(x|w)dx =$$
$$= \frac{\partial}{\partial w_j}\int T_k(x)p(x|w)dx - w_k\int \frac{\partial}{\partial w_j}p(x|w)dx = \frac{\partial}{\partial w_j}w_k = 0$$

Last two derivations imply that $\mathbb{E}_{p(x|w)}[V_1 V_2^T] = I$. Also we note that $\mathbb{E}_{p(x|w)}[V_1 V_1^T] = \mathrm{Cov}_{p(x|w)}[T(x)]$ and $\mathbb{E}_{p(x|w)}[V_2 V_2^T] = J(w)$. Combining all these results and the fact that covariance matrix is always positive semidefinite we get:

$$\mathrm{Cov}_{p(x|w)}[V] = \mathbb{E}_{p(x|w)}[VV^T] = \mathbb{E}_{p(x|w)}\begin{bmatrix} V_1{V_1}^T & V_1{V_2}^T \\ V_2{V_1}^T & V_2{V_2}^T \end{bmatrix} \geq 0$$

Hence:

$$\begin{bmatrix} \mathrm{Cov}_{p(x|w)}[T(x)] & I \\ I & J(w) \end{bmatrix} \geq 0$$

Multiplying this matrix with positive definite matrix of the form:

$$\begin{bmatrix} I & 0 \\ -J^{-1}(w) & I \end{bmatrix}$$

gives:

$$0 \leq \begin{bmatrix} \mathrm{Cov}_{p(x|w)}[T(x)] & I \\ I & J(w) \end{bmatrix}\begin{bmatrix} I & 0 \\ -J^{-1}(w) & I \end{bmatrix} = \begin{bmatrix} \mathrm{Cov}_{p(x|w)}[T(x)] - J^{-1}(w) & I \\ 0 & J(w) \end{bmatrix}$$

Finally, we obtain:

$$\mathrm{Cov}_{p(x|w)}[T(x)] \geq J^{-1}(w)$$

□

Remark. *Derivation of the Cramer-Rao lower bound is based on the assumption that there exists an unbiased estimator. However, models with positive semidefinite Fisher information matrix do not have an unbiased estimator of the entire parameter vector [7]. It has been shown that there exist more general bounds analogous to Cramer-Rao lower bound for estimation of parameters with singular Fisher information matrix if parameters are constrained and/or we allow biased estimation [8].*

Under appropriate conditions, Fisher information matrix represent actually curvature of the Kullback-Leibler distance. We prove this relation in the following theorem.

Theorem 4. *(Quadratic approximation of KL distance, sec. 2.6 in [9]) Let $p(x|w)$ be a statistical model, $x \in \mathbb{R}^N$, $w \in W \subset \mathbb{R}^d$, where W is open and convex parameter space, and let $w_0 \in \mathbb{R}^d$ be the true parameter of the model, i.e. let $K(w_0) = 0$. If the following conditions hold:*

1. *partial derivatives $\frac{\partial \log p(x|w)}{\partial w_j}$, $\frac{\partial^2 \log p(x|w)}{\partial w_j \partial w_k}$ and $\frac{\partial^3 \log p(x|w)}{\partial w_j \partial w_k \partial w_l}$ exist in the neighborhood of w_0 for $j, k, l = 1, 2, \dots, d$,*
2. *there exist functions $F(x)$ and $G(x)$ integrable over the whole space such that $\left|\frac{\partial p(x|w)}{\partial w_j}\right| < F(x)$ and $\left|\frac{\partial^2 p(x|w)}{\partial w_j \partial w_k}\right| < G(x)$. There exists a function $H(x)$ such that $\left|\frac{\partial^3 \log p(x|w)}{\partial w_j \partial w_k \partial w_l}\right| < H(x)$ and $\int p(x|w)H(x)dx < M < \infty$ where M is a constant independent of w,*

3. $\int \frac{\partial p(x|w)}{\partial w_j} dx = 0$ *and* $\int \frac{\partial^2 p(x|w)}{\partial w_j \partial w_k} dx = 0$.

then $K(w)$ *can be approximated in the neighborhood of* w_0 *i.e. for* $w = w_0 + \Delta w$ *holds:*

$$K(w_0 + \Delta w) = \frac{1}{2} \Delta w^T J(w_0) \Delta w + \mathcal{O}(||\Delta w||^3) \tag{24}$$

Proof. By definition of KL-distance:

$$K(w_0 + \Delta w) = -\int p(x|w_0) \log \frac{p(x|w_0 + \Delta w)}{p(x|w_0)} dx$$

Using Taylor expansion with Lagrange form of the remainder we get:

$$\log p(x|w_0 + \Delta w) = \log p(x|w_0) + \sum_{k=1}^{d} \Delta w_k \frac{\partial \log p(x|w_0)}{\partial w_k} + \frac{1}{2!} \sum_{k=1}^{d} \sum_{j=1}^{d} \Delta w_k \Delta w_j \frac{\partial^2 \log p(x|w_0)}{\partial w_j \partial w_k} + \\ + \frac{1}{3!} \sum_{k=1}^{d} \sum_{j=1}^{d} \sum_{l=1}^{d} \Delta w_k \Delta w_j \Delta w_l \frac{\partial^3 \log p(x|w)}{\partial w_k \partial w_j \partial w_l} \Big|_{w = w_0 + t\Delta w} \tag{25}$$

where $t \in [0, 1]$. Note that here we need convexity of W in order to have $p(x|w)$ well-defined for $w = w_0 + t\Delta w$. From condition 2) holds that $K(w)$ is twice differentiable and that the following integrals exist (see Corollary 5.9 in [5]):

$$\frac{\partial K(w_0)}{\partial w_k} = \frac{\partial}{\partial w_k} \int p(x|w_0) \log \frac{p(x|w)}{p(x|w_0)} dx \Big|_{w=w_0} = \int p(x|w_0) \frac{\partial \log p(x|w_0)}{\partial w_k} dx$$

$$\frac{\partial^2 K(w_0)}{\partial w_j \partial w_k} = \frac{\partial^2}{\partial w_j \partial w_k} \int p(x|w_0) \log \frac{p(x|w)}{p(x|w_0)} dx \Big|_{w=w_0} = \int p(x|w_0) \frac{\partial^2 \log p(x|w_0)}{\partial w_j \partial w_k} dx$$

Substituting equation (25) in definition of $K(w_0 + \Delta w)$ and using existence of the integrals above we obtain:

$$K(w_0 + \Delta w) = -\sum_{k=1}^{d} \Delta w_k \underbrace{\int p(x|w_0) \frac{\partial \log p(x|w_0)}{\partial w_k} dx}_{\int \frac{\partial p(x|w_0)}{\partial w_k} dx = 0} - \frac{1}{2!} \sum_{k=1}^{d} \sum_{j=1}^{d} \Delta w_k \Delta w_j \underbrace{\int p(x|w_0) \frac{\partial^2 \log p(x|w_0)}{\partial w_j \partial w_k} dx}_{-J_{j,k}(w_0)} - \\ - \frac{1}{3!} \sum_{k=1}^{d} \sum_{j=1}^{d} \sum_{l=1}^{d} \Delta w_k \Delta w_j \Delta w_l \int p(x|w) \frac{\partial^3 \log p(x|w)}{\partial w_k \partial w_j \partial w_l} \Big|_{w = w_0 + t\Delta w} dx$$

Using condition 2. of the theorem we note that the integral in last term can be upper bounded by constant M and hence we obtain:

$$K(w_0 + \Delta w) = \frac{1}{2} \Delta w^T J(w_0) \Delta w + \mathcal{O}(||\Delta w||^3) \approx \frac{1}{2} \Delta w^T J(w_0) \Delta w, \quad ||\Delta w|| \to 0$$

□

Let's now illustrate this theorem in two examples - example 6 where we show that that K can be represented in the quadratic form, and example 7 where we show that this approximation is not valid for every model.

Example 6. *Using obtained Fisher information matrix of a general neural network (18), we determine the Fisher information matrix of the line regression statistical model from example 3 as follows:*

$$J(a, b) = \begin{bmatrix} \mathbb{E}[x^2] & \mathbb{E}[x] \\ \mathbb{E}[x] & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{26}$$

Statistical model of this form have positive definite Fisher information matrix. Also, using quadratic approximation (24) we have:

$$K(a, b) = \frac{1}{2} \begin{bmatrix} a - a_0 & b - b_0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a - a_0 \\ b - b_0 \end{bmatrix} = \frac{1}{2}(a - a_0)^2 + \frac{1}{2}(b - b_0)^2 \tag{27}$$

which is exactly the same result as obtained in (10).

Example 7. *Let's now consider neural network of the form given in example 4, i.e. let* $\Phi_w(x) = a(e^{bx} - 1)$*. Applying derived expression (18) we get:*

$$J(a,b) = \begin{bmatrix} \mathbb{E}[(e^{bx}-1)^2] & \mathbb{E}[(e^{bx}-1)axe^{bx}] \\ \mathbb{E}[(e^{bx}-1)axe^{bx}] & \mathbb{E}[(axe^{bx})^2] \end{bmatrix} = \begin{bmatrix} e^{2b^2} - 2e^{b^2/2} + 1 & 2abe^{2b^2} - abe^{b^2/2} \\ 2abe^{2b^2} - abe^{b^2/2} & a^2e^{2b^2} \end{bmatrix} \quad (28)$$

Note that $J(a,b)$ *is not positive definite (take for example* $a = 0$*). Also note that KL-distance of this model (equation (12)) cannot be approximated by quadratic form since expansion of equation (4) gives* $K(a,b) = \frac{a^2b^2}{2} + \frac{7a^2b^4}{8} + \cdots$*. This statistical model does not satisfy condition 2) in theorem 4 since there is no constant* $M < \infty$ *such that* $\int p(y|x,w)H(x,y)dxdy < M$ *with* $\left|\frac{\partial^3 \log p(y|x,w)}{\partial w_j \partial w_k \partial w_l}\right| < H(x,y)$ *(consider for example the case* $a = 0$*,* $w_j = w_k = a$ *and* $w_l = b$ *then* $\frac{\partial^3 \log p(y|x,w)}{\partial w_j \partial w_k \partial w_l} \sim xe^{2bx}$ *whereas* $p(y|x,w) \sim e^{-y^2}$ *and hence given integral is not finite for* $b > 0$*).*

1.5 Statistical models

We have seen in previous examples that there are models that have positive-semidefinite Fisher information matrix and do not have quadratic form of the $K(w)$. Models with this type of properties belong to a special group of models called singular models. In this section we define singular models and their properties.

Definition 4. *A statistical model* $p(y|x,w)$ *is identifiable if the map* $p(y|x,\cdot)$ *is injective* $\forall x \in X, y \in Y$*, i.e. if* $p(y|x,w_1) = p(y|x,w_2)$ *implies* $w_1 = w_2$*,* $w_1, w_2 \in W$*.*

One might feel tempted to define an equivalence relation on W such that $w_1 \sim w_2$ if and only if $p(y|x,w_1) = p(y|x,w_2)$ for all $x \in X$ (see Remark 1.6 in [1]). However, even though the map $p(y|x,w)$ is now injective on $W/\sim$, obtained parameter space $W/\sim$ does not have a useful structure on which statistical learning can be developed [1]. In the following sections we will show one of the possible methods to deal with nonidentifiability in models and develop singular learning theory.

Definition 5. *A statistical model* $p(y|x,w)$ *has a positive definite metric if its Fisher information matrix is positive definite for every* $w \in W$*.*

Definition 6. *A statistical model* $p(y|x,w)$ *is regular if it is identifiable and has positive definite metric. A statistical model is singular if it is not regular.*

Remark. *In [1] author makes difference between singular and strictly singular models, where the former contain both regular and strictly singular models. We will use terms introduced in definitions above, but one should be aware that singular statistical learning can be applied also to regular models, hence it can be considered as more general learning theory (see Remark 1.9 in [1]).*

Example 8. *Let's continue analysis from example 6 where we already concluded that line regression model has positive definite metric. In general, condition that* $p(y|x,a_1,b_1) = p(y|x,a_2,b_2)$ *for some* $x \in X$ *and* $y \in Y$ *holds if and only if* $a_1 = a_2$ *and* $b_1 = b_2$*. Hence this model is also identifiable, and we conclude that line regression model is regular.*

Example 9. *We have already provided example of a neural network with positive semidefinite metric in example 7. Now we claim that this model is neither identifiable. Namely, take* $w_1 = (a_1,b_1) = (a_1,0)$ *and* $w_2 = (a_2,b_2) = (0,b_2)$*. Then* $p(y|x,w_1) = p(y|x,w_2)$ *for all* $x \in X$*,* $y \in Y$ *but for* $a_1 \neq 0$ *or* $b_2 \neq 0$ *we have* $w_1 \neq w_2$*. Also, note that in the networks with multiple neurons, sole permutation of neurons in one layer of the network leads to the same output, but with different parameter values. Hence neural networks can be singular models that are both nonidentifiable and without positive definite metric.*

In the following discussion we will always assume that true distribution is realizable by model class as defined below.

Definition 7. *A probability distribution* $q(y|x)$ *is realizable by a statistical model* $p(y|x,w)$ *if there exists* $w_0 \in W$ *such that* $q(y|x) = p(y|x,w_0)$*. A probability distribution that is not realizable is unrealizable.*

In real world usually we do not know the true distribution and hence cannot claim that it is realizable by our model. Moreover, for majority of applications we use models that give only simplifications of the true distribution and realizability is not easily achieved. Discussion about unrealizable true

distributions can be found in [2].

Note that $q(y|x)$ is realizable by $p(y|x,w)$ if and only if there exists $w_0 \in W$ such that $K(w) = 0$. Hence, minimizing $K(w)$ can lead to optimal parameter. However, minimizing $K(w)$ is practically difficult. Therefore we introduce more tractable quantity called log likelihood ratio function as follows. For a given data set $z^{(n)} = \{(x_i, y_i)\}_{i=1}^n$, statistical model $p(y|x,w)$ and true distribution $q(y|x)$, log density ratio function is defined by:

$$f(y|x,w) = \log \frac{q(y|x)}{p(y|x,w)} \tag{29}$$

and the log likelihood ratio function is given by:

$$K_n(w) = \frac{1}{n}\sum_{i=1}^n f(y_i|x_i,w) = \frac{1}{n}\sum_{i=1}^n \log \frac{q(y_i|x_i)}{p(y_i|x_i,w)} \tag{30}$$

Recall that we assumed that true distribution $q(y|x)$ is known. Hence, minimizing $K_n(w)$ is now straightforward. Also, we have that $\mathbb{E}[K_n(w)] = K(w)$ where expectation is over data set $z^{(n)}$. However, minimization of $K_n(w)$ is not equivalent to minimization of $K(w)$ since in general $\min_w K(w) = \min_w \mathbb{E}[K_n(w)] \neq \mathbb{E}[\min_w K_n(w)]$ [1]. It is necessary to examine relation between these two variables in order to deduce when is such minimization indeed useful. We will provide deeper insights in the following sections. If we could estimate $K(w)$ from $K_n(w)$, we could select the model and its parameters such that generalization error $K(w)$ is minimized.

1.6 Statistical estimation

In this section we introduce statistical quantities of our main interest i.e. Bayes generalization and training errors as well as free energy. We start by recalling definition of the likelihood function given samples $\{(x_i, y_i)\}_{i=1}^n$ and model $p(y|x,w)$:

$$L_n(w) = \prod_{i=1}^n p(y_i|x_i,w) \tag{31}$$

In the maximum likelihood estimation we aim to choose w that maximizes equation (31). This approach is equivalent to minimization of $K_n(w)$ since:

$$-\frac{1}{n}\log L_n(w) = K_n(w) + S_n \tag{32}$$

where $S_n = -\frac{1}{n}\sum_{i=1}^n \log q(y_i|x_i)$ is empirical entropy which is independent of w. As we have discussed in the previous section, minimization of $K_n(w)$ is not equivalent to minimization of $K(w)$ in general. Hence we cannot claim that maximum likelihood estimator minimizes $K(w)$.

Recall that predictive distribution defined in (2) satisfies $p(y|x,z^{(n)}) = \mathbb{E}_w[p(y|x,w)]$ where $\mathbb{E}_w$ is expectation with respect to $p(w|z^{(n)})$. Calculating Kullback-Leibler distance between true and predictive distribution we obtain Bayes generalization and training errors.

Definition 8. *For a statistical model $p(y|x,w)$, true distribution $q(x,y)$ and prior $\varphi(w)$ we define the following errors:*

- *The Bayes generalization error:*

$$B_g = \int q(x,y)\log\frac{q(y|x)}{p(y|x,z^{(n)})}dxdy = \int q(x,y)\log\frac{q(y|x)}{\mathbb{E}_w[p(y|x,w)]}dxdy \tag{33}$$

- *The Bayes training error:*

$$B_t = \frac{1}{n}\sum_{i=1}^n \log\frac{q(y_i|x_i)}{p(y_i|x_i,z^{(n)})} = \frac{1}{n}\sum_{i=1}^n \log\frac{q(y_i|x_i)}{E_w[p(y_i|x_i,w)]} \tag{34}$$

- *Free energy:*

$$F_n = -\log Z_n \tag{35}$$

where Z_n is partition function defined in equation (1).

In contrast to functions $K(w)$ and $K_n(w)$ that calculate dissimilarity for given parameter w, variables B_g, B_t are dependent only on number of samples n and not on w. In the following two lemmas we find a relation between $K_n(w)$ and these errors.

Lemma 3. *For a posterior distribution $p(w|z^{(n)})$ of the form (1) and normalized partition function $Z_n^0 = \frac{Z_n}{\prod_{i=1}^n q(y_i|x_i)}$ holds:*

$$p(w|z^{(n)}) = \frac{1}{Z_n^0}\exp(-nK_n(w))\varphi(w) \tag{36}$$

and $Z_n^0 = \int \exp(-nK_n(w))\varphi(w)dw$.

Proof. From equation (1) and using definition of normalized partition function we have:

$$p(w|z^{(n)}) = \frac{1}{Z_n}\varphi(w)\prod_{i=1}^n p(y_i|x_i,w) = \frac{1}{Z_n}\varphi(w)\Big(\prod_{i=1}^n q(y_i|x_i)\Big)\frac{\prod_{i=1}^n p(y_i|x_i,w)}{\prod_{i=1}^n q(y_i|x_i)} = \frac{1}{Z_n^0}\varphi(w)\exp(-nK_n(w))$$

Furthermore:

$$Z_n^0 = \frac{Z_n}{\prod_{i=1}^n q(y_i|x_i)} = \int \varphi(w)\frac{\prod_{i=1}^n p(y_i|x_i,w)}{\prod_{i=1}^n q(y_i|x_i)}dw = \int \varphi(w)\exp(-nK_n(w))dw \tag{37}$$

□

Now we prove a relation between Bayes generalization error and normalized free energy.

Lemma 4. *Let normalized free energy be given by $F_n^0 = -\log Z_n^0$. Then Bayes generalization error is equal to the increase of the normalized free energy:*

$$B_g = \mathbb{E}_{x_{n+1},y_{n+1}}[F_{n+1}^0] - F_n^0 \tag{38}$$

and also:

$$\mathbb{E}[B_g] = \mathbb{E}[F_{n+1}^0] - \mathbb{E}[F_n^0] \tag{39}$$

Proof. From definition of normalized free energy and partition function we get:

$$\begin{aligned} F_{n+1}^0 - F_n^0 &= -\log\frac{Z_{n+1}^0}{Z_n^0} = \log q(y_{n+1}|x_{n+1}) - \log\frac{Z_{n+1}}{Z_n} = \\ &= \log q(y_{n+1}|x_{n+1}) - \log\frac{1}{Z_n}\int p(y_{n+1}|x_{n+1},w)\varphi(w)\prod_{i=1}^n p(y_i|x_i,w)dw \end{aligned}$$

Using expressions (1) and (2) we have:

$$\log p(y_{n+1}|x_{n+1},z^{(n)}) = \log\frac{1}{Z_n}\int p(y_{n+1}|x_{n+1},w)\varphi(w)\prod_{i=1}^n p(y_i|x_i,w)dw$$

and hence:

$$\begin{aligned} F_{n+1}^0 - F_n^0 = \log q(y_{n+1}|x_{n+1}) - \log p(y_{n+1}|x_{n+1},z^{(n)}) = \log\frac{q(y_{n+1}|x_{n+1})}{p(y_{n+1}|x_{n+1},z^{(n)})} \\ \Longrightarrow \mathbb{E}_{x_{n+1},y_{n+1}}[F_{n+1}^0] - F_n^0 = \int q(x,y)\log\frac{q(y|x)}{p(y|x,z^{(n)})}dxdy = B_g \end{aligned}$$

Taking expectation with respect to data $z^{(n)}$ we get expression (39). □

So we could conclude that there exists a pathway from $K_n(w)$ to B_g as follows: $K_n(w) \to Z_n \to Z_n^0 \to F_n^0 \to B_g$.

Remark. *Recall that $Z_n = \int_W \varphi(w)\prod_{i=1}^n p(y_i|x_i,w)dw$ represents likelihood of the pair $(\varphi(w), p(y|x,w))$. Thus F_n is minus log likelihood of the pair and we will be interested in minimizing this variable. Also, note that free energy and normalized free energy are connected as follows:*

$$F_n = -\log Z_n = -\log\Big(Z_n^0\prod_{i=1}^n q(y_i|x_i)\Big) = -\log Z_n^0 - \sum_{i=1}^n \log q(y_i|x_i) = F_n^0 + nS_n \tag{40}$$

where S_n is empirical entropy defined after equation (32).

1.7 Asymptotic normality

For regular statistical models, posterior distribution, MAP and MLE estimators converge to normal distribution as number of samples $n \to \infty$, while this may not be the case for singular models [1]. For a proof of asymptotic normality of posterior distribution see Bernstein-von Mises theorem (e.g. in section 10.2 in [10]). Here we present a theorem that establishes asymptotic normality of maximum likelihood estimators in the case of regular models. Before that we recall definition of convergence in distribution:

Definition 9. *(Convergence in distribution) Let $\{X_n\}$ and X be a sequence of random variables and a random variable on a probability space. Then $\{X_n\}$ converges to X in distribution or law, denoted by $X_n \xrightarrow{d} X$, if for any bounded and continuous function f we have:*

$$\mathbb{E}_X[f(X)] = \lim_{n\to\infty} \mathbb{E}_{X_n}[f(X_n)] \tag{41}$$

Theorem 5. *(Asymptotic normality of maximum likelihood estimator, theorem 6.7 in [11]) Let model be given by $p(x|w)$, $w \in W$. Let $\{X_n\}$ be a sequence of i.i.d random variables distributed according to $p(x|w_0)$, where w_0 is an interior point of W. Define $\widehat{w}$ to be a maximum likelihood estimator from observations $\{X_n\}$. Assume that for every $w \in W$ holds:*

1. *$p(x|w)$ is identifiable i.e. $\forall x\ p(x|w_1) = p(x|w_2) \implies w_1 = w_2$,*
2. *all probability density functions $p(x|w)$ have the same support independent of w,*
3. *the first three partial derivatives of $\log p(x|w)$ with respect to w exist in the support of $p(x|w)$,*
4. *there exists a function $H(x)$ such that in the neighborhood of w:*

$$\forall x \left| \frac{\partial^3 \log p(x|w)}{\partial w_j \partial w_k \partial w_l} \right| < H(x) \qquad \text{and} \qquad \int p(x|w)H(x)dx < \infty, \tag{42}$$

5. *$\mathbb{E}_{p(x|w)}\left[\frac{\partial \log p(x|w)}{\partial w}\right] = 0$ and $J(w)$ is nonsingular.*

If $\widehat{w}$ asymptotically achieves w_0 in probability and $\widehat{w}$ satisfies $\sum_{i=1}^{n} \frac{\partial \log p(X_i|\widehat{w})}{\partial w} = 0$ as $n \to \infty$ then:

$$\sqrt{n}(\widehat{w} - w_0) \xrightarrow{d} \mathcal{N}(0, J(w_0)^{-1}), \quad n \to \infty \tag{43}$$

Since singular models are either nonidentifiable and hence do not satisfy condition 1) or do not have positive definite Fisher matrix and hence condition 5) does not hold. Also, as we showed in example 7, condition 4) is not fulfilled for some singular models. Hence theorem 5 is not applicable to singular models, and we cannot assume that MLE estimator converges to normal distribution, which makes theoretical analysis of statistical models more involved. The similar conclusion can be made about asymptotic normality of posterior distribution and MAP estimator. We illustrate asymptotic normality of posterior distribution for regular models, and lack of it for singular ones, in the following example.

Example 10. *(Experiment is inspired by experiments in section 1.5 in [2]) Let's analyze two statistical models given in examples 3 and 4, i.e. $p(y|x,w) \sim \mathcal{N}(a(e^{bx}-1), 1)$ and $p(y|x,w) \sim \mathcal{N}(ax+b, 1)$. As shown in examples 8 and 9, the first model is regular, whereas the latter is singular. According to theorem 5, posterior distribution of the first model is asymptotically normal, but that of the second one is not. Let true distributions be given by $(a,b) = (0.2, 0.4)$ in both models and let $p(x)$ be uniform distribution over interval $[0,1]$. Assume that the prior of parameters a and b is:*

$$\varphi(a,b) = \begin{cases} 1, & 0 \le a, b \le 1 \\ 0, & \text{otherwise} \end{cases} \tag{44}$$

We sample n independent data points (x_i, y_i) according to true distributions. Then we determine posterior distributions according to equation (1). For three different sizes of sample sets: $n = 100$, $n = 1000$ and $n = 10000$, we show graphically obtained posterior distribution in Figure 3.

As expected, regular model has asymptotically normal posterior distribution, while for singular model that is not the case. It is evident that there exists a curve of possible "true" parameter values and hence it cannot converge to normal distribution.

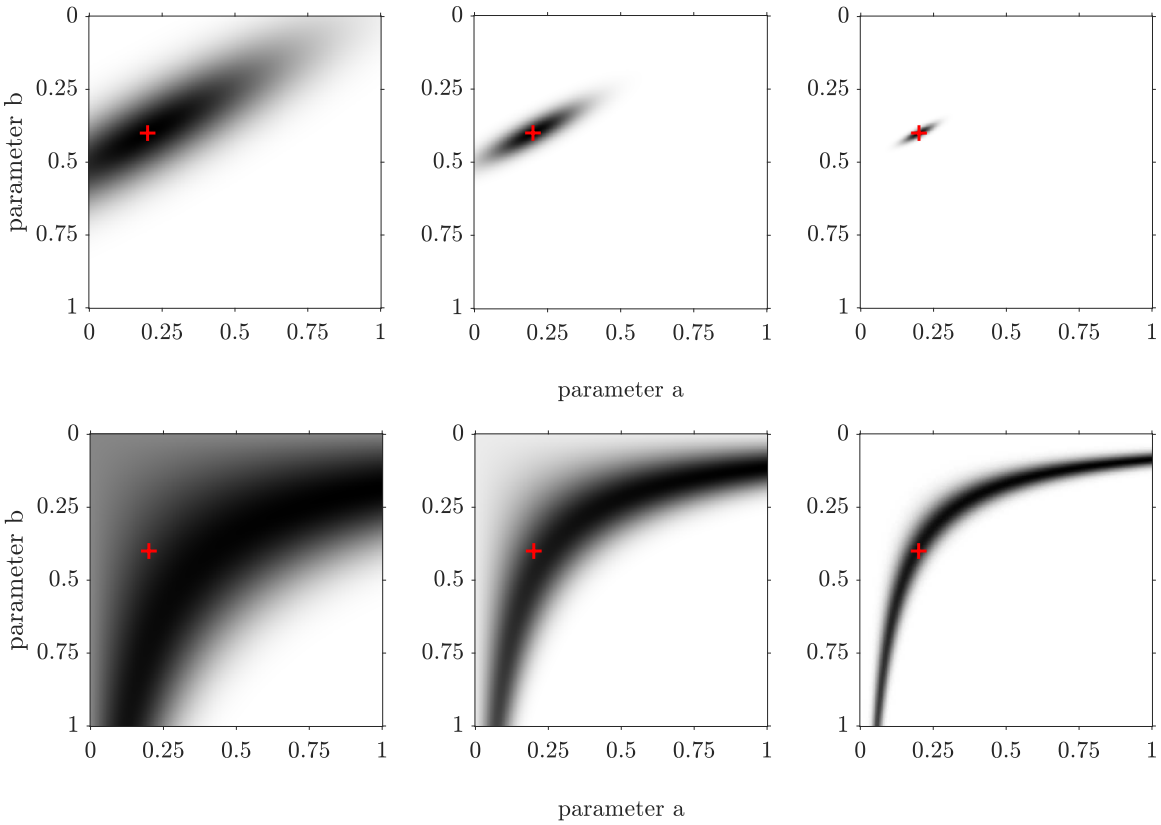


Figure 3: Asymptotic normality of regular (above) and singular (below) models. Posterior distributions for $n = 100$ (left), $n = 1000$ (center) and $n = 10000$ (right) sample points. Red cross shows true parameter values $(a = 0.2, b = 0.4)$.

Remark. *Singular statistical models do not have asymptotic normality even in the case when true parameters are not a singular point. For* $(a, b) = (0.2, 0.4)$ *the Fisher information matrix (28) is not singular, but effects of singularity are evident even for these parameters as shown in Figure 3. Also note that in theorem 5 we require* $J(w)$ *to be nonsingular matrix for any* $w \in W$*, and hence existence of single point at which this is not satisfied can lead to lack of asymptotic normality.*

At the end of this section we give a quick recap of the main messages. We introduced two main classes of models - regular and singular, with difference in the properties of identifiability and positive definiteness of Fisher information matrix. We have introduced Kullback-Leibler distance for measuring distance between predicted and true distributions and proved its relation to Fisher information matrix for the regular models. Then we introduced variables from statistical estimation theory that will be of our interest and commented on some of their connections. At the end we discussed asymptotic normality and the fact that singular models do not have this interesting and useful property.

2 Algebraic geometry and singularities

Kullback-Leibler distance of true distribution of a system and our predicted distribution (7) is one of the main quantities in singular statistical learning. As introduced in section 1.3 this quantity is minimized exactly when we predict the true distribution correctly. This means that our efforts in statistical estimation should be concentrated on minimizing this function on some parameter space W. The quadratic approximation of Kullback-Leibler distance in the case of regular models shows that $K(w)$ has a unique minimum, at least locally. However we have also seen that this approximation is not valid for general singular models and furthermore there can be many points in a arbitrary small neighborhood that all have $K(w) = 0$. As we will see in this section, these points form a set called real algebraic set. In order to examine points at which $K(w) = 0$ we need to characterize properties of real algebraic sets and specially of singularities in such sets. According to Prof. Watanabe "in singular statistical models, the knowledge to be discovered corresponds to singularities in general" [1]. The main idea behind studying singular points is that the behavior of function $K(w)$ near its singularities will determine the asymptotic behavior of singular statistical models.

Before introducing basic terms from algebraic geometry, let's give a short reminder of some of the notions from analysis. Let $U \subset \mathbb{R}^d$ be an open set and let f be a function $f : U \to \mathbb{R}$. In the following section we use multi-index notation, i.e. for $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d) \in \mathbb{R}^d$, $x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$, $b = (b_1, b_2, \ldots, b_d) \in \mathbb{R}^d$, $a_\alpha \in \mathbb{R}$ and simplify writing of power series of function f as follows:

$$f(x) = \sum_{\alpha_1=0}^{\infty} \cdots \sum_{\alpha_d=0}^{\infty} a_{\alpha_1\alpha_2\cdots\alpha_d}(x_1 - b_1)^{\alpha_1} \cdots (x_d - b_d)^{\alpha_d} = \sum_{\alpha} a_\alpha (x - b)^\alpha \tag{45}$$

We denote by $\mathcal{R} = \mathbb{R}[x_1, x_2, \ldots, x_d]$ the set of all polynomials $f : \mathbb{R}^d \to \mathbb{R}$ with real coefficients. Using the standard operations of addition and multiplication of polynomials it can be shown that $\mathcal{R}$ is a ring.

We will also need notion of smoothness of functions. We say that a function $f : U \to \mathbb{R}$ is of class C^r if all partial derivatives of degree $r \in \mathbb{N}$ or smaller are well defined and continuous. If this holds for every partial derivative then function is said to be of class C^∞. Moreover, if function f is real analytic function in U then it is said to be of class C^ω.

2.1 Short tour of algebraic geometry

Now we introduce basic concepts and results from algebraic geometry. Main motivation will be characterization of real algebraic sets, and hence we begin by defining these sets.

Definition 10. *(Real algebraic set) Let $f_1, f_2, \ldots, f_k : \mathbb{R}^d \to \mathbb{R}$ be $k \geq 1$ polynomial functions. Then the set:*

$$V(f_1, f_2, \ldots, f_k) = \{x \in \mathbb{R}^d;\ f_1(x) = f_2(x) = \cdots = f_k(x) = 0\} \tag{46}$$

is called a real algebraic set.

Definition 11. *(Real analytic set) Let $U \subset \mathbb{R}^d$ be an open set and let $f_1, f_2 \ldots, f_d : U \to \mathbb{R}$ be $k \geq 1$ real analytic functions. Then the set:*

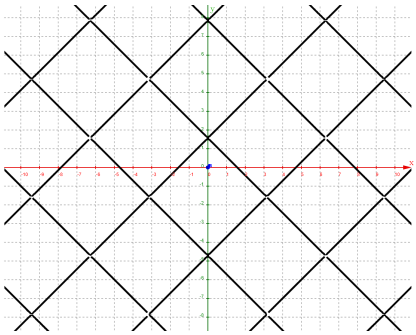
$$A(f_1, f_2, \ldots, f_k) = \{x \in U;\ f_1(x) = f_2(x) = \cdots = f_k(x) = 0\} \tag{47}$$

is called a real analytic set.

Recall that all polynomials are trivially analytic functions and hence every real algebraic set is a real analytic set, while opposite does not hold. Furthermore this implies that every theorem valid for real analytic sets is also correct for any algebraic set.

Example 11. *We illustrate definitions above by two examples. First we define a real algebraic set given by an analytic function $f(x, y) = \cos(x) - \sin(y)$. The set of zeros of function $f(x, y)$ i.e. real analytic set defined by $f(x, y)$ is shown in Figure 4 (left). Then we consider a polynomial function $f(x, y, z) = x^2 + y^2 + z^3 - z^2$. The zero set of this function is a real algebraic set shown in Figure 4 (right) - set of zeros forms a surface called "Dingdong".*

Even though in former definitions we did not impose any relation among functions $f_1, f_2, \ldots, f_k$ it turns out that having a particular type of structure yields fruitful results. That structure is called ideal and let's first recall the formal definition of an ideal:



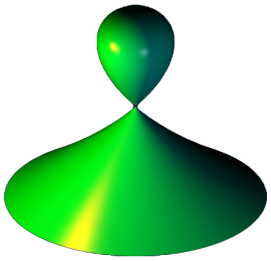


Figure 4: Analytic set $A = \{(x, y) \in \mathbb{R}^2;\ \cos(x) - \sin(y) = 0\}$ (left); Algebraic set $V = \{(x, y, z) \in \mathbb{R}^3;\ x^2 + y^2 + z^3 = z^2\}$ (right) - image is downloaded from Herwig Hauser's gallery[2].

Definition 12. *(Ideal) We say that a subset I of polynomial ring $\mathcal{R}$ is an ideal if $(I, +)$ is a subgroup of $(\mathcal{R}, +)$ and multiplication operation maps $\mathcal{R} \times I$ into I.*

For a set of polynomials $\{f_1, f_2, \ldots, f_r\}$, the minimum ideal containing $f_1, f_2 \ldots, f_k$ is ideal generated by polynomials $f_1, f_2, \ldots, f_k$ defined by:

$$\langle f_1, f_2, \ldots, f_k \rangle = \left\{ \sum_{i=1}^{k} g_i(x) f_i(x);\ g_i \in \mathcal{R} \right\} \tag{48}$$

In a way, ideal generated by polynomials in algebraic geometry is analogous to the span of vectors in linear algebra. One of the differences is that instead of real coefficients in linear algebra, here we use polynomials as coefficients of different elements.

Example 12. *Let $I = \langle x^2 \rangle$. Then $x^3 \in I$ since $x^3 = xx^2$, but $x \notin I$ since there exist no polynomial $g(x)$ such that $x = g(x)x^2$.*

Similarly to definition 10 we define the real algebraic set of an ideal as follows:

Definition 13. *(Real algebraic set of an ideal) Let I be an ideal of $\mathcal{R}$. Then a set:*

$$\mathbb{V}(I) = \{x \in \mathbb{R}^d;\ f(x) = 0\ (\forall f \in I)\} \tag{49}$$

is called a real algebraic set of ideal I.

Simply put a real algebraic set of an ideal is set of common zeros of all functions in the ideal. Note that all nontrivial ideals have infinite number of elements and therefore we need the following proposition to make problem of determining common zeros of an infinite set of functions feasible.

Proposition 1. *Definitions 10 and 13 are congruent, i.e. for any polynomials $f_1, f_2, \ldots, f_k$ holds that:*

$$V(f_1, f_2, \ldots, f_k) = \mathbb{V}(\langle f_1, f_2, \ldots, f_k \rangle) \tag{50}$$

Proof. Let $x \in V(f_1, f_2, \ldots, f_k)$. Then $f_1(x) = f_2(x) = \cdots = f_k(x) = 0$. Since any $f \in I$ is of the form $\sum_{i=1}^{k} g_i(x) f_i(x)$ it holds that $f(x) = 0$ and hence $x \in \mathbb{V}(I)$. On contrary, assume that $x \in \mathbb{V}(I)$. Then $f(x) = 0$, $\forall f \in I$. But since $f_1, f_2, \ldots, f_k \in I$ it holds that $f_1(x) = f_2(x) = \cdots = f_k(x) = 0$ and hence $x \in V(f_1, f_2, \ldots, f_k)$. □

Example 13. *Let $I = \langle x^2 + y^2, x + 1 \rangle$. Then $\mathbb{V}(I) = \{(x, y) \in \mathbb{R}^2;\ x^2 + y^2 = 0, x + 1 = 0\} = \emptyset$.*

Just as we can define a real algebraic set of an ideal, we can also define a defining ideal of a real algebraic set.

Definition 14. *(Defining ideal) Let V be a real algebraic set. The set:*

$$\mathbb{I}(V) = \{f(x) \in \mathcal{R};\ f(x) = 0\ (\forall x \in V)\} \tag{51}$$

is a defining ideal of a real algebraic set V.

[2]link: https://homepage.univie.ac.at/herwig.hauser/bildergalerie/gallery.html

Example 14. *Let $V = \mathbb{V}(I) = V(x^2 + y^2, x + 1)$ as in the previous example. As shown $V = \emptyset$. Then $\mathbb{I}(V(x^2 + y^2, x + 1)) = \mathbb{I}(\emptyset) = \mathbb{R}[x, y]$, i.e. any polynomial function has zeros in the empty set.*

From the previous example we note that $\mathbb{I}(V(I)) \neq I$ in general. Hence, we need to establish suitable relations that hold between all quantities introduced so far. First, we consider relations between I and $\mathbb{V}(I)$, as well as between V and $\mathbb{I}(V)$. Note that algebraic set V, as a zero set of some functions, represent some kind of surface i.e. it is geometric in the nature. On the other hand, polynomials from an ideal I can be analyzed using results from algebra. Hence establishing these relations is at the core of algebraic geometry.

Lemma 5. *For real algebraic sets V_1 and V_2, and ideals I_1 and I_2 holds:*

1. $V_1 \subset V_2 \iff \mathbb{I}(V_1) \supset \mathbb{I}(V_2)$
2. $I_1 \subset I_2 \implies \mathbb{V}(I_1) \supset \mathbb{V}(I_2)$

Proof. See proof of Theorem 3.2 in [1]. □

This lemma seems rather intuitive. If we add zero points to a zero set of functions, we reduce number of functions with such zero points. On the other hand, one should be careful regarding the second implication. Namely, let $I_1 = \langle x \rangle$ and $I_2 = \langle x^2 + 1 \rangle$. Then $I_1 \not\subset I_2$ since $x \in I_1$ but $x \notin I_2$. Simultaneously $\mathbb{V}(I_1) = \{0\}$ but $\mathbb{V}(I_2) = \emptyset$ and hence $\mathbb{V}(I_1) \supset \mathbb{V}(I_2)$.

Lemma 6. *Let I be an ideal and let V be an algebraic set. The map $I \mapsto \mathbb{V}(I)$ is surjective, but not injective. The map $V \mapsto \mathbb{I}(V)$ is injective, but not surjective.*

Let's illustrate this lemma with following examples. Let $I_1 = \langle x \rangle$ and $I_2 = \langle x^2 \rangle$. Obviously $I_1 \neq I_2$ but $\mathbb{V}(I_1) = \mathbb{V}(I_2) = \{0\}$. An example when $V \mapsto \mathbb{I}(V)$ is not surjective is given by $I = \langle x^2 + 1 \rangle$. There exist no V such that $\mathbb{I}(V) = I$.

We are mainly interested in the following process of obtaining ideals from algebraic sets:

$$\underbrace{f_1, \ldots, f_r}_{\text{polynomials}} \to \underbrace{\mathbb{V}(f_1, \ldots, f_r)}_{\text{algebraic set}} \to \underbrace{\mathbb{I}(\mathbb{V}(f_1, \ldots, f_r))}_{\text{ideal}}$$

But we do not know what relations hold between initial polynomials $f_1, \ldots, f_r$ and formed ideal. As we show in theorem 6, there is a simple inclusion relation between these quantities. Before going into this theorem, let's introduce notion of a radical of an ideal.

Definition 15. *(Radical of an ideal) Let $I \subset \mathcal{R}$ be an ideal. The radical ideal of I is:*

$$\sqrt{I} = \{f \in \mathcal{R};\ f^m \in I\ (\exists m \in \mathbb{N})\} \tag{52}$$

Remark. *If $I = \sqrt{I}$ we say that I is a radical ideal. $\mathbb{I}(V)$ is a radical ideal.*

Theorem 6. *For arbitrary real algebraic set V holds $\mathbb{V}(\mathbb{I}(V)) = V$. For arbitrary ideal $I \in \mathcal{R}$ holds $I \subset \sqrt{I} \subset \mathbb{I}(\mathbb{V}(I))$.*

Proof. Let $x \in V$. Then $\mathbb{I}(V)$ contains all polynomials vanishing at x. By definition of $\mathbb{V}(I)$ we have $x \in \mathbb{V}(\mathbb{I}(V))$ i.e. $V \subset \mathbb{V}(\mathbb{I}(V))$. For the opposite direction of the proof assume that $V = V(f_1, f_2, \ldots, f_k)$. Then $f_1, f_2, \ldots, f_k \in \mathbb{I}(V)$ and hence $\langle f_1, f_2, \ldots, f_k \rangle \subset \mathbb{I}(V)$. By lemma 5 we have that $\mathbb{V}(\langle f_1, f_2, \ldots, f_k \rangle) \supset \mathbb{V}(\mathbb{I}(V))$. From (50) we have $V \supset \mathbb{V}(\mathbb{I}(V))$. Combining results we obtain $V = \mathbb{V}(\mathbb{I}(V))$.

Inclusion $I \subset \sqrt{I}$ is obvious since we can take $m = 1$ in definition 15. Now let's prove $\sqrt{I} \subset \mathbb{I}(\mathbb{V}(I))$. Let $f \in \sqrt{I}$. Then there exist $m \in \mathbb{N}$ such that $f^m \in I$. Hence f^m vanishes on $\mathbb{V}(I)$ and therefore f must vanish on $\mathbb{V}(I)$ as well. Hence $f \in \mathbb{I}(\mathbb{V}(I))$. □

Remark. *For polynomial rings with coefficients in algebraic closed field (e.g. $\mathbb{C}$) holds $\sqrt{I} = \mathbb{I}(\mathbb{V}(I))$. This is result of the famous Hilbert's Nullstellensatz which, however, does not hold for polynomial rings with real coefficients.*

We will also need notions of prime ideals and irreducible sets:

Definition 16. *(Prime ideal and irreducible set)*

1. *If an ideal $I \subset \mathcal{R}$ satisfies:*

$$I \neq \mathcal{R} \quad \text{and} \quad f \cdot g \in I \implies f \in I \vee g \in I \tag{53}$$

then I is called a prime ideal.

2. *If a real algebraic set $V \subset \mathbb{R}^d$ satisfies:*

$$V = V_1 \cup V_2 \implies V = V_1 \vee V = V_2 \tag{54}$$

then V is said to be irreducible.

Let's mention a few more useful properties of algebraic sets and ideals [1]:

- A real algebraic set V is irreducible if and only if $\mathbb{I}(V)$ is a prime ideal.
- For every algebraic set V there exist irreducible real algebraic sets V_1, V_2, ..., V_k such that: $V = V_1 \cup V_2 \cup \cdots V_k$.
- A prime ideal is a radical ideal.

Therefore if we determine decomposition of real algebraic set V into its irreducible components, on each of its components we have that $\mathbb{I}(V_k)$ is prime and hence radical ideal, meaning: $I = \sqrt{I} \subset \mathbb{I}(\mathbb{V}(I))$.

In the following section we will employ notions introduced so far for determining singularities of real algebraic sets.

2.2 Singularities

We have established basic notions regarding the singular models. In order to have a better understanding of these models, we need to examine the zero set of function $K(w)$ and analyze its singularities. Hence, in this section we aim to answer the following question:

How to check if a point is singularity of a set and how to find all singularities of a set?

Let's first formally define singular points of an arbitrary set.

Definition 17. *(Singularities of a set) We say that a point P from a nonempty set $A \subset \mathbb{R}^d$ is nonsingular if there exist open sets U and V and an analytic isomorphism $f : U \to V$ such that:*

$$f(A \cap U) = \{(x_1, x_2, \ldots, x_r, 0, 0, \ldots, 0); x_i \in \mathbb{R}\} \cap V \tag{55}$$

for some nonnegative integer r. If a point P in A is not nonsingular, it is singularity of set A.

We illustrate this definition in the Figure 5. We could say that a point is nonsingular if there exists some invertible analytic transformation of its neighborhood which maps it into lower dimensional space.

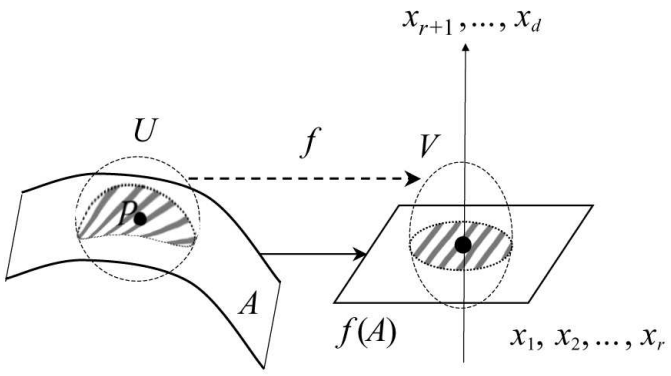


Figure 5: Determining if point of a set is singular. Figure is adapted from Fig. 2.3. in [1].

As we mentioned at the beginning, main goal of this section is determining singular points of a real algebraic set. In theorem 8 we give sufficient conditions for determining nonsingular points of a real algebraic set. In order to prove this theorem we will need so-called inverse function theorem that states the following.

Theorem 7. *(Inverse function theorem, theorem 2.1 in [1]) Let $U \subset \mathbb{R}^d$ be an open set and $f : U \to \mathbb{R}^d$ be a function of class C^r $(1 \leq r \leq \omega)$. Suppose that $x_0 \in U$ and the Jacobian matrix at x_0 is invertible. Then there exists an open set $U' \subset U$, containing x_0, such that f is C^r-isomorphism of U' and $f(U')$.*

Theorem 8. *(A sufficient condition for a nonsingular point, theorem 2.2 in [1]) Let $U = V(f_1, f_2, \ldots, f_r) \subset \mathbb{R}^d$ be arbitrary real algebraic set and let J be Jacobian defined by:*

$$J(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \cdots & \frac{\partial f_1}{\partial x_r}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_r}{\partial x_1}(x) & \cdots & \frac{\partial f_r}{\partial x_r}(x) \end{bmatrix} \tag{56}$$

If a point $x_0 \in U$ satisfies rank $J(x_0) = r$ *then x_0 is a nonsingular point of U.*

Proof. Since $r \leq d$ we define functions $f_i(x) = x_i$ for $i = r+1, \ldots, d$. Then we define the extended Jacobian as follows:

$$J(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \cdots & \frac{\partial f_1}{\partial x_d}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_d}{\partial x_1}(x) & \cdots & \frac{\partial f_d}{\partial x_d}(x) \end{bmatrix} \tag{57}$$

and note that by definition of $f_i(x)$ (for $i = r+1, \ldots, d$) $\frac{\partial f_i}{\partial x_j}(x) = \delta_{ij}$ where δ_{ij} is Kronecker delta function, and hence this part of matrix $J(x)$ forms an identity submatrix. In addition, we have condition that rank $J(x_0) = r$ for $x_0 \in U$, and hence matrix $J(x_0)$ is invertible. By inverse function theorem i.e. theorem 7 there exists an open set $U' \subset U$ such that $x_0 \in U$ and f is analytic isomorphism of U' and $f(U')$. For any point $x \in U \cap U'$ we have $f_1(x) = \cdots = f_r(x) = 0$ and hence $f(x) = (0, 0, \ldots, 0, x_{r+1}, \ldots, x_d) \in f(U')$. According to definition (55) this implies that $x_0 \in U$ is a nonsingular point. □

Example 15. *Let $f(x) = (x-1)^2$ and $V = V(f)$. If we apply theorem 8 for Jacobian $J(x) = \frac{\partial f(x)}{\partial x} = 2(x-1)$, we can conclude that all points $x \neq 1$ are nonsingular. However, we cannot claim that $x = 1$ is a singular point. But note that $V(f) = V((x-1)^2) = V(x-1)$ and hence we can define $f_1(x) = x - 1$ and $U = V(f_1)$. Then the generalized Jacobian is given by $J(x) = \frac{\partial f_1(x)}{\partial x} = 1$ implying that all points, including $x = 1$, are nonsingular.*

Theorem 8 holds for arbitrary polynomials $f_1, \ldots, f_r$. However, in order to obtain theorem that gives both sufficient and necessary conditions for nonsingular points, we need to consider specific polynomials $f_1, \ldots, f_r$, as we show next. In order to gain intuition behind theorem 9, we need to define tangent space as follows.

Definition 18. *(Tangent space, definition 1 in section 9.6 in [12]) Let V be a real algebraic set, let $P = (P_1, P_2, \ldots, P_d) \in V$ be a given point and $f \in \mathbb{R}[x_1, x_2, \ldots, x_d]$ a polynomial. The linear part of f at P is defined by:*

$$d_P(f) = \frac{\partial f(P)}{\partial x_1}(x_1 - P_1) + \cdots + \frac{\partial f(P)}{\partial x_d}(x_d - P_d) = \nabla f(P)^T (x - P) \tag{58}$$

Furthermore, the tangent space of V at P is given by:

$$T_P(V) = \mathbb{V}(d_P(f) : \ f \in \mathbb{I}(V)) \tag{59}$$

Now let $\mathbb{I}(V) = \langle f_1, f_2, \ldots, f_r \rangle$ and $P = (P_1, \ldots, P_d) \in V$. Using definition (59), tangent space at P is given by:

$$T_P(V) = \mathbb{V}(d_P(f) : \ f \in \mathbb{I}(V)) = \mathbb{V}(d_P(f) : \ f \in \langle f_1, f_2, \ldots, f_r \rangle) = \mathbb{V}(d_P(f_i) : \ i = 1, \ldots, r) \tag{60}$$

i.e. it is the zero set of $d_P(f_i)$ for $i = 1, \ldots, r$. Hence we can write:

$$x \in T_P(V) \iff \begin{bmatrix} \frac{\partial f_1(P)}{\partial x_1} & \cdots & \frac{\partial f_1(P)}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_r(P)}{\partial x_1} & \cdots & \frac{\partial f_r(P)}{\partial x_d} \end{bmatrix} \begin{bmatrix} x_1 - P_1 \\ \vdots \\ x_d - P_d \end{bmatrix} = 0 \tag{61}$$

Hence the tangent space of V at point P is determined by the kernel of generalized Jacobian J at point P. Furthermore, we can define dimension of $T_P(V)$ as $\dim T_P(V) =$ Nullity $J(P)$. In order to determine singular points of a real algebraic set, we need to compare dimension of the tangent space at a point with dimension of the set V at this point, defined as follows.

Definition 19. *(Dimension of an irreducible real algebraic set, definition 3.8 in [1]) Let $V \subset \mathbb{R}^d$ be a nonempty irreducible real algebraic set. Let polynomials $f_1, f_2, \ldots, f_r$ satisfy: $\mathbb{I}(V) = \langle f_1, f_2, \ldots, f_r \rangle$. For a generalized Jacobian matrix of the form:*

$$J(x) = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_r(x)}{\partial x_1} & \cdots & \frac{\partial f_r(x)}{\partial x_d} \end{bmatrix} \tag{62}$$

we define dimension of V as:

$$\dim V = \max_{x \in V} \operatorname{rank}\ J(x) \tag{63}$$

Note that we assumed V to be an irreducible real algebraic set in the previous definition. Defining dimension of an irreducible set is convenient since, every point of an irreducible real algebraic set V has the same local dimension equal to $\dim V$ (see theorem 11.25 in [13]). In the case when V consists of multiple irreducible components, one should consider dimension only of that component containing given point (see definition 6 in section 9.6 in [12]). Moreover, in the case of irreducible components with all equal dimensions, one could use definition 19 (as we do in example 16).

The following theorem is sometimes used as definition of nonsingular points [12]. What it actually states is that dimension of a real algebraic set in a nonsingular point is equal to the dimension of tangent space in that point. Moreover, for a singular point P holds that $\dim T_P(V) > \dim V$ (e.g. see theorem 8 in chapter 9.6 in [12]). Recall also the fact that Rank $J(P)$ + Nullity $J(P) = d$ for any point $P \in \mathbb{R}^d$. Hence condition $\dim T_P(V) > \dim V$ can be rewritten as $d - \text{Rank } J(P) > \dim V$ i.e. Rank $J(P) < d - \dim V$.

Theorem 9. *(A necessary and sufficient condition for a nonsingular point, theorem 2.3 in section V in [14]) Let $V \subset \mathbb{R}^d$ be an irreducible real algebraic set of dimension d_0. Let $\mathbb{I}(V) = \langle f_1, f_2, \ldots, f_r \rangle$ be ideal of V for some polynomials $f_1, \ldots, f_r$. Then, a point $x \in V$ is nonsingular if and only if* Rank $J(x) = d - d_0$ *i.e. singular if and only if* Rank $J(x) < d - d_0$*. Furthermore, the set of singular points of V is the common zero set in V of polynomials obtained from the $(d-d_0) \times (d-d_0)$ minors of the generalized Jacobian:*

$$J = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_r}{\partial x_1} & \cdots & \frac{\partial f_r}{\partial x_d} \end{bmatrix} \tag{64}$$

Proof. For the equivalence $x \in V$ is nonsingular $\iff$ Rank $J(x) = d - d_0$ see theorem 2.3 in section V.2. in [14]. The second statement follows from the fact that for all singular points we must have Rank $J(x) < d - d_0$ and hence all $(d - d_0) \times (d - d_0)$ minors of J must be zero. Therefore, any singularity of V is a common zero of polynomials obtained from $(d - d_0) \times (d - d_0)$ minors of J. □

Example 16. *Let $f(x, y) = xy$ and $\mathbb{I}(V) = \langle f \rangle = \langle xy \rangle$. Then $J(x) = \begin{bmatrix} y & x \end{bmatrix}$. Note that V contains two irreducible components $V(x)$ and $V(y)$ of dimension 1. Therefore, singular points are all common zeros of 1×1 minors of $J(x)$ i.e. only point $(x, y) = (0, 0)$ is a singular point of V.*

In conclusion, we showed that in principle one could determine singularities of a algebraic set determined by some finite set of polynomials $\{f_1, \ldots, f_r\}$. Assumption of finite generating set of an ideal is justified in the following section.

2.3 The ideal description problem

In this section we will consider so-called ideal description problem, which answers the following question:

Does every ideal $I \in \mathcal{R}$ have a finite generating set $\{f_1, f_2, \ldots, f_r\}$ (and how to find it)?

The following theorem answers this question affirmatively claiming that there exists a finite generating set for any ideal.

Theorem 10. *(Hilbert basis theorem, theorem 4 in section 2.5 in [12]) Every ideal $I \in \mathcal{R}$ has a finite generating set, i.e. $I = \langle f_1, f_2, \ldots, f_r \rangle$ for some $f_1, f_2, \ldots, f_r \in I$.*

One should note that this theorem is not constructive. However, theory of Groebner bases and algorithms such as Buchberger's algorithm are applicable for constructing such generating sets. For this results as well as proof of Hilbert's basis theorem check section 2 in [12].

Hilbert basis theorem does not hold for ideals from a general ring. Those rings for which Hilbert basis theorem holds are called Noetherian [1]. The following corollary of Hilbert basis theorem will be needed in example 17.

Corollary 1. *(Noetherian ring, remark 3.1 in [1]) For an arbitrary sequence of polynomial ideals $I_i \in \mathcal{R}$ $(i = 1, 2, \dots)$ such that:*

$$I_1 \subset I_2 \subset \cdots \subset I_n \subset \cdots \tag{65}$$

there exists an integer N such that:

$$I_n = I_{n+1} = I_{n+2} = \cdots =, \quad \forall n \geq N \tag{66}$$

Proof. First, $\cup_k I_k$ is an ideal in $\mathcal{R}$ and by Hilbert basis theorem it has a finite generating set $\{f_1, \dots, f_r\}$. Since sequence of ideals is nondecreasing, there must be some ideal I_N that contains all elements of the set $\{f_1, \dots, f_r\}$. Therefore $I_{N+1} = I_{N+2} = \cdots = \cup_k I_k$. □

In order to obtain finite generating sets in general case, results from algebraic geometry are necessary. However, we show in the following example, that in the case of suitably chosen statistical models, finite generating sets can be determined with standard algebra.

Example 17. *Let neural network be of the following form $\Phi_w(x) = \sum_{i=1}^{H} a_i(e^{b_i x} - 1)$, where $w = [a_1\ a_2 \dots a_H\ b_1 b_2 \dots b_H]$ for some positive integer H. Assume that the true distribution is given by setting all parameters to zero. Then according to equation (8):*

$$K(w) = \frac{1}{2} \int q(x) \Big(\sum_{i=1}^{H} a_i (e^{b_i x} - 1) \Big)^2 dx \tag{67}$$

We search for parameters w such that $K(w) = 0$. We can rewrite quadratic term from equation (67) as:

$$\sum_{i=1}^{H} a_i (e^{b_i x} - 1) = \sum_{i=1}^{H} a_i \Big(\Big(\sum_{k=0}^{\infty} \frac{x^k}{k!} b_i^k \Big) - 1 \Big) = \sum_{k=1}^{\infty} \frac{x^k}{k!} \sum_{i=1}^{H} a_i b_i^k \tag{68}$$

Since the set of functions $\{x^k\}_{k=1}^{\infty}$ is linearly independent, in order to have $K(w) \equiv 0$ for any $q(x)$, we need:

$$p_k = \sum_{i=1}^{H} a_i b_i^k = 0, \quad k = 1, 2, \dots \tag{69}$$

Let's define ideals $I_k = \langle p_1, p_2, \dots, p_k \rangle$, for $k = 1, 2, \dots$. Then, since $I_1 \subset I_2 \subset I_3 \subset \cdots$, from corollary 1 follows that there exist an integer N such that $I_n = I_N$, $\forall n \geq N$. Therefore (69) holds if and only if the first N equalities are satisfied. We claim that $N = H$, i.e. that:

$$p_{k+1} = p_k \Big(\sum_{i=1}^{H} b_i \Big) - p_{k-1} \Big(\sum_{\mathcal{I}_2} \prod_{i \in \mathcal{I}_2} b_i \Big) + \cdots + (-1)^{H-2} p_{k-H+2} \Big(\sum_{\mathcal{I}_{H-1}} \prod_{i \in \mathcal{I}_{H-1}} b_i \Big) + (-1)^{H-1} p_{k-H+1} \Big(\prod_{i=1}^{H} b_i \Big)$$

$$\implies p_{k+1} = \sum_{t=1}^{H} (-1)^{H-t} p_{k-H+t} \Big(\sum_{\mathcal{I}_t} \prod_{i \in \mathcal{I}_t} b_i \Big) \tag{70}$$

where $\mathcal{I}_k$ denotes indices of all combinations without repetitions of k elements from $\{b_1, b_2, \dots, b_H\}$. Comparing coefficients of terms multiplying a_i for $i = 1, \dots, H$ we can prove that the equality above holds. To conclude we have that $p_{H+1} \in \langle p_1, p_2, \dots, p_H \rangle$ and by induction for all $k > H$ holds that $p_k \in \langle p_1, p_2, \dots, p_H \rangle$. To sum up, we have that $p_k \in \langle p_1, p_2, \dots, p_H \rangle$ for any nonnegative integer k, i.e. $\{p_1, p_2, \dots, p_H\}$ is finite generating set for any ideal I_k.

3 Resolution of singularities

Now we are ready to discuss our main topic of interest from algebraic geometry - theorem of resolution of singularities. Applying this theorem to the real algebraic set defined by $K(w) = 0$ we will "reduce complexity" of the singularities and obtain an expression for $K(w)$ in so-called normal crossing form. From this type of desingularized set $K(w) = 0$ we will be able to deduce asymptotic behavior of statistical variables.

3.1 Preparation for desingularization

In this section we introduce basic concepts from algebraic geometry - real analytic manifolds and real projective spaces. Motivation behind introducing these terms is the following. We show that analytic sets without singular points form manifolds. Hence we will seek to resolve singularities of real analytic sets in such a way to obtain manifolds as a final result of resolution. We will also show that main theorem for resolving singularities called Hironaka's theorem is in fact a successive procedure of applying transformations called blow-ups. To put simply, blowing-up a space requires adding additional coordinates that actually belong to a projective space. Finally, it will be useful to establish connection between manifolds and projective spaces.

3.1.1 Manifold

In a nutshell, a manifold is an object that looks like Euclidean space $\mathbb{R}^d$ in local neighborhood of each point. A simple motivating example is map of the Earth. Namely, if we consider only map of the Switzerland it does look like it lies in a plane. However, when we consider map of the whole world, like the one on a globe, it is impossible to map all points to a plane (without cutting the globe). Let's now give a more formal introduction to the manifolds. In order to keep highest generality, we will work on topological spaces defined as follows.

Definition 20. *(Topological space) For a set X, we define topology on X as a collection $\mathcal{T}$ of subsets of X satisfying:*

1. *$X, \emptyset \in \mathcal{T}$,*
2. *$\mathcal{T}$ is closed under finite intersections, i.e: $U_1, \ldots, U_n \in \mathcal{T} \implies U_1 \cap \cdots \cap U_n \in \mathcal{T}$,*
3. *$\mathcal{T}$ is closed under arbitrary unions, i.e: $(U_\alpha)_{\alpha \in A} \in \mathcal{T} \implies \cup_{\alpha \in A} U_\alpha \in \mathcal{T}$.*

A pair $(X, \mathcal{T})$ is called a topological space.

Then we define the notion of similarity that we will use to claim that manifolds locally resemble Euclidean spaces.

Definition 21. *(Homeomorphism) Let X and Y be topological spaces. A bijective map $\phi : X \to Y$ such that ϕ and ϕ^{-1} are continuous is called homeomorphism from X to Y. If such map exists, we say that X and Y are homeomorphic (or topologically equivalent).*

Definition 22. *(Locally Euclidean) A topological space M is locally Euclidean of dimension d if every point of M has a neighborhood homeomorphic to an open subset of $\mathbb{R}^d$.*

The rest of the properties manifolds must have are, for our purposes, rather technical and satisfied in all decent cases we will consider. However, in order to give a self-contained presentation of the topic, we shortly introduce these properties.

Definition 23. *(Hausdorff space) A topological space X is called Hausdorff space if any two distinct points can be separated by open subsets, i.e. if for any two distinct points p_1, $p_2 \in X$ there exist neighborhood U_1 of p_1 and U_2 of p_2 such that $U_1 \cap U_2 = \emptyset$.*

Definition 24. *(Basis for the topology) For a topological space X we define basis as a collection $\mathcal{B}$ of subsets of X such that:*

1. *every element of $\mathcal{B}$ is an open subset of X,*
2. *every open subset of X is the union of some collection of elements of $\mathcal{B}$.*

Definition 25. *(Second countable) A topological space is second countable if there exists a countable basis for its topology.*

At last, we are now ready to define manifolds in a precise way.

Definition 26. *(Manifold) An d-dimensional manifold is a second countable Hausdorff space that is locally Euclidean of dimension d.*

Conditions that the space is Hausdorff and second countable are satisfied in decent topological spaces we will consider. Namely, in Euclidean spaces (using standard, Euclidean topology) we can always define small open balls around any two different points such that they do not intersect (i.e. space is Hausdorff). Furthermore, we know that all metric spaces are Hausdorff [15]. Also, taking open balls with rational centers and rational radii gives a countable basis in Euclidean spaces. Hence, in our cases we will mainly be interested in examining whether space is homeomorphic to Euclidean spaces in a neighborhood of every point. Our intuition that we can expect problems only around singular points is confirmed in the following example.

Example 18. *(A nonsingular analytic set is a real analytic manifold [1]) Let's define a real analytic function $f : U \to \mathbb{R}$ for some open set $U \in \mathbb{R}^d$. If it satisfies $\nabla f(x) \neq 0$ on the real analytic set $U_0 = \{x \in U;\ f(x) = 0\}$ then U_0 is a manifold. This statement is for real algebraic sets known as Nash-Tognoli theorem that states:*

Theorem 11. *(Nash-Tognoli theorem, Theorem 2 in [16]) Every smooth, compact manifold is diffeomorphic[3] to a smooth, real algebraic set.*

Therefore, by using procedure of resolution of singularities we expect to obtain smooth, real algebraic set (without singularities) and hence a smooth, compact manifold.

The properties of n-dimensional manifold imply that there exist some open sets $U_1, U_2, \ldots$ which are homeomorphic to open subsets of $\mathbb{R}^d$ and cover the whole manifold M. Moreover, assuming compactness of M there is a finite number of elements in that set: $U_1, U_2, \ldots, U_k$ such that $M \subseteq \cup_{i=1}^{k} U_i$. Sets U_i are called local coordinates (or charts), set $\{U_i\}_{i=1}^{K}$ is a set of local coordinates (or atlas). Moreover, since different local coordinates may have nonempty intersection it is necessary to ensure certain consistency on those sets. Namely, define transition maps $\phi_i : U_i \mapsto \phi_i(U_i)$, where $\phi_i(U_i) \subset \mathbb{R}^d$. Then we enforce consistency on the intersection of local coordinates by defining real analytic manifold in the following manner.

Definition 27. *(Real analytic manifold) Let M be a manifold with local coordinates $\{U_i\}$. If for every set $U^* = U_i \cap U_j \neq \emptyset$ both maps:*

$$\phi_j \circ \phi_i^{-1} : \phi_i(U^*) \to \phi_j(U^*) \tag{71}$$

$$\phi_i \circ \phi_j^{-1} : \phi_j(U^*) \to \phi_i(U^*) \tag{72}$$

are real analytic functions, then M is a real analytic manifold.

In Figure 6 is given illustration of definition 27. Since we expect that $\phi_i(U^*)$ and $\phi_j(U^*)$ are "similar", we need a real analytic function that shows it is indeed possible to transform $\phi_i(U^*)$ into $\phi_j(U^*)$ really smoothly, and vice versa.

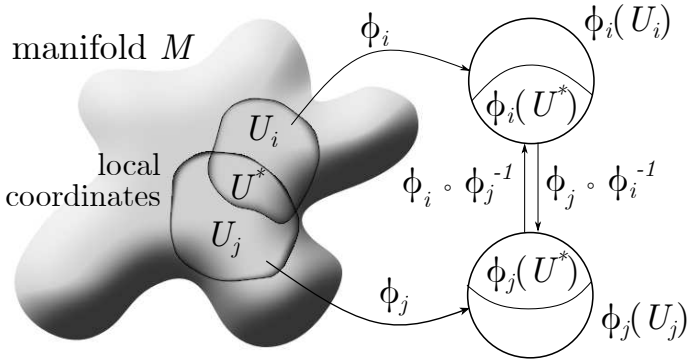


Figure 6: Real analytic manifold - ensuring consistency of overlapping local coordinates. Figure is inspired by Fig. 2.8. in [1].

In the following section we introduce real projective space, which is just a special type of manifold. Therefore, the considerations made in this section remain valid in the following one as well.

[3]Given two manifolds X and Y, a bijective differentiable map $f : X \to Y$ is called diffeomorphism if its inverse is also differentiable.

3.1.2 Real projective space

We start this section by giving some intuition behind projective spaces (for more thorough analysis of this topic see [12]). From elementary geometry we know that intersection of two lines in $\mathbb{R}^2$ is a point, unless lines are parallel - in that case we might say that those lines intersect in a point at ∞. Moreover we would like that different parallel lines intersect at different points at ∞. In order to include those points at infinity in our space, we will introduce the notion of projective spaces.

Note that different points on one line define the same line. In order to remove this type of redundancy, we need to introduce homogeneous coordinates via following equivalence relation.

Definition 28. *Let $\sim$ be an equivalence relation defined for every pair of vectors $(x_0, x_1, \ldots, x_d)$, $(x'_0, x'_1, \ldots, x'_d) \in \mathbb{R}^{d+1} \setminus \{0\}$ as follows:*

$$(x_0, x_1, \ldots, x_d) \sim (x'_0, x'_1, \ldots, x'_d) \iff (\exists c \neq 0)(x'_0, x'_1, \ldots, x'_d) = c(x_0, x_1, \ldots, x_d), \tag{73}$$

A class of equivalence which contains a point $(x_0, x_1, \ldots, x_d)$ is denoted by $(x_0 : x_1 \cdots : x_d)$.

It is straightforward that every point in $(\mathbb{R}^{d+1} \setminus \{0\})/\sim$ defines a line in $\mathbb{R}^{d+1}$. Let's pick an arbitrary point $(x, y, z) \in \mathbb{R}^3 \setminus \{0\}$. This point defines a line going through origin in $\mathbb{R}^3$. First assume that this line is not in the $z = 0$ plane. Then, intersecting this line with plane $z = 1$ we obtain a unique point $(x, y, 1)$ (see Figure 7). Hence a point $(x, y) \in \mathbb{R}^2$ defines a line in $\mathbb{R}^3$ going through origin, and vice versa. Now, assume that given line lies in the plane $z = 0$. Then, it would not intersect plane $z = 1$. Those lines that lie in the plane $z = 0$ correspond exactly to the points at infinity, as we previously introduced. Moreover, note that we can characterize those lines in plane $z = 0$ by using coordinates $(x, y, 0)$. In a nutshell, we have added points at infinity that we missed in standard Euclidean spaces into projective space.

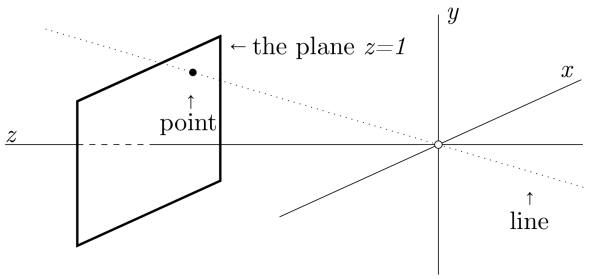


Figure 7: Projective space - identification of lines in $\mathbb{R}^3$ with points on plane $z = 1$. Image is from section 8.1 in [12].

Let's now formally introduce projective space.

Definition 29. *(Projective space) The d-dimensional projective space over $\mathbb{R}$, denoted $\mathbb{P}^d$ is defined as:*

$$\mathbb{P}^d = (\mathbb{R}^{d+1} \setminus \{0\})/\sim \tag{74}$$

Every point $(x_0, x_1, \ldots, x_d) \in \mathbb{R}^{d+1} \setminus \{0\}$ defines a point $p \in \mathbb{P}^d$ in projective space which we denote by $p = (x_0 : x_1 : \cdots : x_d)$.

The following theorem is of great importance to establish connection between projective and Euclidean spaces.

Theorem 12. *Projective space $\mathbb{P}^d$ is a d-dimensional manifold.*

Proof. Here we give proof that projective space $\mathbb{P}^d$ is locally Euclidean of dimension d.

Let's define the following sets:

$$U_k = \{(x_0 : x_1 : \cdots : x_{k-1} : 1 : x_{k+1} : \cdots : x_d) \in \mathbb{P}^d; x_i \in \mathbb{R}, i \in \{0, 1, \ldots, d\} \setminus \{k\}\} \tag{75}$$

for $k = 0, 1, \ldots, d$. Define the following mappings $\phi_k : U_k \to \mathbb{R}^d$ $(k = 0, 1, \ldots, d)$:

$$\phi_k : (x_0 : x_1 : \cdots : x_{k-1} : 1 : x_{k+1} : \cdots : x_d) \mapsto (x_0, x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_d) \tag{76}$$

Mappings ϕ_k are bijective (on U_k), continuous with continuous inverses. Hence U_k and $\mathbb{R}^d$ are homeomorphic. Moreover, since any element in $\mathbb{P}^d$ has at least one nonzero coordinate, we can scale all coordinates by this nonconstant value and hence this point will be a point in one of U_k's i.e. we have $\mathbb{P}^d \subset U_0 \cup U_1 \cup \cdots \cup U_d$. Also since every $U_k \subset \mathbb{P}^d$ we have that $U_0 \cup U_1 \cup \cdots \cup U_d \subset \mathbb{P}^d$. Together this implies that $\mathbb{P}^d = U_0 \cup U_1 \cup \cdots \cup U_d$. Furthermore, we can say that $\mathbb{P}^d$ is locally Euclidean of dimension d. □

We illustrate theorem 12 in the following example, and also show that $\mathbb{P}^2$ is homeomorphic to union of $\mathbb{R}^2$ and a "projective line at infinity".

Example 19. *Consider two-dimensional projective space $\mathbb{P}^2$. By definition:*

$$\begin{aligned}\mathbb{P}^2 &= \{(x:y:z); (x,y,z) \in \mathbb{R}^3 \setminus \{0\}\} = \{(1:y:z); (y,z) \in \mathbb{R}^2\} \cup \{(0:y:z); (y:z) \in \mathbb{P}^1\} \\ &\cong \mathbb{R}^2 \cup \mathbb{P}^1 \end{aligned} \tag{77}$$

where $\cong$ denotes homeomorphism of two spaces. Analogously, one can prove that $\mathbb{P} \cong \mathbb{R}^1 \cup \mathbb{R}^0$, where $\mathbb{R}^0$ is homeomorphic to a point $(0:1)$. Hence the projective space $\mathbb{P}^2$ is homeomorphic to a union of Euclidean spaces: $\mathbb{P}^2 \cong \mathbb{R}^2 \cup \mathbb{R}^1 \cup \mathbb{R}^0$.

Defining functions on projective spaces requires some caution. Namely, for any nonzero constant λ holds that $(x_0 : x_1 : \cdots : x_d) = (\lambda x_0 : \lambda x_1 : \cdots : \lambda x_d)$ and hence we want that scaling of coordinates leads only to the scaling of functions i.e. that:

$$f(\lambda x) = f(\lambda x_0, \lambda x_1, \ldots, \lambda x_d) = g(\lambda) f(x_0, x_1, \ldots, x_d) = g(\lambda) f(x) \tag{78}$$

with $g(\lambda) \neq 0$, or more simply - that $f(\lambda x) = 0 \iff f(x) = 0$. In this way we can ensure that all points on one line in $\mathbb{R}^{d+1}$ are either zeros of function f or not. A class of functions that satisfy this condition is class of homogeneous polynomials, defined as follows.

Definition 30. *(Homogeneous polynomial) If a polynomial $f \in \mathbb{R}[x_0, x_1, \ldots, x_d]$ can be written as:*

$$f(x) = \sum_{|\alpha|=n} a_\alpha x^\alpha \tag{79}$$

then f is a homogeneous polynomial of degree n.

First note that we have used multi-index notation as in equation (45). Now that we can define appropriate functions on projective spaces, it will be crucial in the following discussion to consider zero sets of these homogeneous functions in projective spaces, called real projective varieties.

Definition 31. *(Real projective variety) Let $f_1, f_2, \ldots, f_k \in \mathbb{R}[x_0, x_1, \ldots, x_d]$ be homogeneous polynomials. A set $V \subset \mathbb{P}^d$ defined by $f_1, f_2, \ldots, f_k$:*

$$V = \{(x_0 : x_1 : \cdots : x_d) \in \mathbb{P}^d; f_1(x) = f_2(x) = \cdots = f_k(x) = 0\} \tag{80}$$

is a real projective variety.

A real algebraic set and a real projective variety are special cases of a real algebraic variety [1]. Just as for real algebraic sets, theorem 6 and lemma 5 hold even for real projective varieties [1].

Example 20. *Let $f(x,y,z) = x^3 + xyz + y^2 z$. Since f is a homogeneous polynomial of degree $n = 3$, define a real projective variety V as:*

$$V = \{(x:y:z) \in \mathbb{P}^2; x^3 + xyz + y^2 z = 0\} \tag{81}$$

Let U_0, U_1, U_2 be local coordinates of $\mathbb{P}^2$ as defined in (75). Then:

$$V_0 = V \cap U_0 = \{(1:y:z) \in \mathbb{P}^2; 1 + yz + y^2 z = 0\} = \{(y,z) \in \mathbb{R}^2; 1 + yz + y^2 z = 0\} \tag{82}$$

$$V_1 = V \cap U_1 = \{(x:1:z) \in \mathbb{P}^2; x^3 + xz + z = 0\} = \{(x,z) \in \mathbb{R}^2; x^3 + xz + z = 0\} \tag{83}$$

$$V_2 = V \cap U_2 = \{(x:y:1) \in \mathbb{P}^2; x^3 + xy + y^2 = 0\} = \{(x,y) \in \mathbb{R}^2; x^3 + xy + y^2 = 0\} \tag{84}$$

Hence we see that V_i (for $i = 0, 1, 2$) are now real algebraic sets. Also, each local coordinate U_i and real algebraic set V_i uniquely determine initial real projective variety. Take for example U_0 and V_0, and by scaling coordinates of V_0 we have:

$$1 + \frac{y}{x}\frac{z}{x} + \left(\frac{y}{x}\right)^2 \frac{z}{x} = 0 \implies x^3 + xyz + y^2 z = 0 \tag{85}$$

which is defining equation of real projective variety V.

In the previous sections the main tools from algebraic geometry needed for resolution of singularities have been introduced. Since the work of Heisuke Hironaka [17], Japanese mathematician and Fields Medal winner (1970), it has been known that for any algebraic variety (over a field of characteristic zero[4]) singularities can be resolved in an iterative approach. Every step of this method consists of applying a transformation called blow-up, which will be introduced next.

3.2 Blow-up

In this section we consider a real algebraic set V of the form $V = \{x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d; x_1 = x_2 = \cdots = x_r = 0\}$ with $2 \leq r \leq d$.

Definition 32. *(Blow-up of a real algebraic set) Let $W \supset V$ and $W \subseteq \mathbb{R}^d$ be a real algebraic set. The blow-up of W with center V is given by:*

$$B_V(W) = \overline{\{(x, (x_1 : x_2 : \cdots : x_r)) \in \mathbb{R}^d \times \mathbb{P}^{r-1};\ x \in W \setminus V\}} \tag{86}$$

where closure is taken with respect to the Euclidean topology.

In general, the algebraic set W is a set containing singular points. Moreover, in examples that are of interest to us, we define V as a set of singularities in W. Since in definition 32 at first we take only $x \in W \setminus V$, the points in set V are not used in determining $B_V(W)$. However, by taking closure of the set, we add points $\{(x, y); x \in V, y \in \mathbb{P}^{r-1}\}$. Applying transform to the whole set W yields so-called total transform as introduced next.

Definition 33. *Let $B_V(W)$ be a blow-up of a real algebraic set W with center V. We define a projection map $\pi : B_V(W) \mapsto W$. Then we call $B_V(W)$ a strict transform of W, $\pi^{-1}(W)$ a total transform and $\overline{\pi^{-1}(W) \setminus B_V(W)}$ an exceptional set.*

In the following example, we will show how blow-up works and what are possible methods to determine $B_V(W)$ from given W and V.

Example 21. *(Example 3.13 in [1]) Let $f(x, y) = y^2 - x^3$ and $W = \mathbb{V}(f) = \{(x, y) \in \mathbb{R}^2 : y^2 = x^3\}$. In this example we have that $\mathbb{I}(\mathbb{V}(f)) = \langle f \rangle$. The Jacobian matrix is given by:*

$$J(x, y) = \begin{bmatrix} \frac{\partial f(x,y)}{\partial x} & \frac{\partial f(x,y)}{\partial y} \end{bmatrix} = \begin{bmatrix} -3x^2 & 2y \end{bmatrix} \tag{87}$$

Hence:

$$\text{rank } J(x) = \begin{cases} 0, & (x, y) = (0, 0) \\ 1, & \text{otherwise} \end{cases} \tag{88}$$

We conclude from theorem 9 that point $(0, 0)$ is the only singular point. Therefore we set $V = \{(x, y) \in \mathbb{R}^2; x = y = 0\} = \{(0, 0)\}$ ($r = 2$) and apply blow-up. According to definition 32 blow-up of

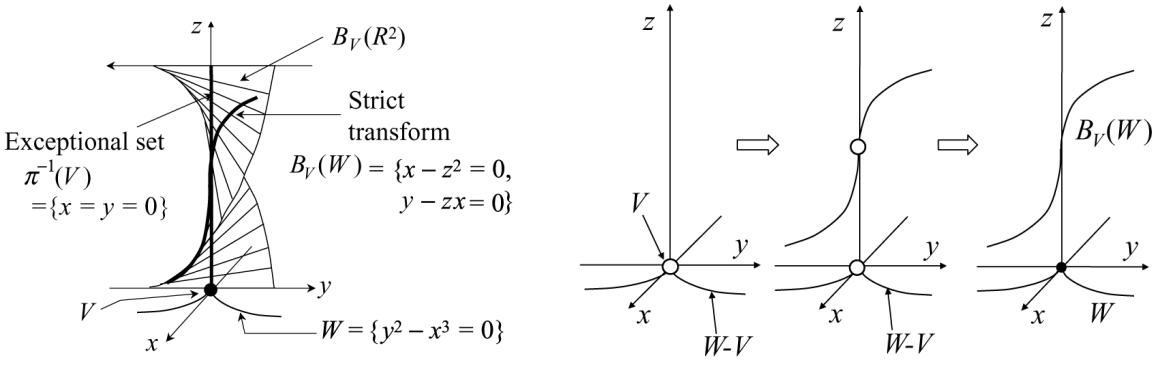


Figure 8: Blow-up of real algebraic set W with center V. Left: obtained strict transform and exceptional set. Right: blow-up process - removing singularity and adding projective space, then adding closure. Figures are from [1] (Fig. 3.2. and Fig. 3.3).

[4]Characteristic of a field is the smallest number of identity elements needed to give neutral element in the sum; if sum never reaches neutral element, then field is of characteristic zero.

W with center V is given by:

$$B_V(W) = \overline{\{(x, y, (x : y)); x \in W \setminus V\}} = \overline{\{(x, y, (x : y)); y^2 = x^3, (x, y) \neq (0, 0)\}} \tag{89}$$

Now introduce a new variable $z \in \mathbb{R}$ such that $(x : y) = (1 : z) \in \mathbb{P}^1$. Using a notation $(0 : 1) = (1 : \infty)$ we can write z as follows:

$$z = \begin{cases} \frac{y}{x}, & x \neq 0 \\ \infty, & x = 0 \end{cases} \tag{90}$$

Hence we can rewrite equation (89) as follows:

$$\begin{aligned} B_V(W) &= \overline{\{(x, y, z); y = zx, y^2 = x^3, (x, y) \neq (0, 0)\}} = \overline{\{(x, y, z); y = zx, x = z^2, (x, y) \neq (0, 0)\}} \\ &= \{(x, y, z); y = zx, x = z^2\} \end{aligned} \tag{91}$$

Obtained blow-up $B_V(W)$ is nonsingular set obtained by resolution of singularities of set W that contains singular point $(0, 0)$. Exceptional set is by definition 33 given by:

$$\overline{\pi^{-1}(W) \setminus B_V(W)} = \pi^{-1}(V) = \pi^{-1}(\{(0, 0)\}) = \{(0, 0)\} \times \mathbb{P}^1 \tag{92}$$

and the total transform is given by:

$$\pi^{-1}(W) = B_V(W) \cup \pi^{-1}(V) \tag{93}$$

In Figure 8 is shown obtained desingularized set $B_V(W)$ together with exceptional set $\pi^{-1}(V)$. Recall that projective space is a manifold by theorem 12, and that manifolds should not contain singularities by theorem 18. Hence in order to add points from projective space $\mathbb{P}^{r-1}$ we have to consider only nonsingular points from W i.e. take $x \in W \setminus V$.

We can also represent this blow-up using two local coordinates and then gluing them. Let $x = x_1$, $y = x_1 y_1$, so we have $f(x_1, y_1) = x_1^2(y_1^2 - x_1)$. On the other hand, let $x = x_2 y_2$ and $y = y_2$ yielding $f(x_2, y_2) = y_2^2(1 - x_2^3 y_2)$.

$$\begin{aligned} B_V(W) &= \overline{\{(x_1, y_1); x_1 = y_1^2, x_1 \neq 0\} \cup \{(x_2, y_2); x_2^3 y_2 = 1, y_2 \neq 0\}} = \\ &= \overline{\{(x_1, y_1); x_1 = y_1^2, x_1 \neq 0\}} \cup \overline{\{(x_2, y_2); x_2^3 y_2 = 1, y_2 \neq 0\}} = \\ &= \{(x_1, y_1); x_1 = y_1^2\} \cup \{(x_2, y_2); x_2^3 y_2 = 1\} \end{aligned} \tag{94}$$

Note that equation $x_1 = y_1^2$ correspond to the equation $x = z^2$ in equation (91) since $z = \frac{y}{x} = y_1$ and hence $x_1 = x = z^2 = y_1^2$. Similarly we conclude that equation $x_2^3 y_2 = 1$ corresponds to the equation $y = zx$ in (91). Gluing of these two local coordinates is achieved by identifying coordinates y_1 and x_2^{-1} since it holds that $z = y_1 = x_2^{-1}$.

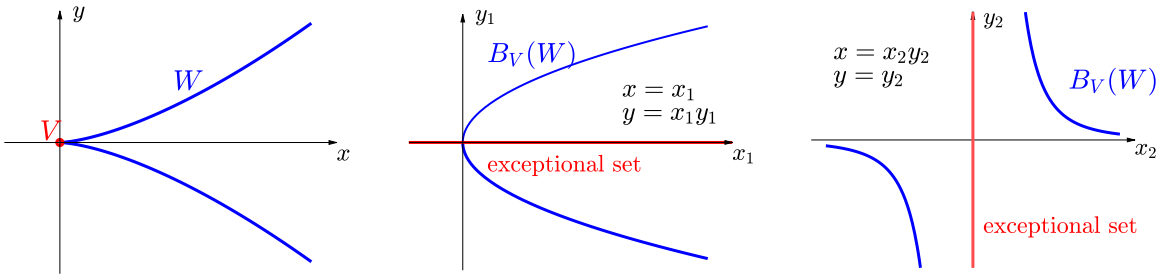


Figure 9: Blow-up of real algebraic set W with center V using local coordinates U_1 and U_2. The result $B_V(W)$ is obtained by gluing coordinates y_1 and x_2 in appropriate way.

Generally, any blow-up of 2-dimensional sets with center $V = \mathbb{V}(x, y)$ can be represented on local coordinates U_1 and U_2 as follows:

$$x = x_1 = x_2 y_2 \tag{95}$$

$$y = x_1 y_1 = y_2 \tag{96}$$

Similar reasoning can be used in higher dimensional spaces.

Remark. *Let $f(x,y,z)$ be a symmetric function in x and y meaning that $f(x,y,z) = 0$ and $f(y,x,z) = 0$ define the same algebraic set. Then doing a blow-up of $W = \mathbb{V}(f)$ with center $\mathbb{V}(x,y)$ gives two local coordinates that are the same up to the permutation of variables. Hence considering "characteristics" of algebraic set on one of those two local coordinates is enough , since the other local coordinate has the same structure. This idea holds for algebraic sets with more degrees of symmetry - one should always consider only those local coordinates that are structurally different from the others.*

Given projective varieties $V \subset W \subset \mathbb{P}^d$, the result of blowing-up $B_V(W)$ is in general not a projective variety, and instead belongs to a larger class called real algebraic variety. Namely, let $U_0, U_1, \ldots, U_d$ be local coordinates of $\mathbb{P}^d$ and define real algebraic sets (as shown in example 20) V_i and W_i:

$$V_i = V \cap U_i \tag{97}$$
$$W_i = W \cap U_i \tag{98}$$

for $i = 0, \ldots, d$. Blow-up of W_i with center V_i is a real projective variety. The blow-up of W with center V is obtained by gluing all local blow-ups $B_{V_i}(W_i)$ and is a real algebraic variety.

Hence in order to use successive blow-ups we will give main theorem of resolution of singularities in the terms of real algebraic varieties, which are invariant under blow-ups.

3.3 Hironaka's method of resolution of singularities

The problem of resolution of singularities consists of finding a space U which under suitable transform can be mapped back to the space W on which function f is defined. Furthermore, any singularities of function f on initial space W should be images of normal crossing singularities on U. This type of singularities is characteristic for normal crossing functions, defined as follows.

Definition 34. *(Normal crossing function) Let $U \subset \mathbb{R}^d$ be an open set and let f be a real analytic function on U. Then we say that f is normal crossing at $x^* = (x_1^*, x_2^*, \ldots, x_d^*) \in \mathbb{R}^d$ if there exist open set $U' \subset U$ such that:*

$$f(x) = a(x) \prod_{i=1}^{d} (x_i - x_i^*)^{k_i}, \quad \forall x \in U' \tag{99}$$

for nonzero real analytic function $a(x)$ and $k_i \in \mathbb{N}_{\geq 0}$ for $i = 1, 2, \ldots, d$.

Example 22. *(Continuing blow-up from example 21) In example 21 by blowing-up function $f(x,y) = y^2 - x^3$ at the origin we obtained the following functions on local coordinates: $f(x_1, y_1) = x_1^2(y_1^2 - x_1)$ and $f(x_2, y_2) = y_2^2(1 - x_2^3 y_2)$. Since $y_1^2 - x_1 = 0$ at the origin, function f is not normal crossing in the first local coordinates. Moreover, it is normal crossing in the second local coordinate. Hence in order to obtain only normal crossing functions in all local coordinates, we will apply blow-up of U_1 once more.*

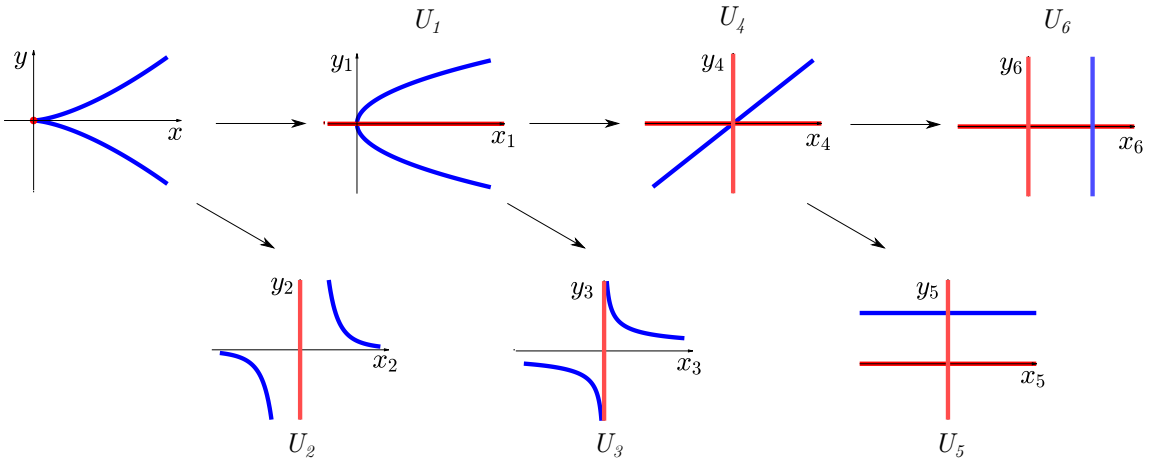


Figure 10: Applying blow-up scheme on an algebraic set to obtain algebraic sets defined by normal crossing functions on each local coordinate.

Similarly as in example 21 we apply blow-up of U_1 with center at the origin, i.e. set $x_1 = x_3, y_1 = x_3 y_3$ and $x_1 = x_4 y_4, y_1 = y_4$. We obtain functions: $f(x_3, y_3) = x_3^3(x_3 y_3^2 - 1)$ and $f(x_4, y_4) = x_4^2 y_4^3(y_4 - x_4)$. The latter function is not normal crossing, hence we apply blow-up of U_4 (local coordinate of (x_4, y_4)) with center at the origin again. We set $x_4 = x_5, y_4 = x_5 y_5$ and $x_4 = x_6 y_6, y_4 = y_6$ and obtain functions: $f(x_5, y_5) = x_5^6 y_5^3(y_5 - 1)$ and $f(x_6, y_6) = x_6^2 y_6^6(1 - x_6)$, both of which are normal crossing at the origin. In Figure 10 are represented all local coordinates on which function f is normal crossing: U_2, U_3, U_5, U_6 as well as intermediate results for local coordinates U_1 and U_4 on which f is not normal crossing at the origin.

As shown in the following theorem, any singular function defined on some space W can be represented on a suitable manifold U as a normal crossing function. This result is known as Hironaka's theorem, and we first give a statement of this theorem which claims existence of space U and a suitable mapping g between the spaces U and W.

Theorem 13. *(Hironaka's theorem 0, theorem 2.3 in [1]) For any non-constant, real analytic function f defined on neighborhood of origin in $\mathbb{R}^d$ and mapping to $\mathbb{R}$ that satisfies $f(0) = 0$, there exists a triple (W, U, g) such that:*

(a) $W \subset \mathbb{R}^d$ is an open set containing 0

(b) U is a d-dimensional real analytic manifold

(c) $g : U \to W$ is a real analytic map

and satisfies:

1. *g is a proper map on U and an analytic isomorphism of $U \setminus U_0$ and $W \setminus W_0$, where $U_0 = \{u \in U;\ f(g(u)) = 0\}$ and $W_0 = \{x \in W;\ f(x) = 0\}$.*

2. *For every point $P \in U_0$ there exists a local coordinate $(u_1, u_2, \ldots, u_d)$ of U such that P is the origin and*
$$f(g(u)) = S u_1^{k_1} u_2^{k_2} \cdots u_d^{k_d} \tag{100}$$
with $S \in \{-1, 1\}$ and $k_1, \ldots, k_d \in \mathbb{N}_{\geq 0}$. The Jacobian determinant of coordinate change $x = g(u)$ is given by:
$$g'(u) = b(u) u_1^{h_1} u_2^{h_2} \cdots u_d^{h_d} \tag{101}$$
for $h_1, \ldots, h_d \in \mathbb{N}_{\geq 0}$ and real analytic function $b(u) \neq 0$.

Even though theorem above considers function f near the origin, one could easily formulate theorem in neighborhood of an arbitrary point in $\mathbb{R}^d$ by shifting coordinate system in the appropriate way. In Figure 11 is given illustration of Hironaka's theorem. What theorem above states is that there always exist manifold U and function g satisfying given properties.

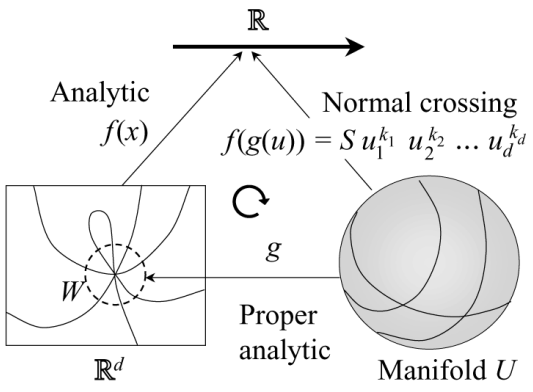


Figure 11: Illustration of Hironaka' theorem of resolution of singularities. Singularities of function f on W are images of normal crossing singularities on U. Figure is taken from [1] - Fig. 2.4.

The fact that g is proper means that g maps compact sets into compact sets. We will usually assume that the parameter space is compact, hence we conclude that manifold U will also be compact. Function g is an analytic isomorphism of $U \setminus U_0$ and $W \setminus W_0$, but not of U and W [1]. This means

that there is no analytic isomorphism between singularities in U and normal crossing singularities in W.

The second point in theorem 13 actually states that f is normal crossing function on manifold U. As we will see, this form of the function f is quite useful for asymptotic analysis of statistical models.

Remark. *(Global resolution of singularities) Note that in theorem 13 we state that for every point P in the set U_0 exist local coordinates on which $f(g(u))$ is normal crossing. Combining all these resolutions for different points in U we get global resolution of singularities defined as follows.*

Let K be a compact set on which function f is defined. Then using Hironaka's theorem above one can find triples (W, U, g) for any point $x \in K$ such that $f(x) = 0$. Since K is compact, there is a finite number of such local coordinates. Gluing these triples together we obtain a global resolution of K.

Remark. *(Remark 2.7 in [1]) Function f does not have to be necessarily of the form given by equation (100). Namely, f can be represented in the following form which will be more suitable for our analysis. If there exists a real analytic function g such that:*

$$f(g(u)) = a(u)u_1^{k_1}u_2^{k_2}\cdots u_d^{k_d} \tag{102}$$

with $|a(u)| > 0$ then using the following coordinate transform: $v_1 = |a(u)|^{1/k_1}u_1$ and $v_i = u_i$ for $i = 2, \ldots, d$, we obtain form (100):

$$f(g(t^{-1}(v))) = \pm v_1^{k_1}v_2^{k_2}\cdots v_d^{k_d} \tag{103}$$

Hence we can use expression (102) interchangeably with equation (100).

Hironaka's theorem as introduced in the theorem 13 guarantees existence of a manifold on which single function f is normal crossing. Furthermore, in the case of multiple functions $f_1, f_2 \ldots, f_k$ similar result can be obtained that guarantees existence of a manifold on which all of these functions are normal crossing. In order to prove theorem of simultaneous resolution of singularities, we will need the following lemma which we state without the proof (proof can be found in [1]).

Lemma 7. *(Theorem 2.7 in [1]) Let f_1, f_2 be real analytic functions defined on an open neighborhood U of the origin. If they satisfy:*

$$f_1(x)f_2(x) = x_1^{k_1}x_2^{k_2}\cdots x_r^{k_r},\ \forall x \in U \tag{104}$$

for some $k_1, \ldots, k_r \in \mathbb{N}$, then there exists an open set $W \subset U$ such that:

$$f_1(x) = a_1(x)x_1^{j_1}x_2^{j_2}\cdots x_r^{j_r} \tag{105}$$

$$f_2(x) = a_2(x)x_1^{h_1}x_2^{h_2}\cdots x_r^{h_r} \tag{106}$$

where a_1 and a_2 are real analytic functions satisfying $a_1(x)a_2(x) = 1$, for $x \in W$, and $j_i, h_i \in \mathbb{N}$ for $i = 1, \ldots, r$.

Theorem 14. *(Simultaneous resolution of singularities, theorem 2.8 in [1]) Let $f_1, f_2, \ldots, f_k$ be functions satisfying conditions of theorem 13. Then there exists a triple (W, U, g) as in theorem 13 such that each f_i is desingularized.*

Proof. Define a function f mapping a neighborhood of the origin to $\mathbb{R}$ as follows:

$$f(x) = f_1(x)f_2(x)\cdots f_k(x) \tag{107}$$

From $f(0) = f_1(0)f_2(0)\cdots f_k(0) = 0$ and the fact that f_i are non-constant real analytic functions follows that f is as well non-constant real analytic functions. Therefore function f satisfies conditions of theorem 13 and there exists a triple (W, U, g) such that f can be represented as:

$$f(g(u)) = u_1^{k_1}u_2^{k_2}\cdots u_d^{k_d} \tag{108}$$

on each local coordinate of U. Note that equation (108) implies that function f is in the form as in equation (104). Hence applying previous lemma $k-1$ times, we obtain that each of the functions f_i can be written as a normal crossing function near the origin. □

Until now, we have only considered existence of a triple (W, U, g) such that f is normal crossing function on U. However, the choice of (W, U, g) is not evident. In the following theorem, we give explicit description of obtaining triple (W, U, g) based on successive blow-ups.

Theorem 15. *(Hironaka's theorem, theorem 3.6 in [1]) For any polynomial $f \in \mathbb{R}[x_1, x_2, \ldots, x_d]$ there exists a sequence of pairs of real algebraic varieties $(V_0, W_0), (V_1, W_1), \ldots, (V_n, W_n)$ such that:*

1. *$V_i \subseteq W_i$ for $i = 1, \ldots, n$,*
2. *$V_0 = \mathbb{V}(f)$, $W_0 = \mathbb{R}^d$,*
3. *W_i for $i = 0, \ldots, n$ is nonsingular algebraic variety,*
4. *V_n is determined by a normal crossing polynomial on each local coordinate of W_n,*
5. *$W_i = B_{C_{i-1}}(W_{i-1})$ for $i = 1, \ldots, n$*
6. *V_i is the total transform of π_i i.e: $V_i = \pi_i^{-1}(V_{i-1})$,*
7. *the center C_i of each blow-up is a nonsingular algebraic variety contained in the set of critical points of $f \circ \pi_1 \circ \pi_2 \circ \cdots \circ \pi_i$ where $\circ$ denotes composition of functions.*

Let's decipher this theorem step by step. First, as we stated already, Hironaka's theorem is based on the iterative blow-ups starting from pair (V_0, W_0) until getting (V_n, W_n). The real algebraic variety $V_0 = \mathbb{V}(f)$ is containing initial singularities and $W_0 = \mathbb{R}^d$ is the whole space. The final pair of varieties (V_n, W_n) is such that V_n contains only normal crossing singularities on each local coordinate of W_n. Hence we determined the manifold with normal crossing singularities that are mapped to singularities of initial variety V_0. In every step V_i represents real algebraic variety containing singularities, whereas W_i are nonsingular algebraic varieties. The center C_i of each blow-up is chosen from the set of possible singularities of V_i. In every blow-up new critical and potentially singular points are added, i.e. the set of singularities is contained in the set of critical points of $f \circ \pi_1 \circ \cdots \circ \pi_i$.

Remark. *((6) in Example 2.6 in [1]) Implication 7) in theorem 15 might seem unusual since we claim that C_i is a nonsingular algebraic variety in the set of critical points. So let's explain on a simple example what are nonsingular algebraic varieties in a singular set. Let's define a set $V = \{(x, y, z); xyz = 0\}$ with set of singularities:*

$$\mathrm{Sing}(V) = \{(x, y, z); x = y = 0 \vee x = z = 0 \vee y = z = 0\} \tag{109}$$

Then a set $\{(x, y, z); x = y = 0\}$ is a nonsingular set (z-line) contained in the set of singularities $\mathrm{Sing}(V)$.

We have already mentioned some analogies between linear algebra and algebraic geometry and from Hironaka's theorem we can note additional one. Namely, it is known from linear algebra that any linear transform can be represented by Jordan matrix under suitable coordinate transform. What Hironaka's theorem claims is that any analytic function can be represented by normal crossing functions on a suitable manifold [1].

For the function f resolved of the singularities and in the form (100) with Jacobian determinant (101) we can define a constant called real log canonical threshold. This constant will be of great importance in establishing asymptotic behavior of singular models.

Definition 35. *(Real log canonical threshold) Let f be a function satisfying theorem 13 and let U be a compact set. Then, the real log canonical threshold (RLCT) is defined by:*

$$\mathrm{RLCT}(U) = \inf_{P \in U_0} \min_{1 \leq j \leq d} \Big(\frac{h_j + 1}{k_j} \Big) \tag{110}$$

Important property of $\mathrm{RLCT}(U)$ is birational invariance, meaning that it does not depend on the triple (W, U, g). This implies that any resolution of singularities of function f will yield the same real log canonical threshold. We prove this property for statistical models in theorem 19 in the following section.

The RLCT of a set corresponds to the deepest singularity of given function [18], and more complicated singularities yield lower RLCT [19]. Hence, we can motivate introduction of RLCT as describing the most complicated singularity in the given set. For more discussion on exact definition of "deepest singularity" check section 3.2.5 in [18].

4 Statistical learning theory for singular models

In this section we apply all theory developed so far for obtaining asymptotic behavior of few important statistical variables for singular models. Let's start with a quick recap of the main properties that differ between regular and singular models. In Table 1 are shown the main properties discussed mainly in section 1.

	regular models	singular models
optimal set	one point	analytic set with singularities
Fisher information matrix	positive definite	positive semi-definite
Cramer-Rao inequality	holds	no meaning
asymptotic normality	yes	no
standard form of $K(w)$	quadratic	normal crossing

Table 1: Fundamental differences between singular and regular models. Inspired by Table 8.2 in [1].

Although differences between these two types of models are evident from Table 1, it is not straightforward how these different properties influence statistical learning of models. In order to demonstrate why regular statistical learning theory cannot be applied to singular models, we first discuss regular models and check which of the conditions we need do not hold for singular models.

As shown in equation (37), normalized partition function is given by:

$$Z_n^0 = \int e^{-nK_n(w)}\varphi(w)dw \tag{111}$$

We have already discussed the importance of partition function Z_n in section 1.1. Similar reasoning can be applied to Z_n^0 as well. In this section we will examine asymptotic behavior of free energy of the model (35), its Bayes generalization (33) and training error (34). Note however that there exist other statistical quantities (such as Gibbs generalization and training errors, see [1]) which can be analyzed in the same way i.e. from behavior of Z_n^0. Hence function (111) will be our function of interest in this section.

From equations (7) and (30) we see that $K(w) = 0$ implies $K_n(w) = 0$, but the opposite does not hold in general. We have seen that in the case of regular models the optimal set consists of a point (see Table 1), i.e. there exists unique $w_0 \in W$ such that $K(w_0) = 0$. However this does not imply that $K_n(w) = 0$ only for $w = w_0$. Let's now define the following function:

$$Z(n) = \int e^{-nK(w)}\varphi(w)dw \tag{112}$$

We will see that this difference in zero sets of $K(w)$ and $K_n(w)$ makes analysis of $Z(n)$ easier than

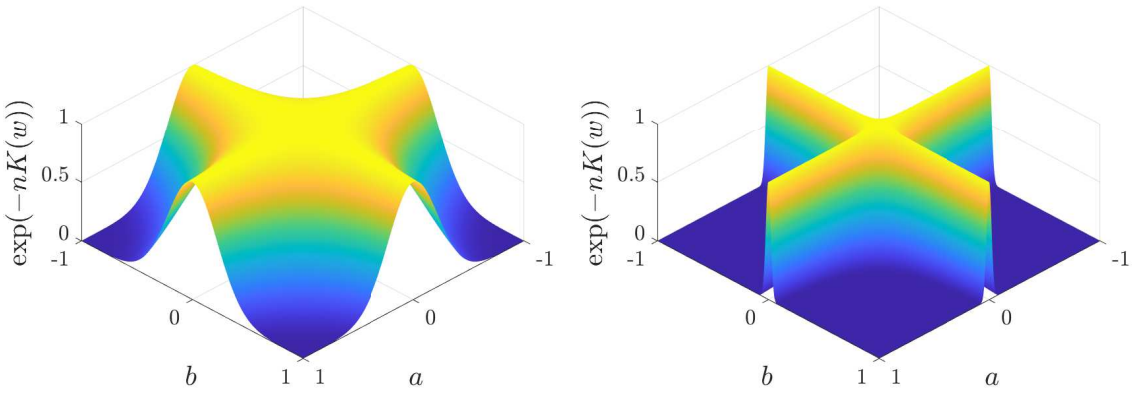


Figure 12: Integrand from singular integral equation (112) depending on the number of samples n - left: $n = 10$, right: $n = 1000$.

that of Z_n^0. Moreover, we will establish connection between $Z(n)$ and Z_n^0. Hence, once we obtain asymptotic behavior of $Z(n)$ we will be able to return to the analysis of Z_n^0.

Example 23. *(Example 4.8 in [1]) Assume that $w = (a, b) \in [-1, 1]^2$ and $K(a, b) = a^2b^2$. Then the term $\exp(-nK(w))$ from integral (111) is shown in Figure 12 for two values of n: $n = 10$ (left) and $n = 1000$ (right). What we obtain is that asymptotically $\exp(-nK(a, b))$ stays nonzero only near the points at which $K(a, b) = 0$ and vanishes elsewhere.*

Integral (112) belongs to a class of integrals called Laplace integrals [18]. Following discussion in section 1.3.1 in [18], for the class of regular models, we introduce the following approximation of Laplace integral:

Theorem 16. *(Laplace approximation, proposition 1.2 in [18]) Let W be a compact subset of $\mathbb{R}^d$ and let functions f and φ be real analytic functions over W. Assume that f has a unique minimum at point $w_0 \in W$ and is defined in small neighborhood of w_0. If the Hessian $H(w_0)$ is positive definite and $\varphi(w_0) > 0$, then:*

$$\int_W e^{-nf(w)}\varphi(w)dw \to e^{-nf(w_0)}\sqrt{\frac{(2\pi)^d}{\det H(w_0)}}\varphi(w_0)n^{-d/2}, \quad \text{when } n \to \infty. \tag{113}$$

Proof. From the condition that f is defined in a small neighborhood of w_0, we know that w_0 is in the interior of W. Hence, since f attains minimum at w_0 we have: $\frac{\partial}{\partial w_k}f(w) = 0$, for $k = 1, \ldots, d$. Now, function f can be written near w_0 as $f(w) = f(w_0) + \frac{1}{2}(w - w_0)^T H(w_0)(w - w_0)$, where $H(w_0)$ is Hessian of f at w_0. Substituting obtained expression into Laplace integral we get:

$$\begin{aligned}
\int_W e^{-nf(w)}\varphi(w)dw &= \int_W e^{-nf(w_0)-\frac{n}{2}(w-w_0)^T H(w_0)(w-w_0)}\varphi(w)dw = \\
&= e^{-nf(w_0)}\int_W e^{-\frac{n}{2}(w-w_0)^T H(w_0)(w-w_0)}\varphi(w)dw = \\
&= e^{-nf(w_0)}\int_{\tilde{W}} e^{-\frac{1}{2}\tilde{w}^T H(w_0)\tilde{w}}\varphi(\tilde{w})n^{-d/2}d\tilde{w} \xrightarrow{n\to\infty} e^{-nf(w_0)}\sqrt{\frac{(2\pi)^d}{\det H(w_0)}}\varphi(w_0)n^{-d/2}
\end{aligned}$$

where we used change of variables $\tilde{w} = \sqrt{n}(w - w_0)$, and properties of Gaussian integral. □

Applying Laplace approximation to integral (112) with regular model we obtain:

$$Z(n) \xrightarrow{n\to\infty} e^{-nK(w_0)}\sqrt{\frac{(2\pi)^d}{\det J(w_0)}}\varphi(w_0)n^{-d/2} = Ce^{-nK(w_0)}n^{-d/2} \tag{114}$$

where C is a constant independent on n. Laplace approximation gives asymptotic expression for $Z(n)$ in the case when function $K(w)$ has a unique minimum. However, as we discussed in the previous sections, for singular models points where $K(w) = 0$ form an analytic set with singularities. Hence applying Laplace approximation for general singular models does not work.

Alternative approach that is applicable to both regular and singular models is shown in Figure 13. In Figure 13 and the rest of the paper, we use notation $\sim$ to denote asymptotic equivalence of two expressions. We will now consider each of the steps from Figure 13 shortly, and suggest consulting [1] for more thorough analysis of this approach.

First step as shown in Figure 13 is resolution of singularities of Kullback-Leibler distance $K(w)$. Under suitable condition (section 4.1), it is possible to apply Hironaka's theorem (theorems 13 and 15) to $K(w)$ as we show in theorem 17 in section 4.2. Simply put, there exist manifold U and appropriate function g such that $K(g(u)) = u^{2k}$, i.e. such that K is normal crossing function on each local coordinate of U. By resolving singularities of $K(w)$ we can also rewrite $K_n(w)$ in a suitable form as follows.

We define a set W_ϵ for some positive constant ϵ as:

$$W_\epsilon = \{w \in W; K(w) \leq \epsilon\} \tag{115}$$

and denote by W_ϵ^c its complement in W. Then, on W_ϵ^c we define a random process $\psi_n(w)$ as follows:

$$\psi_n(w) = \sum_{i=1}^{n} \frac{K(w) - f(X_i, w)}{\sqrt{nK(w)}}, \quad w \in W_\epsilon^c \tag{116}$$

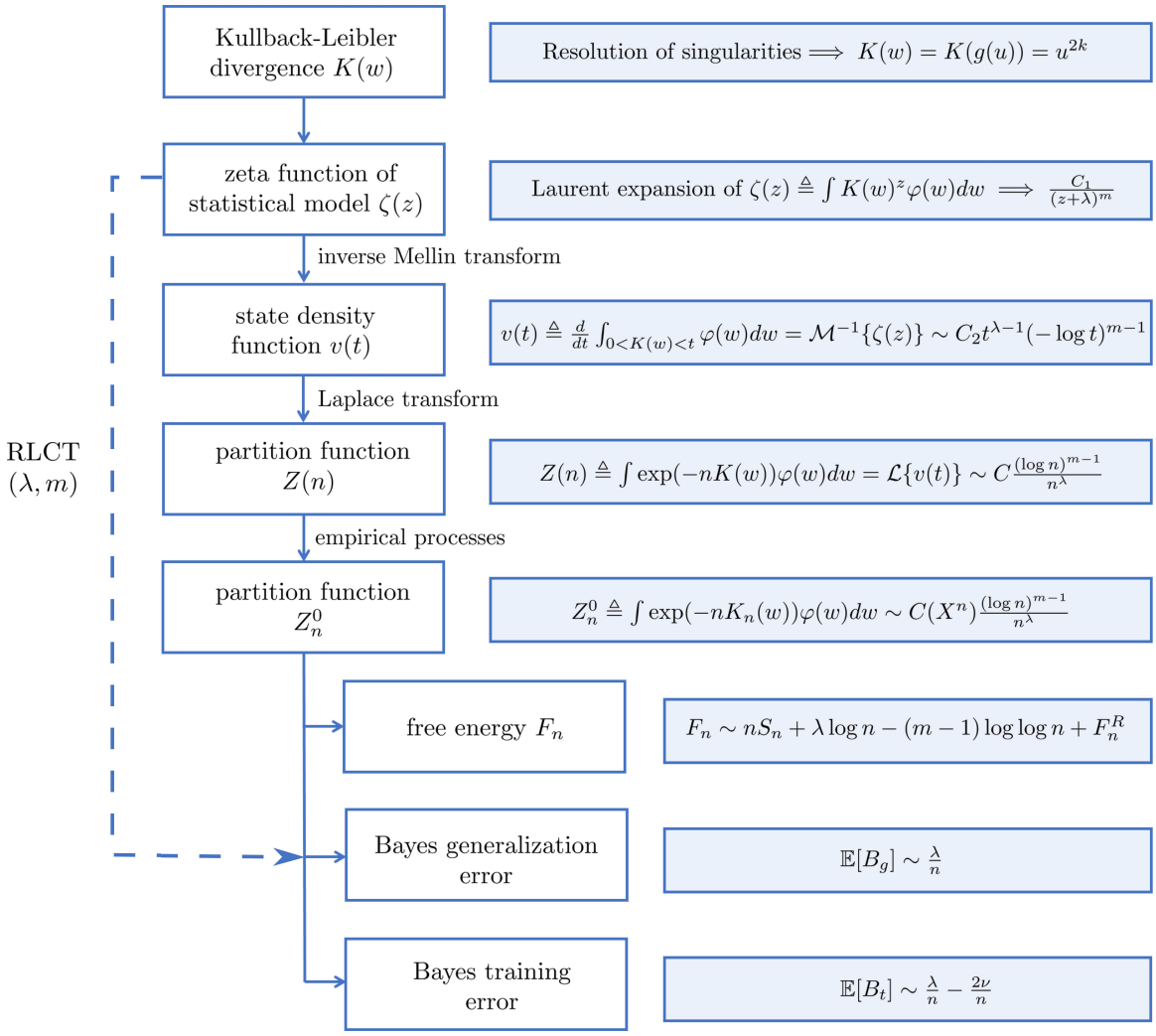


Figure 13: Overview of the framework for determining asymptotic behavior of singular models.

Note that the random process $\psi_n(w)$ is not well defined for zeros of function $K(w)$. Now, using definition of log likelihood ratio function (30), we can write $K_n(w)$ in so-called standard form:

$$K_n(w) = K(w) - \frac{1}{\sqrt{n}}\sqrt{K(w)}\psi_n(w), \quad w \in W_\epsilon^c \tag{117}$$

According to theorem 6.2 in [1], random process $\psi_n(w)$ on W_ϵ converges to a Gaussian process $\psi(w)$. For now, let's keep our attention on regular models, and we discuss standard form of log likelihood function $K_n(w)$ for regular models in the following remark.

Remark. *(Remark 1.14 (3), [1]) For regular statistical models with set of optimal parameters $W_0 = \{w_0\}$, we know that $J(w_0)$ is positive definite and symmetric matrix and applying Cholesky decomposition we get $J(w_0) = J_U^T J_U$, for some invertible, upper triangular matrix J_U. Then we set $w = g(u) = w_0 + J_U^{-1}u$ and hence have:*

$$K(g(u)) = u^2 \tag{118}$$

Similarly as in theorem 4 we can conclude that terms of order higher than three vanish in $K_n(w)$ near w_0. Then, if a triple (W, U, g) yields resolution of singularities of $K(w)$, from equation (117) we can rewrite $K_n(w)$ on local coordinates of manifold U as:

$$K_n(g(u)) = u^2 - \frac{1}{\sqrt{n}}\xi_n^T u \tag{119}$$

where for coordinates of $\xi_n \in \mathbb{R}^d$ holds:

$$\xi_n(k) = -\sqrt{n}\frac{\partial K_n(g(u))}{\partial u_k}\Big|_{u=0} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial u_k}\log p(x_i|g(u))\Big|_{u=0}, \quad k = 1, 2, \ldots, d \tag{120}$$

Note that $\mathbb{E}\left[\frac{\partial}{\partial u_k}\log p(X|g(u))|_{u=0}\right] = 0$ *and by definition 3 we have that* $\mathrm{Cov}\left[\frac{\partial}{\partial u_k}\log p(X|g(u))|_{u=0}\right] = J(u=0) = I$ *(as seen from equations (24) and (118)). Hence by central limit theorem* $\xi_n(k)$ *has standard normal distribution for* $n \to \infty$ *for* $k = 1, 2, \ldots, d$*. Therefore, we conclude that* $\psi_n(w)$ *from equation (117) in the case of regular models has simpler form* ξ_n *independent of parameter* $w \in W$*.*

As shown in the overview of the framework in Figure 13, in order to obtain asymptotic expansion of partition function $Z(n)$ from normal form of $K(w)$ we will use two important transformations defined as follows.

Let $f : [0, \infty) \to \mathbb{C}$ be a measurable function. If the following integrals exist, then we define:

- Mellin transform of function f as:

$$\mathcal{M}\{f(t)\} = \int_0^\infty f(t)t^z dt \tag{121}$$

- Laplace transform of function f as:

$$\mathcal{L}\{f(t)\} = \int_0^\infty f(t)e^{-st} dt \tag{122}$$

Properties of Mellin and Laplace transform are, for the sake of conciseness, not developed any further in this project. Reader is encouraged to check exposition given in [1].

Now, let's consider a statistical model with Kullback-Leibler divergence $K(w)$ and prior probability density function $\varphi(w) \in C^\infty$ defined on a compact set $W \subset \mathbb{R}^d$. Note that this implies that $K(w), \varphi(w) \in \mathbb{R}_{\geq 0}$, $\forall w \in W$. Then we define zeta function of this model as:

$$\zeta(z) = \int_W K(w)^z \varphi(w) dw \tag{123}$$

We devote section 4.3 to examining behavior and properties of function $\zeta(z)$. Furthermore, we define a state density function as a Schwartz distribution given by:

$$v(t) = \frac{d}{dt}\int_{0<K(w)<t} \varphi(w) dw \tag{124}$$

In a more engineering notation we can write $v(t)$ as $v(t) = \int \delta(t - K(w))\varphi(w)dw$. The following lemma connects both $\zeta(z)$ and $v(t)$ as well as $Z(n)$ and $v(t)$ as shown afterwards.

Lemma 8. *(Theorem 4.4 in [1]) Let* $F : \mathbb{R} \to \mathbb{R}$ *be a locally integrable function, i.e. assume that for every compact set* $K \subset \mathbb{R}$*, the integral* $\int_K F(t)dt$ *is well defined and finite. Then, for state density function* $v(t)$ *given by (124) holds:*

$$\int_{\mathbb{R}} F(t)v(t)dt = \int_{\mathbb{R}^d} F(K(w))\varphi(w)dw \tag{125}$$

Let $F(t) = t^z$ in equation (125). Then, combining equations (123) and (125) we have:

$$\zeta(z) = \int K(w)^z \varphi(w)dw = \int t^z v(t)dt = \mathcal{M}\{v(t)\} \tag{126}$$

where last equality follows from definition of Mellin transform (121). Hence we can say that zeta function is Mellin transform of state density function, or equivalently, $v(t)$ is inverse Mellin transform of $\zeta(z)$.

Now let $F(t) = \exp(-nt)$ in equation (125). Once again, combining equations (112) and (125) we get:

$$Z(n) = \int \exp(-nK(w))\varphi(w)dw = \int \exp(-nt)v(t)dt = \mathcal{L}\{v(t)\} \tag{127}$$

with last equality follows from definition of Laplace transform (122). Hence, partition function $Z(n)$ is the Laplace transform of state density function $v(t)$.

Note that we have established all relations depicted in Figure 13. Next we consider how to use any of these variables i.e. $K(w), \zeta(z), v(t), Z(n)$ in order to obtain asymptotic behavior of Z_n^0 and also of important statistical quantities such as F_n, B_g and B_t.

When considering asymptotic expansion of variables, we are mainly interested in first (lowest) order approximation. Determining higher order terms is discussed in [18]. As we discuss in section 4.3, zeta function $\zeta(z)$ has Laurent expansion (of the form (149)) with negative rational poles λ_i. Furthermore the largest pole (i.e. smallest in the absolute value) of $\zeta(z)$ corresponds to RLCT (defined by equation (110)) of $K(w)$ on appropriate manifold. Hence determining largest pole λ of zeta function and its multiplicity m, we obtain RLCT i.e. pair (λ, m). Once we determine this pair, asymptotic expansion of other variables from Figure 13 is determined as follows.

From theorem 4.6 in [1] follows that state density function $v(t)$ has asymptotic expansion:

$$v(t) \sim C_v t^{\lambda-1}(-\log t)^{m-1} \tag{128}$$

for a constant C_v. Theorem 4.7 in [1] states that function $Z(n)$ has asymptotic expansion of the form:

$$Z(n) \sim C\frac{(\log n)^{m-1}}{n^{\lambda}} \tag{129}$$

where C is a constant. According to theorem 6.7 in [1], function Z_n^0 asymptotically converges to:

$$Z_n^0 \sim C(X^n)\frac{(\log n)^{m-1}}{n^{\lambda}} \tag{130}$$

where $C(X^n)$ is a random variable. The asymptotic expansion of free energy, Bayes generalization and training errors will be discussed in section 4.4 and 4.5, respectively.

In the following sections, we first consider two fundamental conditions, which are needed for majority of theorems developed in this chapter. Then we introduce exact theorem which states that $K_n(w)$ can be written in the standard form. In third section we consider properties of zeta function $\zeta(z)$ and determining RLCT for given model. Lastly, in sections 4 and 5 we give explicit dependence of important statistical variables on RLCT pair (λ, m).

4.1 Fundamental conditions

In theorem 19 we will consider what is meant by analytic continuation and show that it exists under certain conditions, which we will call fundamental conditions I and II. These two fundamental conditions will be sufficient (but not always necessary) for establishing all the theory we will need. The following section is heavy on notation and assumptions that cannot be completely explained using the theory we introduced so far. However, for the sake of completeness we give explicit definitions of two conditions used in [1]. Reader is encouraged to come back to this section after reading the ones following it.

First, analogously to the Lebesgue L^p spaces we will introduce $L^s(q)$ spaces. Let $q(x)$ be a probability density function on $\mathbb{R}^N$ and let $s \in \mathbb{N}_{\geq 1}$. Then we define space:

$$L^s(q) = \{\text{measurable } f : \mathbb{R}^N \to \mathbb{C}^1 : \int |f(x)|^s q(x)dx < \infty\} \tag{131}$$

Similarly as Lebesgue L^p spaces, these spaces are Banach spaces with norm defined as $||f||_s = \left(\int |f(x)|^s q(x)dx\right)^{1/s}$ [1]. The inclusion $L^{\tilde{s}}(q) \subset L^s(q)$ holds for every $1 < s < \tilde{s}$. Also, $L^s(q)$ is not a set of functions, but a set of equivalence classes with the following equivalence relation: $f \sim g \iff f(x) = g(x)$ almost surely with respect to the measure $q(x)dx$.

Now, we introduce first of the two fundamental conditions we will need.

Definition 36. *(Fundamental condition I, definition 6.1 in [1]) Assume that a statistical model $p(x|w)$ with $w \in W \subset \mathbb{R}^d$, true distribution $q(x)$ and log density ratio function $f(x, w)$ has a compact parameter space W. Also, assume that $p(x|w)$ and $q(x)$ have the same support and that $W_0 = \{w \in W;\ K(w) = 0\}$ is nonempty set. Then we say that this model satisfies the fundamental condition I with index s ($s \geq 2$) if:*

1. *there exists an open set $W^{(C)} \in \mathbb{C}^d$ such that $W \subset W^{(C)}$*
2. *for every $w \in W^{(C)}$ function $f(x, w) \in L^s(q)$*
3. *$M(x) \equiv \sup_{w \in W^{(C)}} |f(x, w)| \in L^s(q)$*
4. *$\exists \epsilon > 0$ and $Q(x) \equiv sup_{K(w) \leq \epsilon} p(x|w)$ such that $\int M(x)^2 Q(x)dx < \infty$.*

Fundamental condition I is mainly needed for empirical process theory, i.e. to establish relation between $Z(n)$ and Z_n^0. In the following remark we try to convey ideas behind every condition of definition 36.

Remark. *(Remark 6.1, 3 in [1]) Condition 1 determines complex set $W^{(C)}$ which is used in other conditions. We should however be aware that there always exists a real open set $W^{(R)} \subset \mathbb{R}^d$ such that $W \subset W^{(R)} \subset W^{(C)}$. Similarly one could determine open sets $W_\epsilon \subset W_\epsilon^{(R)} \subset W_\epsilon^{(C)}$ and manifolds $\mathcal{M} \subset \mathcal{M}^{(R)} \subset \mathcal{M}^{(C)}$ such that $\mathcal{M} = g^{-1}(W_\epsilon)$, $\mathcal{M}^{(R)} = g^{-1}(W_\epsilon^{(R)})$, $\mathcal{M}^{(C)} = g^{-1}(W_\epsilon)^{(C)}$. Hence satisfying fundamental condition I implies that this conditions are satisfied on $W^{(R)}$ as well.*

Conditions 2 and 3 imply that $f(x,w)$ is represented by the absolutely convergent power series in the neighborhood of any $w_0 \in W$:

$$f(x,w) = \sum_\alpha a_\alpha(x)(w-w_0)^\alpha \tag{132}$$

with bounded functions $a_\alpha(x) \in L^s(q)$:

$$|a_\alpha(x)| \leq \frac{M(x)}{R^\alpha} \tag{133}$$

where R is radius of convergence.
Condition 4 is a rather technical condition used in obtaining theorem 6.3 in [1].

From the fact that $L^{\tilde{s}}(q) \subset L^s(q)$ for $2 \leq s < \tilde{s}$ follows that if fundamental condition I is satisfied with index $\tilde{s}$ it is also satisfied with s. Next we introduce fundamental condition II concerning the parameter space and prior distributions of the model.

Definition 37. *(Fundamental condition II, def 6.3 in [1]) A statistical model satisfies the fundamental condition II if:*

1. *the set of parameters W is compact, and semianalytic i.e:*

$$W = \{w \in \mathbb{R}^d;\ \pi_1(w) \geq 0,\ \pi_2(w) \geq 0, \ldots, \pi_k(w) \geq 0\} \tag{134}$$

   where $\pi_1(w), \pi_2(w), \ldots, \pi_k(w)$ are real analytic functions defined on an open subset of $\mathbb{R}^d$

2. *the a priori probability density function $\varphi(w)$ is nearly analytic function, i.e. there exist functions $\varphi(w)_1$ and $\varphi_2(w)$ such that $\varphi(w) = \varphi_1(w)\varphi_2(w)$ with $\varphi_1(w) > 0$ and $\varphi_1(w) \in C^\infty$, whereas $\varphi_2(w) \geq 0$ and $\varphi_2(w) \in C^\omega$*

Remark. *(Remark 6.6 in [1]) The reason why $\pi_1(w), \pi_2(w), \ldots, \pi_k(w)$ and $\varphi_2(w)$ must be real analytic functions is to avoid pathological examples. For example, if $\varphi_2(w)$ is not analytic, then $\zeta(z)$ may not have poles at all. Also, we need equation (134) in order to have a partitioning of the manifold which can be used for determining poles of the zeta function, as we show in theorem 18.*

4.2 Standard form of log likelihood ratio function

We have already discussed in introduction to section 4 the standard form of log likelihood ratio function $K_n(w)$. Recall that we have determined the standard form for the regular models in equation (119). Moreover, we have mentioned more general form in equation (117), and now we present a theorem claiming validity of such form.

Theorem 17. *(Main theorem 6.1, [1]) Assume that $p(x|w)$ and $q(x)$ satisfy fundamental condition I with index $s = 2$. Then there exists a real analytic manifold $\mathcal{M}^{(R)}$, and a real analytic and proper map $g : \mathcal{M}^{(R)} \to W_\epsilon^{(R)}$ such that:*

$$K(g(u)) = u^{2k} \tag{135}$$

on each local coordinate U of $\mathcal{M}^{(R)}$, for $k \in \mathbb{N}_{\geq 0}^d$. There exists a real analytic function $a(x,u) \in L^2(q)$ such that:

$$f(x,g(u)) = a(x,u)u^k \quad \text{and} \quad \int a(x,u)q(x)dx = u^k, \qquad (u \in U) \tag{136}$$

Furthermore, the standard form of the log likelihood ratio function is given on U by:

$$K_n(g(u)) = u^{2k} - \frac{1}{\sqrt{n}} u^k \xi_n(u) \tag{137}$$

where

$$\xi_n(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big(a(x_i, u) - \mathbb{E}_X[a(X, u)] \Big) \tag{138}$$

is an empirical process converging in law on manifold $\mathcal{M}^{(R)}$ to the Gaussian process $\xi(u)$ (theorem 6.2 in [1]).

Hence by determining analytic function $a(x, u)$, we obtain expression for empirical process $\xi_n(u)$. Note that in contrast to expression (116), expression (138) is well defined even for points with $K(w) = 0$.

4.3 Zeta function of statistical models

We have already introduced zeta function in equation (123), although somewhat simplistically. More accurately, for a given model with prior distribution $\varphi(w)$, parameter space W and KL distance $K(w)$, zeta function of this model is analytic continuation of the following function:

$$\zeta(z) = \int_W K(w)^z \varphi(w) dw, \qquad \mathrm{Re}(z) > 0 \tag{139}$$

into the whole complex space. We consider when this continuation exists in theorem 19. But before that, we prove a theorem showing that an integral on parameter space W can be calculated also on manifold U as defined in Hironaka's theorem of resolution of singularities.

Theorem 18. *(Theorem 2.11 in [1]) Assume that condition 1) of fundamental condition II holds. Let H be an integrable function on W satisfying conditions of Hironaka's theorem 13 with triple (W, U, g). Then:*

$$\int_W H(w) dw = \int_U H(g(u)) |g'(u)| du = \sum_\alpha \int_{U_\alpha} H(g(u)) \sigma_\alpha(u) |g'(u)| du \tag{140}$$

where $\sigma_\alpha(u)$ are supported on U_α.

Proof. First equality follows from Hironaka's theorem 13. Hironaka's theorem can be applied to simultaneously resolve singularities of multiple functions (Theorem 2.8 in [1]). Hence for certain triple (W, U, g) all functions $H(w), \pi_1(w), \ldots, \pi_k(w)$ are normal crossing on U. For any point $p \in U$ we define an open set $(-b, b)^d$ with origin in p, where $b > 0$ is some positive constant, and denote it by $O_p(b)$. Since we define an open set around any point of U we have that $U \subset \cup_p O_p(b)$. From compactness of U, we know that U can be covered by a finite number of open sets $O_p(b)$. Every set $O_p(b) = (-b, b)^d$ can be covered by 2^d sets which are analytically isomorphic to $[0, b)^d$. Each of those sets we denote by U_α. To sum up, we have that $U \subset \cup_\alpha U_\alpha$ and U_α are analytically isomorphic to $[0, b)^d$.

Define $\sigma_\alpha(u)$ in U_α as follows:

$$\sigma_\alpha(u) = \frac{\sigma_\alpha^{(0)}(u)}{\sum_\alpha \sigma_\alpha^{(0)}(u)} \quad \text{where} \quad \sigma_\alpha^{(0)}(u) = \begin{cases} \prod_{i=1}^d \exp(-\frac{1}{1-u_i}), & (0 \le u_i < 1, \text{ for } 1 \le i \le d) \\ 0, & \text{otherwise} \end{cases} \tag{141}$$

Then $\sigma_\alpha(u)$ is analytic in $(0, b)^d$, satisfies $\sigma_\alpha(u) \ge 0$ on every U_α and $\sum_\alpha \sigma_\alpha(u) = 1$ on U. Hence we can partition manifold U into local coordinates U_α by inserting functions $\sigma_\alpha(u)$ inside of the integral (140) (see corollary 2.2 in [1]). The statement of the theorem follows from here. □

In the following theorem we will show how zeta function of a statistical model can be analytically continued to the whole domain. At the end of the proof of the theorem, we will give expression for poles of the zeta function and hence show that real log canonical threshold of zeta function is birational invariant.

Theorem 19. *(Analytic continuation of zeta function, theorem 6.6 in [1]) If the fundamental conditions I (with index $s = 2$) and II hold, then the holomorphic function of $z \in \mathbb{C}$:*

$$\zeta(z) = \int K(w)^z \varphi(w) dw, \qquad Re(z) > 0 \tag{142}$$

has an analytic continuation to the unique meromorphic function on the entire complex plane. Furthemore, poles of this function are all negative rational numbers.

Proof. We will prove that function $\zeta(z)$ is holomorphic except at the poles. Hence, we need to show that $\zeta(z)$ is differentiable as a complex function everywhere except at the poles. Let's split integral (142) into two parts as follows:

$$\zeta(z) = \underbrace{\int_{K(w)<\epsilon} K(w)^z \varphi(w) dw}_{\zeta_1(z)} + \underbrace{\int_{K(w)\geq\epsilon} K(w)^z \varphi(w) dw}_{\zeta_2(z)} \tag{143}$$

Now consider $\zeta_2(z)$. According to fundamental condition II the set of parameters W is compact. From continuity of $K(w)$ follows that it is also bounded on W. Furthermore, on set $K(w) \geq \epsilon$ there exists constant C such that $C = \max_{w \in W \setminus W_0} \{K(w), \frac{1}{K(w)}\}$. Then:

$$\left| \frac{\partial K(w)^z}{\partial z} \right| = |\log K(w)| K(w)^{\mathrm{Re}(z)} \leq |\log C| C^{\mathrm{Re}(z)} \tag{144}$$

Using Lebesgue dominated convergence we obtain that function $\left| \frac{\partial K(w)^z}{\partial z} \right|$ is integrable and furthermore $\zeta_2(z)$ is holomorphic. Now consider $\zeta_1(z)$. According to Hironaka's theorem i.e. theorem 13 there exists a triple $(\tilde{W}, U, g)$ such that $K(g(u))$ is in normal form on U and g maps U into $\tilde{W} \subset W$. Applying theorem 18 to the integral of zeta function we get:

$$\zeta_1(z) = \int_{\tilde{W}} K(w)^z \varphi(w) dx = \sum_\alpha \int_{U_\alpha} K(g(u))^z |g'(u)| \varphi(g(u)) \sigma_\alpha(u) du \tag{145}$$

with local coordinates U_α. Since $K(w) \geq 0$ for any $w \in W$ after applying Hironaka's theorem function $K(w)$ will be of the following form: $K(g(u))) = u_1^{2k_1} u_2^{2k_2} \cdots u_d^{2k_d}$. Substituting this expression into integral (145) yields:

$$\zeta_1(z) = \sum_\alpha \int_{U_\alpha} u^{2k_\alpha z} u^{h_\alpha} \phi_\alpha(u) du \tag{146}$$

where we collected $b_\alpha(u)$, $\varphi(g(u))$ and $\sigma_\alpha(u)$ into a single variable $\phi_\alpha(u)$. Note also that we explicitly denoted dependence of constants k and h on local coordinates α. Assuming that $\varphi(w) > 0$ at singularities, we have that $\phi_\alpha(u) > 0$ at $u = 0$. Then, Taylor expansion of ϕ_α near zero is approximated by a constant. Hence, in order to determine poles of $\zeta_1(z)$ we can rewrite $\zeta_1(z)$ as:

$$\zeta_1(z) = \sum_\alpha \int_{M_\alpha} u^{2k_\alpha z + h_\alpha} du = \sum_\alpha \int_{[0,b]^d} u^{2k_\alpha z + h_\alpha} du \tag{147}$$

On each local coordinate poles are:

$$\int_{[0,b]^d} u^{2kz+h} du = \prod_{i=1}^d \frac{b^{2k_i z + h_i + 1}}{2k_i z + h_i + 1} \implies \lambda_i = -\frac{h_i + 1}{2k_i} \tag{148}$$

and since k_i, h_i are nonnegative integers, we have that λ_i are negative rational numbers. We conclude that function $\zeta(z)$ is holomorphic everywhere except at the poles λ_i and hence it is a meromorphic function in the complex plane. □

Remark. *Assuming that $\zeta(z)$ has poles at $\lambda_1, \lambda_2, \ldots$ such that $0 < -\lambda_1 < -\lambda_2 < \cdots$ with orders m_k ($k = 1, 2, \ldots$), then the zeta function has the Laurent expansion:*

$$\zeta(z) = \zeta_0(z) + \sum_{k=1}^\infty \sum_{m=1}^{m_k} \frac{c_{km}}{(z + \lambda_k)^m} \tag{149}$$

with $\zeta_0(z)$ holomorphic [1]. We are mainly interested in approximation of $\zeta(z)$ that considers only the largest pole λ_1 and its multiplicity m_1.

In the following example we aim to determine the real log canonical threshold of regular models.

Example 24. *(Regular models, example 1 in [20]) Let $p(x|w)$ for $w \in W \subset \mathbb{R}^d$ be a regular statistical model, and let's assume that the true parameters are given by $w_0 = 0$. According to equation (24) Kullback-Leibler distance in the neighborhood of w_0 can be approximated by the quadratic form*

$K(w) = \frac{1}{2} w^T J(w_0) w$. *Also, since the model is regular we know that* $J(w_0)$ *is positive definite and symmetric. Hence, we can apply a suitable coordinate transform (as in (118)) to obtain:*

$$K(w) = w^T w = \sum_{i=1}^{d} w_i^2 \tag{150}$$

Define $W_1 = \{w \in W : \ ||w||_\infty < 1\}$ *and let* $W_{1i} = \{w \in W : \ |w_j| < |w_i| < 1, \ j \in \{1, \dots, d\} \setminus \{i\}\}$ *for* $i = 1, \dots, d$. *Note that* $W_1 = \cup_{i=1}^{d} W_{1i}$ *plus some set of measure zero. Considering statistical zeta function on each of the manifolds* W_{1i} *we get:*

$$\begin{aligned} \zeta(z) &= \int_{W_{1i}} K(w)^z \varphi(w) dw = \int_{W_{1i}} K(w)^z (\varphi(w_0) + \frac{1}{2!} \nabla \varphi(w_0)^T w + \mathcal{O}(||w||^2)) dw \\ &\approx \varphi(w_0) \int_{W_{1i}} K(w)^z dw \end{aligned} \tag{151}$$

where we used assumption that $\varphi(w) > 0$ *on* W. *Since constant* $\varphi(w_0) > 0$ *does not influence poles of the function* $\zeta(z)$ *we will omit* $\varphi(w_0)$ *in the following equations. Substituting expression for* $K(w)$ *from equation (150) into equation (151) yields:*

$$\zeta(z) = \int_{W_{1i}} \left(\sum_{i=1}^{d} w_i^2 \right)^z dw \tag{152}$$

Define $U = \{(u_1, u_2, \dots, u_d); \ |u_i| < 1, \ \forall i = 1, \dots d\}$. *Then, by using the following blow-up map* $g : u \mapsto w$:

$$w_i = u_i, \quad w_j = u_i u_j, \ j \in \{1, \dots, d\} \setminus \{i\} \tag{153}$$

we get:

$$\zeta(z) = \int_U \left(u_i^2 + u_i^2 \sum_{\substack{j=1 \\ j \neq i}}^{d} u_j^2 \right)^z |g'(u)| du_1 du_2 \cdots du_d = \int_{-1}^{1} \int_{-1}^{1} \cdots \int_{-1}^{1} u_i^{2z} \left(1 + \sum_{\substack{j=1 \\ j \neq i}}^{d} u_j^2 \right)^z u_i^{d-1} \prod_{k=1}^{d} du_k \tag{154}$$

The obtained function has a pole λ_i *such that* $2\lambda_i + d - 1 + 1 = 0$, *i.e.* $\lambda_i = -\frac{d}{2}$ *with multiplicity* $m_i = 1$. *The same reasoning can be applied to any other manifold* W_{1i}, $i = 1, \dots, d$. *Hence we conclude that regular models have real log canonical threshold* $\lambda = \frac{d}{2}$ *and* $m = 1$.

Now that we have established real log canonical threshold for regular models, we can compare it with the one for singular models. What the following theorem states is that any singular model is at least as good as any regular model with respect to asymptotic learning capacities.

Theorem 20. *(Theorem 7.2 in [1]) For a given set of parameters* $W \subset \mathbb{R}^d$ *if there exists an open set* $U \subset W$ *such that* $\{w \in U; K(w) = 0, \ \varphi(w) > 0\}$ *is nonempty then* $\lambda \leq \frac{d}{2}$.

Proof. In order to prove this theorem we need the following lemma:

> **Lemma 1.** (Theorem 7.1 1) in [1]) If the set of true parameters $\{w \in W; K(w) = 0, \ \varphi(w) > 0\}$ is nonempty then there exists a constant $c_1 > 0$ such that:
>
> $$\lim_{n \to \infty} \frac{n^\lambda}{(\log n)^{m-1}} \int \exp(-nK(w)) \varphi(w) dw = c_1 \tag{155}$$

According to the assumption there exists some $w_0 \in W$ such that $K(w_0) = 0$ and $\varphi(w_0) > 0$. Without loss of generality we can assume that $w_0 = 0$ (else shift w by w_0). As shown in the proof of theorem 4: $\frac{\partial K(w)}{\partial w_i}\Big|_{w=0} = 0$. Hence Taylor expansion of $K(w)$ around zero yields:

$$K(w) = \frac{1}{2} \sum_{i,j=1}^{d} \frac{\partial^2 K(w)}{\partial w_i \partial w_j}\Big|_{w=0} w_i w_j + \mathcal{O}(|w|^3), \quad |w| < \epsilon \tag{156}$$

where $\epsilon > 0$ is a small constant that might depend on n. Since $\exp(-nK(w))\varphi(w) > 0$ in a neighborhood of w_0 we have:

$$Z(n) = \int \exp(-nK(w)) \varphi(w) dw \geq \int_{|w| < \epsilon} \exp(-nK(w)) \varphi(w) dw \tag{157}$$

$$= \int_{|w|<\epsilon} \exp\Big(-\frac{n}{2}\sum_{i,j=1}^{d}\frac{\partial^2 K(w)}{\partial w_i \partial w_j}\Big|_{w=0} w_i w_j - n\mathcal{O}(|w|^3)\Big)\varphi(w)dw \tag{158}$$

A change of variable $w' = \sqrt{n}w$, with $\epsilon\sqrt{n} < 1$, gives:

$$Z(n)n^{d/2} \geq \int_{|w'|<1} \exp\Big(-\frac{1}{2}\sum_{i,j=1}^{d}\frac{\partial^2 K(w)}{\partial w_i \partial w_j}\Big|_{w=0} w'_i w'_j - \frac{\mathcal{O}(|w'|^3)}{\sqrt{n}}\Big)\varphi\Big(\frac{w'}{\sqrt{n}}\Big)dw' \tag{159}$$

Note that we used expression $dw' = n^{d/2}dw$. Limit of this expression for $n \to \infty$ is given by:

$$\lim_{n\to\infty} Z(n)n^{d/2} \geq \int_{|w'|<1} \exp\Big(-\frac{1}{2}\sum_{i,j=1}^{d}\frac{\partial^2 K(w)}{\partial w_i \partial w_j}\Big|_{w=0} w'_i w'_j\Big)\varphi(0)dw' = c_2 > 0 \tag{160}$$

Comparing equations (155) and (160) asymptotically for $n \to \infty$ we have:

$$n^{d/2} \geq \frac{c_2}{c_1}\frac{n^\lambda}{(\log n)^{m-1}} \sim n^\lambda \implies \frac{d}{2} \geq \lambda \tag{161}$$

□

Recall that for regular models there exists a single point at which $K(w) = 0$. Hence as $n \to \infty$ integral $\int \exp(-nK(w))\varphi(w)dw \to \int \exp(-nK(0))\varphi(0)dw$ and hence inequality in equation (157) would hold with equality. At the end, we get the result as we expected that $\lambda = \frac{d}{2}$ in the regular models.

Remark. *According to section 4 in [19] given some* RLCT *near a point, the effective number of parameters near this point is exactly 2·*RLCT*. Furthermore, the change from $\frac{d}{2}$ to λ may explain why neural networks with huge number of parameters can still achieve high values of partition function. Namely, the singularities of deep neural networks reduce effectively their number of parameters from $\frac{d}{2}$ to λ [19].*

In the following section we will show how obtained values of poles of $\zeta(z)$ function influence statistical learning. The fact that $\lambda = \frac{d}{2}$ for regular, and $\lambda \leq \frac{d}{2}$ for singular models will be enough to show that singular models achieve asymptotically better results.

4.4 Convergence of free energy

In this section we explain asymptotic behavior of free energy as the number of samples goes to infinity. We will use the following theorem to claim that singular models are at least as good as regular models asymptotically.

Theorem 21. *(Corollary 6.1, Main theorem 6.2 [1]) If the fundamental conditions I and II hold with $s = 2$, then for normalized free energy holds the following asymptotic equation:*

$$F_n^0 = \lambda \log n - (m-1)\log\log n + F_n^R \tag{162}$$

where F_n^R is a random variable and sequence $\{F_n^R\}$ converges to a random variable in distribution. Furthermore, if fundamental conditions I and II hold with $s = 4$ then it also holds that:

$$\mathbb{E}[F_n^0] = \lambda \log n - (m-1)\log\log n + \mathbb{E}[F_n^R] \tag{163}$$

where $\mathbb{E}[F_n^R]$ converges to a constant.

Example 25. *We have shown in example 24 that regular models have real log canonical threshold of $\lambda = \frac{d}{2}$ where d is the size of the parameter space i.e. $W \subset \mathbb{R}^d$. We then showed in theorem 20 that any statistical model has real log canonical threshold $\lambda \leq \frac{d}{2}$. Using expression (162), normalized free energy of the regular models is asymptotically given by:*

$$F_n^0 = \frac{d}{2}\log n + o(\log n) \tag{164}$$

and hence free energy of regular models is from equation (40) given by:

$$F_n = nS_n + \frac{d}{2}\log n + o(\log n) \tag{165}$$

where S_n is empirical entropy. Note that obtained formula contains Bayesian information criterion (BIC) as a special case. BIC, also called Schwarz criterion, is defined as [2]:

$$\mathrm{BIC} = nS_n + \frac{d}{2}\log n \tag{166}$$

implying that $F_n = \mathrm{BIC} + o(\log n)$, i.e. that BIC can be derived as a criterion for regular models from approach given in [1]. On other hand, we have that singular models have term $\lambda \log n \leq \frac{d}{2}\log n$. Hence free energy of singular models is at most as high as in the regular models with the same size of the parameter space. Also note that BIC criterion is not applicable to singular models, since it does not describe well the asymptotic behavior of free energy of the model. One alternative that is suggested in [1, 2] is WBIC (widely applicable Bayesian information criterion) that is applicable to both regular and singular models. This criterion does not require to know true distribution $q(y|x)$ and is applicable in practical situations. According to [3], WBIC is defined as:

$$\mathrm{WBIC} = \mathbb{E}_w[-\log L_n(w)] = \mathbb{E}_w\Big[-\sum_{i=1}^{n}\log p(y_i|x_i,w)\Big] \tag{167}$$

where we used definition of likelihood function (31). It can be shown that WBIC *has the same asymptotic expansion as free energy, both for regular and singular models [3].*

4.5 Convergence of Bayes generalization and training errors

Even though $\mathbb{E}[F_n^0]$ has asymptotic expansion of the form (163), and equation (39) holds, Bayes generalization and training error need not have an asymptotic expansion. However, theorem 22 proves that $\mathbb{E}[nB_g]$ exists when $n \to \infty$. Before stating the theorem we introduce notion of empirical variance, which will be needed for convergence of Bayes training error.

Definition 38. *(Empirical variance) For a given model $p(y|x,w)$ and data samples $\{x_i, y_i\}_{i=1}^n$ we define empirical variance by:*

$$V = \sum_{i=1}^{n}\Big(\mathbb{E}_w[(\log p(y_i|x_i,w))^2] - (\mathbb{E}_w[\log p(y_i|x_i,w)])^2\Big) = \sum_{i=1}^{n}\mathrm{Var}_w[\log p(y_i|x_i,w)] \tag{168}$$

Recall that we defined $\mathbb{E}_w[\,\cdot\,] = \int \cdot\, p(w|z^{(n)})dw$ and $\mathrm{Var}_w[\,\cdot\,] = \mathbb{E}_w[\,(\cdot)^2\,] - (\mathbb{E}_w[\,\cdot\,])^2$. Now we state a theorem that characterizes asymptotic behavior of Bayes generalization and training error.

Theorem 22. *(Theorem 6.8 and Theorem 6.10 in [1]) Let a statistical model be given by $p(x|w)$, $\varphi(w)$ and let true distribution be $q(x)$. If fundamental conditions I (with index $s = 6$) and II hold then:*

1. *there exist random variables B_g^* and B_t^* such that:*

$$nB_g \xrightarrow{d} B_g^* \tag{169}$$

$$nB_t \xrightarrow{d} B_t^* \tag{170}$$

2. *expectation values converge:*

$$\mathbb{E}[nB_g] \to \mathbb{E}[B_g^*] \tag{171}$$

$$\mathbb{E}[nB_t] \to \mathbb{E}[B_t^*] \tag{172}$$

3. *there exists a constant $\nu > 0$ such that $\lim_{n\to\infty}\mathbb{E}[V] = 2\nu$ and:*

$$\mathbb{E}[B_g^*] = \lambda \tag{173}$$

$$\mathbb{E}[B_t^*] = \lambda - 2\nu \tag{174}$$

Bayes training error is determined by learning coefficient λ, but also by an additional birational invariant constant ν called singular fluctuation (see remark below). We skip definition of singular fluctuation (see definition 5.9 in [1]), since it requires a more thorough introduction to empirical processes. We mention that ν can be determined from numerical experiments [1], and encourage

interested reader to check remark 6.11 in [1] for determining ν in the regular model case.
Once we establish that $\mathbb{E}[B_g^*]$ exists, we can derive equation (174) in a simplified manner as follows:

$$\begin{aligned}
\mathbb{E}[B_g] &= \mathbb{E}[F_{n+1}^0] - \mathbb{E}[F_n^0] = \lambda \log(n+1) - (m-1)\log\log(n+1) + \mathbb{E}[F_{n+1}^R] - \lambda \log n + \\
&+ (m-1)\log\log n - \mathbb{E}[F_n^R] = \lambda \log \frac{n+1}{n} - (m-1)\log\frac{\log(n+1)}{\log n} + \underbrace{\mathbb{E}[F_{n+1}^R] - \mathbb{E}[F_n^R]}_{\to 0 \text{ as } n\to\infty} = \\
&= \lambda \log\Big(1+\frac{1}{n}\Big) - (m-1)\log\frac{\log\Big(n(1+\frac{1}{n})\Big)}{\log n} = \lambda \log\Big(1+\frac{1}{n}\Big) - (m-1)\log\frac{\log n + \log(1+\frac{1}{n})}{\log n} = \\
&= \frac{\lambda}{n} + \mathcal{O}\Big(\frac{1}{n^2}\Big) - (m-1)\log\frac{\log n + \frac{1}{n} + \mathcal{O}(\frac{1}{n^2})}{\log n} = \\
&= \frac{\lambda}{n} + \mathcal{O}\Big(\frac{1}{n^2}\Big) - (m-1)\log\Big(1 + \frac{1}{n\log n} + \mathcal{O}\Big(\frac{1}{n^2\log n}\Big)\Big) = \\
&= \frac{\lambda}{n} - \frac{m-1}{n\log n} + \mathcal{O}\Big(\frac{1}{n^2}\Big) \qquad (175)
\end{aligned}$$

Remark. *Note that we considered only the case when the partition function is given by equation (1), i.e. in the form:* $Z_n = \int_W \varphi(w)\prod_{i=1}^n p(y_i|x_i,w)dw$. *However, one could in general consider partition functions of the form* $Z_n = \int_W \varphi(w)\prod_{i=1}^n p(y_i|x_i,w)^\beta dw$ *with inverse temperature* $\beta > 0$. *In that case obtained asymptotic errors would depend on* β, *but also on a quantity* $\nu(\beta)$ *called singular fluctuation. In the theorem above, we considered only case* $\beta = 1$ *and hence* $\nu(\beta) = \nu = \text{const}$. *Together with the real log canonical threshold* λ, *the singular fluctuation is important for describing asymptotic behavior of singular models and statistical learning. For more details on this topic see [1].*

5 Statistical learning for neural networks

Singular statistical learning as presented in the previous chapters has been applied to different statistical models, among others to layered neural networks [21] and reduced rank regression [22]. Also, in the recent years there have been more practical approaches to this theory [19].

In this section we will apply established results from the previous sections to statistical learning of simple neural networks. Let's consider neural networks with single hidden layer, and let model be given by:

$$p(y|x,w) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}\Big(y - \Phi_w(x)\Big)\Big) \tag{176}$$

We will consider two simple cases as follows:

1. single hidden layer neural network with single neuron: $\Phi_w(x) = a(e^{bx} - 1)$ Let true parameters be set to $a_0 = b_0 = 0$ and assume that $\varphi(w) > 0$ around origin. Then, according to equation (8):

$$K(w) = \frac{1}{2}\int q(x)\Big(a(e^{bx}-1)\Big)^2 dx = \frac{1}{2}\int q(x)\Big(a\sum_{k=1}^{\infty}\frac{(bx)^k}{k!}\Big)^2 dx = \\ = \int \frac{1}{2}q(x)(abx + \frac{1}{2!}ab^2x^2 + \frac{1}{3!}ab^3x^3 + \cdots)^2 dx \tag{177}$$

   We have that $K(w) \equiv 0$ if and only if $ab = ab^2 = \cdots = ab^k = \cdots = 0$. This implies that $ab = 0$ and we see that $f(a,b) = ab$ is already a normal crossing function. Let $W_1 \subset W$ be the neighborhood of the origin and assume that $\varphi(w) > 0$ for $w \in W_1$. Using Taylor expansion of φ around origin, for zeta function (123) of this model, we have (up to a constant factor):

$$\zeta(z) = \int_{W_1} a^{2z}b^{2z}dadb \tag{178}$$

   and hence poles are given by $\lambda = -\frac{1}{2}$ with multiplicity $m = 2$. Furthermore, real log canonical threshold (110) is equal to $-\lambda = \frac{1}{2}$. Compared to the regular case with two parameters where RLCT $= \frac{d}{2} = 1$ (see example 24) , we see that here is real log canonical threshold is half of RLCT in regular case.

2. single hidden layer neural network with two neurons (see example 7.1 in [1]). Let neural network map be given by $\Phi_w(x) = a_1(e^{b_1x} - 1) + a_2(e^{b_2x} - 1)$. Assume that true parameters are equal to zero, and that $\varphi(w) > 0$ at singularities. Then, according to equation (8):

$$K(w) = \frac{1}{2}\int q(x)(a_1(e^{b_1x}-1)+a_2(e^{b_2x}-1))^2dx = \frac{1}{2}\int q(x)\Big(\sum_{k=1}^{\infty}\frac{x^k}{k!}(a_1b_1^k+a_2b_2^k)\Big)^2 dx \tag{179}$$

   Define the following polynomials:

$$p_k = a_1b_1^k + a_2b_2^k, \quad k = 1,2,\ldots \tag{180}$$

   According to example 17 we have $p_k \in \langle p_1, p_2\rangle$ for $k \geq 3$ (since $H = 2$ in example 17). Therefore if $p_1 \equiv 0$ and $p_2 \equiv 0$ holds then we have $p_k \equiv 0$, for all positive integers k. Moreover, $p_1(a,b) = p_2(a,b) = 0$ holds if and only if:

$$(a_1b_1 + a_2b_2)^2 + (a_1b_1^2 + a_2b_2^2)^2 = 0 \tag{181}$$

   We will provide different desingularization scheme than in example 3.19 in [1]. Instead of starting with blow-up at center $\mathbb{V}(a_1, b_1)$ as in [1], we start with $\mathbb{V}(a_1, a_2)$. Note that even though this approach yields a desingularization scheme that is different, parameters we want to determine λ and m are birational invariant i.e. independent of the desingularization scheme.

   Let's first show the whole desingularization scheme starting with $\mathbb{R}^4$. Scheme in equation (182) consists of five sequential blow-ups. Manifolds $U_3, U_6, U_7, U_9, U_{10}$ contain only normal crossing singularities (denoted by (+) in the scheme below). Let's now show that this scheme indeed

desingularizes function f.

$$\mathbb{R}^4 \xrightarrow{\mathbb{V}(a_1,a_2)} \begin{cases} U_1 \xrightarrow{\mathbb{V}(a_1,b_2)} \begin{cases} U_3\ (+) \\ U_4 \xrightarrow{\mathbb{V}(b_1,b_2)} \begin{cases} U_5 \xrightarrow{\mathbb{V}(b_1,b_2)} \begin{cases} U_7\ (+) \\ U_8 \xrightarrow{\mathbb{V}(a_1+1,b_2)} \begin{cases} U_9\ (+) \\ U_{10}\ (+) \end{cases} \end{cases} \\ U_6\ (+) \end{cases} \end{cases} \\ U_2\ (\text{sym.}) \end{cases} \tag{182}$$

In order to make notation less cumbersome we will use the same symbols for variables before and after blow-up, e.g. in $a_1 = a_1 a_2$ the variable a_1 on the left hand side is the old variable, whereas a_1 is a new variable (that only shares the same symbol with the previous one). Also, since we need to keep track of terms obtained through change of variables, i.e. terms from determinants of Jacobians of transformations, we will use differential form of function f (as in [21]):

$$f = \Big[(a_1 b_1 + a_2 b_2)^2 + (a_1 b_1^2 + a_2 b_2^2)^2\Big]^z da_1 da_2 db_1 db_2 \tag{183}$$

Furthermore, since we use the same variables a_1, a_2, b_1, b_2 in the whole process, we will leave out the symbols for differentials in equation (183).

i) We start by blowing up $\mathbb{R}^4$ with center $\mathbb{V}(a_1, a_2)$. We obtain two local coordinates U_1 and U_2:

- on U_1 we get:

$$\left.\begin{aligned} a_1 &= a_1 a_2 \\ a_2 &= a_2 \end{aligned}\right\} \implies f = \Big[(a_1 b_1 + a_2 b_2)^2 + (a_1 b_1^2 + a_2 b_2^2)^2\Big]^z =$$
$$= a_2^{2z}\Big[(\underbrace{a_1 b_1 + b_2}_{=:b_2})^2 + (a_1 b_1^2 + b_2^2)^2\Big]^z a_2$$
$$= a_2^{2z}\Big[b_2^2 + (a_1 b_1^2 + (b_2 - a_1 b_1)^2)^2\Big]^z a_2$$

  which is not a normal crossing function since the term inside the brackets vanishes near the origin (i.e. function $a(\cdot)$ in definition 34 is not nonzero). Hence we will need to continue applying blow-ups. Obtained function has a singularity at point $(a_1, b_2) = (0, 0)$ (can be checked from Jacobian of f and theorem 9). Hence in the following step ii) we apply blow-up of U_1 with center $\mathbb{V}(a_1, b_2)$.
- on U_2: by symmetry in a_1 and a_2, normal crossing singularities in this local coordinates are the same as in U_1. Hence it is enough to consider only U_1.

ii) after blowing-up U_1 with center $\mathbb{V}(a_1, b_2)$, on two local coordinates U_3 and U_4, we get:

- on U_3

$$\left.\begin{aligned} a_1 &= a_1 b_2 \\ b_2 &= b_2 \end{aligned}\right\} \quad \implies \quad f = a_2^{2z}\Big[b_2^2 + (a_1 b_1^2 b_2 + (b_2 - a_1 b_1 b_2)^2)^2\Big]^z a_2 b_2$$
$$= a_2^{2z} b_2^{2z}\Big[1 + (a_1 b_1^2 + b_2(1 - a_1 b_1)^2)^2\Big]^z a_2 b_2$$

  Obtained function is normal crossing since the term between the brackets is strictly positive and hence does not contain singularities. Real log canonical threshold on U_3 is $\lambda_{U_3} = \min\{\frac{1+1}{2}, \frac{1+1}{2}\} = 1$.
- on U_4

$$\left.\begin{aligned} a_1 &= a_1 \\ b_2 &= a_1 b_2 \end{aligned}\right\} \implies f = a_2^{2z}\Big[a_1^2 b_2^2 + (a_1 b_1^2 + (a_1 b_2 - a_1 b_1)^2)^2\Big]^z a_1 a_2 =$$
$$= a_1^{2z} a_2^{2z}\Big[b_2^2 + (b_1^2 + a_1(\underbrace{b_1 - b_2}_{=:b_1})^2)^2\Big]^z a_1 a_2$$
$$= a_1^{2z} a_2^{2z}\Big[b_2^2 + ((b_1 + b_2)^2 + a_1 b_1^2)^2\Big]^z a_1 a_2$$

  Obtained function has a singularity at $(b_1, b_2) = (0, 0)$ and hence we will apply blow-up of U_4 with center $\mathbb{V}(b_1, b_2)$ in the next step.

iii) after blowing-up U_4 with center $\mathbb{V}(b_1, b_2)$, on two local coordinates U_5 and U_6, we get:

- on U_5

$$\left.\begin{array}{l} b_1 = b_1 \\ b_2 = b_1 b_2 \end{array}\right\} \implies f = a_1^{2z} a_2^{2z} \left[b_1^2 b_2^2 + ((b_1 + b_1 b_2)^2 + a_1 b_1^2)^2\right]^z a_1 a_2 b_1 =$$
$$= a_1^{2z} a_2^{2z} b_1^{2z} \left[b_2^2 + b_1^2((1 + b_2)^2 + a_1)^2\right]^z a_1 a_2 b_1$$

Obtained function has a singularity at $(b_1, b_2) = (0, 0)$ and hence we will apply blow-up of U_5 with center $\mathbb{V}(b_1, b_2)$ in the next step.

- on U_6

$$\left.\begin{array}{l} b_1 = b_1 b_2 \\ b_2 = b_2 \end{array}\right\} \implies f = a_1^{2z} a_2^{2z} \left[b_2^2 + ((b_1 b_2 + b_2)^2 + a_1 b_1^2 b_2^2)^2\right]^z a_1 a_2 b_2 =$$
$$= a_1^{2z} a_2^{2z} b_2^{2z} \left[1 + (b_2(b_1 + 1)^2 + a_1 b_1^2 b_2)^2\right]^z a_1 a_2 b_2$$

Obtained function is normal crossing since the term inside of the brackets is strictly positive. Real log canonical threshold on U_6 is $\lambda_{U_6} = \min\{\frac{1+1}{2}, \frac{1+1}{2}, \frac{1+1}{2}\} = 1$.

iv) $\mathbb{V}(b_1, b_2)$:

- on U_7

$$\left.\begin{array}{l} b_1 = b_1 b_2 \\ b_2 = b_2 \end{array}\right\} \implies f = a_1^{2z} a_2^{2z} b_1^{2z} b_2^{2z} \left[b_2^2 + b_1^2 b_2^2((1 + b_2)^2 + a_1)^2\right]^z a_1 a_2 b_1 b_2^2 =$$
$$= a_1^{2z} a_2^{2z} b_1^{2z} b_2^{4z} \left[1 + b_1^2((1 + b_2)^2 + a_1)^2\right]^z a_1 a_2 b_1 b_2^2$$

Obtained function is normal crossing since the term inside the brackets is positive. Real log canonical threshold on U_6 is $\lambda_{U_7} = \min\{\frac{1+1}{2}, \frac{1+1}{2}, \frac{1+1}{2}, \frac{2+1}{4}\} = \frac{3}{4}$.

- on U_8

$$\left.\begin{array}{l} b_1 = b_1 \\ b_2 = b_1 b_2 \end{array}\right\} \implies f = a_1^{2z} a_2^{2z} b_1^{2z} \left[b_1^2 b_2^2 + b_1^2((1 + b_1 b_2)^2 + a_1)^2\right]^z a_1 a_2 b_1^2 =$$
$$= a_1^{2z} a_2^{2z} b_1^{4z} \left[b_2^2 + ((1 + b_1 b_2)^2 + a_1)^2\right]^z a_1 a_2 b_1^2 =$$
$$= a_1^{2z} a_2^{2z} b_1^{4z} \left[b_2^2 + (b_1 b_2(b_1 b_2 + 2) + a_1 + 1)^2\right]^z a_1 a_2 b_1^2$$

Obtained function is not normal crossing and has a singularity at $(a_1, b_2) = (-1, 0)$ and hence in the next step we will apply blow-up of U_8 with center $\mathbb{V}(a_1 + 1, b_2)$.

v) $\mathbb{V}(a_1 + 1, b_2)$:

- on U_9

$$\left.\begin{array}{l} a_1 + 1 = a_1 b_2 \\ b_2 = b_2 \end{array}\right\} \implies f = (a_1 b_2 - 1)^{2z} a_2^{2z} b_1^{4z} \left[b_2^2 + (b_1 b_2(b_1 b_2 + 2) + a_1 b_2)^2\right]^z (a_1 b_2 - 1) a_2 b_1^2 b_2 =$$
$$= a_2^{2z} b_1^{4z} b_2^{2z} \left[(a_1 b_2 - 1)(1 + (b_1(b_1 b_2 + 2) + a_1)^2)\right]^z (a_1 b_2 - 1) a_2 b_1^2 b_2$$

Obtained function is normal crossing since the term inside of the brackets defines an algebraic set $\mathbb{V}(a_1 b_2 - 1)$ and by checking Jacobian of the whole term, we can see that it does not have singular points. Real log canonical threshold on U_9 is $\lambda_{U_9} = \min\{\frac{1+1}{2}, \frac{2+1}{4}, \frac{1+1}{2}\} = \frac{3}{4}$.

- on U_{10}

$$\left.\begin{array}{l} a_1 = a_1 \\ b_2 = (a_1 + 1) b_2 \end{array}\right\} \implies$$
$$f = a_1^{2z} a_2^{2z} b_1^{4z} \left[(a_1 + 1)^2 b_2^2 + (b_1(a_1 + 1) b_2^2(b_1(a_1 + 1) b_2 + 2) + a_1 + 1)^2\right]^z a_1(a_1 + 1) a_2 b_1^2 =$$
$$= a_1^{2z}(a_1 + 1)^{2z} a_2^{2z} b_1^{4z} \left[b_2^2 + (b_1 b_2^2(b_1(a_1 + 1) b_2^2(b_1(a_1 + 1) b_2 + 2) + 1)^2\right]^z a_1(a_1 + 1) a_2 b_1^2$$

Obtained function is normal crossing since the term inside of the brackets is strictly positive. Real log canonical thresholdon U_{10} is given by $\lambda_{U_{10}} = \min\{\frac{1+1}{2}, \frac{1+1}{2}, \frac{1+1}{2}, \frac{2+1}{4}\} = \frac{3}{4}$.

Final resolved manifold is given by $U = U_3 \cup U_6 \cup U_7 \cup U_9 \cup U_{10}$, and real log canonical threshold is hence given by $\lambda = \min\{\lambda_{U_3}, \lambda_{U_6}, \lambda_{U_7}, \lambda_{U_9}, \lambda_{U_{10}}\} = \frac{3}{4}$. Compared to the regular models with $d = 4$ parameters and RLCT= $\frac{d}{2} = 2$, we see that RLCT obtained for neural network with $d = 4$ parameters is smaller. Using obtained RLCT, we can determine asymptotic values of free energy and Bayes generalization error as discussed in sections 4.4 and 4.5, respectively.

References

[1] Sumio Watanabe. *Algebraic geometry and statistical learning theory*. Number 25. Cambridge university press, 2009.

[2] Sumio Watanabe. *Mathematical theory of Bayesian statistics*. CRC Press, 2018.

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

[4] Dennis Elbrächter, Dmytro Perekrestenko, Philipp Grohs, and Helmut Bölcskei. Deep neural network approximation theory. *arXiv preprint arXiv:1901.02220*, 2019.

[5] Robert G Bartle. *The elements of integration and Lebesgue measure*. John Wiley & Sons, 2014.

[6] Athanasios Papoulis and H Saunders. Probability, random variables and stochastic processes. 1989.

[7] Petre Stoica and Thomas L Marzetta. Parameter estimation problems with singular information matrices. *IEEE Transactions on Signal Processing*, 49(1):87–90, 2001.

[8] Zvika Ben-Haim and Yonina C Eldar. On the constrained cramér–rao bound with a singular fisher information matrix. *IEEE Signal Processing Letters*, 16(6):453–456, 2009.

[9] Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

[10] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[11] Dennis D Boos and Leonard A Stefanski. *Essential statistical inference: theory and methods*, volume 120. Springer Science & Business Media, 2013.

[12] David Cox, John Little, and Donal OShea. *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra*. Springer Science & Business Media, 2013.

[13] Michael Francis Atiyah and Ian Grant Macdonald. *Introduction to commutative algebra*. CRC Press, 2018.

[14] Daniel Perrin. Tangent spaces and singular points. *Algebraic Geometry: An Introduction*, pages 87–99, 2008.

[15] John Lee. *Introduction to topological manifolds*, volume 202. Springer Science & Business Media, 2010.

[16] János Kollár. Nash's work in algebraic geometry. *Bulletin (New Series) of the American Mathematical Society*, 54(2), 2017.

[17] Heisuke Hironaka. Resolution of singularities of an algebraic variety over a field of characteristic zero. *Annals of Mathematics*, pages 109–326, 1964.

[18] Shaowei Lin. *Algebraic methods for evaluating integrals in Bayesian statistics*. PhD thesis, UC Berkeley, 2011.

[19] Daniel Murfet, Susan Wei, Mingming Gong, Hui Li, Jesse Gell-Redman, and Thomas Quella. Deep learning is singular, and that's good. *arXiv preprint arXiv:2010.11560*, 2020.

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

[21] Miki Aoyagi and Sumio Watanabe. Resolution of singularities and the generalization error with bayesian estimation for layered neural network. *IEICE Trans*, pages 2112–2124, 2005.

[22] Miki Aoyagi and Sumio Watanabe. Stochastic complexities of reduced rank regression in bayesian estimation. *Neural Networks*, 18(7):924–933, 2005.