REVIEW **Communicated by Keisuke Yamazaki**

Parameter Identifiability in Statistical Machine Learning: A Review

Zhi-Yong Ran
ranzy@cqupt.edu.cn
Chongqing Key Laboratory of Computational Intelligence, School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China

Bao-Gang Hu
hubg@nlpr.ia.ac.cn
NLPR & LIAMA, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China

This review examines the relevance of parameter identifiability for statistical models used in machine learning. In addition to defining main concepts, we address several issues of identifiability closely related to machine learning, showing the advantages and disadvantages of state-of-the-art research and demonstrating recent progress. First, we review criteria for determining the parameter structure of models from the literature. This has three related issues: parameter identifiability, parameter redundancy, and reparameterization. Second, we review the deep influence of identifiability on various aspects of machine learning from theoretical and application viewpoints. In addition to illustrating the utility and influence of identifiability, we emphasize the interplay among identifiability theory, machine learning, mathematical statistics, information theory, optimization theory, information geometry, Riemann geometry, symbolic computation, Bayesian inference, algebraic geometry, and others. Finally, we present a new perspective together with the associated challenges.

1 Introduction

The main focus of this review is parameter identifiability in statistical machine learning in which the modeling approach is a statistical one. In addition to providing a formal description of basic concepts, key theories, typical techniques relevant to machine learning and parameter identifiability, we give a comprehensive review on the criteria for examining identifiability and present a broad analysis of the influence of identifiability in various aspects of machine learning. We also review some advanced topics and ongoing research.

Neural Computation **29, 1151–1203** (2017)
doi:10.1162/NECO_a_00947

Table 1: Variables, Meanings, and Equation Numbers.

Variable	Meaning	Equation Number
$p(\mathbf{z})$	PDF of $\mathbf{z} = (\mathbf{x}, \mathbf{y})$	1.1
$\mathbf{f}(\mathbf{x}, \mathbf{w}), \mathbf{w} \in \mathcal{W}$	Functions implemented by learning machine	1.2
$\mathcal{D}_n$	Training set of size n	1.3
$R(\mathbf{w})$	Risk functional	1.5
$\mathbf{w}_{\text{opt}}$	Optimal parameter	1.6
$R_{\text{emp}}(\mathbf{w})$	Empirical risk functional	1.7
$R_{\text{reg}}(\mathbf{w})$	Regularized risk functional	1.8
$\widehat{\mathbf{w}}_n$	Minimizer of error function $E(\mathbf{w})$	1.9
$[\mathbf{w}]$	Equivalent class of $\mathbf{w}$	1.14
$\mathcal{W}_c$	Constrained parameter space	2.1
$\text{FIM}(\mathbf{w})$	Fisher information matrix of PDF $p(\mathbf{z}\|\mathbf{w})$	2.2
$\text{KL}(\mathbf{w}_0, \mathbf{w})$	KLD between $p(\mathbf{z}\|\mathbf{w}_0)$ and $p(\mathbf{z}\|\mathbf{w})$	2.9
$\mathbf{H}(\mathbf{w})$	Hessian matrix of $\text{KL}(\mathbf{w}_0, \mathbf{w})$	2.12
$\varphi(\mathbf{w})$	Identifying function	2.17
$\mathcal{C}$	Critical set	3.1

This introduction has two parts: the statistical formulation of machine learning, and the general description of parameter identifiability and its major concerns.

Table 1 lists the important variables, meanings, and equation numbers used throughout the review.

1.1 Machine Learning

1.1.1 Machine Learning: Learning from Data. The problem of searching for regularities (or dependencies) in data is a fundamental one and has a long and successful history. For instance, the extensive astronomical observations of Tycho Brahe in the 16th century allowed Johannes Kepler to discover the empirical laws of planetary motion, which provided a springboard for the development of classical mechanics (Bishop, 2006). Similarly, the discoveries of regularities in atomic spectra played a key role in the development and verification of quantum physics in the early 20th century (Bishop, 2006). With the advent of the era of big data, the deluge of data calls for automatic and powerful methods of data analysis, which is what machine learning provides.

The goal of machine learning is to build learning machines (see section 1.1.2) that are capable of employing learning algorithms to automatically discover regularities in data and, with the uncovered regularities, predict future data or perform other kinds of tasks. The predictive accuracy is known as *generalization capability*: the capability of the learned model to provide accurate estimation for further data (Murphy, 2012).

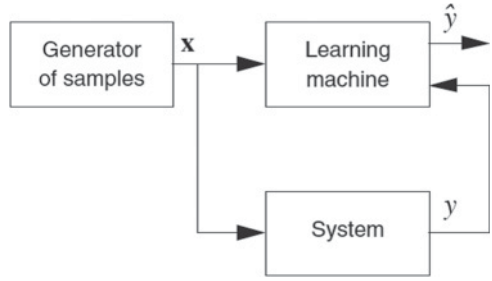

Figure 1: Schematic illustration of statistical machine learning (Cherkassky & Mulier, 2007).

1.1.2 Formulation of the Learning Problem. This review adopts the viewpoint that the best way to solve the problems of machine learning is to use the tools of statistical theory, as the statistical theory can be applied to any problem involving uncertainty (Murphy, 2012).

The general formulation of a learning problem consists of three components (see Figure 1) (Cherkassky & Mulier, 2007):

- *Generator*. The generator produces random inputs $\mathbf{x} \in \mathcal{X}$ drawn from a fixed but unknown probability density function (PDF) $p(\mathbf{x})$, where $\mathcal{X}$ is the input space.
- *System*. The system returns an output $\mathbf{y} \in \mathcal{Y}$ to every input $\mathbf{x}$, according to the conditional PDF $p(\mathbf{y}|\mathbf{x})$, also fixed but unknown, where $\mathcal{Y}$ is the output space. The probabilistic dependency between $\mathbf{x}$ and $\mathbf{y}$ is summarized by $p(\mathbf{y}|\mathbf{x})$. Let $\mathbf{z} = (\mathbf{x}, \mathbf{y})$; the PDF of $\mathbf{z}$ is

$$p(\mathbf{z}) = p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x}). \tag{1.1}$$

- *Learning machine*. The learning machine is capable of implementing a set of functions

$$\{\mathbf{f}(\mathbf{x}, \mathbf{w}), \mathbf{w} \in \mathcal{W}\} \tag{1.2}$$

indexed by a parameter $\mathbf{w} \in \mathcal{W}$, and eventually yields an output $\widehat{\mathbf{y}}$ for each input $\mathbf{x}$. Here $\mathcal{W}$ is the admissible parameter space. For simplicity, we assume that $\mathbf{w}$ is a finite-dimensional vector $\mathbf{w} = (w_1, \ldots, w_k) \in \mathbb{R}^k$.

The problem of learning is that of choosing from $\{\mathbf{f}(\mathbf{x}, \mathbf{w}), \mathbf{w} \in \mathcal{W}\}$ the one that best approximates the system's response (Vapnik, 1996). The selection of desired function is based on a training set of n independent and identically distributed (i.i.d.) samples,

$$\mathcal{D}_n = \{(\mathbf{x}_j, \mathbf{y}_j),\ j = 1, \ldots, n\}, \tag{1.3}$$

drawn according to equation 1.1. Note that we use the same font, such as $\mathbf{x}_j$, $\mathbf{y}_j$, to denote a random variable and its realization; their meaning is clear from the context.

1.1.3 Types of Machine Learning. The spectrum of machine learning is rather broad and can be roughly subdivided into two main types. In *supervised* learning, the goal is to learn an unknown dependency (or mapping) from input $\mathbf{x}$ to output $\mathbf{y}$, given a set of labeled input-output pairs, equation 1.3.

In the simplest situation, the input $\mathbf{x}$ is a finite-dimensional real vector. In general, however, $\mathbf{x}$ could be a complex structured object, such as an image, a sentence, or a time series. Similarly, the form of the output $\mathbf{y}$ in principle can be anything, but most methods assume that $\mathbf{y}$ is a real vector or a categorical or nominal variable from a certain finite set. When $\mathbf{y}$ is real valued, the problem is known as *regression* (Bishop, 2006; Cherkassky & Mulier, 2007), and when it is categorical, the problem is known as *classification* or *pattern recognition* (Bishop, 1995; Duda, Hart, & Stork, 2001).

In *unsupervised* learning, we have only input samples $\mathbf{x}_j$, $j = 1, \ldots, n$, and there is no output during learning. The goal of unsupervised learning may be to detect interesting structures such as clusters in the data. This practice is sometimes called *knowledge discovery* (Cherkassky & Mulier, 2007).

For simplicity, we consider only regression problems and assume that the inputs and outputs dwell in the Euclidean spaces $\mathbb{R}^d$, $\mathbb{R}^m$, respectively.

1.1.4 Modeling Approach. Mathematical models have become another sensing channel for humans to perceive, describe, and understand natural and virtual worlds. Thus, a large number of models have been generated for a vast variety of applications. Their modeling approaches are of course different in various aspects. Typical modeling approaches include the following:

- *Knowledge-driven modeling*. Traditional science and engineering are based on using a first-principle modeling approach to describe physical, biological, and social phenomena. Such an approach starts with a basic scientific model (e.g., Newton's law of mechanics or Maxwell's theory of electromagnetism) and then builds various applications on them. Under this approach, experimental data are used to verify the underlying model or estimate its parameters that are difficult to measure directly. This modeling approach is called a *knowledge-driven* or *mechanistic-based* manner. The associated inference methodology is a deductive one, that is, progress from general (dependency) to particular (observational data).
- *Data-driven modeling*. In some applications, the underlying first principles are unknown or the systems under study are too complex to be mathematically described. Fortunately, the available data can be used

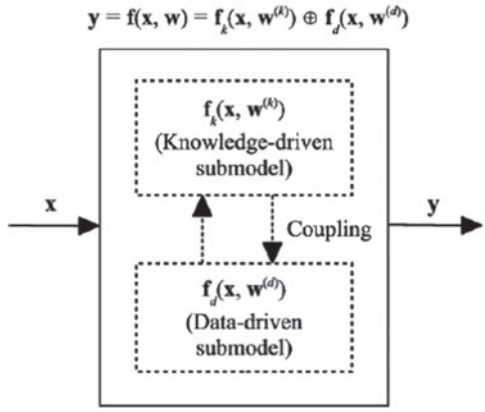

Figure 2: Schematic diagram of the GC model including the KD and DD submodels. Two parameters, $\mathbf{w}^{(k)}$ and $\mathbf{w}^{(d)}$, are associated with the two submodels, respectively (Ran & Hu, 2014a).

to derive useful relationships between a system's input and output variables. This modeling approach is a data-driven fashion. The associated inference methodology is an inductive one, that is, a progress from particular to general.

- *Knowledge- and data-driven modeling*. In order to take advantage of the two previous approaches, a study of integrating two approaches is reported (Todorovski & Dzeroski, 2006; Hu, Qu, & Yang, 2009). Hence, Psichogios and Ungar (1992) call these hybrid models. For a stress on a mathematical description, the term *generalized constraint* (GC) (Zadeh, 2005) is used to refer to these models. Figure 2 schematically depicts a GC model, which basically consists of two modules: knowledge driven (KD) submodel $\mathbf{f}_k(\mathbf{x}, \mathbf{w}^{(k)})$ and data driven (DD) submodel $\mathbf{f}_d(\mathbf{x}, \mathbf{w}^{(d)})$. The complete GC model function is formulated as

$$\mathbf{f}(\mathbf{x}, \mathbf{w}) = \mathbf{f}_k(\mathbf{x}, \mathbf{w}^{(k)}) \oplus \mathbf{f}_d(\mathbf{x}, \mathbf{w}^{(d)}), \tag{1.4}$$

where $\mathbf{w} = (\mathbf{w}^{(k)}, \mathbf{w}^{(d)})$, and the symbol $\oplus$ represents a coupling operation between the two submodels. Generally the KD submodel contains physically interpretable parameter $\mathbf{w}^{(k)}$ whose identifiability is of particular importance to understanding the system. (For a detailed description of GC models, see Hu, Wang, Yang, & Qu, 2007; Hu et al., 2009; Qu & Hu, 2011; Ran & Hu, 2014a.)

1.1.5 Goal of Learning. In order to obtain the best approximation to the system's response, one measures the loss $L(\mathbf{x}, \mathbf{y}, \mathbf{f}(\mathbf{x}, \mathbf{w}))$ between the output

$\mathbf{y}$ of the system and the response $\mathbf{f}(\mathbf{x}, \mathbf{w})$ provided by the learning machine. The expectation of the loss is called the *risk functional* (Schölkopf & Smola, 2002; Cherkassky & Mulier, 2007):

$$R(\mathbf{w}) = \int L(\mathbf{x}, \mathbf{y}, \mathbf{f}(\mathbf{x}, \mathbf{w}))p(\mathbf{x}, \mathbf{y})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}. \tag{1.5}$$

The goal of learning is to find the optimal parameter $\mathbf{w}_{\text{opt}}$ that minimizes the risk functional. This gives rise to the following optimization problem:

$$\mathbf{w}_{\text{opt}} = \arg\min_{\mathbf{w}\in\mathcal{W}} R(\mathbf{w}). \tag{1.6}$$

1.1.6 Inductive Principles. In a classical parametric setting, the statistical model is specified first, and then its parameters are estimated from the data. However, in the formulation of machine learning, the underlying model is unknown, and it is estimated using a large (infinite) number of candidate models to describe available data. The main issue is choosing a model of the right complexity to describe the training data, as stated qualitatively by Occam's razor principle: "Entities are not to be multiplied beyond necessity" (Duda et al., 2001). Several inductive principles, such as the regularization (penalization) inductive principle, the early stopping rule, structural risk minimization (SRM), Bayesian inference, and minimum description length (MDL) (Cherkassky & Mulier, 2007), provide different quantitative interpretations of Occam's principle. (For more details, see Cherkassky & Mulier, 2007.)

In the classical parametric setting, the commonly used empirical risk minimization (ERM) principle (Vapnik, 1996, 1998) recommends that the risk functional should be replaced by the empirical risk functional:

$$R_{\text{emp}}(\mathbf{w}) = \frac{1}{n}\sum_{j=1}^{n} L(\mathbf{x}_j, \mathbf{y}_j, \mathbf{f}(\mathbf{x}_j, \mathbf{w})). \tag{1.7}$$

When n is sufficiently large, the law of large numbers ensures that it is a good approximation to $R(\mathbf{w})$. From a statistical viewpoint, this constitutes a particular case of an M-estimator (Van der Vaart & Wellner, 1996). Estimators of this type are studied in detail in the field of empirical processes (Van der Vaart & Wellner, 1996).

The structural risk minimization (SRM) principle, an essential part of statistical learning theory attributed to Vapnik (1996, 1998), shows that it is imperative to restrict the parameter space $\mathcal{W}$ from which $\mathbf{w}$ is chosen, that is, to restrict the complexity of the set of learning machines to avoid overfitting. One popular way for implementing the SRM principle is to minimize the regularized empirical risk,

$$R_{\text{reg}}(\mathbf{w}) = R_{\text{emp}}(\mathbf{w}) + \gamma||\mathbf{w}||, \tag{1.8}$$

where $||\mathbf{w}||$ is a certain norm of $\mathbf{w}$ and the parameter $\gamma > 0$ controls the trade-off between the goodness of fit $R_{\text{emp}}(\mathbf{w})$ and model complexity $||\mathbf{w}||$.

The regularization method expresses our prior belief that the type of function we seek exhibits a certain smooth behavior and can usually be cast in the Bayesian framework (Bishop, 2006; Murphy, 2012).

No matter what kind of inductive principle is used, we are eventually led to minimizing an error function $E(\mathbf{w})$. From a statistical viewpoint, the process of learning from data is in fact the process of computing a particular statistic.

In what follows, we use $E(\mathbf{w})$ to denote an empirical risk $R_{\text{emp}}(\mathbf{w})$ or a regularized empirical risk $R_{\text{reg}}(\mathbf{w})$ and denote

$$\widehat{\mathbf{w}}_n = \arg\min_{\mathbf{w}\in\mathcal{W}} E(\mathbf{w}). \tag{1.9}$$

Here, $\widehat{\mathbf{w}}_n$, shorthand for $\widehat{\mathbf{w}}(\mathbf{z}_1, \ldots, \mathbf{z}_n)$, can be understood as an estimator of $\mathbf{w}_{\text{opt}}$.

1.1.7 Learning Algorithm. As noted, an inductive principle tells us what to do with the finite data; this results in the problem of minimizing an error function $E(\mathbf{w})$. Nevertheless, a learning algorithm specifies how to solve this minimization problem; that is, it concerns a constructive method to obtain a good approximation $\widehat{\mathbf{w}}_n$ for $\mathbf{w}_{\text{opt}}$. Such a mapping

$$\mathcal{D}_n \to \widehat{\mathbf{w}}_n \tag{1.10}$$

is called a learning algorithm (Watanabe, 2009). In machine learning, the parametric form of function (and hence the error function) is nonlinear in parameters, giving rise to a nonlinear optimization problem whose global closed-form solutions are generally unavailable. Hence, one needs to consider numerical algorithms, which consist of a succession of steps,

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \Delta\mathbf{w}_t, t = 1, 2, \cdots \tag{1.11}$$

where $\Delta\mathbf{w}_t$ is a general form of update at step t. Different algorithms involve different choices for the increment $\Delta\mathbf{w}_t$ and the starting point $\mathbf{w}_1$. Since a thorough discussion of nonlinear optimization theory and method is beyond the scope of this review, we mainly consider the most basic techniques, which make use of the gradient $\nabla E(\mathbf{w})$ of error function.

Geometrically, the $E(\mathbf{w})$ can be viewed as a surface sitting above parameter space $\mathbb{R}^k$, as shown in Figure 3 (Bishop, 1995). The learning process takes place by successively modifying the parameters. These parameters

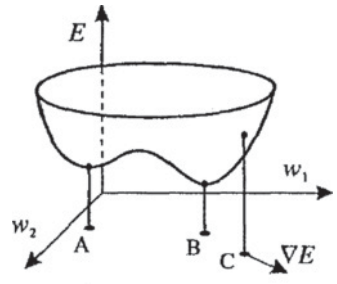

Figure 3: Geometrical view of the error function $E(\mathbf{w})$ as a surface sitting over $\mathbb{R}^k$. Point $\mathbf{w}_A$ is a local minimum, and $\mathbf{w}_B$ is the global minimum. At any point $\mathbf{w}_C$, the gradient of the error surface is given by the vector ∇E (Bishop, 1995).

change during learning, forming a trajectory in $\mathbb{R}^k$. It is therefore important to study the learning trajectory to elucidate the dynamic process of learning.

To sum up, the final success of a learning task depends on appropriately selecting a set of parametric functions, an appropriate risk functional, a theoretically sound inductive principle, and an effective optimization strategy.

1.2 Parameter Identifiability

1.2.1 Parameter Identifiability: A General Introduction. Parameter identifiability in statistical learning is concerned with the theoretical uniqueness of model parameters determined from an underlying statistical family (Rothenberg, 1971; Paulino & Pereira, 1994; Dasgupta, Self, & Gupta, 2007). In a general sense the identifiability study is just one aspect of a larger problem, the inverse problem (Tikhonov & Arsenin, 1977), which basically encompasses identifiability and identification (e.g., objective function, regularization and algorithm). For a full treatment of the identifiability problem, one should distinguish between the concepts of theoretical uniqueness and numerical uniqueness. Roughly, in a theoretically unidentifiable model, a subset of these parameters cannot be uniquely determined even if infinitely many data are accessible. For example, this can occur if two parameters are confounded and appear only as a product (Hu et al., 2009; Dimattina & Zhang, 2010; Cole, Morgan, & Titterington, 2010). However, in a theoretically identifiable model, numerical nonuniqueness can occur due to a lack of data or collinearity of data, which is discussed in classical linear regression analysis (Shao, 1999). In this review, identifiability means theoretical uniqueness.

Identifiability analysis is important not only for models whose parameters have physically interpretable meaning, but also for models whose parameters have no physical implications, because identifiability has a significant influence on many aspects of a learning problem, such as estimation

theory, hypothesis testing, model selection, learning algorithm, learning dynamic, and Bayesian inference (see section 3).

The problem of parameter identifiability arises in many scientific areas, so the literature on identifiability analysis is found in journals in a large variety of fields, including mathematical statistics (Rothenberg, 1971; Paulino & Pereira, 1994; Shao, 1999; Dasgupta et al., 2007), machine learning (Amari, Park, & Ozeki, 2006; Bishop, 2006; Watanabe, 2009), system identification (Ljung, 1999), signal processing (Moore & Sadler, 2004; Yao & Giannakis, 2005; Fortunati et al., 2012), mathematical biosciences (Jacquez & Greif, 1985; Vajda, Godfrey, & Rabitz, 1989; Cole, Morgan, & Titterington, 2010), carcinogenesis models (Little, Heidenreich, & Li, 2009, 2010), and dynamic control (Xia & Moog, 2003; Miao, Xia, Perelson, & Wu, 2011). This diversity justifies several formulations and meanings that have been given to the term *identifiability*. Historically, the study of identifiability can be traced back to Koopmans and Reierøsl (1950), where a specific statistical model, a linear simultaneous equation system, is considered. As they emphasized, the identifiability problem is "a general and fundamental problem arising, in many fields of inquiry, as a concomitant of the scientific procedure that postulates the existence of a structure." Identifiability when perfect records are assumed has been the subject of much research; this analysis was set out by Bellman and Astrom (1970) in the context of a dynamic control model, where the concept of identifiability is alternatively called *structural identifiability*. At the same time, the term *a priori identifiability* has also been quite widely used on the ground that identifiability analysis should be addressed before a proposed experiment is carried out (Saccomani, Audoly, Bellu, & D'Angio, 2010). For clarity, we will use the term *parameter identifiability* because we restrict ourselves to models completely determined by a finite-dimensional parameter vector.

1.2.2 Parameter Identifiability: Basic Concepts. To put the above description into a formal mathematical framework, we consider a statistical space $\{\mathcal{Z}, \mathcal{A}, \mathcal{P}\}$, where $\mathcal{Z}$ is the sample space, $\mathcal{A}$ is the σ-algebra defined in $\mathcal{Z}$, $\mathcal{P} = \{P_{\mathbf{w}} : \mathbf{w} \in \mathcal{W}\}$ is a family of probability measures on $\{\mathcal{Z}, \mathcal{A}\}$ and $P_{\mathbf{w}}$ is a probability measure indexed by $\mathbf{w}$ (Paulino & Pereira, 1994; Shao, 1999; Dasgupta et al., 2007). Parameter identifiability is concerned with the theoretical uniqueness of model parameter $\mathbf{w}$ determined from the statistical family $\mathcal{P}$.

Definition 1. *Let $p(\boldsymbol{x})$ be the PDF of input $\boldsymbol{x}$ and $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})$ be the PDF of output $\boldsymbol{y}$ conditioned on $\boldsymbol{x}$ and a parameter $\boldsymbol{w}$. Let $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$, and $p(\boldsymbol{z}|\boldsymbol{w}) = p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) = p(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})$ be the PDF of $\boldsymbol{z}$; then parameter identifiability is defined as the mapping $\boldsymbol{w} \mapsto p(\boldsymbol{z}|\boldsymbol{w})$ being one-to-one.*

For many application scenarios, the PDF $p(\mathbf{z}|\mathbf{w})$ can be replaced by the characteristic function of $\mathbf{z}$ (the Fourier transformation of $p(\mathbf{z}|\mathbf{w})$) since

the correspondence between a PDF and its characteristic function is bijective (Shao, 1999). Hereafter, the probability measure $P_{\mathbf{w}}$ will be interchangeably used with its PDF $p(\mathbf{z}|\mathbf{w})$.

The relationship between the conditional PDF $p(\mathbf{y}|\mathbf{x}, \mathbf{w})$ and the function $\mathbf{f}(\mathbf{x}, \mathbf{w})$ is context dependent. As an example, consider the regression model $\mathbf{y} = \mathbf{f}(\mathbf{x}, \mathbf{w}) + \epsilon$. Here $\epsilon \sim \mathcal{N}(\epsilon|0, \mathbf{\Sigma})$; that is, ϵ is a gaussian PDF with mean 0 and covariance $\mathbf{\Sigma}$. We have

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}) = \mathcal{N}(\mathbf{y}|\mathbf{f}(\mathbf{x}, \mathbf{w}), \mathbf{\Sigma}). \tag{1.12}$$

To proceed, we introduce the following definition (Rothenberg, 1971):

Definition 2. *If two distinct parameters $\boldsymbol{w}_1$, $\boldsymbol{w}_2$ in $\mathcal{W}$ define the same probability measure, we say $\boldsymbol{w}_1$ is observationally equivalent to $\boldsymbol{w}_2$, and denote $\boldsymbol{w}_1 \sim \boldsymbol{w}_2$, that is,*

$$\boldsymbol{w}_1 \sim \boldsymbol{w}_2 \Leftrightarrow P_{\boldsymbol{w}_1} = P_{\boldsymbol{w}_2}. \tag{1.13}$$

For actual samples, the likelihood function takes a constant value on all observationally equivalent parameters. It is clear that the relation $\sim$ is a proper equivalent relation (reflectivity, symmetry and transitivity) (Paulino & Pereira, 1994; Dasgupta et al., 2007), so it breaks $\mathbb{R}^k$ into disjoint subsets called *equivalent classes*. For $\mathbf{w} \in \mathbb{R}^k$, we denote the corresponding equivalence class by

$$[\mathbf{w}] = \{\mathbf{w}' \in \mathbb{R}^k : \mathbf{w}' \sim \mathbf{w}\}. \tag{1.14}$$

Then the mapping $[\mathbf{w}] \to P_{\mathbf{w}}$ is bijective. However, the quotient space $\mathcal{W}/\sim = \{[\mathbf{w}] : \mathbf{w} \in \mathcal{W}\}$ is neither a Euclidean space nor a manifold (Amari et al., 2006; Watanabe, 2009). Therefore, it is generally difficult to construct statistical theory in $\mathcal{W}/\sim$.

Definition 3. *Let $\boldsymbol{w}_0$ be a fixed parameter in $\mathcal{W}$. If $[\boldsymbol{w}_0] = \{\boldsymbol{w}_0\}$ (a singleton), we say $\boldsymbol{w}_0$ is globally identifiable. If there exists a neighborhood $N(\boldsymbol{w}_0)$ of $\boldsymbol{w}_0$ such that $[\boldsymbol{w}_0] \bigcap N(\boldsymbol{w}_0) = \{\boldsymbol{w}_0\}$, we say $\boldsymbol{w}_0$ is locally identifiable.*

In definition 3, parameter identifiability is checked with respect to a specific point $\mathbf{w}_0$. In usual cases, this fixed $\mathbf{w}_0$ can be viewed as the real value (or critical value) of the model parameter.

Definition 4. *A statistical model $P_{\boldsymbol{w}}$ is globally identifiable if for any fixed $\boldsymbol{w}_0 \in \mathcal{W}$, $[\boldsymbol{w}_0] = \{\boldsymbol{w}_0\}$. $P_{\boldsymbol{w}}$ is locally identifiable if for any fixed $\boldsymbol{w}_0 \in \mathcal{W}$, there is an open neighborhood $N(\boldsymbol{w}_0)$ of $\boldsymbol{w}_0$ such that $[\boldsymbol{w}_0] \bigcap N(\boldsymbol{w}_0) = \{\boldsymbol{w}_0\}$.*

One readily obtains the following:

- Global identifiability implies local identifiability.
- A model is globally (locally) identifiable if and only if it is globally (locally) identifiable at every fixed parameter.

In the context of GC models, identifiability is a critical issue since one is interested in estimating physically meaningful parameters. More specifically, the subvector $\mathbf{w}^{(k)}$ in the KD submodel and the subvector $\mathbf{w}^{(d)}$ in the DD submodel are not equally important in the parameter vector $\mathbf{w}$. We are interested in the subvector $\mathbf{w}^{(k)}$ rather than $\mathbf{w}^{(d)}$ due to the consideration of physical meaning. Hence, we are led to the following definition:

Definition 5. *A subvector $\boldsymbol{w}^{(k)}$ is globally identifiable if for any fixed $\boldsymbol{w}_0^{(k)}$, $P_{\boldsymbol{w}} = P_{\boldsymbol{w}_0} \Rightarrow \boldsymbol{w}^{(k)} = \boldsymbol{w}_0^{(k)}, \forall \boldsymbol{w} \in \mathbb{R}^k$, where $\boldsymbol{w}_0^{(k)}$ and $\boldsymbol{w}^{(k)}$ are the KD subvectors of $\boldsymbol{w}_0$ and $\boldsymbol{w}$, respectively. A subvector $\boldsymbol{w}^{(k)}$ is locally identifiable if for any fixed $\boldsymbol{w}_0^{(k)}$, there is a neighborhood $N(\boldsymbol{w}_0^{(k)})$ such that for any $\boldsymbol{w}$ with $\boldsymbol{w}^{(k)} \in N(\boldsymbol{w}_0^{(k)})$, $P_{\boldsymbol{w}} = P_{\boldsymbol{w}_0} \Rightarrow \boldsymbol{w}^{(k)} = \boldsymbol{w}_0^{(k)}$, where $N(\boldsymbol{w}_0^{(k)})$ is the neighborhood with respect to $\boldsymbol{w}_0^{(k)}$.*

From definitions 3 to 5, we can see that identifiability is a theoretical or intrinsic property of the model and that the presence (or absence) of identifiability is a feature of the model specification, and so is independent of the experimental data or the numerical estimation procedures (Paulino & Pereira, 1994). Yet the choice of learning algorithms as well as the learning dynamic are strongly affected by the nonidentifiability, as we will see in section 3.

Typically, if a model has the following features, it may be unidentifiable:

- Hierarchical structures (Sussmann, 1992; Chen, Lu, & Hecht-Nielsen, 1993; Bishop, 2006)
- Latent variables (Shapiro, 1986; Bishop, 2006; Henao & Winther, 2011)
- Unobservable state variables (Ljung, 1999; Miao et al., 2011)
- Nuisance parameters (Fortunati et al., 2012)
- Coupled submodels (Yang, Hu, & Cournde, 2008; Hu et al., 2009)
- Grammatical rules (Watanabe, 2007, 2009)

Generally nonidentifiability occurs if the model can be equivalently rewritten in terms of a smaller set of parameters, which is termed *parameter redundancy* (PR; Catchpole & Morgan, 1997). The concept that is intimately related to nonidentifiability and redundancy is *parameter dependence* (PD) in the sense that a certain subvector of a parameter can be expressed as the function of the remaining one (Ran & Hu, 2015).

1.2.3 Importance of Identifiability in Machine Learning. The issue of parameter identifiability has a deep influence in almost all stages of the learning procedure. Besides being an important way for enhancing model

transparency and comprehensibility (Hu et al., 2007, 2009), identifiability is also a necessary prerequisite for system modeling and parameter estimation (Ljung, 1999). As a result, the identifiability problem should be addressed before any experimental data have been collected, because the difficulties associated with parameter estimation usually result from improper parameter structure rather than inappropriate experiment design or poor data. In other words, in an unidentifiable model, no matter how carefully we design the experiment or how good the observations are, one will definitely fail to get a reasonable estimation (Yang et al., 2008). In the community of machine learning, the identifiability issue has a close connection with a wide range of subjects, such as variational Bayesian matrix factorization (Nakajima & Sugiyama, 2010), low-rank matrix completion (Király & Tomioka, 2012), latent factor model (Shapiro, 1986; Henao & Winther, 2011), probabilistic PCA (Bishop, 2006), and Bayesian network (Whiley, 1999).

In theoretic neuroscience, an open problem is to determine when different ANNs having different synaptic weights implement identical input-output transformation (Dimattina & Zhang, 2010). Such networks are called *functionally equivalent*. Determining the exact conditions for structurally distinct yet functionally equivalent networks may shed light on the theoretical constraints on how diverse neural circuits might develop and be maintained to serve identical functions. Such consideration also imposes practical limits on our ability to uniquely infer the structure of underlying neural circuits from stimulus-response measurements (Dimattina & Zhang, 2010). Hence, if the function implemented by a neural network does not require a unique network structure, then when one synapse in a network is damaged, other synapses can be used to compensate for the damage and restore the original input-output function. Dimattina & Zhang (2010) introduced a biologically inspired mathematical method for determining when the structure of a neural network can be perturbed gradually while preserving functionality.

In a nutshell, the utility and importance of parameter identifiability for machine learning can be recognized in at least the following aspects:

- *Knowledge-based models*. In these models, all parameters have physically interpretable meaning. Identifiability analysis is the first step for estimating unknown parameters because a lack of identifiability implies that the interpretability of the learned model will be severely limited and the obtained parameter estimation is meaningless to understand the real system, which is a critical problem if decisions are to be taken on the basis of their numeric values. Especially in the situation of causal inference, one would not select a model if its parameters cannot be uniquely determined. For instance, Henao and Winther (2011) considered a sparse and identifiable linear latent factor and linear Bayesian network for parsimonious analysis of multivariate data, showing that parameter identifiability is a necessary prerequisite for capturing the correlations between the latent factors.

- *Partially knowledge-based models*. Within the context of nonlinear system identification, a common practice is to build a *black box* model in order to achieve accurate prediction or control. However, the full black box model may be too generic for some situations where there is evidence to include a knowledge-based submodel in the complete model. The practical rule of "do not estimate what you already know" would require us to define an ad hoc model structure if we know that the real system contains a prior known part (Espinoza, Suykens, & Moor, 2005). Thus, a certain subvector of the model has physically interpretable meaning. In the field of machine learning, knowledge- and data-driven models are typical examples, and identifiability is of special importance for the knowledge-based submodel. (For more details about this type of model, see Espinoza et al., 2005; Yang et al., 2008; Hu et al., 2009; Qu & Hu, 2011; Ran & Hu, 2014a; and Fan, Kang, Reffye, Heuvelink, & Hu, 2015.)
- *Singular learning theory*. The concept of local identifiability is closely related to singular learning theory (SLT; Watanabe, 2007, 2009). More specifically, if a model is not locally identifiable, its Fisher information matrix (FIM) degenerates (Rothenberg, 1971); then it is a singular learning machine (Watanabe, 2007, 2009). Due to the universal existence of singularities in machine learning, Watanabe (2007) pointed out that "almost all learning machines are singular." Therefore, it is imperative to check singularity for SLT. (For a systematic study on SLT, see Watanabe, 2009. We review this aspect in section 3.7.)
- *Statistical inference*. Parameter identifiability is a fundamental assumption in nearly all classical statistical models (Dasgupta et al., 2007) because it is a necessary prerequisite for statistical inference. In singular models, the likelihood function cannot be approximated by any quadratic form (Watanabe, 2009), resulting in conventional model selection criteria such as Akaike information criterioin (AIC; Akaike, 1974), Bayesian information criterion (BIC) (Schwarz, 1978), minimum description length (MDL; Rissanen, 1983), and hypothesis testing methods cannot be properly applied (Amari et al., 2006; Watanabe, 2007, 2009). In addition, singularity is closely related to the convergence of a range of estimators, including maximum likelihood estimator (MLE) and Bayesian estimator (Watanabe, 2009). For instance, in a singular model, the standard formulation of the Cramér-Rao bound does not hold (Amari et al., 2006) and the MLE and the Bayesian posterior distribution are no longer subject to gaussian distributions even asymptotically (Amari et al., 2006).
- *Learning algorithm and learning dynamics*. In singular models, the trajectories of dynamics of learning generated by a standard gradient descent algorithm are strongly affected by the singularities, causing plateaus or slow manifolds (Amari et al., 2006). To overcome

such slow convergence phenomenon, Amari (1998) proposed a natural gradient descent (NGD) algorithm, demonstrating that the NGD method works more efficiently in such singular models. For example, Amari et al. (2006), Cousseau, Ozeki, and Amari (2008), and Wei and Amari (2008) studied the dynamical behaviors of learning in gaussian mixture model (GMM), multilayer perceptron (MLP), and radial basis function (RBF), respectively, and showed that once parameters are attracted to singualr points, the learning trajectories are very slow to move away from them.

Considering that the identifiability issue has not yet been fully recognized, it is of great value to give a comprehensive review in machine learning. As noted before, in this review, we place a special focus on identifiability problems with statistical models. (The study of identifiability for dynamic ordinary differential equation models is beyond the scope of this review. Interested readers are referred to Walter, 1982; Walter & Pronzato, 1997; Audoly, D'Angio, Saccomani, & Cobelli, 1998; Xia & Moog, 2003; and Miao et al., 2011.)

1.3 Parameter Identifiability: Main Issues. This review is intended for readers with varying interests, including statistics, machine learning theory/algorithm, and system identification. We aim at providing an overview of the fundamental elements and recent advances on identifiability in machine learning. The main problems we address are:

- *Parameter identifiability analysis*:
  Criteria for examining parameter identifiability
  Criteria for examining parameter redundancy
  Methods for reparameterizing a parameter-redundant model so as to make it an identifiable one
- *Influence of identifiability in machine learning*:
  Statistical analysis, which mainly includes estimation theory, hypothesis testing, and model selection
  The information geometry framework for singular learning, which mainly involves learning algorithm and learning dynamics
  The algebraic geometry framework for singular learning, which mainly concerns Bayesian inference

1.4 Structure of This Review. The remainder of this review is organized as follows. In section 2, we provide a review of various representative criteria for examining identifiability and parameter redundancy, as well as some results regarding reparameterization. Section 3 reviews the deep influence of identifiability in various aspects of machine learning, and section 4 concludes with a brief summary and perspective.

2 Criteria for Examining Identifiability

This section reviews some well-established criteria for examining whether a model is locally (or globally) identifiable. We mainly consider two types of models: unconstrained and parameter-constrained models.

In unconstrained models, the admissible parameter space is the entire Euclidean space $\mathbb{R}^k$. In the field of machine learning, a vast majority of research has been done in the context of MLPs. It has long been known that the parameter space of hierarchical systems such as MLPs contains singularities due to symmetry and the degeneration of hidden units (Amari et al., 2006). Chen et al. (1993) proved that for a three-layer network with h hidden units having tanh activation functions and full connectivity in both layers, there will be an overall weight space symmetry factor of $h!2^h$. Sussmann (1992) proved that for feedforward networks with a single hidden layer, a single output, and tanh activation functions, the net is uniquely determined by its input-output, up to an obvious finite group of symmetries, provided that the net is irreducible.

If the parameter in the unconstrained model is unidentifiable, one can change the modeling approach to make it identifiable. Typically, there are two approaches to achieve this purpose. The first is to introduce a priori distribution on the unknown parameter and cast the estimation problem into a Bayesian framework (Berger, 1985; Bishop, 2006; Murphy, 2012). The second is to impose some deterministic constraints, such as functional constraints (Rothenberg, 1971; Stoica & Ng, 1998), sparsity constraints (Henao & Winther, 2011), or monotonicity constraints (Qu & Hu, 2011) on the unknown parameter, resulting in a parameter estimation problem with smaller parameter space (Rothenberg, 1971; Stoica & Ng, 1998; Moore, 2010). In these parameter-constrained models, we assume the admissible parameter space

$$\mathcal{W}_c = \{\mathbf{w} \in \mathbb{R}^k : \Lambda(\mathbf{w}) = (\lambda_1(\mathbf{w}), \ldots, \lambda_c(\mathbf{w})) = 0\}, \tag{2.1}$$

is constrained by a set of c equation constraints.

In the following, we assume that $\mathcal{W}_c \neq \emptyset$ which means the constraints are consistent. We also assume that the rank of the Jacobian matrix $\mathbf{J}_\Lambda(\mathbf{w})$ of $\Lambda(\mathbf{w})$ is c for all $\mathbf{w}$. This implies the constraints are nonredundant; otherwise, certain constraints are redundant and can be removed. Incorporating these additional constraints into an original unconstrained model results in an alteration of the PDF's dependence on the unknown parameter.

The identifiability problem of linear models and mixture models is well studied, and there are a number of methods to perform such a task (Tallis & Chesson, 1982; Paulino & Pereira, 1994). However, there are only a few methods for testing the identifiability of nonlinear models. In addition, most previous work on identifiability problems focused mainly on the special

features of particular model structures such as a linear model (Paulino & Pereira, 1994; Shao, 1999), gaussian model (Hochwald & Nehorai, 1997), or exponential family (Catchpole & Morgan, 1997). This tends to obscure the fact that the identifiability problem is a general one arising in machine learning. Despite extensive literature and a number of criteria that exist for various specific models, the identifiability issue has not yet been resolved completely.

In this section, we give a concise review of several representative approaches for dealing with identifiability problems and analyze the requirements, advantages, and disadvantages of those approaches.

2.1 An FIM Approach for Examining Identifiability. The FIM is a well-established criterion in identifiability analysis. Before formally presenting the result, we introduce the following concept:

Definition 6 *(constant-rank matrix). Let* $\boldsymbol{M}(\boldsymbol{w}) = (M_{ij}(\boldsymbol{w}))$ *be a matrix whose elements* $M_{ij}(\boldsymbol{w})$ *are functions of* $\boldsymbol{w}$*. If* $\boldsymbol{M}(\boldsymbol{w})$ *has the same rank for all* $\boldsymbol{w} \in U$*, where* $U \subseteq \mathbb{R}^k$*, we call* $\boldsymbol{M}(\boldsymbol{w})$ *a constant-rank matrix in* U*.*

Rothenberg (1971) presented the following theorem:

Theorem 1. *Under constant-rank condition, a PDF* $p(\boldsymbol{z}|\boldsymbol{w})$ *is locally identifiable at* $\boldsymbol{w}_0$ *if and only if the* $k \times k$ *matrix FIM,*

$$FIM(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{z}}\left[\left(\frac{\partial \log p(\boldsymbol{z}|\boldsymbol{w})}{\partial \boldsymbol{w}}\right)\left(\frac{\partial \log p(\boldsymbol{z}|\boldsymbol{w})}{\partial \boldsymbol{w}^{\mathrm{T}}}\right)\right], \tag{2.2}$$

is positive definite at $\boldsymbol{w}_0$*, where* $\mathbb{E}_{\boldsymbol{z}}$ *denotes the expectation operation with respect to* $\boldsymbol{z}$*.*

The important implication of this theorem is that it establishes the connection between local identifiability and singularity (Watanabe, 2007, 2009).

Definition 7. *A statistical model* $p(\boldsymbol{z}|\boldsymbol{w})$ *is called regular if its FIM* $FIM(\boldsymbol{w})$ *is positive definite for all* $\boldsymbol{w}$*. Otherwise, it is called singular.*

Therefore, it is evident that if a model is not locally identifiable, it is a singular learning machine. Rothenberg (1971) proved that this criterion is a global one in the case of exponential family. Hochwald and Nehorai (1997) studied the connection between identifiability and regularity of the FIM for gaussian PDF and established a tool to check regularity with the help of holomorphic functions.

From classical statistics, the FIM is positive definite if and only if

$$\frac{\partial \log p(\mathbf{z}|\mathbf{w})}{\partial w_1}, \ldots, \frac{\partial \log p(\mathbf{z}|\mathbf{w})}{\partial w_k} \tag{2.3}$$

are linearly independent as functions of $\mathbf{z}$ on the support of $p(\mathbf{z}|\mathbf{w})$ (Watanabe, 2009). Applying this result to the Multiple-input multiple-output (MIMO) nonlinear regression model,

$$\mathbf{y} = \mathbf{f}(\mathbf{x}, \mathbf{w}) + \epsilon, \tag{2.4}$$

where $\mathbf{f}(\mathbf{x}, \mathbf{w}) = (f_1(\mathbf{x}, \mathbf{w}), \ldots, f_m(\mathbf{x}, \mathbf{w}))$ is a vector-valued mapping, $\epsilon \sim \mathcal{N}(\epsilon|0, \mathbf{\Sigma})$, one obtains the following result (Ran & Hu, 2014a):

Theorem 2. *The MIMO nonlinear regression model, equation 2.4, is locally identifiable if and only if* $\frac{\partial f(x,w)}{\partial w_i}$, $i = 1, \ldots, k$, *are linearly independent as functions of* $\boldsymbol{x}$, *where*

$$\frac{\partial \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{w})}{\partial w_i} = \left(\frac{\partial f_1(\boldsymbol{x}, \boldsymbol{w})}{\partial w_i}, \ldots, \frac{\partial f_m(\boldsymbol{x}, \boldsymbol{w})}{\partial w_i} \right). \tag{2.5}$$

As for parameter-constrained models, Stoica and Ng (1998) presented an identifiability criterion that needs to compute the orthogonal complement of the Jacobian $\mathbf{J}_\Lambda(\mathbf{w})$ of $\Lambda(\mathbf{w})$. More specifically, since $\mathbf{J}_\Lambda(\mathbf{w})$ has rank c for all $\mathbf{w}$, there exists a $k \times (k - c)$ matrix $\mathbf{U}(\mathbf{w})$ such that

$$\mathbf{J}_\Lambda(\mathbf{w})\mathbf{U}(\mathbf{w}) = 0 \quad \text{and} \quad \mathbf{U}^{\mathrm{T}}(\mathbf{w})\mathbf{U}(\mathbf{w}) = \mathbf{I}_{k-c}; \tag{2.6}$$

then $\mathbf{w}$ is locally identifiable if and only if

$$|\mathbf{U}^{\mathrm{T}}(\mathbf{w})\mathrm{FIM}(\mathbf{w})\mathbf{U}(\mathbf{w})| \neq 0. \tag{2.7}$$

The difficulty with this method is that one needs to compute the orthogonal complement of a functional matrix $\mathbf{J}_\Lambda(\mathbf{w})$, making it a hard task to perform in nonlinear cases.

2.2 A Kullback-Leibler Divergence Approach for Examining Identifiability. Essentially, nonidentifiability is the consequence of the lack of enough "information" to discriminate among admissible parameter values. Hence, it is natural to examine identifiability with the help of Kullback-Leibler divergence (KLD), defined as (Cover & Thomas, 1991)

$$\mathrm{KL}(p, q) = \mathbb{E}_p \left(\log \frac{p(\mathbf{z})}{q(\mathbf{z})} \right) = \int p(\mathbf{z}) \log \frac{p(\mathbf{z})}{q(\mathbf{z})} \mathrm{d}\mathbf{z}, \tag{2.8}$$

where $p(\mathbf{z})$ and $q(\mathbf{z})$ are two PDFs. Historically, the KLD was first defined by Boltzmann and Gibbs in statistical physics in the 19th century and called

the *relative entropy* in physics literature (Watanabe, 2009). In information theory, the KLD is used to measure the dissimilarity between two PDFs $p(\mathbf{z})$ and $q(\mathbf{z})$ (Cover & Thomas, 1991). In classical statistics, the KLD arises as an expected logarithm of the likelihood ratio and is a measure of the inefficiency of assuming that the distribution is $p(\mathbf{z})$ while the true distribution is $q(\mathbf{z})$ (Shao, 1999). The KLD plays a central role in the theoretical development of identifiability study.

Bowden (1973) presented the following theorem for testing identifiability:

Theorem 3. *In a statistical model $p(\boldsymbol{z}|\boldsymbol{w})$, $\boldsymbol{w}_0 \in \mathbb{R}^k$ is globally (locally) identifiable if and only if $\boldsymbol{w}_0$ is the unique solution of the equation $KL(\boldsymbol{w}_0, \boldsymbol{w}) = 0$ in $\mathbb{R}^k$ (a neighborhood of $\boldsymbol{w}_0$), where*

$$KL(\boldsymbol{w}_0, \boldsymbol{w}) = \int p(\boldsymbol{z}|\boldsymbol{w}_0) \log \frac{p(\boldsymbol{z}|\boldsymbol{w}_0)}{p(\boldsymbol{z}|\boldsymbol{w})} d\boldsymbol{z}. \tag{2.9}$$

The proof can be easily seen from the two properties: (1) $\mathrm{KL}(\mathbf{w}_0, \mathbf{w}) \geq 0$ for all $\mathbf{w} \in \mathbb{R}^k$ and (2) $\mathrm{KL}(\mathbf{w}_0, \mathbf{w}) = 0$ if and only if $P_{\mathbf{w}_0} = P_{\mathbf{w}}$ (Cover & Thomas, 1991).

Generally the FIM method can deal only with a local identifiability problem. This is because $\mathrm{FIM}(\mathbf{w})$ is a function of a single argument, and the positive definiteness of $\mathrm{FIM}(\mathbf{w}_0)$ guarantees local identifiability only with respect to $\mathbf{w}_0$. However, the KLD is a function of two arguments, which makes it able to deal simultaneously with global and local identifiability problems (see theorem 3). For simplicity, the KLD equation approach in Bowden (1973) is abbreviated as KLDE.

Theorem 3 can be easily extended to deal with parameter-constrained models (Ran & Hu, 2014b).

Theorem 4. *In a statistical model $p(\boldsymbol{z}|\boldsymbol{w})$ with constrained parameter space, equation 2.1, $\boldsymbol{w}_0 \in \mathcal{W}_c$ is globally (locally) identifiable if and only if $\boldsymbol{w}_0$ is the unique solution of the following system of equations in $\mathcal{W}_c$ (a neighborhood $N(\boldsymbol{w}_0) \cap \mathcal{W}_c$ of $\boldsymbol{w}_0$):*

$$\begin{cases} KL(\boldsymbol{w}_0, \boldsymbol{w}) = 0 \\ \Lambda(\boldsymbol{w}) = 0 \end{cases}. \tag{2.10}$$

However, this approach requires solving a set of $c + 1$ nonlinear equations whose close-form solution is exceedingly difficult to obtain, resulting in the use of numerical methods and the loss of the formal nature of the solution.

2.3 An Optimization Theory Approach for Examining Identifiability. To overcome the difficulties of the KLDE method, by making use of the KLD, Ran and Hu (2014b) cast the identifiability problem into the optimization theory framework, and derived the corresponding identifiability criteria. To proceed, we recall some basic concepts from optimization theory (Sundaram, 1996).

Definition 8. *A point $\boldsymbol{w}_0 \in \mathbb{R}^k$ is said to be a local minimum point of $f(\boldsymbol{w})$ if there is a neighborhood $N(\boldsymbol{w}_0)$ of $\boldsymbol{w}_0$ such that $f(\boldsymbol{w}) \geq f(\boldsymbol{w}_0)$ for all $\boldsymbol{w} \in N(\boldsymbol{w}_0)$. If $f(\boldsymbol{w}) > f(\boldsymbol{w}_0)$ for all $\boldsymbol{w} \in N(\boldsymbol{w}_0)$, $\boldsymbol{w} \neq \boldsymbol{w}_0$, then $\boldsymbol{w}_0$ is said to be a strict local minimum point.*

Definition 9. *A point $\boldsymbol{w}_0 \in \mathbb{R}^k$ is said to be a global minimum point of $f(\boldsymbol{w})$ if $f(\boldsymbol{w}) \geq f(\boldsymbol{w}_0)$ for all $\boldsymbol{w} \in \mathbb{R}^k$. If $f(\boldsymbol{w}) > f(\boldsymbol{w}_0)$ for all $\boldsymbol{w} \in \mathbb{R}^k$, $\boldsymbol{w} \neq \boldsymbol{w}_0$, then $\boldsymbol{w}_0$ is said to be a strict global minimum point.*

In the language of optimization theory, theorem 3 can be equivalently rewritten as follows (Ran & Hu, 2014b):

Theorem 5. *In a statistical model $p(\boldsymbol{z}|\boldsymbol{w})$, a parameter point $\boldsymbol{w}_0 \in \mathbb{R}^k$ is globally (locally) identifiable if and only if $\boldsymbol{w}_0$ is the strict global (local) minimum point of the following unconstrained optimization problem:*

$$\min KL(\boldsymbol{w}_0, \boldsymbol{w}). \tag{2.11}$$

In fact, the statement of theorem 5 can be regarded as a dual interpretation of that in theorem 3. Specifically, theorem 3 formulates the identifiability problem as the task of seeking the roots of a nonlinear equation $\mathrm{KL}(\mathbf{w}_0, \mathbf{w}) = 0$, while theorem 5 formulates the identifiability problem as an unconstrained optimization problem. The following theorem (Ran & Hu, 2014b) is derived from the optimization theory:

Theorem 6. *Suppose that $p(\boldsymbol{z}|\boldsymbol{w})$ is a statistical model, $\boldsymbol{w}_0 \in \mathbb{R}^k$, and that the Hessian matrix,*

$$\boldsymbol{H}(\boldsymbol{w}) = \frac{\partial^2 KL(\boldsymbol{w}_0, \boldsymbol{w})}{\partial \boldsymbol{w}^2}, \tag{2.12}$$

of $KL(\boldsymbol{w}_0, \boldsymbol{w})$ has constant rank in a neighborhood $N(\boldsymbol{w}_0)$ of $\boldsymbol{w}_0$; then $\boldsymbol{w}_0$ is locally identifiable if and only if $\boldsymbol{H}(\boldsymbol{w}_0)$ is positive definite.

The necessity can be readily derived as follows. The Taylor expansion of $\mathrm{KL}(\mathbf{w}_0, \mathbf{w})$ with respect to $\mathbf{w}_0$ is (Ran & Hu, 2014b)

$$\mathrm{KL}(\mathbf{w}_0, \mathbf{w}) = \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^{\mathrm{T}} \mathbf{H}(\mathbf{w}_0)(\mathbf{w} - \mathbf{w}_0) + o(||\mathbf{w} - \mathbf{w}_0||_2^2). \tag{2.13}$$

Since $\mathbf{H}(\mathbf{w}_0)$ is positive definite, there exists a neighborhood $N(\mathbf{w}_0)$ of $\mathbf{w}_0$ such that $\mathrm{KL}(\mathbf{w}_0, \mathbf{w}) > 0$ for all $\mathbf{w} \in N(\mathbf{w}_0)$, $\mathbf{w} \neq \mathbf{w}_0$. This implies that $\mathbf{w}_0$ is a strict local minimum point of $\mathrm{KL}(\mathbf{w}_0, \mathbf{w})$. The sufficiency can be seen in Ran and Hu (2014b).

Based on theorem 5, Ran and Hu (2014b) proposed a global result as follows. Compared with the global result in Rothenberg (1971), this result is valid for any statistical model without restricting it to exponential family.

Theorem 7. *Suppose that the Hessian matrix $\boldsymbol{H}(\boldsymbol{w})$ of $KL(\boldsymbol{w}_0, \boldsymbol{w})$ is positive definite for all $\boldsymbol{w} \in \mathbb{R}^k$, $\boldsymbol{w} \neq \boldsymbol{w}_0$; then $\boldsymbol{w}_0$ is globally identifiable.*

Following the same line as the unconstrained cases, the identifiability problem of parameter-constrained models can be equivalently formulated as a constrained optimization problem (Ran & Hu, 2014b).

Theorem 8. *In a statistical model $p(\boldsymbol{z}|\boldsymbol{w})$ with constrained parameter space, equation 2.1, a parameter $\boldsymbol{w}_0 \in \mathcal{W}_c$ is globally (locally) identifiable if and only if $\boldsymbol{w}_0$ is the strict global (local) minimum of the constrained optimization problem:*

$$\begin{aligned} &\min KL(\boldsymbol{w}_0, \boldsymbol{w}) \\ &s.t. \quad \boldsymbol{w} \in \mathcal{W}_c \end{aligned} \tag{2.14}$$

This gives rise to the following theorem (Ran & Hu, 2014b):

Theorem 9. *Suppose the parameter space of the statistical model $p(\boldsymbol{z}|\boldsymbol{w})$ is restricted to $\mathcal{W}_c$ (see equation 2.1), $\boldsymbol{w}_0 \in \mathcal{W}_c$, $\boldsymbol{A}(\boldsymbol{w})$ is a block matrix of the form*

$$A(\boldsymbol{w}) = \begin{pmatrix} \boldsymbol{H}(\boldsymbol{w}) \\ \boldsymbol{J}_{\Lambda}(\boldsymbol{w}) \end{pmatrix}, \tag{2.15}$$

where $\boldsymbol{H}(\boldsymbol{w})$ is the Hessian of $KL(\boldsymbol{w}_0, \boldsymbol{w})$, and $\boldsymbol{J}_{\Lambda}(\boldsymbol{w})$ is the Jacobian of $\Lambda(\boldsymbol{w})$. If $\boldsymbol{A}(\boldsymbol{w})$ has constant rank in a neighborhood $N(\boldsymbol{w}_0)$ of $\boldsymbol{w}_0$, then the following three conditions are equivalent:

1. *$\boldsymbol{w}_0 \in \mathcal{W}_c$ is not locally identifiable.*
2. *$\boldsymbol{A}(\boldsymbol{w}_0)$ is column rank deficient.*
3. *$\boldsymbol{H}(\boldsymbol{w}_0) + \boldsymbol{J}_{\Lambda}^{\mathrm{T}}(\boldsymbol{w}_0)\boldsymbol{J}_{\Lambda}(\boldsymbol{w}_0)$ is rank deficient.*

The benefit gained from the optimization theory framework is that when information theory through KLD is the link, the interplay between identifiability theory and optimization theory is derived.

In some practical applications, it is of interest to study the problem of how many constraints are needed to guarantee identifiability. The following

result (Ran & Hu, 2014b) can be applied as a guideline to quantitative experiment design.

Theorem 10. *Suppose rank$\boldsymbol{H}(\boldsymbol{w}_0) = r, r < k$; the minimum number of constraints needed to achieve local identifiability is $k - r$.*

A direct result from the convex optimization theory (Boyd & Vandenberghe, 2004) is that if the objective function $\mathrm{KL}(\mathbf{w}_0, \mathbf{w})$ is strictly convex in $\mathbf{w}$ and the constraint $\Lambda(\mathbf{w})$ is convex, then the local identifiability criterion in theorem 9 becomes a global one. However, the $\mathrm{KL}(\mathbf{w}_0, \mathbf{w})$ is not generally convex in $\mathbf{w}$ although $\mathrm{KL}(p, q)$ is convex in the second argument q (Principe, 2010), as q is nonlinear in $\mathbf{w}$. Thus, one cannot cast the identifiability problem into the convex optimization theory framework, making it difficult to derive a global criterion. Therefore, the global identifiability problem remains a challenging subject in identifiability theory.

2.4 An Identifying Function Approach for Examining Identifiability. This part first recalls the definitions and notations required for the necessary mathematical background:

Definition 10 *(diffeomorphism). Let U and U' be two open sets in $\mathbb{R}^k$. A bijective mapping $\boldsymbol{f} : U \mapsto U'$ in $C^1(U, U')$ is a diffeomorphism from U to U' if $\boldsymbol{f} \in C^1(U, U')$ and $\boldsymbol{f}^{-1} \in C^1(U', U)$, where $C^1(U, U')$ is the function space expanded by all continuously differentiable mappings from U to U', and $\boldsymbol{f}^{-1}$ is the inverse mapping of $\boldsymbol{f}$.*

Definition 11 *(C^1 equivalence). $\varphi \in C^1(U, V)$ and $\psi \in C^1(U', V')$ are two diffeomorphisms, where $U, U' \subset \mathbb{R}^k$, $V, V' \subset \mathbb{R}^q$. If $\varphi = \boldsymbol{g}^{-1} \circ \psi \circ \boldsymbol{f}$, then φ and ψ are said to be C^1 equivalent, where $\circ$ denotes the function composition.*

The following rank theorem in Riemann geometry (Gallot, Hulin, & Lafontaine, 2008) describes nonlinear constant-rank mappings. It essentially says that locally, they behave just like linear projection.

Theorem 11. *Locally, a constant-rank mapping $\varphi(\boldsymbol{w}) \in C^1(\mathbb{R}^k, \mathbb{R}^q)$ is C^1 equivalent to a linear projection*

$$\boldsymbol{P}_r(w_1, \ldots, w_k) = (w_1, \ldots, w_r, 0, \ldots, 0), \tag{2.16}$$

where r is the rank of the Jacobian matrix $\boldsymbol{J}_\varphi(\boldsymbol{w})$ of $\varphi(\boldsymbol{w})$.

To give an efficient method for testing the identifiability of models, we introduce the concept of identifying function (Ran & Hu, 2015), which can be viewed as a reduced-form parameter of the original model:

Definition 12. *Suppose $P_{\boldsymbol{w}}$ is a statistical model; a vector-valued function $\varphi(\boldsymbol{w})$ in $C^1(\mathbb{R}^k, \mathbb{R}^q)$ is an identifying function (IF) if*

$$\boldsymbol{w}_1 \sim \boldsymbol{w}_2 \Leftrightarrow \varphi(\boldsymbol{w}_1) = \varphi(\boldsymbol{w}_2), \forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^k. \tag{2.17}$$

It is easy to see that the value of an IF φ is unchanged within each equivalent class $[\mathbf{w}]$, while it takes different values among different equivalent classes. That is, only the parameters that are not observationally equivalent can be identified by the values of IF. This is the origin of the terminology *identifying function* (Paulino & Pereira, 1994).

Combining equation 1.13 with 2.17, we have

$$\varphi(\mathbf{w}_1) = \varphi(\mathbf{w}_2) \Leftrightarrow P_{\mathbf{w}_1} = P_{\mathbf{w}_2}. \tag{2.18}$$

This implies that the parameter structure of $P_{\mathbf{w}}$ is completely determined by its IF, so the identifiability problem of the original model is transformed into the problem of checking the injectivity of mapping $\varphi(\mathbf{w})$.

Based on the rank theorem, (Ran and Hu, 2015) derive the following theorem:

Theorem 12. *Suppose $P_{\boldsymbol{w}}$ is a statistical model and $\varphi(\boldsymbol{w})$ is an IF. If the Jacobian $\boldsymbol{J}_{\varphi}(\boldsymbol{w})$ of $\varphi(\boldsymbol{w})$ is of column full rank at $\boldsymbol{w}_0$, the model is locally identifiable at $\boldsymbol{w}_0$.*

It should be noted that the converse of theorem 12 is not true. Consider the simple model $y = w^2 + \epsilon$, where the noise $\epsilon \sim \mathcal{N}(\epsilon|0, 1)$. It is obvious that the IF $\varphi(w) = w^2$ since the model is fully determined by the mean w^2. The Jacobian $\mathbf{J}_{\varphi}(w) = 2w$ vanishes at $w = 0$, yet $w = 0$ is locally identifiable.

With the IF approach, the following theorem can be used to test global identifiability (Ran & Hu, 2015):

Theorem 13. *Suppose $P_{\boldsymbol{w}}$ is a statistical model and $\varphi(\boldsymbol{w})$ is an IF. The model is globally identifiable at $\boldsymbol{w}_0 \in \mathbb{R}^k$ if and only if $\boldsymbol{w}_0$ is the unique solution of the equation $\varphi(\boldsymbol{w}) = \varphi(\boldsymbol{w}_0)$ for $\boldsymbol{w} \in \mathbb{R}^k$.*

We use univariate PDF $p(z|\mathbf{w})$ for illustrating the procedure of obtaining IF, since the extension to the case of multivariate is rather straightforward (Shao, 1999). Recall that the correspondence between a PDF $p(z|\mathbf{w})$ and its associated characteristic function,

$$\psi(t|\mathbf{w}) = \mathbb{E}_z(e^{\mathrm{i}tz}) = \int e^{\mathrm{i}tz} p(z|\mathbf{w}) \mathrm{d}z, \tag{2.19}$$

is bijective; where $\mathrm{i} = \sqrt{-1}$, the IF can be formed from the collection of Taylor expansion coefficients of $\psi(t|\mathbf{w})$ in terms of t,

$$\psi(t|\mathbf{w}) = \sum_{k=0}^{\infty} \psi^{(k)}(0|\mathbf{w}) \frac{(\mathrm{i}t)^k}{k!} = \sum_{k=0}^{\infty} \mathbb{E}_z(z^k) \frac{(\mathrm{i}t)^k}{k!}, \tag{2.20}$$

where $\psi^{(k)}(0|\mathbf{w})$ is the kth derivative of $\psi(t|\mathbf{w})$ evaluated at $t = 0$. This is because $\psi(t|\mathbf{w})$ is an analytical function of t, so $\psi(t|\mathbf{w})$ (or $p(z|\mathbf{w})$) is uniquely determined by the collection of the Taylor expansion coefficients. In this case, the kth component of $\varphi(\mathbf{w})$ is in fact the kth moment of the PDF $p(z|\mathbf{w})$.

Note that the dimension q of the IF is assumed to be finite. This assumption does hold for certain parametric models—for example, a gaussian model (Ran & Hu, 2014a), exponential family (Catchpole & Morgan, 1997), moving average model (Ran & Hu, 2015), gaussian process (Ran & Hu, 2014a), autoregressive model (Ran & Hu, 2014a), linear or rational dynamic model (Walter & Pronzato, 1997)—in real-world scenarios. For instance, the IF for an m-dimensional gaussian PDF,

$$p(\mathbf{z}|\mathbf{w}) = \frac{1}{(2\pi)^{\frac{m}{2}}|\mathbf{\Sigma}_{\mathbf{w}}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{z}-\mu_{\mathbf{w}})^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{w}}^{-1}(\mathbf{z}-\mu_{\mathbf{w}})\right), \tag{2.21}$$

can be formed from the m elements in the mean $\mu_{\mathbf{w}}$ plus the $\frac{m(m+1)}{2}$ elements in the (symmetric) covariance $\mathbf{\Sigma}_{\mathbf{w}}$. However, this is not always the case for complicated models. For generic nonlinear models, if one takes the components of an IF to be the Taylor expansion coefficients of the characteristic function, the dimension of the IF may be infinite, thus imposing the problem of determining the upper bound of the dimension of the IF for a guaranteed identifiability result (Magaria, Riccomagno, Chappell, & Wynn, 2001). This challenging problem has not yet been fully resolved.

2.5 A Statistical Approach for Examining Identifiability. Most previous studies have been concerned mainly with local identifiability. Few investigations have been reported on how to examine global identifiability. However, in some cases, such as knowledge-based models, we are more interested in global identifiability than simple local identifiability. Unfortunately, it is very difficult to obtain global results in generic statistical settings.

In this section, we review the statistical method that explores the interface between identifiability and various statistics. Recall that a statistic is a measurable function of samples (Shao, 1999). The major merit of the statistical approach is that the derived results are capable of examining global identifiability.

Definition 13. *The $\widehat{\boldsymbol{w}}_n$ is a consistent estimator for $\boldsymbol{w}$, written as $\widehat{\boldsymbol{w}}_n \overset{a.s.}{\to} \boldsymbol{w}$, if $\widehat{\boldsymbol{w}}_n$ converges to $\boldsymbol{w}$ almost surely (a.s.) for all $\boldsymbol{w} \in \mathcal{W}$.*

A basic argument between identifiability and the existence of a consistent estimator is given in Lehmann (1983) as follows:

Theorem 14. *The global identifiability of the parameter is a necessary condition for the existence of a consistent estimator.*

The relationship of identifiability and sufficient statistics is given by Picci (1977). Paulino and Pereira (1994) proved that global identifiability is a necessary condition for the existence of an unbiased estimator. This result can be extended to asymptotically unbiased estimators (Ernesto & Fernando, 2002). To date, there is no general theoretical result about how to construct a consistent (or unbiased) estimator for arbitrary statistical distribution. However, the following case appears often in practice. Suppose that the parameter w_i can be interpreted as characteristic of the PDF. For instance, the w_i might be moment of the PDF; then distinct values of $\mathbf{w}$ imply distinct distributions, resulting in global identifiability. This is given in Rothenberg (1971):

Theorem 15. *Suppose there exist k functions* $\tau_1(z), \ldots, \tau_k(z)$*, such that for all* $\boldsymbol{w} \in \mathbb{R}^k$*,*

$$w_i = \mathbb{E}_z \tau_i(z). \tag{2.22}$$

Then $\boldsymbol{w}$ *is globally identifiable.*

2.6 A Critical Comparison of Various Identifiability Criteria. In Table 2, we summarize the requirements, advantages and disadvantages of the FIM (Rothenberg, 1971), KLDE (Paulino & Pereira, 1994), IF (Ran & Hu, 2015), and statistical method (Rothenberg, 1971).

2.7 Parameter Redundancy. One of the most relevant concepts related to identifiability is parameter redundancy (PR). In parameter-redundant models, the intrinsic parameter dimension is strictly less than the number of model parameters. Another related concept is parameter dependence (PD), in the sense that a certain subvector of parameter can be expressed as a function of the remaining one.

Definition 14. *A statistical model* $P_{\boldsymbol{w}}$ *is parameter redundant if it can be expressed in terms of a smaller parameter vector* $\beta = \beta(\boldsymbol{w})$*, where* $dim\beta < dim\boldsymbol{w}$*. Otherwise, the model is of full rank.*

The following concept of PD is an extension of that in Yang et al. (2008) and Hu et al. (2009), where PD is defined in a pairwise manner.

Definition 15. *A statistical model* $P_{\boldsymbol{w}}$ *is parameter dependent if a certain subvector* $\boldsymbol{w}^{(1)}$ *of* $\boldsymbol{w}$ *can be expressed as the function of the remaining subvector* $\boldsymbol{w}^{(2)}$*, where* $\boldsymbol{w}^{(1)} \cup \boldsymbol{w}^{(2)} = \boldsymbol{w}$*,* $\boldsymbol{w}^{(1)} \cap \boldsymbol{w}^{(2)} = \emptyset$*.*

Table 2: A Comparison of Various Criteria for Examining Identifiability.

FIM	Requirements	Constant-rank condition
		Calculate the explicit formulation of FIM
	Advantages	Theoretically workable for general statistical models
	Disadvantages	Computationally complex
		Can deal only with local identifiability problems
KLDE	Requirements	Calculate the explicit formulation of KLD
		Seek all the roots from the KLD equation
	Advantages	Theoretically workable for general statistical models
		Can simultaneously deal with global and local identifiability problems
	Disadvantages	Computationally complex
IF	Requirements	Dimension of IF is finite
	Advantages	Can simultaneously deal with global and local identifiability problems
		Computationally simple for testing local identifiability
	Disadvantages	Computationally complex for testing global identifiability
Statistic	Requirements	Construct suitable statistics
	Advantages	Can deal with global identifiability problems
	Disadvantages	There are no unified methods for constructing statistics

The relationship of PR and PD is stated in Ran and Hu (2015).

Theorem 16. *If a model is parameter dependent, the model is parameter redundant. The converse is not true.*

Catchpole and Morgan (1997) introduced the concept of PR in a special exponential family and proved that PR and nonidentifiability are equivalent in that case. More recently, Hu (2014) analyzed the redundancy of Bayesian classifiers whose parameters are given in the form of functionals and proved that for an M-class classification problem with reject option, the number of independent parameters in a cost matrix is M.

For linear models, the problem of processing PR and reparameterization is well understood (see example 1 in Ran & Hu, 2015). Nevertheless, such a task cannot be easily tackled in nonlinear models. This section reviews the representative methods for processing PR and reparameterization in nonlinear settings.

According to Jacquez and Greif (1985), there are two levels of parameters: structural and observational. The structural parameter $\mathbf{w}$ is a basic one associated with a model, and so explicitly appears in a model equation. In contrast, the observational parameter is the function of $\mathbf{w}$ that is inherent in the model, and so implicitly appears in original model. The reparameterized $\beta(\mathbf{w})$ in definition 14 is in spirit similar to an observational parameter, so it is at a level different from $\mathbf{w}$.

By definition 14, the choice of $\beta(\mathbf{w})$ does not need to be unique. Obviously, if a statistical model is of full rank, the only feasible reparameterization (up to a bijective mapping) is $\beta(\mathbf{w}) = \mathbf{w}$. Based on the IF approach, Ran and Hu (2015) presented the following theorem:

Theorem 17. *Suppose $P_{\boldsymbol{w}}$ is a statistical model and $\varphi(\boldsymbol{w})$ is an IF. Then $P_{\boldsymbol{w}}$ is parameter redundant if and only if the Jacobian $\boldsymbol{J}_{\varphi}(\boldsymbol{w})$ of $\varphi(\boldsymbol{w})$ is symbolically column deficient, that is, there is a nontrivial vector $\boldsymbol{v}(\boldsymbol{w})$ such that*

$$\boldsymbol{J}_{\varphi}(\boldsymbol{w})\boldsymbol{v}(\boldsymbol{w}) = 0, \forall \boldsymbol{w} \in \mathbb{R}^{k}. \tag{2.23}$$

This result can be viewed in two complementary ways. In an algebraic viewpoint, a model is parameter redundant if and only if the columns of the Jacobian $\mathbf{J}_{\varphi}(\mathbf{w})$ are linearly dependent. From a geometrical viewpoint, by equation 2.23, we can see that

$$\nabla\varphi_{l}^{\mathrm{T}}(\mathbf{w})\mathbf{v}(\mathbf{w}) = 0, l = 1, \ldots, q. \tag{2.24}$$

This implies that each $\nabla\varphi_{l}(\mathbf{w})$ is orthogonal to the vector field $\mathbf{v}(\mathbf{w})$. Hence, $\varphi_{l}(\mathbf{w})$ has ridges along $\mathbf{v}(\mathbf{w})$. In other words, all the parameters giving rise to the vector field $\mathbf{v}(\mathbf{w})$ are completely contained in the same equivalent class, so that the value of IF is unchanged. This highlights the fact that the IF is capable of distinguishing between different equivalent classes, yet it fails to distinguish parameters within the same equivalent class.

We now briefly review the problem of software implementation for checking for PR. The problem of detecting PR boils down to that of examining the symbolic rank of the Jacobian of IF. In principle, this test can be tackled by symbolic computation softwares such as Maple, Mathematica (Bekker, Merchens, & Wansbeek, 1994). (For specific examples, see Catchpole, Morgan, & Viallefont, 2002; Gimenez, Viallefont, Catchpole, Choquet, & Morgan, 2004; Yang et al., 2008; Hu et al., 2009; and Cole et al., 2010.) However, calculating the symbolic rank of a functional matrix is still an extremely challenging task since its complexity increases very quickly with the number of parameters, the dimension of input and output vectors, and the nonlinear degree of models (e.g., nonlinear dependency on parameters, input variables). Therefore, it calls for more powerful software that effectively combines a symbolic with a numerical approach for dealing with PR in sophisticated models.

2.8 Reparameterization. In parameter-redundant models, although some obvious reparameterization can be visually inspected (Dasgupta et al., 2007), relatively little consideration has been made concerning the procedural reparameterization approach. Dasgupta et al. (2007) presented a local reparameterization approach based on the FIM. The drawback of

this approach is that it starts by examining local identifiability, so all the sequential reparameterization procedures inherently possess a local nature. However, in many practical applications, we are more interested in global reparameterization than local reparameterization. Therefore, it is highly desirable to seek an effective method for processing global reparameterization.

In order to obtain the most parsimonious formulation of original model, any functional dependence in $\beta(\mathbf{w})$ should be removed. We introduce the following definition from classical mathematical analysis (Zorich, 2004):

Definition 16. *A set of functions $\beta_1(\boldsymbol{w}), \ldots, \beta_r(\boldsymbol{w})$ is functionally independent in $\mathbb{R}^k$ if for any continuous function $F(y_1, \ldots, y_r)$, the equation $F(\beta_1(\boldsymbol{w}), \ldots, \beta_r(\boldsymbol{w})) \equiv 0$ holds for $\boldsymbol{w} \in \mathbb{R}^k$ only when $F(y_1, \ldots, y_r) \equiv 0$ in $\mathbb{R}^r$.*

It is obvious that functional independence is the nonlinear generalization of linear independence for which dependency is defined with respect to linear combination $F(y_1, \ldots, y_r) = \lambda_1 y_1 + \cdots + \lambda_r y_r$.

Functional independence means that any redundancy in $\beta(\mathbf{w})$ is removed, so it is of special importance to characterize the intrinsic parameter dimension of models. To formalize this, we introduce the following definition (Ran & Hu, 2015), which is more stringent in comparison with definition 14:

Definition 17. *Suppose $P_{\boldsymbol{w}}$ is a statistical model. A vector $\beta = \beta(\boldsymbol{w})$ is a minimal reparameterization of $P_{\boldsymbol{w}}$ if β satisfies the following two conditions:*

1. *$P_{\boldsymbol{w}}$ can be rewritten in terms of $\beta = \beta(\boldsymbol{w})$ with $dim\beta < dim\boldsymbol{w}$.*
2. *The $\beta = \beta(\boldsymbol{w})$ has a minimality property, that is, there exists no other vector $\kappa(\boldsymbol{w})$ such that $dim\kappa(\boldsymbol{w}) < dim\beta(\boldsymbol{w})$, and $P_{\boldsymbol{w}}$ can be rewritten in terms of $\kappa(\boldsymbol{w})$.*

Based on the IF approach, Ran and Hu (2015) presented the following theorem:

Theorem 18. *Let $P_{\boldsymbol{w}}$ be a statistical model. $\varphi(\boldsymbol{w})$ is an IF and $\boldsymbol{J}_{\varphi}(\boldsymbol{w})$ is the Jacobian of $\varphi(\boldsymbol{w})$. Suppose that $\boldsymbol{J}_{\varphi}(\boldsymbol{w})$ is symbolically column deficient, and that the null subspace*

$$\mathcal{V} = \{\boldsymbol{v}(\boldsymbol{w}) \in \mathbb{R}^k : \boldsymbol{J}_{\varphi}(\boldsymbol{w})\boldsymbol{v}(\boldsymbol{w}) = 0\} \tag{2.25}$$

of $\boldsymbol{J}_{\varphi}(\boldsymbol{w})$ is spanned by a maximal system of d linearly independent vectors $\boldsymbol{v}_s(\boldsymbol{w}) = (v_{s1}(\boldsymbol{w}), \ldots, v_{sk}(\boldsymbol{w}))^{\mathrm{T}}$, $1 \le s \le d$. Let $r = k - d$. Then $\beta = \beta(\boldsymbol{w}) \in \mathbb{R}^r$ is the minimal reparameterization if and only if $\beta = \beta(\boldsymbol{w}) \in \mathbb{R}^r$ satisfies the following Lagrange first-order linear partial differential equation:

$$v_{si}(\boldsymbol{w})\frac{\partial \beta}{\partial w_1} + \cdots + v_{sk}(\boldsymbol{w})\frac{\partial \beta}{\partial w_k} = 0, 1 \le s \le d. \tag{2.26}$$

Moreover, if the matrix $\mathbf{J}_{\varphi}(\beta)$ is of column full rank, the model is locally identifiable with respect to the minimal reparameterization β.

The entire process of determining parameter structure can be implemented sequently using the following steps:

1. Solve the IF $\varphi(\mathbf{w})$ from original model $P_{\mathbf{w}}$ (see section 2.4).
2. Form the Jacobian $\mathbf{J}_{\varphi}(\mathbf{w})$ of IF and check its rank. Specifically, if the Jacobian is of column full rank at $\mathbf{w}_0$, the model is locally identifiable at $\mathbf{w}_0$ (see theorem 12). If the Jacobian is symbolically column deficient, the model is parameter redundant and can be reparameterized (see theorem 18).
3. Solve the maximal independent vectors in the null space $\mathcal{V}$ (see equation 2.25) of $\mathbf{J}_{\varphi}(\mathbf{w})$ if PR is detected.
4. Solve equation 2.26 for minimal reparameterization.

In previous sections, we discussed the parameter-constrained method for alleviating nonidentifiability problem and the reparameterization method for alleviating the parameter redundance problem. There is also major research on approaches to alleviating data-driven ill-conditioned problems—for instance, adopting newer learning settings, such as a transductive setting, a universum setting, and learning under privileged information (Vapnik, 2006), different from inductive setting, and modifying underlying statistical inference problems (Vapnik & Izmailov, 2015a). In many cases, such settings have been adopted for an inductive setting, which leads to new data-driven regularization techniques. (Interested readers can consult Collobert, Sinz, Weston, & Botton, 2006; Cherkassky, Dhar, & Dai, 2011; and Vapnik & Izmailov, 2015b for more details.)

3 The Influence of Parameter Identifiability in Machine Learning

Identifiability is a primary assumption in almost all classical statistical models (Dasgupta et al., 2007). However, such an assumption may be violated in a large variety of models. It has been difficult to study the learning theory of singular models because there has been no mathematical theory for such problems. Up to now, this difficulty has not been fully resolved in machine learning.

In a wide variety of statistical models, the critical set

$$\mathcal{C} = \{\mathbf{w} \in \mathcal{W} : |\mathrm{FIM}(\mathbf{w})| = 0\} \tag{3.1}$$

at which the FIM degenerates is a zero-measure subset in $\mathbb{R}^k$. Hence, one might argue that in generic cases, the true parameter $\mathbf{w}_{\text{true}}$ is seldom contained in $\mathcal{C}$ and that the learning theory assuming $|\mathrm{FIM}(\mathbf{w}_{\text{true}})| > 0$ is sufficient in practical applications. Here $\mathbf{w}_{\text{true}}$ means that the model can express

the true distribution. However, this consideration is generally wrong for the following reasons:

- In some cases, we have to optimize a statistical model by comparing several candidate models or hyperparameters. Thus, one always examines models under the condition that the optimal parameter lies in a neighborhood of $\mathcal{C}$. Especially in model selection, hyperparameter optimization, or hypothesis testing, one needs the theoretical results in the case $\mathbf{w}_{\text{true}} \in \mathcal{C}$ because we have to determine whether $\mathbf{w}_{\text{true}} \in \mathcal{C}$ or not (Watanabe, 2009).
- The learning process takes place in the entire parameter space, so even if the true parameter is regular, the singular structure strongly affects the choice of learning algorithms as well as the dynamics of learning. It has been shown that once parameters are attracted to singular points, the learning trajectory is very slow to move away from them (Amari et al., 2006; Cousseau et al., 2008; Wei & Amari, 2008).

From the viewpoint of information geometry (Amari & Nagaoka, 2000), the parameter space of statistical models forms a geometrical manifold, called the *neuromanifold* in the case of ANNs (Amari et al., 2006). Such a model is endowed with a statistical structure, and a Riemann metric is given by the FIM. However, the FIM degenerates at singularities. Such a singular structure is ubiquitous not only in MLPs but also in RBFs, GMMs, and ARMA time series models, linear systems whose transfer functions are rational functions (Amari et al., 2006), and many other cases. In singular models, the standard statistical paradigm of the Cramér-Rao theorem (see theorem 19) does not hold, and the singularity gives rise to unusual behaviors in parameter estimation, hypothesis testing, model selection, learning algorithm, the dynamic of learning, and Bayesian inference, for example.

In this section, we review the significant influence of singularity on various aspects of machine learning. First, we review results concerning special parameter structures of singular models. Second, we briefly survey the influence of singularity on estimation theory, hypothesis testing, model selection, and so on. Third, we review the issues of learning algorithms and learning dynamics within the information geometry framework due to Amari and Nagaoka (2000). Finally, we review the SLT developed by Watanabe within the algebraic geometry framework (Shafarevich, 1974); the SLT has a profound theoretic impact in Bayes inference.

3.1 Parameter Structure of Singular Models. This section focuses on the problem regarding the geometrical structure of parameter space. In a singular model, when we summarize observationally equivalent parameters, the model is known to have a generic cone-type singularity embedded in a finite-dimensional, sometimes infinite-dimensional, regular manifold (Castelle & Gassiat, 1997). In machine learning, such a structure was

described in the pioneering work of Brockett (1976) in the case of linear systems and in the case of MLPs by Sussmann (1992), Chen et al. (1993), Kurková and Kainen (1994), Fukumizu (1996), and Ruger and Ossen (1997). More recently, extensive research has focused on this problem, and theories are now being established.

The local parameter structure of regular models is represented by the tangent space of the parameter manifold, where the first-order asymptotic theory is well formulated. The concepts of affine connections and the related *e*- and *m*-curvatures are necessary for high-order asymptotic theory (Amari & Nagaoka, 2000). Nevertheless, the singular model does not have tangent space at singularities; instead, the tangent cone is useful for analyzing its local structure. Castelle and Gassiat (1997) showed the generic cone structure and elucidated why unusual behaviors emerge in singular models. Also, the local conic structure and the related random gaussian field play a fundamental role in analyzing the behaviors of likelihood ratio statistics (Fukumizu, 2003) as well as the MLE and its generalization capability (Amari et al., 2006). Specifically, the quotient space $\mathcal{W}/\sim$ forms a cone (Amari et al., 2006). This is called a *tangent cone* and is different from the tangent space in regular models. (For more details, to refer Amari et al., 2006.) The geometry of equivalent class $[\mathbf{w}]$ is discussed in Dasgupta et al., 2007.

3.2 Estimation Theory. Despite our ignorance of $p(\mathbf{z}|\mathbf{w}_{\text{true}})$, the ability to make repeated measurements on $\mathbf{z}$ enables us to obtain empirical knowledge about $\mathbf{w}_{\text{true}}$. By minimizing an error function $E(\mathbf{w})$, we obtain an estimator $\widehat{\mathbf{w}}_n$ of $\mathbf{w}_{\text{true}}$. Just as with any other statistic, the statistical behavior of $\widehat{\mathbf{w}}_n$ is completely determined by its probabilistic law, which is prohibitively difficult to obtain for a training set of size n. We first use the MLE $\widehat{\mathbf{w}}_{\text{ML}}$ as a typical case to illustrate the influence of singularity in estimation theory.

3.2.1 Maximum Likelihood Estimator. In classical statistics, the MLE is known to be asymptotically optimal, and the FIM is extensively used to measure the average amount of information included in a single measurement $\mathbf{z}_n$ (Amari et al., 2006).

Theorem 19 *(Cramér-Rao). Let $\widehat{\boldsymbol{w}}_n$ be an unbiased estimator of $\boldsymbol{w}$ from n examples in a globally identifiable model. Then the error covariance of $\widehat{\boldsymbol{w}}_n$ satisfies*

$$\mathbb{E}_{\boldsymbol{z}}[(\widehat{\boldsymbol{w}}_n - \boldsymbol{w})(\widehat{\boldsymbol{w}}_n - \boldsymbol{w})^T] \geq \frac{1}{n} FIM^{-1}(\boldsymbol{w}), \tag{3.2}$$

where the inequality holds in the sense of positive definiteness of matrices. The equality holds asymptotically for the MLE $\widehat{\boldsymbol{w}}_{ML}$,

$$\lim_{n\to\infty} n\mathbb{E}_z[(\widehat{\boldsymbol{w}}_{ML} - \boldsymbol{w})(\widehat{\boldsymbol{w}}_{ML} - \boldsymbol{w})^T] = FIM^{-1}(\boldsymbol{w}). \tag{3.3}$$

Moreover, $\widehat{\boldsymbol{w}}_{ML}$ is asymptotically subject to the gaussian PDF with mean $\boldsymbol{w}$ and covariance matrix $\frac{1}{n}FIM^{-1}(\boldsymbol{w})$.

In contrast to regular models in which the MLE maximizes the likelihood function, the MLE in singular models is solved by making use of random field theory (Amari et al., 2006). Hence, it is generally very difficult to calculate the MLE and analyze its properties.

3.2.2 Large Sample Behavior. It is difficult to obtain the precise probability law of $\widehat{\mathbf{w}}_n$. However, approximation to this probability law for large n can be obtained by using standard tools, including the law of large numbers and the central limit theorem. These approximations reveal that the probability law of $\widehat{\mathbf{w}}_n$ collapses into a particular well-defined set, becoming more and more concentrated around this set as n increases (White, 1989a). This increasing concentration property is referred to as consistency, and the probability law is known as limiting distribution or asymptotic distribution of $\widehat{\mathbf{w}}_n$. We briefly review these two issues below.

Consistency. Consistency means that the estimator $\widehat{\mathbf{w}}_n$ converges almost surely to the optimal $\mathbf{w}_{\text{opt}}$. White (1994) showed that consistency holds under a collection of regularity conditions; among these conditions is parameter identifiability. Unfortunately, this condition can be easily violated in a considerable variety of learning machines. For instance, in MLPs, if $\mathcal{W}$ is restricted to a single cone described by Nielsen (1989), then multiple minima can be to some extent be alleviated; a heuristic explanation is that it eliminates the interchangeability of hidden units. Nevertheless, this restriction to a Nielsen cone cannot guarantee identifiability if $\mathbf{w}_{\text{opt}}$ happens to lie at the singularities. In the case of a single hidden-layer feedforward network, there are two reasons for this possibility (White, 1989a). The first is referred to as the case of *redundant inputs* and the second as the case of *irrelevant hidden units*. Specifically, the former occurs when one or more of the network inputs is a linear combination of the other inputs, while the latter occurs when identical optimal network performance can be achieved with fewer hidden units. Both cases generate the manifolds on which $E(\mathbf{w})$ is flat and minimal. This coincides with the observations noticed by Amari (see section 3.5).

In singular models, instead of $\widehat{\mathbf{w}}_n \overset{\text{a.s.}}{\to} \mathrm{w}_{\text{opt}}$, we have (White, 1989a)

$$\inf_{\mathbf{w}^*\in\mathcal{W}^*} ||\widehat{\mathbf{w}}_n - \mathbf{w}^*||_2 \overset{\text{a.s.}}{\to} 0, \tag{3.4}$$

where $\mathcal{W}^*$ is the set of all minimizers of $E(\mathbf{w})$, that is,

$$\mathcal{W}^* = \{\mathbf{w}^* \in \mathcal{W} : E(\mathbf{w}^*) \leq E(\mathbf{w}), \forall \mathbf{w} \in \mathcal{W}\}. \tag{3.5}$$

This gives an answer to the question of what is learned when one seeks the MLE: the learned parameters collapse into the set of parameters that deliver minimal error.

Asymptotic distribution. The formal concept for studying the limiting distribution of $\widehat{\mathbf{w}}_n$ is that of convergence in distribution (White, 1989a).

Definition 18. *Let $\boldsymbol{z}_n$, $n = 1, 2, \ldots$ be a sequence of random variables having distribution functions $F_n(\boldsymbol{a}) = P(\boldsymbol{z}_n \leq \boldsymbol{a})$, $n = 1, 2, \ldots$, and let $\boldsymbol{z}$ be another random variable having distribution function $F(\boldsymbol{a}) = P(\boldsymbol{z} \leq \boldsymbol{a})$. $\boldsymbol{z}_n$ is said to converge to $\boldsymbol{z}$ in distribution, if $|F_n(\boldsymbol{a}) - F(\boldsymbol{a})| \to 0$ at all $\boldsymbol{a}$ for which F is continuous.*

The limiting distribution of $\widehat{\mathbf{w}}_n$ depends on the nature of $\mathcal{W}^*$. In general, $\mathcal{W}^*$ may consist of isolated points or isolated flats, or both. If convergence to a flat occurs, then $\widehat{\mathbf{w}}_n$ has a limiting distribution that can be analyzed using the theory of partially identified models (Phillips, 1989). This distribution belongs to the limiting mixed gaussian (LMG) family introduced by Phillips (1989). When $\mathbf{w}_{\text{opt}}$ is a regular point, the $\widehat{\mathbf{w}}_n$ has a limiting gaussian distribution. This result is also a consequence of Phillips's results but can be derived as well from the conventional central limit theorem.

Hence, the important influence caused by singularity is that we work with an LMG family rather than a gaussian family when analyzing the limiting distribution of $\widehat{\mathbf{w}}_n$.

3.3 Hypothesis Testing. Hypothesis testing is an important method to judge from data whether the true distribution lies at the singularities. In the context of ANNs, because many questions about the precise form of the optimal network architecture can be formulated as hypothesis testing regarding $\widehat{\mathbf{w}}_n$, these questions can be solved by calculating some standard statistics. It is therefore important to address the effect of singularity from the aspect of hypothesis testing.

In regular models, the log-likelihood ratio statistic,

$$\lambda = 2\sum_{j=1}^{n} \log \frac{p(\mathbf{z}_j|\widehat{\mathbf{w}}_{\text{ML}})}{p(\mathbf{z}_j|\mathbf{w}_{\text{true}})}, \tag{3.6}$$

obeys the χ^2 distribution with degree of freedom k when the data size n is large enough. However, when the model is singular, the λ may not be subject to χ^2 and may diverge to infinity in proportion to n. This study can be dated back to the work of Weyl (1939) and Hotelling (1939), and a precise asymptotic form of the λ in singular models is given in Liu and Shao (2003). However, it is unfortunate that such tangled problems have usually been

excluded as pathological cases and have not been well studied. In fact, such problems emerge ubiquitously in machine learning.

We consider the following statistical test:

$$H_0 : \mathbf{w} = \mathbf{w}_{\text{true}} \quad \text{vs.} \quad H_1 : \mathbf{w} \neq \mathbf{w}_{\text{true}}. \tag{3.7}$$

When $\mathbf{w}_{\text{true}}$ is regular, one expands λ in the Taylor series, yielding

$$\lambda = n(\widehat{\mathbf{w}}_{\text{ML}} - \mathbf{w}_{\text{true}})^{\text{T}} \text{FIM}^{-1}(\mathbf{w}_{\text{true}})(\widehat{\mathbf{w}}_{\text{ML}} - \mathbf{w}_{\text{true}}). \tag{3.8}$$

Hence, λ obeys the χ^2 with degree of freedom k when n is large enough. The expectation of λ is $\mathbb{E}_{\mathbf{z}}\lambda = k$. However, when $\mathbf{w}_{\text{true}}$ is singular, this situation changes. The $\text{FIM}^{-1}(\mathbf{w}_{\text{true}})$ diverges, so the similar expansion is no longer valid. The expectation of λ is asymptotically written as $\mathbb{E}_{\mathbf{z}}\lambda = c(n)k$, where the term $c(n)$ takes various forms depending on the nature of singularities. With the help of the gaussian random field, Fukumizu (2003) proved that $c(n) = \log n$ in the case of MLP, while in the GMM, $c(n) = \sqrt{\log \log n}$.

In the case of MLPs, the limiting distribution of $\widehat{\mathbf{w}}_n$ can be used to test hypotheses about $\mathbf{w}_{\text{true}}$. This technique permits statistical inference to questions regarding the precise form of optimal ANN architectures. Two hypotheses of particular interest for MLPs are the irrelevant inputs hypothesis and the irrelevant hidden units hypothesis. The former can be solved by making use of standard statistical tools. Nevertheless, there are some difficulties in the limiting distribution of $\widehat{\mathbf{w}}_n$ under the null hypothesis that the hidden units are irrelevant. Problems arise because when the null hypothesis is true, the optimal weights from input units to the irrelevant hidden units contain singularities. This problem is known in the statistics literature as "nuisance parameters are identifiable only under alternative hypothesis" (White, 1989a). The LMG family plays an essential and unavoidable role in this case. Hypothesis testing in MLPs has been studied by Davies (1987). As one should expect from the LMG family, the distribution of the statistic is no longer χ^2. However, certain techniques can be adopted to avoid these difficulties, yielding a χ^2 statistic. (For such tests, see White, 1989b, for more details.)

In addition to hypothesis testing, the generalization error is related to the log-likelihood ratio. The generalization error has so far been evaluated based on the Cramér-Rao paradigm, so we need a new method to attack this problem in singular models. Fukumizu used a simple linear model and showed that the generalization error of MLPs with singularities is different from that of the regular models (Fukumizu, 1999), and this problem is further studied in Fukumizu (2003).

3.4 Model Selection. To obtain an adequate model, one should select a model from many alternatives based on the data. In classical statistics,

this issue is known as *bias-variance dilemma* (Hastie, Tibshirani, & Friedman, 2001). For instance, in MLPs, one needs to determine the preferred model size, that is, the number h of hidden units. Conventional AIC (Akaike, 1974), BIC (Schwarz, 1978), and MDL (Rissanen, 1983) have been extensively used as model selection criteria.

In pioneering work, Akaike (1974) proposed the well-known AIC:

$$\text{AIC} = -2\log(\text{likelihood}) + 2(\text{number of independent parameters}), \tag{3.9}$$

which basically consists of the badness-of-fit term $-2\log(\text{likelihood})$ and the complexity term $2(\text{number of independent parameters})$. This is derived from asymptotic statistical analysis, where the MLE is subject to gaussian with covariance $\frac{1}{n}\text{FIM}^{-1}(\mathbf{w})$.

MDL is a criterion to minimize the length of encoding for the observed data by using a family of parametric models. It is given asymptotically by the minimizer of

$$\text{MDL} = \text{training error} + \frac{\log n}{2n}k. \tag{3.10}$$

The BIC gives the same criterion as MDL. Both MDL and BIC are derived from the same assumption regarding the gaussianity of the MLE.

The AIC is a criterion that minimizes the generalization error. Several authors, such as Murata, Yoshizawa, and Amari (1994), Bozdogan (2000), and Amari et al. (2006) have questioned whether the complexity term is a sufficient measure to capture the overparameterization phenomenon in parameter-redundant models and have proposed several variants. For instance, Hagiwara, Toda, and Usui (1993) noticed this problem when they used AIC to determine the size of MLPs and found that AIC did not work well. They observed that this is caused by the singularity of the hierarchical structures, and investigated ways to overcome this difficulty (Hagiwara, 2002a, 2002b). Akaho and Kappen (2000) also noticed this in the GMMs. Hence, one should evaluate the log-likelihood ratio carefully in such cases. For ANNs, the network information criterion as a modified version of AIC has been proposed (Murata et al., 1994).

Many comparisons of AIC and MDL by computer simulations have been reported. Sometimes AIC works better, while MDL does better in other cases. Such confusing reports seem to stem from the differences between regular and singular models, as well as the differences in the nature of singularities (Amari et al., 2006). Watanabe (2010) proposed the widely applicable information criterion (WAIC) for model selection in singular models. More recently, Watanabe (2013) presented the widely applicable

Bayesian information criterion (WBIC), a generalized version of BIC for singular models.

3.5 Learning Algorithm. We now consider how the problem of minimizing $E(\mathbf{w})$ might be solved in practical situations and give an overview of the influence of singularities on learning algorithms. Among the large variety of numerical methods, we merely consider some gradient-based techniques. If $E(\mathbf{w})$ is differentiable, a typical iteration can be constructed as

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha_t \mathbf{G}(\mathbf{w}_t)\nabla E(\mathbf{w}_t), t = 1, 2, \ldots, \tag{3.11}$$

where $\mathbf{G}(\mathbf{w}_t)$ is a positive-definite matrix and $\alpha_t > 0$ is the learning rate. Different choices for α_t and $\mathbf{G}(\mathbf{w}_t)$ implement different gradient descent algorithms. Under appropriate conditions, $\mathbf{w}_t$ converges to $\mathbf{w}^\dagger$, a parameter solving the equation $\nabla E(\mathbf{w}) = 0$, which is the necessary first-order condition for a local minimum of $E(\mathbf{w})$ (Sundaram, 1996).

When the parameter space contains singularities, by using standard gradient descent algorithm where $\mathbf{G}(\mathbf{w}_t) = \mathbf{I}$, it has been shown that if a parameter reaches a local minimum, it will remain there for a long period of time because there is no mechanism to escape (as this would require a temporary increase in $E(\mathbf{w})$) (Amari et al., 2006). Also, the presence of saddle points, or regions where the error surface is very flat, can cause the algorithm to become "stuck" in the flat domains. In singular models, this slow convergence or flat plateau becomes extremely severe since the error surface has completely flat valleys, causing the learning algorithm to be extremely slow. This is the plateau phenomenon observed in back propagation (BP) learning for MLPs (Duda, Hart, & Stork, 2001).

To overcome the slow convergence phenomenon, Amari proposed a *natural (or Riemannian) gradient descent (NGD)* method from the perspective of information geometry (Amari, 1998). In statistical models, the parameter space $\mathbb{R}^k$ is endowed with a Riemann metric given by

$$||\mathbf{w}||_{\mathrm{FIM}} = \sqrt{\mathbf{w}^{\mathrm{T}}\mathrm{FIM}(\mathbf{w})\mathbf{w}}. \tag{3.12}$$

The $||\mathbf{w}||_{\mathrm{FIM}}$ plays the role of Riemann metric, so that the gradient $\nabla E(\mathbf{w})$ does not represent the steepest direction, but the natural (or contravariant) gradient $\mathrm{FIM}^{-1}(\mathbf{w})\nabla E(\mathbf{w})$ does (Amari, 1998).

For MLPs, the weight $\mathbf{w}$ can be updated in sequential, stochastic, or batch mode (Duda et al., 2001). In order to demonstrate the learning dynamic in section 3.6, we consider only a sequential (or online) algorithm, in which the data are assumed one at a time and the parameter updates after each presentation. In contrast, the batch techniques deal with all of the data in one go. Thus, we obtain the sequential algorithm by applying the techniques

of stochastic gradient descent, also known as sequential gradient descent (Bishop, 2006), as follows. If the error function comprises a sum over data points

$$E(\mathbf{w}) = \sum_{j=1}^{n} E_j(\mathbf{w}), \tag{3.13}$$

then after presentation of observation $\mathbf{z}_j$, the stochastic gradient descent algorithm updates the parameter $\mathbf{w}$ using

$$\mathbf{w}_{j+1} = \mathbf{w}_j - \alpha_j \nabla E_j(\mathbf{w}_j),\ j = 1, 2, \ldots, n, \tag{3.14}$$

where the α_j needs to ensure that the algorithm converges. This learning process is Markovian, since at step $j+1$, the estimator $\mathbf{w}_j$ is modified to give a new estimator $\mathbf{w}_{j+1}$ based on the current observation $\mathbf{z}_j$, and the old observations $\mathbf{z}_1, \ldots, \mathbf{z}_{j-1}$ cannot be reused to obtain $\mathbf{w}_{j+1}$. This method is a specific version of stochastic approximation proposed by Robbins and Monro (1951).

The stochastic gradient descent algorithm was proposed by Amari (1967) and was named the BP algorithm in the ANNs literature (Bishop, 1995; Haykin, 1998). Specifically, in the case of squared loss function, the stochastic gradient learning is known as the least-mean-squares (LMS) algorithm (Haykin, 1998). For an input-output example $(\mathbf{x}_j, \mathbf{y}_j)$, the squared loss function is

$$E(\mathbf{w}) = \frac{1}{2}\sum_{j=1}^{n} E_j(\mathbf{w}) = \frac{1}{2}\sum_{j=1}^{n} ||\mathbf{y}_j - \mathbf{f}(\mathbf{x}_j, \mathbf{w})||_2^2. \tag{3.15}$$

This is easily recognized as the BP algorithm when applied to MLPs. Thus, the sequential gradient descent mode of a BP algorithm can be viewed as an application of the Robbins-Monro stochastic approximation procedure to solve the first-order condition in a nonlinear least-squared regression problem.

By using natural gradient, the online NGD algorithm can be written as (Amari, 1998)

$$\mathbf{w}_{j+1} = \mathbf{w}_j - \alpha_j \mathrm{FIM}^{-1}(\mathbf{w}_j)\nabla E_j(\mathbf{w}_j),\ j = 1, 2, \ldots, n. \tag{3.16}$$

By using a simple MLP model, it has been shown that the performance of NGD is remarkably good (Amari et al., 2006), and it is sometimes free from being trapped in plateaus, which gives rise to slow convergence in the BP algorithm. This suggests that the Riemannian structure might eliminate

such plateaus or might make them not so severe. (For more applications, such as blind source separation, blind multichannel deconvolution and Boltzmann machine, about NGD, see Amari, 1998.)

Theoretically, the asymptotic performance of online NGD cannot be better than the optimal batch procedure in which all the examples can be reused again and again. Thus the online NGD is at best capable of near-optimal performance. However, Amari (1998) proved that the online NGD gives a Fisher asymptotically efficient estimator, so that it is asymptotically equivalent to the optimal batch method, while the BP is not Fisher efficient.

The main drawback of NGD is that it is difficult to calculate $\mathrm{FIM}(\mathbf{w}_j)$, because the PDF $p(\mathbf{x})$ of input is generally unknown. Moreover, the calculation of $\mathrm{FIM}^{-1}(\mathbf{w}_j)$ is costly. For alleviating this difficulty, two further improvements about the learning rate α_i and adaptive $\mathrm{FIM}^{-1}(\mathbf{w}_j)$ were presented in Amari (1998).

3.6 Learning Dynamics. Through the NGD method, slow convergence is alleviated because the NGD takes the geometrical structure of parameter manifold into account. However, the FIM (or Riemann metric) degenerates at singular points, so one needs to study its dynamic behavior of learning—in particular, the effects of singularities in the BP and NGD methods. Work done regarding this aspect includes that of Fukumizu and Amari (2000) as well as the statistical-mechanical approaches taken by Saad and Solla (1995) and Rattray, Saad, and Amari (1998); Rattray and Saad (1999).

Although there exists no unified theory concerning learning dynamics in singular models, problems caused by particular hierarchical models have been addressed by many researchers, and various approaches have been proposed. Kang, Oh, Kwon, and Park (1993) used a special perceptron model and found that the parameters are attracted to the critical set $\mathcal{C}$ and are very slow to move away from it. Saad and Solla (1995) analyzed the learning dynamics in a more general case and showed that such a phenomenon is universal. They argued that the slow convergence (or the plateau phenomenon) in the BP algorithm, is caused by this singularity. Through a simple MLP model, Fukumizu and Amari (2000) calculated the Hessian along the lines in the critical set. Specifically, when the Hessian is positive definite, the lines are attractive; when the Hessian has negative eigenvalues, the learning trajectory eventually escapes from these lines. They further showed that in some cases, part of the line can be truly attractive, although it is not a usual asymptotically stable equilibrium but has directions of escape (even though the derivative along the line vanishes) in other parts. This is not a usual saddle point, but belongs to the special type called the Milnor attractor (Milnor, 1985). In such a case, the trajectory is truly attracted to the line and stays inside the line, fluctuating around it due to the random noise, until it finds a place from which it can escape. This explains the flat phenomenon. Inoue, Park, and Okada (2003) demonstrated

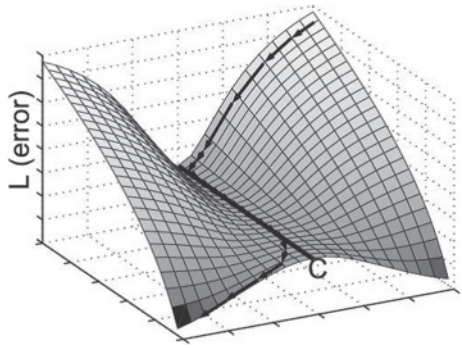

Figure 4: Learning trajectory near the singularities (Amari et al., 2006).

this by using a specific perceptron. Saad and Solla (1995) argued that the problem of the flat phenomenon cannot be resolved by simply increasing learning rate α, because even when the trajectory goes outside the line due to a large α, it may again return to it. Park, Inoue, and Okada (2003) illustrated the flat phenomenon via an MLP model whose true parameters are on the singularities. Once the trajectory reaches $\mathcal{C}$ in Figure 4, all points are observationally equivalent and suboptimal.

The NGD algorithm enables the influence of the singularity to be reduced, and the trajectory is not trapped in the plateaus. Rattray et al. (1998) analyzed the dynamics of NGD by means of statistical physics and showed that it is almost ideal. Fukumizu and Amari (2000) showed that the NGD imposes an efficiently strong repulsive force on the directions of escape from $\mathcal{C}$, so the trajectory moves away without being trapped in $\mathcal{C}$.

3.7 Singular Learning Theory: An Algebraic Geometry Framework. Watanabe (2001a, 2001b) was the first to study the effect of singularity in Bayesian inference. He and his colleagues introduced the approaches of algebraic geometry by using Hironaka's theorem of singularity resolution (Hironaka, 1964) and Sato's formula (Sato & Shintani, 1974) to evaluate the asymptotic performance of the Bayesian predictive distribution in various hierarchical models; remarkable results have been derived (Watanabe & Amari, 2003; Watanabe, 2001a, 2001b, 2009, 2013).

The Bayesian inference is used in many cases where a prior distribution $p(\mathbf{w})$ about $\mathbf{w}$ is available. When the prior distribution penalizes complex models, it plays a role equivalent to the regularization term in equation 1.8. When the data $\mathcal{D}_n$, equation 1.3, are generated, the posterior distribution of the parameter $\mathbf{w}$ is written as

$$p(\mathbf{w}|\mathcal{D}_n) = \frac{1}{Z_n} p(\mathbf{w}) p(\mathcal{D}_n|\mathbf{w}), \tag{3.17}$$

where Z_n is a normalized factor such that $p(\mathbf{w}|\mathcal{D}_n)$ is a proper PDF with respect to $\mathbf{w}$:

$$Z_n = \int p(\mathbf{w})p(\mathcal{D}_n|\mathbf{w})\mathrm{d}\mathbf{w}. \tag{3.18}$$

Note that Z_n is a measurable function of $\mathcal{D}_n$, so it is a random variable called the *evidence* or the *marginal likelihood* (Watanabe, 2009). The normalized evidence is given by

$$Z_n^0 = \frac{Z_n}{\prod_{j=1}^n p(\mathbf{z}_j|\mathbf{w}_{\text{true}})}. \tag{3.19}$$

The stochastic complexity, the minus log marginal likelihood, or the free energy is defined by

$$F_n = -\log Z_n. \tag{3.20}$$

Similarly, the normalized free energy is defined by

$$F_n^0 = -\log Z_n^0. \tag{3.21}$$

The Bayesian predictive distribution is the distribution of a new sample $\mathbf{z}$ based on $\mathcal{D}_n$. It is given by averaging the $p(\mathbf{z}|\mathbf{w})$ over the posterior $p(\mathbf{w}|\mathcal{D}_n)$:

$$p(\mathbf{z}|\mathcal{D}_n) = \int p(\mathbf{z}|\mathbf{w})p(\mathbf{w}|\mathcal{D}_n)\mathrm{d}\mathbf{w}. \tag{3.22}$$

The Bayesian estimation is then defined by

$$p_{Bay}(\mathbf{z}) = p(\mathbf{z}|\mathcal{D}_n). \tag{3.23}$$

In other words, the Bayesian estimation is the mapping $\mathcal{D}_n \to p_{Bay}(\mathbf{z})$. It is known that if $p(\mathbf{w})$ is smooth, its influence decreases as the sample size n increases, and it approaches the MLE, which is regarded as the maximum a posterior (MAP) under the uniform prior (Berger, 1985). However, unusual behaviors occur in singular models. A smooth prior on $\mathcal{W}$ is singular in the equivalent class $\mathcal{W}/\sim$ because a singular point in $\mathcal{W}/\sim$ includes infinitely many equivalent parameters in $\mathcal{W}$. Hence, the prior density is infinitely large on the singular points compared with that on regular points. The consequence is that the Bayesian smooth prior is in favor of singular points with an infinitely large factor (e.g., MLPs with a smaller number of hidden units). This amounts to imposing a prior $p(\mathbf{w}) = \infty$ in singular points. Hence, the Bayesian method works well in such a manner to avoid

overfitting. When a very large perceptron with a smooth Bayesian prior is used, and an adequately small model will be selected, although no theory exists that explains how to choose the specific form of prior.

For a full study of SLT, we introduce the following concepts (Watanabe, 2009):

Definition 19. *Suppose $p(\boldsymbol{z}|\boldsymbol{w})$ is a statistical model and $p(\boldsymbol{z}|\boldsymbol{w}_{true})$ is the true PDF. The log density ratio function $r(\boldsymbol{z}|\boldsymbol{w})$, the KLD $KL(\boldsymbol{w})$, and the log-likelihood ratio function $KL_n(\boldsymbol{w})$ of $p(\boldsymbol{z}|\boldsymbol{w})$ are, respectively, defined by*

$$r(\boldsymbol{z}|\boldsymbol{w}) = log\frac{p(\boldsymbol{z}|\boldsymbol{w}_{true})}{p(\boldsymbol{z}|\boldsymbol{w})}, \tag{3.24}$$

$$KL(\boldsymbol{w}) = \int p(\boldsymbol{z}|\boldsymbol{w}_{true}) r(\boldsymbol{z}|\boldsymbol{w}) d\boldsymbol{z}, \tag{3.25}$$

$$KL_n(\boldsymbol{w}) = \frac{1}{n}\sum_{j=1}^{n} r(\boldsymbol{z}_j|\boldsymbol{w}), \tag{3.26}$$

where $KL_n(\boldsymbol{w})$ is referred as an empirical KLD.

Here, by notational convention in SLT (Watanabe, 2010), we use the simpler notation $\mathrm{KL}(\mathbf{w})$ to denote the KLD between the PDFs $p(\mathbf{z}|\mathbf{w}_{\mathrm{true}})$ and $p(\mathbf{z}|\mathbf{w})$ instead of the more precise form, $\mathrm{KL}(\mathbf{w}_{\mathrm{true}}, \mathbf{w})$. This convention will be used in the rest of this review. For analyzing the performance of various estimation methods, the following concepts are necessary:

Definition 20. *For a given learning algorithm $\mathcal{D}_n \to \widehat{\boldsymbol{w}}_n$, the KLD*

$$KL(\widehat{\boldsymbol{w}}_n) = KL(\boldsymbol{w}_{true}, \widehat{\boldsymbol{w}}_n) = \int p(\boldsymbol{z}|\boldsymbol{w}_{true}) log\frac{p(\boldsymbol{z}|\boldsymbol{w}_{true})}{p(\boldsymbol{z}|\widehat{\boldsymbol{w}}_n)} d\boldsymbol{z} \tag{3.27}$$

between the true parameter $\boldsymbol{w}_{true}$ and the learned parameter $\widehat{\boldsymbol{w}}_n$ is called the generalization error. The training error is defined by

$$KL_n(\widehat{\boldsymbol{w}}_n) = KL_n(\boldsymbol{w}_{true}|\widehat{\boldsymbol{w}}_n) = \frac{1}{n}\sum_{j=1}^{n} log\frac{p(\boldsymbol{z}_j|\boldsymbol{w}_{true})}{p(\boldsymbol{z}_j|\widehat{\boldsymbol{w}}_n)}. \tag{3.28}$$

In general, the $\mathrm{KL}(\widehat{\mathbf{w}}_n)$ and $\mathrm{KL}_n(\widehat{\mathbf{w}}_n)$ are measurable functions of the random samples $\mathcal{D}_n$; hence, they are real-valued random variables. One of the main goals of SLT is to clarify the probability distributions of the generalization error and training error for a given training algorithm. The expectation values $\mathbb{E}_{\mathcal{D}_n}\mathrm{KL}(\widehat{\mathbf{w}}_n)$ and $\mathbb{E}_{\mathcal{D}_n}\mathrm{KL}_n(\widehat{\mathbf{w}}_n)$ are called the *mean generalization error* and *mean training error*, respectively. If the mean generalization error is

small, the learning algorithm is more appropriate. One other goal of SLT is to establish a relationship between the generalization error and training error. If the generalization error can be estimated from the training error, one can select the suitable model or hyperparameter among several candidate models.

If a model is regular, then the Bayes a posterior distribution can be approximated by the gaussian distribution (Watanabe, 2009). Also, the MLE and MAP estimators are asymptotically subject to gaussian distribution. Such a property is called asymptotic gaussianity. However, singular models do not have such a property, so we need a new theoretical tool that enables us to analyze such singular models.

Let $\mathbb{E}_{\widetilde{\mathbf{w}}}[.]$ be the expectation value using the posteriori distribution $p(\mathbf{w}|\mathcal{D}_n)$. In Bayesian estimation, the true distribution is estimated by the predictive distribution $\mathbb{E}_{\widetilde{\mathbf{w}}}[p(\mathbf{z}|\mathbf{w})]$. In Gibbs estimation, a parameter $\mathbf{w}$ is randomly chosen from $p(\mathbf{w}|\mathcal{D}_n)$; then the true distribution is estimated by $p(\mathbf{z}|\mathbf{w})$ (Watanabe, 2009). Gibbs estimation depends on a random choice of the parameter $\mathbf{w}$; hence, to study its generalization error, we need the expectation value over random choices of $\mathbf{w}$. The Bayes and Gibbs estimations have generalization and training errors, respectively. The set of four errors is referred as the *Bayes quartet* (Watanabe, 2009).

Definition 21 *(Bayes quartet). For a statistical model $p(\boldsymbol{z}|\boldsymbol{w})$ and a prior PDF $p(\boldsymbol{w})$, the four errors are defined as follows:*

1. *The Bayes generalization error,*

$$B_g = \mathbb{E}_{\boldsymbol{z}}\left[log\,\frac{p(\boldsymbol{z}|\boldsymbol{w}_{true})}{\mathbb{E}_{\widetilde{\boldsymbol{w}}}[p(\boldsymbol{z}|\boldsymbol{w})]}\right], \tag{3.29}$$

   is the KLD from $p(\boldsymbol{z}|\boldsymbol{w}_{true})$ to the predictive distribution $\mathbb{E}_{\widetilde{\boldsymbol{w}}}[p(\boldsymbol{z}|\boldsymbol{w})]$.
2. *The Bayes training error,*

$$B_t = \frac{1}{n}\sum_{j=1}^{n} log\,\frac{p(\boldsymbol{z}_j|\boldsymbol{w}_{true})}{\mathbb{E}_{\widetilde{\boldsymbol{w}}}[p(\boldsymbol{z}_j|\boldsymbol{w})]}, \tag{3.30}$$

   is the empirical KLD from $p(\boldsymbol{z}|\boldsymbol{w}_{true})$ to the predictive distribution $\mathbb{E}_{\widetilde{\boldsymbol{w}}}[p(\boldsymbol{z}|\boldsymbol{w})]$.
3. *The Gibbs generalization error,*

$$G_g = \mathbb{E}_{\widetilde{\boldsymbol{w}}}\left[\mathbb{E}_{\boldsymbol{z}}\left[log\,\frac{p(\boldsymbol{z}|\boldsymbol{w}_{true})}{p(\boldsymbol{z}|\boldsymbol{w})}\right]\right], \tag{3.31}$$

   is the mean KLD from $p(\boldsymbol{z}|\boldsymbol{w}_{true})$ to $p(\boldsymbol{z}|\boldsymbol{w})$.

4. *The Gibbs training error,*

$$G_t = \mathbb{E}_{\tilde{\mathbf{w}}}\left[\frac{1}{n}\sum_{j=1}^{n} log\,\frac{p(\mathbf{z}_j|\mathbf{w}_{true})}{p(\mathbf{z}_j|\mathbf{w})}\right], \tag{3.32}$$

is the mean empirical KLD from $p(\mathbf{z}|\mathbf{w}_{true})$ *to* $p(\mathbf{z}|\mathbf{w})$.

To evaluate how good the statistical model $p(\mathbf{z}|\mathbf{w})$ and prior $p(\mathbf{w})$ are for a given data set $\mathcal{D}_n$, we have to study the case when the set of true parameters,

$$\begin{aligned}\mathcal{W}_{\text{true}} &= \{\mathbf{w} \in \mathcal{W} : p(\mathbf{z}|\mathbf{w}) = p(\mathbf{z}|\mathbf{w}_{\text{true}})\} \\ &= \{\mathbf{w} \in \mathcal{W} : \text{KL}(\mathbf{w}) = 0\},\end{aligned} \tag{3.33}$$

consists of not one single point but a union of several manifolds. If $\text{KL}(\mathbf{w})$ is a polynomial, then $\mathcal{W}_{\text{true}}$ is called an algebraic set; if $\text{KL}(\mathbf{w})$ is an analytic function, then $\mathcal{W}_{\text{true}}$ is called an analytic set (Watanabe, 2009).

The basic term in SLT is the empirical KLD $\text{KL}_n(\mathbf{w})$, which is a function of $\mathbf{w}$. For $\mathbf{w} \in \mathcal{W}\backslash\mathcal{W}_{\text{true}}$, a random process

$$\psi_n(\mathbf{w}) = \sum_{j=1}^{n} \frac{\text{KL}(\mathbf{w}) - r(\mathbf{z}_j|\mathbf{w})}{\sqrt{n\text{KL}(\mathbf{w})}} \tag{3.34}$$

is well defined. The log-likelihood ratio is rewritten as

$$n\text{KL}_n(\mathbf{w}) = n\text{KL}(\mathbf{w}) - \sqrt{n\text{KL}(\mathbf{w})}\psi_n(\mathbf{w}). \tag{3.35}$$

This expression has two problems (Watanabe, 2009):

1. *Geometrical problem*. In a singular model, $\mathcal{W}_{\text{true}}$ is not one single point but a real analytical set; hence, the log-likelihood ratio cannot be treated locally. Moreover, since the set $\mathcal{W}_{\text{true}}$ contains complicated singularities, it is difficult to analyze its behavior even in each local neighborhood of $\mathcal{W}_{\text{true}}$.
2. *Probabilistic problem*. When $n \to \infty$, $\psi_n(\mathbf{w})$ converges in law to a gaussian process $\psi(\mathbf{w})$ on the set $\mathcal{W}\backslash\mathcal{W}_{\text{true}}$. However, neither $\psi_n(\mathbf{w})$ nor $\psi(\mathbf{w})$ is well defined on the set $\mathcal{W}_{\text{true}}$. Therefore, it is difficult to analyze such a stochastic process near the true parameters.

Before introducing the main results, we recall some materials from algebraic geometry (Shafarevich, 1974). For real analytic function $\text{KL}(\mathbf{w})$, the fundamental theorem in algebraic geometry ensures that there exists a real k-dimensional manifold $\mathcal{M}$ and a real analytic mapping,

$$g:\ \mathcal{M} \ni \mathbf{m} \to \mathbf{w} \in \mathcal{W}, \tag{3.36}$$

such that for each coordinate $\mathcal{M}_\alpha$ of $\mathcal{M}$, $\mathrm{KL}(g(\mathbf{m}))$ is a direct product,

$$\mathrm{KL}(g(\mathbf{m})) = m_1^{2s_1} m_2^{2s_2} \dots m_k^{2s_k}, \tag{3.37}$$

where $\mathbf{m} = (m_1, \dots, m_k)$ and $s_1, \dots, s_k$ are nonnegative integers. Moreover, there exists a function $\rho(\mathbf{m}) > 0$ and nonnegative integers $h_1, \dots, h_k$ such that

$$p(g(\mathbf{m}))|g'(\mathbf{m})| = \rho(\mathbf{m})|m_1^{h_1} m_2^{h_2} \cdots m_k^{h_k}|, \tag{3.38}$$

where $|g'(\mathbf{m})|$ is the Jacobian determinant of $\mathbf{w} = g(\mathbf{m})$. By using the following compact notations,

$$\mathbf{m} = (m_1, \dots, m_k), \tag{3.39}$$

$$\mathbf{s} = (s_1, \dots, s_k), \tag{3.40}$$

$$\mathbf{h} = (h_1, \dots, h_k), \tag{3.41}$$

the function $\mathrm{KL}(g(\mathbf{m}))$ and the prior PDF $p(g(\mathbf{m}))|g'(\mathbf{m})|$ are, respectively, expressed as

$$\mathrm{KL}(g(\mathbf{m})) = \mathbf{m}^{2\mathbf{s}}, \tag{3.42}$$

$$p(g(\mathbf{m}))|g'(\mathbf{m})| = \rho(\mathbf{m})|\mathbf{m}^{\mathbf{h}}|. \tag{3.43}$$

This theorem that ensures the existence of such a real analytic manifold $\mathcal{M}$ and a real analytic mapping $\mathbf{w} = g(\mathbf{m})$ is called *Hironaka's theorem* or *resolution of singularities* (Hironaka, 1964). The function $\mathbf{w} = g(\mathbf{m})$ is called a *resolution mapping*. Since it can be proved that there exists a real analytic function $a(\mathbf{z}, \mathbf{m})$ such that

$$r(\mathbf{z}|g(\mathbf{m})) = a(\mathbf{z}, \mathbf{m})\mathbf{m}^{2\mathbf{s}}, \tag{3.44}$$

one can introduce a well-defined stochastic process on $\mathcal{M}$:

$$\xi_n(\mathbf{m}) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left\{ \mathbf{m}^{\mathbf{s}} - a(\mathbf{z}_j, \mathbf{m}) \right\}. \tag{3.45}$$

The four main formulas in SLT (Watanabe, 2009) are summarized as follows:

Theorem 20 *(standard form of log-likelihood ratio function). For a singular statistical model, there exists a k-dimensional real analytic manifold $\mathcal{M}$ and a real analytic mapping $\boldsymbol{g}: \mathcal{M} \to \mathcal{W}$ such that the log-likelihood ratio function is represented by*

$$KL_n(g(\boldsymbol{m})) = \boldsymbol{m}^{2s} - \frac{1}{\sqrt{n}}\boldsymbol{m}^s \xi_n(\boldsymbol{m}), \tag{3.46}$$

where $\xi_n(\boldsymbol{m})$ converges in law to a gaussian with mean zero and variance 2.

This theorem states that by algebraic geometrical transformation, the log-likelihood ratio function of any singular statistical model can be changed to the standard form, which allows $|g'(\mathbf{m})| = 0$.

The following theorem concerns the convergence of stochastic complexity:

Theorem 21. *Let $(-\lambda)$ and c be, respectively, the largest pole and order of the zeta function*

$$\zeta(z) = \int KL^z(\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}, \tag{3.47}$$

where $z \in \mathbb{C}$ is a complex number. The normalized stochastic complexity has the following expansion,

$$F_n^0 = \lambda \log n - (c-1)\log\log n + F^R(\xi) + o_p(1), \tag{3.48}$$

where $F^R(\xi)$ is a random variable and $o_p(1)$ is a random variable that converges in probability to 0. Therefore, the stochastic complexity has the following asymptotic expansion,

$$F_n = nS_n + \lambda \log n - (c-1)\log\log n + F^R(\xi) + o_p(1), \tag{3.49}$$

where S_n is the empirical entropy defined by

$$S_n = -\frac{1}{n}\sum_{j=1}^{n} \log p(\boldsymbol{z}_j|\boldsymbol{w}_{true}). \tag{3.50}$$

This theorem claims that the stochastic complexity is asymptotically determined by the algebraic geometrical birational invariant. If a model is regular, the $\mathrm{KL}(\mathbf{w})$ is equivalent to $||\mathbf{w}||_2^2$; hence, $\lambda = k/2$ and $c = 1$. The asymptotic expansion of F_n in a regular model is well known as the BIC, or the MDL.

By using the convergence in law $\xi_n(\mathbf{m}) \to \xi(\mathbf{m})$, one can prove the convergence in law of the Bayes quartet:

$$nB_g \to B_g^*, nB_t \to B_t^*, \tag{3.51}$$

$$nG_g \to G_g^*, nG_t \to G_t^*. \tag{3.52}$$

In real-world problems, the true distribution is generally unknown. The following theorem is therefore especially useful because it holds independent of the true distribution:

Theorem 22 *(equations of states in statistical estimation). There are two universal relations in Bayes quartet:*

$$\mathbb{E}_{\mathcal{D}_n}[B_g^*] - \mathbb{E}_{\mathcal{D}_n}[B_t^*] = 2(\mathbb{E}_{\mathcal{D}_n}[G_t^*] - \mathbb{E}_{\mathcal{D}_n}[B_t^*]), \tag{3.53}$$

$$\mathbb{E}_{\mathcal{D}_n}[G_g^*] - \mathbb{E}_{\mathcal{D}_n}[G_t^*] = 2(\mathbb{E}_{\mathcal{D}_n}[G_t^*] - \mathbb{E}_{\mathcal{D}_n}[B_t^*]). \tag{3.54}$$

The remarkable merit of these equations is that one is capable of estimating the Bayes and Gibbs generalization errors from the Bayes and Gibbs training errors without any knowledge of the true distribution.

The last formula concerns the ML method or a posterior method in singular models.

Theorem 23 *(symmetry of generalization and training errors). If the maximum likelihood method or maximum a posterior method is applied, the symmetry of generalization and training errors holds:*

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{D}_n}[nR_g] = -\lim_{n\to\infty} \mathbb{E}_{\mathcal{D}_n}[nR_t]. \tag{3.55}$$

4 Summary and Perspective

We live in a world where massive amounts of data are collected and recorded on nearly every aspect of human endeavor. Understanding and making sense of the complex and information-rich data is the main concern of machine learning. We have presented a review on the relevance of parameter identifiability for statistical machine learning. First, we review various approaches for determining parameter structure from the existing literature. This involves three interrelated issues: parameter identifiability, parameter redundancy, and reparameterization. Second, we review the significant influence of identifiability on various aspects of machine learning. In addition to demonstrating the influence and utility of identifiability, we show the interplay among, for example, identifiability theory, machine learning, mathematical statistics, information theory, optimization theory, information geometry, Riemann geometry, symbolic computation, Bayesian inference, and algebraic geometry.

In the following, we present a new perspective of parameter identifiability in machine learning; we expect that the discussion will be helpful for understanding new directions and possible challenges in the future. A new perspective that relies on the modeling approach should show great potential toward human-like machines. The current machine learning

approaches, including the deep learning models (LeCun, Bengio, & Hinton, 2015), are mostly based on a data-driven manner. The future learning machines, however, will evolve into knowledge- and data-driven models. In other words, human-like machines will use both knowledge and data maximally. When these machines emulate functions and behaviors of human beings, they will outperform an average human being in utilization of knowledge and data. The knowledge- and data-driven models will bring us new challenges in the study of parameter identifiability. We list three challenges.

The first challenge is to add transparency (or interpretability) to a model. The study of adding interpretation to black box models is a broad field. For example, Hand (2006) showed that a simplistic interpretable approach that argues superiority of more sophisticated methods may be something of an illusion. Breiman (2001) argued that if the goal is to use data to solve problems, then one needs to adopt an algorithmic modeling approach and make use of a more diverse set of tools. This will become one of the most important issues for machines in the future. It is particularly true for studies on complex systems, such as biological, ecological, economic, financial, or social ones. These systems are largely black box in nature. Cherkassky and Dhar (2015) argued that model interpretation cannot be achieved by theoretical analysis of predictive models. That is, any meaningful interpretation should incorporate application domain knowledge outside data analysis. We wish to reveal their governing mechanisms, or physical insights, of the processes through artificial models. Parameter identifiability will be the key issue for achieving the transparency of a model, especially for the physical parameters in the knowledge-driven submodel.

The second challenge is to use any type of prior information. This challenge is derived from the first one. However, there exists no unified framework to embed any type of priors into a model because priors may impose one or a combination of limitations in a modeling approach (Hu et al., 2009), such as the diversity or lack of structures in the representation of knowledge. Hence, machine learning encounters generalized constraints (Zadeh, 2005). When a prior boils down to a finite-dimensional parameter, identifiability analysis will play a key role.

The third challenge is to select coupling operation between a knowledge submodel and a data submodel. The knowledge- and data-driven submodels present a new degree of freedom in the modeling approach: the selection of various coupling operations between the two submodels. The implementation of parameter identification can be changed with the coupling operations and with the criteria for model design.

While parameter identifiability is a classic subject in economics, dynamic control, machine learning, and other model-related fields, the new perspective from these three challenges will enlarge our understanding of this subject. The benefit from studying knowledge- and data-driven models is their unified framework, which includes traditional modeling approaches,

either first-principle model or black box model, as a special case. Much more work is expected to emerge for theoretical and practical aspects in the identifiability study from the new perspective. Mathematical statistics, machine learning, information theory, optimization theory, and other relevant areas must work together, hand in hand.

Acknowledgments

Many thanks to the reviewers for their valuable comments and advice. This work is supported in part by NSFC 61273196, 61620106003.

References

Akaho, S., & Kappen, H. J. (2000). Nonmonotonic generalization bias of gaussian mixture models. *Neural Computation*, *12*, 1411–1428.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *6*, 716–723.

Amari, S. (1967). Theory of adaptive pattern classifiers. *IEEE Transactions on Electronic Computers*, *16*, 299–307.

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, *10*, 251–276.

Amari, S., & Nagaoka, H. (2000). *Methods of information geometry*. New York: Oxford University Press.

Amari, S., Park, H., & Ozeki, T. (2006). Singularities affect dynamics of learning in neuromanifolds. *Neural Computation*, *18*, 1007–1065.

Audoly, S., D'Angio, L., Saccomani, M. P., & Cobelli, C. (1998). Global identifiability of biokinetic models of linear compartment models, a computer algebra algorithm. *IEEE Transactions on Biomedical Engineering*, *45*, 36–47.

Bekker, P. A., Merchens, A., & Wansbeek, T. J. (1994). *Identification, equivalent models and computer algebra*. Boston: Academic Press.

Bellman, R., & Astrom, K. J. (1970). On structural identifiability. *Mathematical Bioscience*, *7*, 329–339.

Berger, J. O. (1985). *Statistical decision theory and bayesian analysis* (2nd ed.). New York: Springer.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. New York: Oxford University Press.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Bowden, R. (1973). The theory of parametric identification. *Econometrica*, *41*, 1069–1074.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*, Cambridge: Cambridge University Press.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information conplexity. *Journal of Mathematical Psychology*, *44*, 62–91.

Brockett, R. W. (1976). Some geometric questions in the theory of linear systems. *IEEE Transactions on Automatic Control*, *21*, 449–455.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–231.

Castelle, D. D., & Gassiat, E. (1997). Testing in locally conic models, and application to mixture models. *Probability and Statistics*, *1*, 285–317.

Catchpole, E. A., & Morgan, B. J. T. (1997). Detecting parameter redundancy. *Biometrika*, *84*, 187–196.

Catchpole, E. A., Morgan, B. J. T., & Viallefont, A. (2002). Solving problems in parameter redundancy using computer algebra. *Journal of Applied Statistics*, *29*, 625–636.

Chen, A. M., Lu, H., & Hecht-Nielsen, R. (1993). On the geometry of feed-forward neural network error surfaces. *Neural Computation*, *5*, 910–927.

Cherkassky, V., & Dhar, S. (2015). *Interpretation of black-box predictive models: Measures of complexity*. New York: Springer.

Cherkassky, V., Dhar, S., & Dai, W. (2011). Practical conditions for effectiveness of the universum learning. *IEEE Transactions on Neural Networks*, *22*(8), 1241–1255.

Cherkassky, V., & Mulier, F. (2007). *Learning from data* (2nd ed.). Hoboken, NJ: Wiley.

Cole, D. J., Morgan, B. J. T., & Titterington, D. M. (2010). Determining the parametric structure of models. *Mathematical Biosciences*, *228*, 16–30.

Collobert, R., Sinz, F., Weston, J., & Botton, L. (2006). Large scale transductive SVMs. *Journal of Machine Learning Research*, *7*, 1687–1712.

Cousseau, F., Ozeki, T., & Amari, S. I. (2008). Dynamics of learning in multilayer perceptrons near singularities. *IEEE Transcations on Neural Networks*, *19*, 1313–1328.

Cover, T. M., & Thomas, J. A. (1991). *Element of information theory* (2nd ed.). Chichester: Wiley.

Dasgupta, A., Self, S. G., & Gupta, S. D. (2007). Nonidentifiable parametric probability models and reparameterization. *Journal of Statistical Planning and Inference*, *137*, 3380–3393.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, *74*, 33–43.

Dimattina, C., & Zhang, K. C. (2010). How to modify a neural network gradually without changing its input-output functionality. *Neural Computation*, *22*, 1–47.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). Hoboken, NJ: Wiley.

Ernesto, S. M., & Fernando, Q. (2002). Consistency and identifiability, revisited. *Brazilian Journal of Probability and Statistics*, *16*, 99–106.

Espinoza, M., Suykens, J. A. K., & Moor, B. D. (2005). Kernel based partially linear models and nonlinear identification. *IEEE Transactions on Automatic Control*, *50*, 1602–1606.

Fan, X. R., Kang, M. Z., Reffye, P. D., Heuvelink, E., & Hu, B. G. (2015). A knowledge-and-data-driven modeling approach for simulating plant growth: A case study on tomato growth. *Ecological Modelling*, *312*, 363–373.

Fortunati, S., Gini, F., Greco, M. S., Farina, A., Graziano, A., & Giompapa, S. (2012). On the identifiability problem in the presence of random nuisance parameters. *Signal Processing*, *92*, 2545–2551.

Fukumizu, K. (1996). A regularity condition of the information matrix of a multilayer perceptron network. *Neural Networks*, *9*, 871–879.

Fukumizu, K. (1999). Generalization error of linear neural networks in unidentifiable cases. In O. Watanabe & T. Yokomori (Eds.), *Algorithmic learning theory: Proceedings of the 10th International Conference on Algorithmic Learning Theory* (pp. 51–62). Berlin: Springer-Verlag.

Fukumizu, K. (2003). Likelihood ratio of unidentifiable models and multilayer neural networks. *Annals of Statistics*, *31*, 833–851.

Fukumizu, K., & Amari, S. (2000). Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural Networks*, *13*, 317–327.

Gallot, S., Hulin, D., & Lafontaine, J. (2008). *Riemann geometry* (3rd ed.). Berlin: Springer-Verlag.

Gimenez, O., Viallefont, A., Catchpole, E. A., Choquet, R., & Morgan, B. J. T. (2004). Methods for investigating parameter redundancy. *Animal Biodiversity and Conservation*, *27*, 1–12.

Hagiwara, K. (2002a). On the problem in model selection of neural network regression in overrealizable scenario. *Neural Computation*, *14*, 1979–2002.

Hagiwara, K. (2002b). Regularization learning, early stopping and biased estimator. *Neurocomputing*, *48*, 937–955.

Hagiwara, K., Toda, N., & Usui, S. (1993). On the problem of applying AIC to determine the structure of a layered feed-forward neural network. In *Proceedings of the International Joint Conference on Neural Networks* (pp. 2263–2266). Piscataway, NJ: IEEE.

Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science*, *21*(1) 1–14.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer-Verlag.

Haykin, S. (1998). *Neural networks: A comprehensive foundation* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Henao, R., & Winther, O. (2011). Sparse linear identifiable multivariate modeling. *Journal of Machine Learning Research*, *12*, 863–905.

Hironaka, H. (1964). Resolution of singularities of an algebraic variety over a field of characteristic zero. *Annals of Mathematics*, *79*, 109–326.

Hochwald, B., & Nehorai, A. (1997). On identifiability and information-regularity in parameterized normal distributions. *Circuits Systems Signal Processing*, *16*, 83–89.

Hotelling, H. (1939). Tubes and spheres in *n*-spaces, and a class of statistical problems. *American Journal of Mathematics*, *61*, 440–460.

Hu, B.-G. (2014). What are the differences between Bayesian classifiers and mutual-information classifiers? *IEEE Transactions on Neural Networks and Learning Systems*, *25*, 249–264.

Hu, B.-G., Qu, H.-B., & Yang, S.-H. (2009). A generalized-constraint neural network model: Associating partially known relationships for nonlinear regressions. *Information Sciences*, *179*, 1929–1943.

Hu, B.-G., Wang, Y., Yang, S.-H., & Qu, H.-B. (2007). How to add transparency to artificial neural networks. *Pattern Recognition and Artificial Intelligence*, *20*, 72–84.

Inoue, M., Park, H., & Okada, M. (2003). Online learning theory of soft committee machines with correlated hidden units: Steepest gradient descent and

natural gradient descent. *Journal of the Physical Society of Japan*, *72*, 805–810.

Jacquez, J. A., & Greif, P. (1985). Numerical parameter identifiability and estimability: Integrating identifiability, estimability, and optimal sampling design. *Mathematical Biosciences*, *77*, 201–227.

Kang, K., Oh, J. H., Kwon, S., & Park, Y. (1993). Generalization in a two-layer neural networks. *Physical Review*, *48*, 4805–4809.

Király, F., & Tomioka, R. (2012). A combinatorial algebraic approach for the identifiability of low-rank matrix completion. In *Proceedings of the International Conference on Machine Learning*. Piscataway, NJ: IEEE.

Koopmans, T. C., & Reierøsl, O. (1950). The identification of structural characteristics. *Annuals of Mathematical Statistics*, *21*, 165–181.

Kurková, V., & Kainen, P. C. (1994). Functionally equivalent feedforward neural networks. *Neural Computation*, *6*, 543–558.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Lehmann, E. L. (1983). *Theory of point estimation*. New York: Springer-Verlag.

Little, M. P., Heidenreich, W. F., & Li, G. (2009). Parameter identifiability and redundancy in a general class of stochastic carcinogenesis models. *PLoS ONE*, *4*, 1–6.

Little, M. P., Heidenreich, M. F., & Li, G. (2010). Parameter identifiability and redundancy, a theoretical considerations. *PLoS ONE*, *5*, 1–6.

Liu, X., & Shao, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *Annals of Statistics*, *31*, 807–832.

Ljung, L. (1999). *System identification: Theory for the user* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Magaria, G., Riccomagno, E., Chappell, M. J., & Wynn, H. P. (2001). Differential algebra methods for the study of the structural identifiability of rational function state-space models in bioscience. *Mathematical Biosciences*, *174*, 1–26.

Miao, H., Xia, X., Perelson, A. S., & Wu, H. (2011). On identifiability of nonlinear ODE models and applications in viral dynamics. *SIAM Review*, *53*, 3–39.

Milnor, J. (1985). On the concept of attractor. *Communications in Mathematical Physics*, *99*, 177–195.

Moore, T. J. (2010). *A theory of Cramér-Rao bounds for constrained parametric models*. Doctoral diss., University of Maryland.

Moore, T. J., & Sadler, B. M. (2004). Sufficient conditions for regularity and strict identifiability in MIMO systems. *IEEE Transactions on Signal Processing*, *52*, 2650–2655.

Murata, N., Yoshizawa, S., & Amari, S. I. (1994). Network information criterion-determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks*, *5*, 865–872.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. Cambridge, MA: MIT Press.

Nakajima, S., & Sugiyama, M. (2010). Implicit regularization in variational Bayesian matrix factorization. In *Proceedings of the International Conference on Machine Learning*. Piscataway, NJ: IEEE.

Nielsen, R. H. (1989). Theory of the back-propagation neural network. In *Proceedings of the International Joint Conference on Neural Networks*. Piscataway, NJ: IEEE.

Park, H., Inoue, M., & Okada, M. (2003). On-line learning dynamics of multilayer perceptrons with unidentifiable parameters. *Journal of Physics*, *36*, 11753–11764.

Paulino, C. D. M., & Pereira, C. A. D. B. (1994). On identifiability of parametric statistical models. *Journal of the Italian Statistical Society*, *3*, 125–151.

Phillips, P. C. B. (1989). Partially identified econometric models. *Econometric Theory*, *5*, 181–240.

Picci, G. (1977). Some connections between the theory of sufficient statistics and the identifiability problem. *SIAM Journal on Applied Mathematics*, *33*, 383–398.

Principe, J. C. (2010). *Information theoretic learning: Renyi's entropy and kernel perceptive*. New York: Springer.

Psichogios, D., & Ungar, L. H. (1992). A hybrid neural network: First principles approach to process modeling. *AIChE Journal*, *38*, 1499–1511.

Qu, Y. J., & Hu, B. G. (2011). Generalized constraint neural network regression model subject to linear priors. *IEEE Transactions on Neural Networks*, *22*, 2447–2459.

Ran, Z.-Y., & Hu, B.-G. (2014a). Determining structural identifiability of parameter learning machines. *Neurocomputing*, *127*, 88–97.

Ran, Z.-Y., & Hu, B.-G. (2014b). Determining parameter identifiability from the optimization theory framework: A Kullback-Leibler divergence approach. *Neurocomputing*, *142*, 307–317.

Ran, Z.-Y., & Hu, B.-G. (2015). An identifying function approach for determining parameter structure of statistical learning machines. *Neurocomputing*, *162*, 209–217.

Rattray, M., & Saad, D. (1999). Analysis of natural gradient descent for multilayer neural networks. *Physical Review*, *59*, 4523–4532.

Rattray, M., Saad, D., & Amari, S. (1998). Natural gradient descent for on-line learning. *Physical Review Letters*, *81*, 5461–5464.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, *11*, 416–431.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, *22*, 400–407.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica*, *39*, 577–591.

Ruger, S. M., & Ossen, A. (1997). The metric of weight space. *Neural Processing Letters*, *5*, 63–72.

Saad, D., & Solla, A. (1995). On-line learning in soft committee machines. *Physical Review*, *52*, 4225–4243.

Saccomani, M. P., Audoly, S., Bellu, G., & D'Angio, L. (2010). Examples of testing global identifiability of biological and biomedical models with the DAISY software. *Computers in Biology and Medicine*, *40*, 402–407.

Sato, M., & Shintani, T. (1974). On zeta functions associated with prehomogeneous vector space. *Annals of Mathematics*, *100*, 131–170.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization and beyond*. Cambridge, MA: MIT Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Shafarevich, I. R. (1974). *Basic algebraic geometry*. Berlin: Springer-Verlag.

Shao, J. (1999). *Mathematical statistics*. New York: Springer.

Shapiro, A. (1986). Asymptotical theory of overparameterized structure models. *Journal of the American Association*, *81*, 142–149.

Stoica, P., & Ng, B. C. (1998). On the Cramér-Rao bound under parametric constraints. *IEEE Signal Processing Letters*, *5*, 177–179.

Sundaram, R. K. (1996). *A first course in optimization theory*. Cambridge: Cambridge University Press.

Sussmann, H. J. (1992). Uniqueness of the weights for minimal feed-forward nets with a given input-output map. *Neural Networks*, *5*, 589–593.

Tallis, G. M., & Chesson, P. (1982). Identifiability of mixtures. *Journal of the Australian Mathematical Society*, *32*, 339–348.

Tikhonov, A. N., & Arsenin, V. Y. (1977). *Solution of ill-posed problems*. New York: Wiley.

Todorovski, L., & Dzeroski, S. (2006). Integrating knowledge-driven and data-driven approaches to modeling. *Ecological Modeling*, *194*, 3–13.

Vajda, S., Godfrey, K. R., & Rabitz, H. (1989). Similarity transformation approach to identifiability analysis of nonlinear compartmental models. *Mathematical Biosciences*, *93*, 217–248.

Van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes*. New York: Springer.

Vapnik, V. (1996). *The nature of statistical learning theory*. New York: Springer.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Vapnik, V. (2006). *Estimation of dependencies based on empirical data* (2nd ed.). New York: Springer-Verlag.

Vapnik, V., & Izmailov, R. (2015a). V-matrix method of solving statistical inference problems. *Journal of Machine Learning Research*, *16* (2015), 1683–1730.

Vapnik, V., & Izmailov, R. (2015b). Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research*, *16*, 2023–2049.

Walter, E. (1982). *Identifiability of state space models*. Berlin: Springer.

Walter, E., & Pronzato, L. (1997). *Identification of parametric models from experimental data*. London: Springer.

Watanabe, S. (2001a). Algebraic analysis for non-identifiable learning machines. *Neural Computation*, *13*, 899–933.

Watanabe, S. (2001b). Algebraic geometrical methods for hierarchical learning machines. *Neural Networks*, *14*, 1049–1060.

Watanabe, S. (2007). Almost all learning machines are singular. In *Proceedings of the IEEE Symposium on Foundations of Computational Intelligence*. Piscataway, NJ: IEEE.

Watanabe, S. (2009). *Algebraic geometry and statistical learning theory*. Cambridge: Cambridge University Press.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

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

Watanabe, S., & Amari, S. (2003). Learning coefficients of layered models when the true distribution mismatches the singularities. *Neural Computation*, *15*, 1013–1033.

Wei, H. K., & Amari, S. I. (2008). Dynamics of learning near singularities in radial basis function networks. *Neural Networks*, *21*, 989–1005.

Weyl, H. (1939). On the volume of tubes. *American Journal of Mathematics*, *61*, 461–472.

Whiley, M. (1999). *Aspects of the interface between statistics and neural networks*. Doctoral diss., University of Glasgow, Scotland.

White, H. (1989a). Learning in artificial neural networks: A statistical perspective. *Neural Computation*, *1*, 425–464.

White, H. (1989b). An additional hidden unit test for neglected nonlinearity in multilayer feedforward networks. In *Proceedings of the International Joint Conference on Neural Networks*. Piscataway, NJ: IEEE.

White, H. (1994). *Estimation, inference and specification analysis*. Cambridge: Cambridge University Press.

Xia, X., & Moog, C. H. (2003). Identifiability of nonlinear systems with application to HIV/AIDS models. *IEEE Transactions on Automatic Control*, *48*, 330–336.

Yang, S.-H., Hu, B.-G., & Cournde, P. H. (2008). Structural identifiability of generalized-constraint neural network models for nonlinear regression. *Neurocomputing*, *72*, 392–400.

Yao, Y. W., & Giannakis, G. (2005). On regularity and identifiability of blind source separation under constant modulus constraints. *IEEE Transactions on Signal Processing*, *53*, 1272–1281.

Zadeh, L. A. (2005). The concept of a generalized constraint—a bridge from natural languages to mathematics. In *Proceedings of the Annual Meeting of the North American Fuzzy Information Processing Society* (pp. 1–2). Piscataway, NJ: IEEE.

Zorich, V. A. (2004). *Mathematical analysis*. Berlin: Springer-Verlag.

Received September 23, 2016; accepted December 18, 2016.