INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

NOTE TO USERS

This reproduction is the best copy available.

UMI®

Université d'Ottawa • University of Ottawa

Uncertainty Management in the Activated Sludge Process

Innovative Applications of Computational Learning Theory

by
Abdelaziz Guergachi

A thesis submitted
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in
Engineering

Ottawa-Carleton Institute of Civil Engineering
University of Ottawa
Ottawa, Canada
K1N 6N5
July, 2000

Copyright © 2000 by A. Guergachi

National Library of Canada
Bibliothèque nationale du Canada

Acquisitions and Bibliographic Services
Acquisitions et services bibliographiques

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-58278-7

Canada

À

mon père, l'architecte de ma formation académique
depuis les classes primaires.

et

ma mère, pour qui mon bien-être passe devant tout.

Acknowledgments

It would not be correct to not start these acknowledgments by thanking the Creator of the Universe, the One who knows every thing about microorganisms and for whom uncertainty is meaningless. He has provided me with all the required energy to carry out this work.

This thesis is highly interdisciplinary and would not have become possible without the enlightening guidance of my supervisor Professor Gilles G. Patry. Owing to his wide knowledge in engineering, computer science and mathematics, I was able to go with my research work where no one has ever gone before in the area of wastewater engineering. I am deeply indebted to Gilles and I would like to express my sincere gratitude to him.

I also want to thank my co-supervisor Professor Eric J. Schiller for his valuable and pertinent comments, and help throughout my research.

My first contact with the subject of artificial intelligence was through Professor Emil Petriu's courses in fuzzy logic systems and neural networks. My discussion with Emil on applying these new technologies to engineering systems analysis was very enlightening. Emil was the one who offered me the book where I first discovered the theory that demystifies the power of neural networks. I would like to take this opportunity to thank him.

When I first started working on the applications of computational learning theory, my only source of information about this theory was published papers until I met Professor Mario Marchand. Through my discussion with Mario, I could get live information about the progress of the research on computational learning theory in the world. Mario has also provided me with a great deal of valuable comments with regard to my research work. I am much indebted to him and I wish to thank him sincerely.

I would also like to thank Professor Nasir Uddin Ahmed for his valuable comments and explanations with regard to the model identifiability issue. His thorough analysis of this issue has convinced me that model identifiability is not an essential criterion for model evaluation and selection.

I am also grateful to Dr. Vladimir N. Vapnik and Dr. David Haussler for their prompt responses to my e-mails.

Finally, I thank my wife, parents and other members of my family for their support, patience and care, as well as all my office mates with whom I had spent some of the best moments of my life while I was writing this thesis.

The work has been supported financially by CIDA and NSERC. This support is gratefully acknowledged.

Abstract

In this thesis, the foundations of a new area of research regarding mathematical modelling of biological wastewater treatment (WWT) processes are set. The main feature of this area is the introduction of innovative concepts and tools from emerging information modelling technologies into the traditional field of WWT process modelling. The model identification procedure is viewed as a learning problem or, equivalently, an information transfer from a set of real data into the process model.

An innovative mathematical framework for the identification and validation of dynamic mechanistically-based WWT process models is developed. Within this framework, a relationship between the model identification procedure and the computational machine learning methodology is established at the foundational level. The deviation $\mathcal{D}$ between model prediction and the real process behaviour is characterized mathematically in terms of some simple variables that govern model performance — namely:

- the size of the data set used for model identification
- the quality of these data
- the model complexity
- the empirical measure of $\mathcal{D}$ computed on the basis of the foregoing data set

The development of the relationship between $\mathcal{D}$ and these variables is based on a principle called "*Inductive Principle of Empirical Risk Minimization*" ($\mathcal{IPERM}$). The conditions of applicability of $\mathcal{IPERM}$ are thoroughly examined in the case of the activated sludge process being described by a simple mechanistic model denoted $\mathcal{M}$. The Vapnik-Chervonenkis (VC) dimension of this model is estimated and two uncertainty models are developed for the activated sludge process (ASP).

These two uncertainty models are compared and the differences between them accounted for. The following result is established: empirical data *cannot* compensate for our limited

knowledge of process mechanisms, even if an infinite amount of data and computing power are made available during the model identification procedure.

Measures of process model maximal and marginal improvements are developed. It is established that 80% of the model ($\mathcal{M}$) maximal improvement occurs at a number of data points of about $N_{80\%} \approx 15$ to 18. To achieve the other 20%, N has to be increased from the relatively small number $N_{80\%}$ to infinity.

Procedures for computing the marginal cost of process model improvement and the guaranteed prediction accuracy of the identified model are developed.

A new approach to modelling the activated sludge process itself and dealing with the almost-infinite degree of complexity of the ASP behaviour is developed. The basic idea of this approach is to construct an infinite series $\mathcal{NS}$ of nested mechanistic models of increasing complexity. This nested series is developed using the *multi-substrate hypothesis*. Both the Monod and the Tiessier models are considered in developing this nested series.

Another principle called "*Inductive Principle of Structural Risk Minimization*" ($\mathcal{IPSRM}$) is introduced and implemented to determine the optimal model structure complexity, for a fixed and small number N of data points. Computer simulations are carried out to confirm the theory and illustrate the use of the $\mathcal{IPSRM}$ and that of the nested series $\mathcal{NS}$ of ASP models.

Contents

Part I

THESIS MOTIVATION AND CONTRIBUTIONS

Chapter 1

General Introduction

1.1 Motivation and Problem Statement

Every dynamic system generates wastes in the course of its activity, and urban systems are no exception. They produce liquid, solid and gaseous wastes. The production of all these forms of wastes has increased significantly in the western world during the last 50 years. In particular, the daily volume of waste-water has recently reached 14 Mm^3 in Canada (Environment Canada, 1997), about 100 Mm^3 in the United States and 40 Mm^3 in Europe (Lens and Verstraete, 1992).

By their nature, wastewaters are obnoxious and their presence leads to unhealthy environments. If they are allowed to accumulate, the risk of water supply contamination would become very high and waterborne infections would propagate very fast among the population. If they are discharged without treatment into receiving waters, the latter would become unsuitable for the survival and propagation of fish and other aquatic life, and even navigation. For these reasons, the immediate and nuisance-free removal of wastewaters from their sources of generation, followed by treatment and disposal is not only desirable but necessary for the sustainable development of any community. In the foregoing process of controlling the impact of liquid wastes — *collection, treatment and disposal* —, municipal wastewater treatment (WWT) plants play a key role. They provide an important buffer between the concentrated wastewater from urban areas and the natural environment. All together, they constitute the largest industry in terms of treatment of raw materials in the world. However, numerous studies have reported that even well attended (operated) wastewater treatment plants sometimes fail to comply with the regulatory effluent standards. For example:

- The U.S. Environmental Protection Agency estimated that one third of existing municipal wastewater disposal facilities failed to meet the required effluent quality standards (Novotny and Capodaglio, 1992).
- Gall and Patry (1988) reported that some studies found that 50% to 87% of plants investigated regularly violate regulatory treatment standards.
- Berthouex *et al.* (1985) reported that identifiable upsets occurred up to 9% of the time in 15 well operated plants.
- Huang (1981) reported that as many as 50% of newly constructed biological wastewater treatment plants do not meet effluent standards.

While the wastewater engineering community may agree that the main reasons for this lack of compliance are closely related to plant operation, there is a great deal of confusion as to what should be done to resolve this issue.

Because of the dynamic nature of the influent to WWT plants, engineers dealing with the WWT process have one single objective: *controlling the process.* This means that engineers want to be able to operate (i.e., manipulate) the plant in order to meet environmental and economical requirements. They want to keep the effluent contaminant concentrations *less* than those specified by the governmental laws, but at the same time high enough to accomplish treatment at minimum cost. To achieve this, it is necessary to understand the *dynamical behaviour* of the WWT process that is being used in the plant. Information about this behaviour is obtained from three possible sources:

- fundamental principles of physics, chemistry and biochemistry
- experimental data
- empirical qualitative and quantitative knowledge

Information from these sources is combined, processed and presented in various forms such as *traditional mechanistic models, neural network models* or *fuzzy logic models.* A control strategy for the considered WWT process is then developed based on one of such models. It is however a fact that, among all the WWT process models that have been investigated by researchers, no one model can be considered to describe the process in its entirety. In other words, available models do not explain fully all mechanisms that govern the dynamics of the system and, as a result, their predictions do not always match with the reality. As a consequence, a fundamental question arises: *how far (or close) is a process model from the true dynamical behaviour of the process?*

This is an ambitious and very challenging question that has never been dealt with in the area of wastewater engineering. Addressing it for the case of mechanistic models is one of the main subjects of this thesis.

1.2 Objectives and Contributions of the Thesis

Since the development of the IAWPRC model (Henze *et al.*, 1987), research in the area of biological WWT process mathematical modelling has focused on:

- model identification and identifiability:

  Rationale: the design of process control strategies requires models that are uniquely identifiable, i.e., models for which a unique set of parameters can be determined through the model identification procedure (Jeppsson, 1996). Most WWT process models do not meet the identifiability criterion.

- model verifiability:

  Rationale: For a model to be truly verifiable, all its state variables have to be directly measurable (Jeppsson, 1996). This is not the case for most existing models which are then considered to be only partly verifiable.

- model reduction:

  Rationale: Complex models with many parameters are generally difficult to identify uniquely, hence the need for reduced-order models that may not describe the full dynamics of the system (Jeppsson, 1996).

In its broad perspective, the goal of this thesis is to set the foundations of a new area of research regarding mathematical modelling of biological WWT processes. The main feature of this area of research is the introduction of innovative concepts and tools from emerging information modelling technologies into the traditional field of WWT process modelling. The model identification procedure is viewed as a learning problem or, in other words, an information transfer from a set of real data to the process model. The hypothesis is that model identifiability is not considered an essential criterion for model evaluation and selection. In this thesis, model evaluation is based on only one criterion: performance. Model performance is measured by the mathematical deviation $\mathcal{D}$ between reality and model prediction. The issue of *"simple models versus complex ones"* is resolved into the determination of an optimal model structure complexity. Finally, for the model verifiability problem, it is suggested to replace the condition *"measurability of all model state variables"* by a less stringent observability criterion. This criterion does not require that all state variables be measurable directly and separately, as long as the unmeasurable ones are uniquely determinable from those that are measurable.

In more specific terms, this thesis is about developing an innovative mathematical framework for the identification and validation of dynamic mechanistically-based WWT process models. Within this framework, a relationship between the model identification procedure and computational machine learning methodology is to be established at the foundational level. The deviation $\mathcal{D}$ between model prediction and the real process behaviour is to be characterized mathematically in terms of some simple variables that govern model performance — namely:

- the size of the data set used for model identification
- the quality of these data
- the model complexity
- the empirical measure of $\mathcal{D}$ computed on the basis of the foregoing data set.

This characterization will be carried out under the assumption of *"maximum uncertainty"*, also called *"worst case scenario"*. Probability theory, mathematical statistics and stochastic processes analysis will be used extensively. The effect of an increase in the size of the data set on model improvement will be evaluated quantitatively.

Optimal model structure complexity will be determined in the case where a fixed amount of data is available for model identification. To illustrate the use of the proposed framework, an application for the case of heterotrophic biodegradation will be developed.

The contributions of this thesis are numerous. They are summarized in Chapter 11, along with suggestions of topics for further research. Here, we report on the major ones:

1. *In the area of wastewater engineering:*

   - before the publication of this thesis, there was no general approach for a true model validation in wastewater engineering. There were only empirical approaches for what we can call *"history-matching"* and *"benchmarking"*. With the publication of this thesis, an innovative, simple and well-founded methodology for model quality evaluation has become available to wastewater researchers and practicing engineers (WRPE).
   - With the publication of this thesis, WRPE are now able to determine the minimum amount of data required for WWT process model identification under "maximum uncertainty".
   - WRPE are now able to determine the optimal model structure complexity from a "nested" family of WWT process models, using a simple equation based on model performance.

2. *In the area of computational machine learning theory:*

   - this thesis represents the first study that develops an engineering application of computational learning theory to a dynamical system.
   - this is the first attempt to use the solutions of a parameterized differential equation for the definition of the learning machine.

This thesis relies heavily on some advanced and fairly complex mathematical tools. To make these tools easily understandable to the reader, efforts were made to simplify the mathematical component of the thesis. Numerous examples are presented to illustrate the new concepts and concrete explanations are provided whenever it is possible. However a minimum of mathematical knowledge is still required of the reader to fully understand the subject of this thesis. The reader is referred to the graduate textbook: *"Probability, Random Processes, and Estimation Theory for Engineers"*, by Stark and Woods (1994) for the necessary basic concepts.

In science and engineering, one often hears comments and assertions such as "applied mathematics is bad mathematics" or "the only really useful mathematics is the elementary mathematics". The history of science and engineering has proven, however, that this is not always the case. A famous counterexample to these assertions is the Kalman-Bucy filter: this tool was developed in the early sixties by R.E. Kalman and R.S. Bucy on the basis of some advanced mathematical theories of stochastic differential equations, and despite the complexity of the theoretical framework, the Kalman-Bucy filter has found almost immediately applications in aerospace engineering (Ranger, Mariner, Apollo etc.). Another recent counterexample to the foregoing assertions is the Fermat's last theorem in number theory. This theorem was stated by Pierre de Fermat in 1630, but remained unproven until October 1994. The result of this theorem as well as the tools that were developed for its proof have found applications in the area of cryptography. The results of this thesis represent also a counterexample to the same assertions: it relies on advanced mathematics, yet it should have widespread applications in several different situations of WWT process modelling, operation and control. Here are two practical cases where the results of this thesis can be extremely useful:

- **Practical Case n° 1**: *Consider the case where the manager of an activated sludge wastewater treatment plant has formulated a request to his consulting engineer, asking him to develop a model for the plant. Also the manager specified that he is able to supply as much data as the consultant may need and that he wants to obtain the most accurate model with the lowest risk possible. Because this model is going to be part of a comprehensive control system using the most recent control technologies and because the manager is aware of the difficulty in modelling the complete dynamics of the process, he asked the consultant to provide him with the* ***prediction accuracy*** *and the* ***degree of confidence*** *associated with this accuracy. The manager needs this information (accuracy and confidence) to compare the performance of the model of his consultant to other empirical models that have been developed for his plant by his own employees.*

- **Practical Case n° 2**: *Consider the case where the manager of the wastewater treatment plant cannot supply an unlimited amount of data. He responds that actually only a sequence of about* $N_0 = 10$ *input-output empirical data are available. Still, the consultant is asked to develop the best possible* ***model structure*** *based on this limited amount of data. The manager is aware of the problem of developing accurate models with very few data, especially for a highly complex and uncertain system such as the ASP. He knows that the risk associated with this model is likely to be high, but he is expecting the consultant to extract maximum information from the small amount of data supplied to him, so that the lowest possible risk under such circumstances can be achieved.*

Table 1.1: Objective of each part of this thesis

Part number	Objective
Part I (Chapter 1)	Defining motivation and contributions
Part II (Chapters 2 and 3)	Explaining the ideas that distinguish the thesis approach to mathematical modelling
Part III (Chapters 4, 5 and 6)	Developing the mathematical framework
Part IV (Chapters 7, 8, 9 and 10)	Applications to process uncertainty management
Part V (Chapter 11)	Summarizing the results

1.3 Thesis Outline

In order to help the reader locate the areas of interest to him in this thesis, the latter has been organized in five different parts. The objective of each part is described in Table 1.1. Parts I and V are the introduction and conclusion, respectively. Parts II, III and IV represent the core of the thesis. Part II explains why a mathematical framework is needed for the ASP, Part III develops this framework and Part IV applies it to the ASP case. The scope of each chapter is presented in Table 1.2.

The third column of Table 1.2 indicates the degree of abstraction and "amount of theory" utilized in each chapter. Chapter 6 (with the darkest shade of gray) is the most theoretical one. It presents the mathematical framework for uncertainty modelling and management. As we move away from this Chapter towards either the introduction or the conclusion, the thesis becomes less theoretical and more concrete. As part of the introduction side, the reader will find a series of practical questions that are crucial to the task of ASP modelling and operation. As part of the conclusion side, the reader will find practical and systematic answers to these questions.

Table 1.2: Scope of the thesis chapters

Part	**Chapter**	**Scope**	**Degree of abstraction**
I	Chapter 1	Introduction	
II	Chapter 2	Identifiability, verifiability, observability, validation	
	Chapter 3	Definition of research needs for ASP modelling	
III	Chapter 4	Intuitive explanation of uncertainty models	
	Chapter 5	Overfitting, process noise, learning machine, process response function	
	Chapter 6	Inductive principle of empirical risk minimization	
IV	Chapter 7	Uncertainty models for ASP	
	Chapter 8	Uncertainty management when process model is fixed and N variable	
	Chapter 9	A family of ASP mechanistic models	
	Chapter 10	Uncertainty management when N is fixed and process model is variable	
V	Chapter 11	Conclusion	

Key: *the darker the shade of gray, the higher the degree of abstraction.*

Part II

MATHEMATICAL MODELLING FUNDAMENTALS

Chapter 2

Process Modelling: a New Perspective

This Chapter lays out the fundamental ideas that drove the development of this thesis, namely:

- Model identifiability is not a major issue in process mathematical modelling.
- Model verifiability is a very demanding criterion that can be replaced by a less stringent one: model observability.
- The issue of "complex models versus reduced-order models" is to be resolved by introducing a new concept: optimal model complexity.
- The traditional procedures of model validation are not adequate and a mathematical framework for model quality evaluation is needed.

2.1 Definitions

Consider a system $\mathcal{S}$ whose state space $\mathcal{X}$ is a finite dimensional one and assume that this system is described by a mathematical model $\mathcal{M}_{\mathcal{S}}$ of the form:

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}, \mathbf{p}) \tag{2.1}$$

where $\mathbf{x}$ is the system state vector, $\mathbf{p}$ is the parameter vector, t is the time and $\mathbf{f}$ is a mathematical function which is generally nonlinear.

A fundamental problem in system modelling is the determination of the values of model parameters such that the corresponding response of the model equation approximates as closely as possible the actual response of the physical system. Assume that the response of the physical system is given in the form of a set of real data:

$$\Upsilon_N : \quad \mathbf{x}^{data}(t_1), \mathbf{x}^{data}(t_2), \ldots, \mathbf{x}^{data}(t_N) \tag{2.2}$$

The mathematical procedure for determining the model parameters on the basis of a set Υ_N of data is called *model identification* (or calibration). Traditionally, it consists in minimizing an objective function $J(\mathbf{p})$ such that:

$$J(\mathbf{p}) = \sum_{k=1}^{N} \|\mathbf{x}(\mathbf{p}, t_k) - \mathbf{x}^{data}(t_k)\|^2 \tag{2.3}$$

where $\mathbf{x}(\mathbf{p}, t)$ represents the solution to the model equation 2.1. If there is only one unique minimum for J then the system is defined as *identifiable* (Jeppsson, 1996). A system model is said to be *verifiable* if all its state variables are directly measurable (Jeppsson, 1996). After a model is developed and identified, we need to know how well it mimics the true system behaviour. The procedure of verifying this property is called *model validation*.

2.2 The Model Identifiability Issue

The lack of identifiability has been considered a handicap for process models in the area of wastewater engineering. The most complete work on WWT process model identifiability is that of Jeppsson (1993, 1996). Using a simple example, Jeppsson showed that models that use the Monod equation are not identifiable. He then developed a set of reduced models for which he investigated the identifiability using computer simulations. He did not provide, however, a formal proof of the identifiability of these models.

In this thesis, it is argued that model identifiability is not a major problem in mathematical modelling: a model that is not identifiable can still be useful if it produces a good performance. The arguments in favour of this view are two-fold:

1. **Models of complex systems are practically impossible to uniquely identify.** This is a fact. When Beck (1986) pointed out the lack of identifiability of the IAWPRC model, he immediately added that *"there is nothing unusual*

in this, for the same problem is widespread in the environmental sciences and in the adjacent disciplines of pharmacokinetics and biomedical system analysis. ... It is well known that there are difficulties with structural identifiability of biochemical process models, specifically in association with the use of the Monod expression". The lack of model identifiability is due to the fact that available models explain just a portion of the behaviour of highly complex systems. The other portion which is *not* accounted for by those models shows itself through the variability of model parameters. Lack of identifiability is then an inherent feature of complex systems.

Systems that are identifiable are usually associated with some unique values of parameters called *universal constants*. Other identifiable systems do not give rise to universal constants, but their model parameters are always reported to take the same unique values by all researchers. There are numerous systems with such property in physical sciences. Here is a list of a few of them:

- *Interaction of two electric charges:* the magnitude of the force F resulting from the interaction in a free space of two charges q_1 and q_2 separated a distance r is expressed as:

$$F = p\frac{q_1 q_2}{r^2}$$

  where p is a parameter with a unique numerical value: $p = 1/4\pi\epsilon_0$, ϵ_0 being equal to 8.854×10^{-12} SI

- *Interaction of two bodies:* the magnitude of the force F of attraction between any two bodies is given by:

$$F = p\frac{m_1 m_2}{r^2}$$

  where p is a parameter with a unique numerical value: $p = 6.670 \times 10^{-11}$ SI, m_1 and m_2 are the masses of the two bodies and r is the distance between them.

- *equation of state of gases:* For ideal gases, this equation is:

$$PV = pnT$$

  where p is a parameter with a unique numerical value: $p = 8.31434 \times 10^3$ SI, P is the pressure, V is the volume, T is the absolute temperature and n is the number of moles. As the pressure gets higher and temperature close to the gas boiling point, the gas becomes non-ideal and governed by the equation:

$$\left(P + p_1 \frac{n^2}{V^2}\right)(V - p_2 n) = p n T$$

where $p = 8.31434 \times 10^3$ SI and p_1 and p_2 are two constant parameters that are gas-specific.

- *The hydrogen-bromine system:* The reaction rate model for this system ($H_2 + Br_2 \rightarrow 2HBr$) is as follows:

$$r_{HBr} = \frac{1}{2}\left(\frac{p_1[H_2][Br_2]^{1/2}}{1 + p_2([HBr]/[Br_2])}\right)$$

  where p_1 and p_2 are two uniquely determinable parameters: $p_1 \propto e^{E_a/RT}$ with $E_a = 175$ SI and $R = 8.31434 \times 10^3$ SI, and $p_2 = 0.1$ (temperature independent)

Identifiable systems, such as the ones presented above, have all one common property that can be expressed qualitatively in the following way.

Similarly to what Cohen and Stewart did in their book *The Collapse of Chaos* (1994), let us imagine that the information content of a system $\mathcal{S}$ can be measured by one single number $\mathcal{I}(\mathcal{S})$. Considering the model $\mathcal{M}_\mathcal{S}$ that is used to describe $\mathcal{S}$ as a mathematical system, its information content can also be measured by a number $\mathcal{I}(\mathcal{M}_\mathcal{S})$. Identifiable systems have the property that the two quantities of information $\mathcal{I}(\mathcal{S})$ and $\mathcal{I}(\mathcal{M}_\mathcal{S})$ are almost equal. In more concrete terms, identifiable systems have models that explain practically all mechanisms governing the system behaviour. This is the case of all the foregoing examples of identifiable systems (*electrostatic interaction, gravitation, equation of state of gases, chemical system H_2/Br_2, ...*). However, in the case of a highly complex system, the quantity of information $\mathcal{I}(\mathcal{M}_\mathcal{S})$ is always strictly less than $\mathcal{I}(\mathcal{S})$, meaning that the model $\mathcal{M}_\mathcal{S}$ does not account for all modes of the system behaviour. Jeppsson (1996) has expressed this fact very rightly for the case of the activated sludge WWT process: *"Though available models are quite complex they are still greatly simplifying the representation of many species of organisms. As the microbial population changes this needs to be reflected in changing* *kinetic parameters and even adding new state variables"*. It is therefore the existence of a significant portion of the system behaviour not accounted for by the model $\mathcal{M}_S$ that renders the latter non-identifiable. Model identifiability is then practically unavoidable in the case of complex systems. We have to live with it.

2. **Model identifiability is not needed for systems control anyway.** What is wrong with a system model that is not identifiable, but produces an acceptable performance in predicting the system behaviour? Nothing, if there is a mathematical guarantee on the model performance. This thesis is about developing an innovative framework to help us derive and establish such a guarantee. With this framework developed and the guarantee established, model identifiability becomes irrelevant.

   Moreover, it should be noted that some emerging modelling technologies have also showed that model identifiability is not essential (Haykin, 1994). Neural networks, for instance, are fundamentally non-identifiable, yet they have been used extensively and successfully in several areas such as pattern recognition.

2.3 Verifiability versus Observability

Model verifiability requires that all state variables be directly measurable. This is very demanding and not at all practical. Most systems, and especially complex ones, have indeed variables that are easily and directly measurable and others that are difficult or even impossible to measure. Because of that, it is suggested here to introduce a less stringent criterion called *model observability*. With this criterion, the system state variables do not have to be measured directly and separately. Rather, they are considered as "hidden variables" and have somehow to be inferred from what can be measured (output).

Textbooks have defined the observability concept in several different ways which are all equivalent to each other (Maybeck, 1979; Ahmed, 1988; Borrie, 1992). Here we will consider the following definition:

Re-write the model equation 2.1 of the system $\mathcal{S}$ in a more general form:

$$\begin{cases} \dot{\mathbf{x}} &= \mathbf{f}(t, \mathbf{x}, \mathbf{u}, \mathbf{p}) \\ \mathbf{y} &= \mathbf{h}(\mathbf{x}) \end{cases} \tag{2.4}$$

where $\mathbf{u}$ is a vector of input control variables (usually called either "input" or "control"), $\mathbf{y}$ is a vector of measured outputs (usually called just "output") and $\mathbf{h}$ is a function of the state vector $\mathbf{x}$. The system $\mathcal{S}$ model is said to be *observable* if, given $\mathbf{y}(t)$ and $\mathbf{u}(t)$ for all $t \in [t_0, t_1]$, it is possible to uniquely determine the state vector $\mathbf{x}(t)$ for all $t \in [t_0, t_1]$. Thus, the system model is observable if any state variable $x_i(t)$ can be determined for $t \in [t_0, t_1]$ from the knowledge of only the input and output over the interval $[t_0, t_1]$. The structure of an observable model must then be such that the output $\mathbf{y}(t)$ is affected in some manner by the change of any single state variable. In addition, the effect of any one state variable on the output must be distinguishable from the effect of any other state variable.

Here is a simple example to illustrate the concept of observability intuitively (adapted from Ahmed, 1988):

Example 1. Consider a system governed by the model equations:

$$\begin{cases} \dot{x}_1 &= -p_1 x_3 + u_1 \\ \dot{x}_2 &= u_2 \\ \dot{x}_3 &= p_1 x_1 + u_3 \end{cases}$$

where x_1, x_2 and x_3 are the state variables, u_1, u_2 and u_3 are the inputs and p_1 is the model parameter ($\mathbf{x} = (x_1\ x_2\ x_3)^T$, $\mathbf{u} = (u_1\ u_2\ u_3)^T$, $\mathbf{p} = (p_1)$). Consider now the two following cases:

- **Case i:** the only variable that is measured is the sum of x_1 and x_2. In other words, the output $\mathbf{y}$ is a scalar and equal $x_1 + x_2$. The model equations are then:

$$\begin{cases} \dot{x}_1 &= -p_1 x_3 + u_1 & (I) \\ \dot{x}_2 &= u_2 & (II) \\ \dot{x}_3 &= p_1 x_1 + u_3 & (III) \\ y &= x_1 + x_2 & (IV) \end{cases}$$

Can we reconstruct x_1, x_2 and x_3 from the knowledge of the values of only y and the controls u_1, u_2 and u_3? Yes. This is how we can proceed:

Add equations (I) and (II). We get:

$$\dot{y} = -p_1\, x_3 + u_1 + u_2$$

Assuming of course that $p_1 \neq 0$, then:

$$x_3 = \frac{u_1 + u_2 - \dot{y}}{p_1}$$

From equation (III), we get x_1:

$$x_1 = \frac{1}{p_1}(\dot{x}_3 - u_3) = \frac{1}{p_1}\left(\frac{\dot{u}_1 + \dot{u}_2 - \ddot{y}}{p_1} - u_3\right)$$

From equation (IV), we obtain x_2:

$$x_2 = y - x_1 = y - \frac{1}{p_1}\left(\frac{\dot{u}_1 + \dot{u}_2 - \ddot{y}}{p_1} - u_3\right)$$

Thus, this example shows that there is no need to measure directly and separately all the three state variables x_1, x_2 and x_3. If just the sum of x_1 and x_2 is measured (and the values of the control variables are assumed to be known to the operator, because she manipulates them), it is possible to uniquely reconstruct estimates $\hat{x}_1$, $\hat{x}_2$ and $\hat{x}_3$ for all the three state variables, using the system input and output as the basis of this estimation. Because of this, the foregoing system (equations (I), (II), (III), (IV)) is observable. In practice, to check the observability criterion for linear systems, we just determine the rank of one matrix called the observability matrix (see for instance Ahmed, 1988). There is no need to go systematically through the above algebraic calculations.

- **Case ii:** the only variable that is measured is x_1. In other words, the output $\mathbf{y}$ is a scalar and equal x_1. The model equations are then:

$$\begin{cases} \dot{x}_1 &= -p_1\, x_3 + u_1 & (I) \\ \dot{x}_2 &= u_2 & (II) \\ \dot{x}_3 &= p_1\, x_1 + u_3 & (III) \\ y &= x_1 & (IV) \end{cases}$$

As said previously, it is easy to check the observability criterion of this linear system by computing the rank of the observability matrix (Ahmed, 1988) and establish that the system is not observable. In this example, the observability criterion is again examined intuitively:

From equations (I) and (IV), we determine x_3:

$$x_3 = \frac{u_1 - \dot{y}}{p_1}$$

The variable x_1 is also uniquely determinable from equation (IV):

$$x_1 = y$$

But, the variable x_2 is however is not uniquely determinable; any function:

$$x_2 = x_{2_0} + \int u_1$$

with $x_{2_0} \in \Re$ is acceptable. Therefore, the model is not observable.

When a system model is observable, a *state observer* can usually be designed to generate an estimate of $\mathbf{x}$ using $\mathbf{u}$ and $\mathbf{y}$ as the basis for that estimation (Borrie, 1992).

The concepts of observability and observer were first introduced by Kalman in the early sixties, but they have never been implemented in the area of WWT process mathematical modelling. The study of observability of linear systems is quite straightforward. It is however a challenging mathematical subject in the case of nonlinear systems such as WWT processes.

2.4 Complex models or reduced-order models?

"An 'optimal' model incorporates all of the important dynamic effects, is no more complicated in its structure than necessary, ..." (Jeppsson, 1996). This is just another statement of the celebrated principle of simplicity commonly attributed to William of Ockham (1290? - 1349?) and known as Occam's razor: "If there are alternative explanations for a phenomenon, then, all other things being equal, we should select the simplest one" (Li and Vitányi, 1993).

In the case of the behaviour of biological WWT processes, however, we are faced with a complexity that is unparalleled in the chemical industry (Jeppsson, 1996). The reactions occuring in a bioreactor involve indeed hundreds of different types of microorganisms biodegrading a multitude of different organic waste compounds. A simple bacterial cell in this bioreactor, *E. coli* for instance, has about 2 500 different *kinds* of macromolecules and contains about 24 million individual molecules (Madigan *et al.*, 1997). *E. coli*, as well as the other microorganisms present in the bioreactor, have to synthesize all of these molecules from the organic wastes so that they can grow and generate other organisms. In the course of this synthesis process, a large number of reactions take place involving the use of a multitude of different kinds of enzymes. It is the author's opinion that a description with *scientific* accuracy of the bioreactor dynamics is unlikely to result in a model with a finite number of state variables and parameters.

However, even if such accurate and highly complex model were possible to develop, it would be useless from an engineering viewpoint. The reason for this is not the identifiability problem (as it is suggested by Jeppsson (1996)), but it is the *data scarcity.* If the size of the data set used for model identification is small while the number of model parameters is large (i.e., the model is highly complex), then the problem of *data overfitting* by the model would occur. On the other hand, if the model is too simple and, therefore, the number of parameters is sufficiently low compared to the size of the data set, then the explanatory power of the model would be so low that the value of the objective function 2.3 would become very high, meaning the model prediction of the true process behaviour is of a low quality. Consequently, the degree of complexity of a process model has to be adjusted to the amount of data that is available for the identification of this model. For any fixed amount of data, there is an *optimal* model complexity that has to be determined. Models that are more complex would cause overfitting, and models that are less complex would lead to a low prediction quality.

The mathematical framework that is developed in this thesis defines all the necessary concepts and tools that help determine the optimal structure complexity of a WWT process model, corresponding to a fixed amount of data. In the following paragraphs, a qualitative explanation of the idea behind this framework is presented using some simple metaphors.

The identification procedure is viewed as an information transfer from a set of real data into the process model to be identified. Any data set Υ carries a certain amount of information $\mathcal{I}(\Upsilon)$ about the true process behaviour. This amount of information is characterized by the quantity and quality of the data. The quantity can be measured by the size N of the data set. The quality can be evaluated from a statistical point of view: the higher the statistical dependence among the elements of the data set, the less information Υ carries. For a fixed size N of the data set Υ, the latter contains maximum information when its elements are statistically independent. The process model $\mathcal{M}$ can be viewed as an information container. Its size is denoted $\mathcal{I}(\mathcal{M})$. The more complex this model, the more information can be "poured" into it from a real data set during the identification procedure.

Now consider a model $\mathcal{M}$ of the studied process and a fixed data set Υ carrying an amount of information $\mathcal{I}(\Upsilon)$. If $\mathcal{M}$ is too simple, the information container it represents will overflow during the identification procedure and, therefore, some of the information carried in the set Υ will pour out and be lost. If, however, $\mathcal{M}$ is too complex, then the available amount of information $\mathcal{I}(\Upsilon)$ will not be enough to fill up the model container completely. We will end up with an information container which is impressively large, but carrying very little information about the true process behaviour. Consequently, the best solution is to choose a degree of complexity for $\mathcal{M}$ such that $\mathcal{I}(\Upsilon)$ matches $\mathcal{I}(\mathcal{M})$.

2.5 The illusion of model validation

When a model is developed and identified, it needs to be validated. The procedures that are used for model validation have, however, been criticized not only in the area of WWT process mathematical modelling, but also in several other engineering areas. Jeppsson (1996) pointed out that, "*in strict sense, model validation is impossible*". Similarly, Zheng and Bennett (1995) noted that "process *models, like any scientific hypothesis, cannot be validated in the absolute sense ... They can only be invalidated*". Konikow and Bredehoeft (1992) suggested that terms like model verification and model validation convey a false sense of truth and accuracy and thus should be abandoned in favour of more realistic assessment descriptors such as history-matching and benchmarking.

In this thesis, an original approach to deal with this question of model validation is proposed. An innovative framework for model quality evaluation is developed. The basic idea of this framework is as follows:

> Let $\mathcal{M}$ be a model of the system $\mathcal{S}$ and suppose we are interested in the model predictions of the i_0-th state variable x_{i_0} of $\mathcal{S}$. We need to know the prediction accuracy of the model and also the risk of getting significant deviations between model predictions and the system response.
>
> During the identification procedure, the model $\mathcal{M}$ "sees" only a finite number of examples:
>
> $$x_{i0}^{data}(t_1), x_{i0}^{data}(t_2), \ldots, x_{i0}^{data}(t_N)$$
>
> (the elements of the data set are called examples). However, the user expects the model to produce good predictions not only for the situations that it has seen before, but also for the other unseen situations that will occur in the real-world operation of the system. Consequently, the system modeler should strive to make sure that, by minimizing the objective function:
>
> $$J_{i0}(\mathbf{p}) = \sum_{k=1}^{N} \mid x_{i0}(\mathbf{p}, t_k) - x_{i0}^{data}(t_k) \mid^2 \qquad (2.5)$$
>
> or equivalently the arithmetic mean value:
>
> $$R_{emp}(\mathbf{p}) = \frac{J_{i0}(\mathbf{p})}{N} = \frac{1}{N} \sum_{k=1}^{N} \mid x_{i0}(\mathbf{p}, t_k) - x_{i0}^{data}(t_k) \mid^2 \qquad (2.6)$$
>
> the expected time-average:
>
> $$R(\mathbf{p}) = \lim_{n \to +\infty} \frac{1}{n} \sum_{k=1}^{n} \mid x_{i0}(\mathbf{p}, t_k) - x_{i0}^{data}(t_k) \mid^2 \qquad (2.7)$$
>
> will also get minimized. This is because the *true* measure of the model performance is not the empirical objective function $J_{i0}(\mathbf{p})$ or the arithmetic mean $R_{emp}(\mathbf{p})$, but the expected average $R(\mathbf{p})$ of the infinite time

sequence:

$$\mid x_{i_0}(\mathbf{p}, t_1) - x_{i_0}^{data}(t_1) \mid^2, \mid x_{i_0}(\mathbf{p}, t_2) - x_{i_0}^{data}(t_2) \mid^2, \mid x_{i_0}(\mathbf{p}, t_3) - x_{i_0}^{data}(t_3) \mid^2, \ldots$$

However, the value of $R(\mathbf{p})$ is not known. Therefore, we end up with the following situation:

- $R_{emp}(\mathbf{p})$ is merely an empirical measure of the model performance, but its numerical value is accessible to us.
- $R(\mathbf{p})$ is the exact measure of the model performance, but its value is inaccessible to us.

Consequently, the whole question here is how to infer information about the exact model performance measure $R(\mathbf{p})$ from the knowledge of the value of the empirical measure $R_{emp}(\mathbf{p})$. Addressing this question and developing a methodology for model quality evaluation is one of the subjects of this thesis.

Chapter 3

Activated Sludge Process Modelling: Research Needs

This thesis is about developing innovative tools to manage the uncertainty that underlies the behaviour of the activated sludge process. Development of these tools is based on the most recent results of computational learning theory.

The author has reviewed more than 300 papers in the areas of activated sludge process modelling and computational learning theory (CLT). None of the papers reviewed has dealt with the application of CLT to ASP modelling.

The objective of this Chapter is to place the thesis topic in its proper perspective. First, general research needs in the area of ASP modelling are defined, based on the conclusions of Chapter 2. It is suggested to address these needs in three steps (short, medium and long terms). Then, in the second section, the topic of this thesis is described briefly in relation to the aforementioned research needs.

3.1 Research Needs for the Activated Sludge Process Modelling

3.1.1 Process Observability and Uncertainty Management

Two different studies have to be carried out independently in the short term:

1. **Process Observability:**

First the state space of the process has to be defined. This amounts to defining the number and nature of the variables that will be used to describe the ASP state. The IAWPRC model (Henze *et al.*, 1987) uses 13 state variables. Jeppsson (1996) reduced them to 5 variables in order to improve model identifiability. It is of course obvious that fewer state variables provide less information about the exact state of the process. Several parts of the process have indeed to be lumped together to reduce the number of state variables. However, a large number of state variables is the cause of the problem known as the "curse of dimensionality", i.e., as the number of dimensions grows, the computational effort required for process analysis increases extremely fast. Up until now, there is no well-founded methodology to determine the optimal number of variables to be used for process state characterization. It is suggested however to use the preliminary design of the overall process control strategy to determine the number and nature of the process state variables.

When the state space is selected and the relationships among the state variables (i.e., the process model) are developed, an extensive process observability study can then be carried out. The objective is to determine the most economical combination of measurable quantities that can be associated with the process model and produce an observable process. Here is an example to illustrate this concept:

Example 3.1. Consider the IAWPRC model for the ASP and assume that it is written in the compact form:

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}, \mathbf{u}, \mathbf{p})$$

where $\mathbf{p}$ is the model parameter vector, $\mathbf{u}$ is the input vector, $\mathbf{x}$ is the state vector and t is the time. Using the same notation as in the IAWPRC report *"Activated Sludge Process Model N° 1"* (Henze *et al.*, 1987), the state vector components are:

$$x_1 = S_I \quad x_2 = S_S \quad x_3 = X_I \quad x_4 = X_S \quad x_5 = X_{B,H}$$
$$x_6 = X_{B,A} \quad x_7 = X_P \quad x_8 = S_O \quad x_9 = S_{NO} \quad x_{10} = S_{NH}$$
$$x_{11} = S_{ND} \quad x_{12} = X_{ND} \quad x_{13} = S_{ALK}$$

Very few of these variables are directly and separately measurable. Some of the most commonly measured quantities (called outputs) are:

COD_T : Total Chemical Oxygen Demand

COD_F : Filtered Total Chemical Oxygen Demand

NH_4-N : Ammonia Nitrogen

NO_3-N : Nitrates Nitrogen

TKN_T : Total Kjeldahl Nitrogen

TKN_F : Filtered Total Kjeldahl Nitrogen

OUR : Oxygen Uptake Rate

Each of these outputs can be expressed as a function of the state vector components:

$$
\begin{array}{lll}
COD_T = \sum_{i=1}^{7} x_i & \text{, i.e.,} & y_1 = COD_T = h_1(\mathbf{x}) \\
COD_F = S_S + S_I = x_1 + x_2 & \text{, i.e.,} & y_2 = COD_F = h_2(\mathbf{x}) \\
NH_4\text{-}N = S_{NH} = x_{10} & \text{, i.e.,} & y_3 = NH_4\text{-}N = h_3(\mathbf{x}) \\
NO_3\text{-}N = S_{NO} = x_9 & \text{, i.e.,} & y_4 = NO_3\text{-}N = h_4(\mathbf{x}) \\
TKN_T = x_{10} + x_{11} + i_{XB}(x_5 + x_6) + i_{XP}x_7 + x_{12} & \text{, i.e.,} & y_5 = TKN_T = h_5(\mathbf{x}) \\
TKN_F = S_{NH} + S_{ND} = x_{10} + x_{11} & \text{, i.e.,} & y_6 = TKN_F = h_6(\mathbf{x}) \\
OUR = \frac{1-Y_H}{Y_H}\frac{\mu_H x_2}{K_S + x_2} x_5 + \frac{4.57-Y_A}{Y_A}\frac{\mu_A x_{10}}{K_S + x_{10}} x_6 & \text{, i.e.,} & y_7 = OUR = h_7(\mathbf{x})
\end{array}
$$

The goal of the observability study here is to find the smallest and most economical output combination $y_{i_1}, y_{i_2}, \ldots, y_{i_m}$ that leads to an observable process governed by:

$$
\begin{cases}
\dot{\mathbf{x}} & = & \mathbf{f}(t, \mathbf{x}, \mathbf{u}, \mathbf{p}) \\
y_{i_1} & = & h_{i_1}(\mathbf{x}) \\
y_{i_2} & = & h_{i_2}(\mathbf{x}) \\
\vdots & & \\
y_{i_m} & = & h_{i_m}(\mathbf{x})
\end{cases}
$$

Qualitatively, "observable process" means that:

- The structure of the process model is such that the output:

$$\mathbf{y}(t) = (y_{i_1}(t), y_{i_2}(t), \ldots, y_{i_m}(t))$$

  is affected in some manner by the change of any single state variable.

- The effect of any one state variable on the output is distinguishable from the effect of any other state variable.

Mathematically, the observability study aims at finding whether $x_1, x_2, \ldots, x_{13}$ are uniquely determinable from the knowledge of measurable quantities $y_{i_1}, y_{i_2}, \ldots, y_{i_m}$ and the inputs $\mathbf{u}$.

When the observability criterion is ensured, a state observer to generate estimates $\hat{x}_1, \hat{x}_2, \ldots \hat{x}_{13}$ of the state variables on the basis of $\mathbf{y}$ and $\mathbf{u}$ has to be developed.

2. **Uncertainty Management:**

The previous study is a deterministic one. It does not deal with the uncertainty underlying the ASP behaviour. It just defines an intelligent way of obtaining measurements from the real system and processing them in order to produce estimates of the internal state variables. Another study of process uncertainty management is needed.

At first, this study has to be carried out separately from all process observability issues. To do this, the study is to be placed in the same context as that used by Jeppsson (1996): all the process state variables are assumed to be measurable directly and separately. The merger of process observability and uncertainty management can take place at later stages of the research (see section 3.2.2).

Some of the major questions that need to be addressed in this study are the following:

- How do we define process uncertainty?
- How can we measure it?
- Can we model process uncertainty?
- What makes uncertainty high or low? What governs its variations?
- Can uncertainty be controlled and to what extent can it be reduced?
- How does uncertainty relate to process model performance?

A sound mathematical framework that handles all these questions needs to be developed.

3.1.2 Uncertainty Management for a Partially Observable Process

Process observability and uncertainty management studies have to be combined. The state variables are not measured directly anymore. They are inferred from outputs and inputs. The model identification procedure becomes based on a sequence of output measurements:

$$\mathbf{y}^{meas}(t_1), \mathbf{y}^{meas}(t_2), \ldots, \mathbf{y}^{meas}(t_N)$$

instead of a series of state variables measurements $\mathbf{x}^{data}(t_1), \ldots, \mathbf{x}^{data}(t_N)$. The identification objective function becomes:

$$J(\mathbf{p}) = \sum_{k=1}^{N} \|\mathbf{y}(\mathbf{p}, t_k) - \mathbf{y}^{meas}(t_k)\|^2 \tag{3.1}$$

instead of the equation 2.3.

The framework for uncertainty management developed under the conditions of the previous sub-section needs be refined and adapted to this case which is a more realistic one.

3.1.3 Information Content, Transfer and Conversion

Research in process modelling must aim at conceiving new fundamental approaches to modelling the behaviour of complex systems in general, of which the activated sludge process is a particular case. It would consist of two parts:

1. **Concepts of Information Content, Transfer and Conversion**

The author believes that the complexity of ASP modelling requires a shift from the traditional *state space* approach to a new *information-based* approach. The state-space approach is suitable for systems with mild non-linearities that can be described adequately by a finite-dimensional model. In the case of complex systems, however, a new approach based on the concepts of *system information content, transfer* and *conversion* needs to be developed. These concepts will be implemented qualitatively in next Chapter to define process uncertainty management intuitively. But, in order to make them more useful in the area of complex systems modelling, a mathematical framework to unify and quantify all these concepts is needed. In particular, a universal measure $\mathcal{I}$ of system information content should be developed. The use of the energy analogy, as will be illustrated in next Chapter, can be helpful in that respect. The framework that defines the information measure should consider *uncertainty* as a particular type of information (for example, "uncertainty is the inaccessible part of the system total information content"). In that way, system uncertainty can be quantified using the same measure $\mathcal{I}$. Here, also, similarity with the "energy" concept can be quite staggering:

> The total energy content of a thermodynamic system is the enthalpy H (in joules). It is composed of two parts: G and TS with:
>
> $$H = G + TS \qquad (3.2)$$
>
> This equation is usually known as $G = H - TS$.
>
> The energy G represents the part of the system total energy that is available to perform "useful work". It is called *free energy*. The other part, product TS of temperature (in Kelvin) and entropy (in joules/Kelvin), is an internal manifestation of energy that is not available for useful work.

Using this analogy, the total information content $\mathcal{I}$ of a system may be expressed as:

$$\mathcal{I} = i + u \qquad (3.3)$$

where i is the part of useful information for system management and control and u is the system uncertainty (inaccessible information).

A good starting point to develop the concepts of system information content, transfer and conversion is the textbook *"An Introduction to Kolmogorov Complexity and its Applications"* by Li and Vitányi (1993).

Once a measure $\mathcal{I}$ of system information content is developed and the concepts of information transfer and conversion defined, then a new ASP modelling methodology can be investigated: if the total information content of the real process is $\mathcal{I}_{ASP}$ (when this process was just discovered, $\mathcal{I}_{ASP} = u_{ASP}$, that is: $i_{ASP} = 0$), then the ultimate objective of modelling becomes to design a mathematical system $\mathcal{S}$ — which can be a combination of, for example, a mechanistic model, a fuzzy logic model and a neural network model — such that:

$$\begin{cases} i_{ASP} &= i_S \\ u_{ASP} &= 0 \end{cases}$$

or, equivalently:

$$\mathcal{I}_{ASP} = i_{ASP} = i_S = \mathcal{I}_S$$

2. <u>Over-arching Framework to Reconcile Existing Modelling Technologies</u>

At least *eight* different modelling technologies have been used to approach the ASP dynamical behaviour: stochastic differential equations, ARMA, chaos analysis, fuzzy modelling, neural network modelling, expert systems, qualitative modelling and mechanistic modelling. Each of these technologies has advantages and disadvantages and, because of that, there is a strong need to develop an over-arching framework — preferably at the foundational level — to reconcile all these technologies and design *one* all-advantages mathematical system to predict ASP behaviour.

One attempt to combine mechanistic and neural network models was presented in Côté *et al.* (1995). A preliminary study to combine neural networks and fuzzy logic systems for ASP control was carried out by the author in a first research proposal report (Guergachi, 1995). At a more general level of systems science, several attempts to combine the foregoing modelling technologies are presented in the following books:

- "*Neural network learning and expert systems*" by Gallant (1993).
- "*Artificial intelligence and neural networks : steps toward principled integration*" by Honavar and Uhr (1994).
- "*Fuzzy sets, neural networks, and soft computing*" by Yager and Zadeh (1994).

3.2 Thesis Topic

This thesis focuses exclusively on ASP uncertainty modelling and management, that is, the second part of the research work outlined in sub-section 3.1.1. It is composed of two parts:

1. Development of a General Mathematical Framework

An innovative framework for uncertainty analysis and modelling is developed. This framework is based on the work of Vapnik (1982, 1995, 1998) and that of Vapnik and Chervonenkis (1968, 1981, 1991) in computational and statistical learning theory.

First a measure of process uncertainty is defined both intuitively and mathematically. Then two measures of model performance are defined:

- an *empirical* measure $R_{emp}(\mathbf{p})$ (equation 2.6) which is readily computable on the basis of a data set from process operation.
- the *exact* measure $R(\mathbf{p})$ (equation 2.7) whose value is not accessible.

The first measure is known as *empirical risk* and the second as *expected risk.*

In order to infer information about $R(\mathbf{p})$ from the knowledge of $R_{emp}(\mathbf{p})$, a principle called (Vapnik, 1998):

"*$\mathcal{I}$nductive $\mathcal{P}$rinciple of $\mathcal{E}$mpirical $\mathcal{R}$isk $\mathcal{M}$inimization*"

($\mathcal{IPERM}$) is introduced and its applicability examined. With the introduction of this principle, the model identification procedure is rationalized.

Then a measure of model complexity q, known as the Vapnik-Chervonenkis (VC) dimension, is defined. This measure is used to develop what is called in this thesis process uncertainty models. These models relate the uncertainty measure to the process empirical risk and other parameters that govern uncertainty variations. They are used to investigate some important issues such as process model maximal and marginal improvements.

2. Applications

The framework developed in this thesis has numerous applications. This work will, however, focus on only two of them:

- **Case 1**: *the amount of data to be used for process model identification is not a limiting factor in uncertainty management* (data available in abundance)

  An application of the framework is developed for a simple mechanistic model of the ASP. The applicability conditions of the $\mathcal{IPERM}$ are thoroughly examined. Then a handy and simple equation for process model quality evaluation is developed in terms of the size of the data set and the value of the empirical risk.

- **Case 2**: *only a small amount of data is available for process model identification* (data scarcity)

  A family of mechanistic models, ranging from a simple model to a complex one (with large number of parameters), is developed on the basis of the following idea: an entire set of biodegradation processes with time constants ranging, in a continuous fashion, from *fast to slow biodegradability* is at work in the ASP bioreactor. The process uncertainty models developed previously are then generalized to the case of this family of process models, limited to the heterotrophic biodegradation. On the basis of these uncertainty models, a new methodology of tuning process model complexity to match the small amount of available data is then developed. This methodology uses another principle called (Vapnik, 1998):

"$\mathcal{I}$*nductive* $\mathcal{P}$*rinciple of* $\mathcal{S}$*tructural* $\mathcal{R}$*isk* $\mathcal{M}$*inimization*"

($\mathcal{IPSRM}$). This methodology is illustrated using a simulated ASP wastewater treatment plant.

Part III

DEVELOPING AN INNOVATIVE FRAMEWORK

Chapter 4

Intuitive Introduction to Uncertainty Analysis and Modelling

In what follows, $\mathcal{M}_{ASP}$ denotes a model of the activated sludge process.

Uncertainty represents the 'information gap' that exists between model predictions and reality. Assume that we are interested in the prediction of one state variable of the ASP. This state variable is fixed, but can be any component of the process state vector (substrate concentration, biomass concentration, ...). It will be denoted as x_{i_0} (index $_i$ indicates that x_{i_0} is a generic state variable, and index $_0$ indicates that it is fixed).

Process uncertainty, with respect to this state variable, arises in the form of discrepancy between the model prediction $x_{i_0}^{\mathcal{M}_{ASP}}$ for x_{i_0} and the actual process response function denoted here as g. The closer $x_{i_0}^{\mathcal{M}_{ASP}}$ to g, the lower are the discrepancy and the uncertainty underlying the ASP. A fundamental question arises here: how to measure the closeness of two functions such as $x_{i_0}^{\mathcal{M}_{ASP}}$ and g ? A definition of such a measure is crucial to this study, because it will define a way of quantifying process uncertainty.

4.1 Closeness of Two Functions

In the case of two points a and b on the real line, the notion of closeness can be defined naturally as the distance $d(\, . \, , \, . \,)$ between these two points, that is:

$$d(a, b) = |a - b|$$

In the case of two vectors:

$$\vec{v}_1 = (v_{1,1}, v_{1,2}, \ldots, v_{1,n})^T \text{ and } \vec{v}_2 = (v_{2,1}, v_{2,2}, \ldots, v_{2,n})^T$$

closeness can also be defined as the distance $d(\, . \, , \, . \,)$ between these two vectors, that is the norm of the difference vector $\vec{v}_1 - \vec{v}_2$:

$$d(\vec{v}_1, \vec{v}_2) = \|\vec{v}_1 - \vec{v}_2\| = \sqrt{|v_{1,1} - v_{2,1}|^2 + |v_{1,2} - v_{2,2}|^2 + \ldots + |v_{1,n} - v_{2,n}|^2}$$

Similarly, there is a need to develop a measure to evaluate the distance between two functions such as $x_{i_0}^{\mathcal{M}_{ASP}}$ and g. One way to develop this measure is as follows:

- Compute the mean-square deviation between the values of $x_{i_0}^{\mathcal{M}_{ASP}}$ and g at times $t_0, t_1, t_2, \ldots$:

$$\frac{|x_{i_0}^{\mathcal{M}_{ASP}}(t_0) - g(t_0)|^2 + |x_{i_0}^{\mathcal{M}_{ASP}}(t_1) - g(t_1)|^2 + \ldots + |x_{i_0}^{\mathcal{M}_{ASP}}(t_N) - g(t_N)|^2}{N}$$

  where N is a large whole number.

- Then, to get a mean with the same unit as the state variable's, apply the square root to the previous expression. We obtain a measure $\mathcal{D}$ of the deviation between the functions $x_{i_0}^{\mathcal{M}_{ASP}}$ and g:

$$\mathcal{D}(x_{i_0}^{\mathcal{M}_{ASP}}, g) = \sqrt{\frac{|x_{i_0}^{\mathcal{M}_{ASP}}(t_0) - g(t_0)|^2 + \ldots + |x_{i_0}^{\mathcal{M}_{ASP}}(t_N) - g(t_N)|^2}{N}}$$

  or:

$$\mathcal{D}(x_{i_0}^{\mathcal{M}_{ASP}}, g) = \sqrt{\frac{\sum_{k=0}^{N} |x_{i_0}^{\mathcal{M}_{ASP}}(t_N) - g(t_N)|^2}{N}} \quad (4.1)$$

The fact that N is a large number can be expressed mathematically using the notion of "limit". The previous expression becomes then:

$$\mathcal{D}(x_{i_0}^{\mathcal{M}_{ASP}}, g) = \sqrt{\lim_{N \to +\infty} \frac{\sum_{k=0}^{N} |x_{i_0}^{\mathcal{M}_{ASP}}(t_k) - g(t_k)|^2}{N}} \quad (4.2)$$

Intuitively, $\mathcal{D}(x_{i0}^{\mathcal{M}_{ASP}}, g)$ represents the average deviation between the values of $x_{i0}^{\mathcal{M}_{ASP}}$ and those of g. This average is understood in the *quadratic* sense. But it could have been considered in any other sense such as the arithmetic, geometric or harmonic. The reason why the quadratic average was selected here is that it has a precise statistical meaning. It measures the standard deviation of the model predictions about the true process response function (which is of course totally unknown to the process modeler). A large value of $\mathcal{D}(x_{i0}^{\mathcal{M}_{ASP}}, g)$ indicates a poor approximation of the process behaviour by $\mathcal{M}_{ASP}$, while a small value indicates a better approximation. In addition, the deviation $\mathcal{D}$ defines a true metric on the space of functions. It can be used as a mathematical measure of distance between two functions. To see this property clearly, a transformation of the expression 4.2 is needed. The objective of the following reasoning is to establish equation 4.8.

Let's denote the positive variable:

$$|x_{i0}^{\mathcal{M}_{ASP}}(t_k) - g(t_k)|^2$$

as $\xi(k)$ for each number k, and assume for now that the following hypothesis is true:

$$\left\{ \begin{array}{l} \xi(k) \text{ takes only a finite number of values} \\ \quad \text{when } k \text{ varies from 0 to infinity.} \\ \text{Denote these values as } \xi_{i_0}, \xi_{i_1}, \ldots, \xi_{i_s}. \end{array} \right. \tag{4.3}$$

All terms of the summation:

$$\sum_{k=0}^{N} |x_{i0}^{\mathcal{M}_{ASP}}(t_k) - g(t_k)|^2 = \sum_{k=0}^{N} \xi(k) = \xi(0) + \xi(1) + \xi(2) + \ldots \tag{4.4}$$

would then belong to the set:

$$\{\xi_{i_0}, \xi_{i_1}, \ldots, \xi_{i_s}\}$$

Let's denote the number of times that ξ_{i_k} appears in the previous summation as n_k. The summation 4.4 can then be re-arranged as follows:

$$\sum_{k=0}^{N} |x_{i0}^{\mathcal{M}_{ASP}}(t_k) - g(t_k)|^2 = \sum_{k=0}^{s} n_k \xi_{i_k}$$

And, therefore, the expression of the deviation $\mathcal{D}$ can be re-written as:

$$\mathcal{D}(x_{i0}^{\mathcal{M}_{ASP}}, g) = \sqrt{\lim_{N \to +\infty} \frac{\sum_{k=0}^{s} n_k \xi_{i_k}}{N}} = \sqrt{\lim_{N \to +\infty} \sum_{k=0}^{s} \frac{n_k}{N} \xi_{i_k}} \tag{4.5}$$

In the summation $\sum_{k=0}^{s} \frac{n_k}{N} \xi_{i_k}$, the limit concerns only the fractions:

$$\frac{n_k}{N}$$

because they are the only values that depend on N. From the frequency definition of probability, one can write that:

$$\lim_{N \to +\infty} \frac{n_k}{N} = \mathbf{Pr}[\xi = \xi_{i_k}] = \mathbf{Pr}(\xi_{i_k}) \tag{4.6}$$

where $\mathbf{Pr}[\xi = \xi_{i_k}]$ or $\mathbf{Pr}(\xi_{i_k})$ denote the probability of the event of having the variable ξ equal the number ξ_{i_k}. Consequently, from equation 4.5, the expression of $\mathcal{D}$ becomes:

$$\mathcal{D}(x_{i0}^{\mathcal{M}_{ASP}}, g) = \sqrt{\sum_{k=0}^{s} \xi_{i_k} \mathbf{Pr}(\xi_{i_k})} \tag{4.7}$$

In probability, the summation $\sum_{k=0}^{s} \xi_{i_k} \mathbf{Pr}(\xi_{i_k})$ represents the expected or average value of the variable ξ, which is denoted as $\mathbf{E}(\xi)$. Replacing ξ by its original value

$$|x_{i0}^{\mathcal{M}_{ASP}} - g|^2$$

the expression of $\mathcal{D}$ becomes then:

$$\mathcal{D}(x_{i0}^{\mathcal{M}_{ASP}}, g) = \sqrt{\mathbf{E}(|x_{i0}^{\mathcal{M}_{ASP}} - g|^2)} \tag{4.8}$$

This equation is established under the hypothesis 4.3. This hypothesis is obviously not true in the case of the ASP. A priori, the variable $\xi(k)$ can take any value from the set $\Re+$. However, it is possible to develop a similar reasoning to deduce the equation 4.8 in the general case. This reasoning is presented below for the sake of completeness.

Denote the interval from which ξ takes values as I (this interval can be the whole set $\Re+$). Let x be an element of I. From all the terms of the summation 4.4, denote the number of those that fall inside the infinitesimal interval $[x, x+dx[$ as $n_{[x,x+dx[}$. The summation 4.4 can then be re-arranged as follows:

$$\sum_{k=0}^{N} |x_{i0}^{\mathcal{M}_{ASP}}(t_k) - g(t_k)|^2 = \sum_{k=0}^{N} \xi(k) = \int_I x \, n_{[x,x+dx[}$$

Consequently, the expression of the deviation $\mathcal{D}$ can be re-written as:

$$\mathcal{D}(x_{i_0}^{\mathcal{M}_{ASP}}, g) = \sqrt{\lim_{N \to +\infty} \frac{\int_I x\, n_{[x,x+dx[}}{N}} = \sqrt{\lim_{N \to +\infty} \int_I x \frac{n_{[x,x+dx[}}{N}} \tag{4.9}$$

As $N \to +\infty$, the fraction:

$$\frac{n_{[x,x+dx[}}{N}$$

converges to the probability $\mathbf{Pr}[x < \xi < x + dx]$ of the event "$\xi \in [x, x + dx[$". This probability is equal, by definition, to the product $\mathbf{Pr}_\xi(x)\, dx$ of the probability density function $\mathbf{Pr}_\xi(x)$ of ξ and the infinitesimal distance dx. Therefore, from equation 4.9, the expression of $\mathcal{D}$ becomes:

$$\mathcal{D}(x_{i_0}^{\mathcal{M}_{ASP}}, g) = \sqrt{\int_I x \mathbf{Pr}_\xi(x)\, dx} \tag{4.10}$$

The integral $\int_I x\mathbf{Pr}_\xi(x)\, dx$ being equal to the expected value $\mathbf{E}(\xi)$ of the variable:

$$\xi = |x_{i_0}^{\mathcal{M}_{ASP}} - g|^2$$

the equation 4.10 can then written as the equation 4.8.

Remark: Note that the foregoing reasonings are valid only if the random sequence $\xi(k), k = 1, 2, 3, \ldots$ is ergodic. Ergodicity is discussed extensively in next Chapters and, since it is a very weak condition, it is assumed to hold true for the sequence $\xi(k)$.

4.2 Modelling Process Uncertainty

It is proposed in this thesis to shift the attention from modelling the activated sludge process itself to modelling the uncertainty that underlies its behaviour. The aim is to answer questions such as: what makes uncertainty high or low? How can it be controlled and to what extent can it be reduced?

The metric $\mathcal{D}$ defined in the previous section provides a way of quantifying process uncertainty. Indeed, the larger the value of $\mathcal{D}$, the more important the 'information gap' between the model predictions and the actual process response function and, therefore, the higher the uncertainty underlying the process behaviour. If the deviation $\mathcal{D}$ can be controlled and reduced, so will be the process uncertainty. Consequently, modelling the uncertainty amounts to determining the main variables that influence $\mathcal{D}$ and the relationship among these variables and $\mathcal{D}$.

The most direct approach to carry out this modelling task would be to establish an equation of the form:

$$\mathcal{D}(x_{i0}^{\mathcal{M}_{ASP}}, g) = \varphi(v_1, v_2, \ldots, v_l) \tag{4.11}$$

where v_1, v_2, ...,v_l are the variables governing the variation of $\mathcal{D}$ and φ is a function relating $\mathcal{D}$ to these variables. In order for equation 4.11 to be useful from an engineering point of view, the following condition must be satisfied:

$$\left\{ \begin{array}{c} \text{the variables } v_i \text{ and the function } \varphi \text{ are readily} \\ \text{determinable/computable, and not just} \\ \text{abstract mathematical objects.} \end{array} \right. \tag{4.12}$$

Equation 4.11, if it is possible to develop under the condition 4.12, would mean that uncertainty, as it is measured by the deviation $\mathcal{D}$, can be totally eliminated in a highly complex process such as the ASP. Equation 4.11 suggests indeed that uncertainty is readily transformable into useful information — namely "$\varphi(v_1, v_2, \ldots, v_l)$" — that can eventually be exploited to exactly determine the unknown function g on the basis of a just finite-dimensional model $\mathcal{M}_{ASP}$. This is not realistic and, because of this, it is not recommended to pursue investigation of process uncertainty analysis and modelling using this approach.

An alternative and more promising approach, however, consists in developing an inequality of the form:

$$\mathcal{D}(x_{i0}^{\mathcal{M}_{ASP}}, g) \leq \varphi(v_1, v_2, \ldots, v_l) \tag{4.13}$$

that *satisfies* condition 4.12. Since the deviation $\mathcal{D}$ is always positive by definition, inequality 4.13 should actually be written as:

$$0 \leq \mathcal{D}(x_{i0}^{\mathcal{M}_{ASP}}, g) \leq \varphi(v_1, v_2, \ldots, v_l) \tag{4.14}$$

Inequality 4.13 is realistic and also useful. The objective of uncertainty management is indeed not to eliminate uncertainty (which is usually not possible), but to reduce it and control it. In that respect, inequality 4.13 is useful: manipulating the variables v_1, v_2, ..., v_l (whose values are accessible by virtue of 4.12) allows the control of $\varphi(v_1, v_2, \ldots, v_l)$, which in turn allows the control of $\mathcal{D}$. In particular, low values of $\varphi(v_1, v_2, \ldots, v_l)$ yield low values of $\mathcal{D}$ and, therefore, low uncertainty.

4.3 Developing Uncertainty Models

The two key functions of the uncertainty model 4.13 are the process response function g and the model prediction $x_{io}^{\mathcal{M}_{ASP}}$. The latter function $x_{io}^{\mathcal{M}_{ASP}}$ results from the model identification procedure. This procedure can be viewed as an information transfer from a *finite* set of process data:

$$\Upsilon_N : \quad x_{io}^{data}(t_1), x_{io}^{data}(t_2), \ldots, x_{io}^{data}(t_N) \tag{4.15}$$

to the process model $\mathcal{M}_{ASP}$. The model in turn is viewed as an information container. The size of this container is characterized in part by the number of process model parameters: the larger the number of parameters, the more information can be transferred to the model.

The foregoing metaphoric description of the identification procedure has highlighted the two first variables that affect the value of $\mathcal{D}$:

- **Variable** v_1: *the total amount of information* $\mathcal{I}(\Upsilon_N)$ *contained in the process data set* Υ_N. This amount of information is function of the size N of the data set and the degree of statistical dependence among the elements of this set. The stronger the dependence, the less information Υ_N carries. It is suggested to take the product ςN as a numerical characterization of $\mathcal{I}(\Upsilon_N)$, where $\varsigma \in [0, 1]$ is a coefficient accounting for the data set statistical dependence. The stronger this dependence, the lower will be the value of ς. For illustration, consider for instance the following cases:

  - Case 1: the sequence Υ_N is statistically independent.
  - Case 2: the sequence Υ_N is α-mixing (mixing is a condition on sequence dependence; see White (1984) for definition).
  - Case 3: the sequence Υ_N is ergodic.
  - Case 4: the sequence Υ_N is such that:

  $$\begin{array}{llll} x_{io}^{data}(t_1) = 10.1 & x_{io}^{data}(t_2) = 10.1 & x_{io}^{data}(t_3) = 10.2 & x_{io}^{data}(t_4) = 10.1 \\ x_{io}^{data}(t_5) = 10.2 & x_{io}^{data}(t_6) = 10.1 & \ldots & x_{io}^{data}(t_N) = 10.1 \end{array}$$

  If ς_1, ς_2, ς_3 and ς_4 are the dependence coefficients corresponding to these 4 cases respectively, then we would have:

  $$0 = \varsigma_4 < \varsigma_3 \leq \varsigma_2 \leq \varsigma_1 = 1$$

This is because "statistical independence" implies "α-mixing" and "α-mixing" implies "ergodicity" (White, 1984).

- **Variable** v_2: *the maximum amount of information* $\mathcal{I}(\mathcal{M}_{APS})$ *that can potentially be transferred to the model* $\mathcal{M}_{APS}$. This amount of information is dictated by model complexity. The number n_p of process model parameters can be considered a measure of it. However, in some cases, this measure can prove to be inconsistent. Take for instance the two following simple models where the state variables are x and y and the parameters are p_1 and p_2:

$$y = p_1\, x + p_2 \qquad (n_\mathrm{p} = 2) \tag{4.16}$$

and

$$y = p_1\, \sin(p_2\, x) \qquad (n_\mathrm{p} = 2) \tag{4.17}$$

If n_p is taken as a measure of $\mathcal{I}(\mathcal{M}_{APS})$, we would have to conclude that these two models have the same level of complexity, which is obviously not true: model 4.16 is a family of simple straight lines, while model 4.17 is a complex family of curves that can sweep across the whole plane. Consequently, a more adequate and consistent measure of model complexity needs to be developed. In what follows, this measure will be denoted as q.

Now, during the model identification procedure, the information transfer:

$$\Upsilon_N \longrightarrow \mathcal{M}_{APS}$$

may not always be complete for various reasons such as:

- a too low model complexity (meaning a small size of the information container $\mathcal{M}_{APS}$).
- local minima problem that are usually encountered in the course of minimizing the identification objective function (p being the parameter vector of $\mathcal{M}_{ASP}$):

$$J_{i_0}(\mathbf{p}) = \sum_{k=1}^{N} \mid x_{i_0}^{\mathcal{M}_{ASP}}(t_k) - x_{i_0}^{data}(t_k) \mid^2 \tag{4.18}$$

Here, the reader may be misled by the previous metaphorical discussion and think that, since $\mathcal{I}(\Upsilon_N)$ and $\mathcal{I}(\mathcal{M}_{APS})$ can be represented by the numbers ςN and q respectively, it suffices to ensure that $\varsigma N = q$ to guarantee a complete information transfer from Υ_N to $\mathcal{M}_{ASP}$. The problem is however more complex and, because of this, another more accurate metaphoric image needs to be introduced here to illustrate the author's idea about information conversion. This image is "*energy*".

Energy can exist in various forms: mechanical, calorific, chemical, electrical, ... Using different procedures, it can be converted from one form to the other. Take for instance the conversion:

$$mechanical\ work \longrightarrow heat \tag{4.19}$$

When James Prescott Joule (1818 - 1878) carried out his famous experiment (1847) on the work-heat equivalency, the results were that an amount x (in joules) of mechanical work was not converted into exactly the same amount x (in calories) of heat. There was a conversion factor of

$$c = 4.18\,J/cal$$

In the case of the reverse conversion:

$$heat \longrightarrow mechanical\ work \tag{4.20}$$

other studies have shown that things are even more complex. An amount x (in calories) of heat never generates the expected amount:

$$x \times c = x \times 4.18\,J/cal$$

of mechanical work (in joules). There is a conversion efficiency factor η that is always less than 1 for the transformation 4.20. It is a law of nature. An amount x (in calories) of heat generates $x \times 4.18\,J/cal \times \eta$ (in joules) of mechanical work, meaning that there is always a loss of energy of $x \times 4.18\,J/cal \times (1-\eta)$ (in joules) during the conversion 4.20. In real-world situations, the exact value of η, which determines the degree of completeness of energy conversion, has to be determined experimentally.

Similarly, "*information*" can be expressed in various forms such as:

- real data from the process operation:

$$\Upsilon_N: \quad x_{io}^{data}(t_1), x_{io}^{data}(t_2), \ldots, x_{io}^{data}(t_N)$$

- the parameter values of an *identified* process model:

$$\mathbf{p} = (p_1, p_2, \ldots, p_m)$$

- the set of symbolic rules of an expert or a fuzzy logic system.
- the values of the weights of a trained neural network:

$$\mathbf{w} = (w_{1,1}, w_{1,2}, \ldots, w_{n,m})$$

Also, "information" can be converted from one of these forms to the other, using different mathematical procedures. Here are some examples of information conversion:

- information in a sequence of data is converted into a set of model parameter values using the procedure of model identification.
- information contained in a process model, $\mathcal{M}_{APS}$ with a parameter vector:

$$\mathbf{p} = (p_1, p_2, \ldots, p_m)^T$$

  can be converted into IF-THEN symbolic rules to be used in process expert systems (Nolasco *et al.*, 1989).

- information contained in a sequence of data is converted into the weight values of a neural network using a training procedure such as back-propagation.

Since there is no proof that information conversion from one form to the other is always complete, there is a need of a measure of the degree of completeness of such transformation. In the case of "energy", this measure was the empirical value of the conversion efficiency factor η. In the case of this study — ASP model identification —, it will be the inverse $\vartheta(\mathbf{p})$ of the empirical function ($\mathbf{p}$ is the parameter vector of $\mathcal{M}_{ASP}$):

$$R_{emp}(\mathbf{p}) = \frac{1}{N} \sum_{k=1}^{N} \mid x_{i_0}^{\mathcal{M}_{APS}}(t_k) - x_{i_0}^{data}(t_k) \mid^2 \qquad (4.21)$$

Note that the latter function $R_{emp}(\mathbf{p})$, called *empirical risk*, is equal to $J_{i_0}(\mathbf{p})/N$, where $J_{i_0}(\mathbf{p})$ is the objective function that has to be minimized during the model identification procedure. Its inverse $\vartheta(\mathbf{p})$, called the information conversion efficiency function, will represent the third **variable** v_3 to be considered in the uncertainty model 4.13.

As a recapitulation of the foregoing discussion, three variables to be considered in the uncertainty model 4.13 have been defined. These are:

$$v_1 = \varsigma N \; ; \quad v_2 = q \; ; \quad v_3 = \vartheta(\mathbf{p})$$

The uncertainty model 4.13 would then look like:

$$\mathcal{D}(x_{i_0}^{\mathcal{M}_{ASP}}, g) \leq \varphi(\varsigma N, q, \vartheta(\mathbf{p})) \tag{4.22}$$

Do these three variables provide adequate description of the uncertainty underlying the ASP behaviour, so that an inequality of the type of 4.22 can be mathematically developed? The answer is that they do not, because no one of them can describe process model *suitability*. With v_1, v_2 and v_3 only, inequality 4.22 treats neural network models, polynomial models and mechanistically-based models, for instance, in the same way, which is obviously not reasonable. Let's view this question from another angle.

For the uncertainty model 4.22 to be meaningful, the function

$$\varphi(\varsigma N, q, \vartheta(\mathbf{p}))$$

must decrease as the value of N increases. This is because process uncertainty reduces when more data are made available for model identification. The rate of decrease can be expressed as:

$$\frac{\partial \varphi}{\partial N}(\varsigma N, q, \vartheta(\mathbf{p}))$$

It is common sense to believe that this rate of decrease must *not* be the same for any type of ASP model. A well thought-out model such as:

$$\left\{ \begin{array}{lllllll} \dot{S} & = & \frac{(Q_{in}S_{in}+Q_rS_r)}{V} & - & \frac{(Q_{in}+Q_r)}{V}S & - & \frac{kS}{K_S+S}X \\ \dot{X} & = & \frac{(Q_{in}X_{in}+Q_rX_r)}{V} & - & \frac{(Q_{in}+Q_r)}{V}X & + & \frac{\mu_H S}{K_S+S}X - bX \end{array} \right. \tag{4.23}$$

for instance, should require much less data to start giving meaningful predictions than a black-box model such as the neural network model ($\mathbf{w}$ represents the weight vector):

$$t \longrightarrow S_t \tag{4.24}$$

In the case of model 4.23, indeed, a substantial amount of information about the ASP has been put in the model equations beforehand:

- the structure of the equations results from a fundamental physical concept: *mass balance.*
- biochemical kinetics are based on Monod equation, which is empirical indeed, but its structure resembles the proven equation of Michaelis-Menten.
- the model parameters k, K_S, μ_H, b are known to vary within limited ranges.

On the other hand, a neural network is a general model that can be used to approximate the behaviour of *any* system. Its parameters (weights $\mathbf{w}$, that is) can, a priori, take any value from the real line $]-\infty, +\infty[$. The equations structure is arbitrary. At the beginning of the training procedure, the neural network model is actually quite far from the very behaviour of the activated sludge process. As a result, a lot more data are needed for the neural network to just build-up in it the foregoing basic information that exists beforehand in the model 4.23. Consequently, if the rate of decrease of:

$$\varphi(\varsigma N, q, \vartheta(\mathbf{p}))$$

is, for instance, logarithmic (that is $1/\log N$) for neural network models, the uncertainty model 4.22 must predict a faster rate of decrease (power N^{-a} or exponential a^{-N} or the like) for a model such as 4.23. This is why, a fourth **variable** v_4 expressing a *weak prior information* about the ASP model suitability needs to be added. This variable will be denoted as $v_4 = \mathcal{WPI}$ and a numerical measure of it needs to be developed.

With the introduction of v_4, the ASP uncertainty model 4.22 can then be re-written as:

$$\mathcal{D}(x_{i0}^{\mathcal{M}_{ASP}}, g) \leq \varphi(\varsigma N, q, \vartheta(\mathbf{p}), \mathcal{WPI}) \tag{4.25}$$

Now that the main variables that affect process uncertainty have been defined, a mathematical framework needs to be developed to establish the relationships among these variables and the uncertainty measure $\mathcal{D}$.

Chapter 5

Model Identification: a New Perspective

5.1 Stochastic Nature of Experimental Data

The first ASP model developed by the IAWPRC in 1987 uses 13 state variables to characterize the ASP state. The second model, developed in 1994, uses 20 state variables, without including other general state variables such as temperature, pH and redox potential. Despite this large number of state variables, the foregoing models "*are still greatly simplifying the representation of many species of organisms. As the microbial population changes this needs to be reflected in changing kinetic parameters and even by adding new state variables*" (Jeppsson, 1996). To paraphrase Jeppsson's statement, a complete description of the ASP dynamical behaviour requires a much larger number of state variables and an extremely complex model. Designate the vector of *all* variables that are necessary for a complete description of the ASP behaviour as:

$$\mathbf{x}^{\infty} = (x_1, x_2, \ldots, x_n, x_{n+1}, \ldots, x_{\infty}) \tag{5.1}$$

In practice, only a few components of $\mathbf{x}^{\infty}$ can be used to model ASP dynamics. Jeppsson (1996), for instance, proposed to reduce the number of ASP state variables to not more than five. Other researchers such as Marsili-Libelli (1989) advocated similar principles. In any case, our current knowledge of the ASP behaviour is at such a level that a reliable model based on all components of $\mathbf{x}^{\infty}$ is impossible to develop.

Denote then the few state variables to be used practically in process modelling as:

$$x_1, x_2, \ldots, x_l$$

and the vector composed of these variables as $\mathbf{x}$. It is of course assumed that the components of $\mathbf{x}^\infty$ have been re-arranged in such a way that the l variables that are used in practice are listed first in the vector:

$$(x_1, x_2, \ldots, x_n, x_{n+1}, \ldots, x_\infty)$$

Now consider the following imaginary situation where the process modeler collects experimental data regarding the ASP behaviour and, more specifically, the variations of one state variable x_{i_0} ($i_0 \in \{1, 2, \ldots, l\}$): at time t_0, the modeler sets the state variables x_1, x_2, ..., x_l at fixed values x_1^0, x_2^0, ..., x_l^0, respectively. That is:

$$\text{at time } t_0 : \quad x_1 = x_1^0 \ , \ x_2 = x_2^0 \ , \ \ldots \ , \ x_l = x_l^0$$

or, using the vector notation:

$$\text{at time } t_0 : \quad \mathbf{x} = \mathbf{x}_0, \qquad \text{with } \mathbf{x}_0 = (x_1^0, x_2^0, \ldots, x_l^0)$$

He then lets input perturbations $\mathbf{u}_{\Delta t}$ act on the process from t_0 to time $t_1 = t_0 + \Delta t$. At time t_1, he measures the value of x_{i_0}. For this purpose, the modeler uses a sophisticated automatic device that can take a *large* number of samples at the same time and carry out the measurements immediately for all samples. In that way, he obtains a long series of measurements:

$$x_{i_0}^{meas_1} \ , \ x_{i_0}^{meas_2} \ , \ x_{i_0}^{meas_3} \ , \ \ldots$$

of the same variable x_{i_0} at time t_1. He then computes the arithmetic average of this series and obtains a highly *accurate* estimate of x_{i_0} at time t_1. Denote this estimate as $\hat{x}_{i_0}$.

Imagine now that the process modeler repeats exactly the same measurement procedure a second time:

- First, at time t'_0, he sets the state variables at the *same* values as before:

$$\text{at time } t'_0 : \quad \mathbf{x} = \mathbf{x}_0$$

- He then applies the *same* input perturbations $\mathbf{u}_{\Delta t}$ to the process from t'_0 to time $t'_1 = t'_0 + \Delta t$.
- A large number of samples are then taken at time t'_1 and measurements of x_{i_0} are carried out.
- Finally, the process modeler computes the arithmetic average of the x_{i_0} measurement series and obtains a highly accurate estimate of x_{i_0} at time t'_1, denoted as $\tilde{x}'_{i_0}$

The measurement devices are assumed to be highly accurate and reliable and the experimental error eliminated due to the use of a large number of sample measurements.

Question: Is it guaranteed that $\tilde{x}'_{i_0}$ will equal $\hat{x}_{i_0}$?

Despite all the precautions that the modeler has taken in both experiments, the answer to this question is negative. Here is why:

In both experiments, the modeler has set the state vector $\mathbf{x}$ at the same value and applied the same input during the time interval of $t_1 - t_0 = t'_1 - t'_0 = \Delta t$. But the real process state is governed by $\mathbf{x}^\infty$ and not the vector $\mathbf{x}$. When $\mathbf{x}$ is set at a 'fixed value', it does not mean that $\mathbf{x}^\infty$ will also be set at a 'fixed value'. The variables $x_1, x_2, \ldots, x_l$ that are used in practice do not necessarily control the values of the other state variables of $\mathbf{x}^\infty$. This would happen only if the components x_k, with $k > l$ (that is the elements of the set $\{x_{l+1}, x_{l+2}, \ldots, x_n, x_{n+1}, \ldots, x_\infty\}$), are observable from the measurements of $x_1, x_2, \ldots, x_l$. This observability property is, however, not guaranteed and there is no operational way to verify it. It is actually unlikely to hold true because the size of the set:

$$\{x_{l+1}, x_{l+2}, \ldots, x_n, x_{n+1}, \ldots, x_\infty\}$$

of the variables to be inferred is much larger than that of the set:

$$\{x_1, x_2, \ldots, x_l\}$$

of the variables that are measurable. Consequently, in both experiments, the variables $x_1, x_2, \ldots, x_l$ will have the same values ($x_1^0, x_2^0, \ldots, x_l^0$ respectively) at times

t_0 and t'_0, but the other components of $\mathbf{x}^\infty$ will take, a priori, any arbitrary values. Therefore, if $\mathbf{x}^\infty(t_0)$ and $\mathbf{x}^\infty(t'_0)$ designate the real process state at times t_0 and t'_0, it is likely that $\mathbf{x}^\infty(t_0) \neq \mathbf{x}^\infty(t'_0)$. As a result, the input perturbations $\mathbf{u}_{\Delta t}$ is likely to steer the process from these two different states to two other different states, $\mathbf{x}^\infty(t_1)$ and $\mathbf{x}^\infty(t'_1)$ at times t_1 and t'_1 respectively, and then lead to different values of x_{i_0}. Thus, it is not guaranteed that the estimate $\hat{x}'_{i_0}$ obtained in the second experiment will equal $\hat{x}_{i_0}$ that is obtained in the first experiment.

Because of this, the activated sludge process is called a *stochastic system*, as opposed to deterministic. The process response to external perturbations shows a certain degree of randomness that is expressed in the stochastic nature of the process experimental data (as illustrated in the foregoing two imaginary experiments). The process *unknown* phenomena that cause randomness in the process response and experimental data are called *process noise*. They are the main source of process uncertainty.

Imagine now that we have a very large number of x_{i_0} estimates and not just $\hat{x}_{i_0}$ and $\hat{x}'_{i_0}$. Denote these estimates as:

$$\hat{x}_{i_0}^{[1]}\ ,\ \hat{x}_{i_0}^{[2]}\ ,\ \hat{x}_{i_0}^{[3]}\ ,\ \hat{x}_{i_0}^{[4]}\ ,\ \ldots \tag{5.2}$$

The values $\hat{x}_{i_0}^{[j]}$ with $j = 1, 2, 3, \ldots$ are called *realizations* of the variable x_{i_0}. They are dispersed around their expected (arithmetic) average which is to be determined under the conditions that $\mathbf{x}_0$ and $\mathbf{u}_{\Delta t}$ are given. Denote such average as $\mathbf{E}(\hat{x}_{i_0} \mid \mathbf{x}_0, \mathbf{u}_{\Delta t})$ (read it as: "expected value of $\hat{x}_{i_0}$ given $\mathbf{x}_0$ and $\mathbf{u}_{\Delta t}$") . The stronger is the process noise, the higher will be the dispersion around $\mathbf{E}(\hat{x}_{i_0} \mid \mathbf{x}_0, \mathbf{u}_{\Delta t})$. This fact is usually expressed in an equation of the form:

$$\hat{x}_{i_0} = \mathbf{E}(\hat{x}_{i_0} \mid \mathbf{x}_0, \mathbf{u}_{\Delta t}) + \epsilon_{x_{i_0}} \tag{5.3}$$

where $\epsilon_{x_{i_0}}$ is a numerical measure of the process noise. Mathematically, $\epsilon_{x_{i_0}}$ is a random variable (with an expected average $\mathbf{E}(\epsilon_{x_{i_0}} \mid \mathbf{x}_0, \mathbf{u}_{\Delta t})$ equal 0) that characterizes the degree of randomness of $\hat{x}_{i_0}$ around its average. In what follows, it will also be called process noise with respect to the variable x_{i_0}, or simply process noise.

5.2 Data Overfitting: a Fundamental Problem

5.2.1 What is Data Overfitting?

Consider a sequence of experimental data for the process state variable x_{i_0}:

$$\Upsilon_N : \quad x_{i_0}^{data}(t_1), x_{i_0}^{data}(t_2), \ldots, x_{i_0}^{data}(t_N) \tag{5.4}$$

generated under the following conditions:

- the process initial state $\mathbf{x}(t_1)$ is set at a fixed vector $\mathbf{x}_1$:

$$\mathbf{x}(t_1) = \mathbf{x}_1 \tag{5.5}$$

- the input perturbation that acts on the process during the interval $[t_1, t_N]$ is represented by the time function:

$$\mathbf{u}_{t_1,t_N} : t \quad \mapsto \quad \mathbf{u}_{t_1,t_N}(t) \tag{5.6}$$

Each term $x_{i_0}^{data}(t_k)$ ($k \in \{1, 2, \ldots, N\}$) of this sequence represents one realization of the state variable x_{i_0} at time t_k. If, at a certain time t'_1, the process state is set again at $\mathbf{x}_1$ and the input perturbation $\mathbf{u}_{t_1,t_N}$ is applied to the process from t'_1 to $t'_N = t'_1 + \Delta t$ (with $\Delta t = t_N - t_1$), then the process response would generate another time sequence:

$$\Upsilon'_N : \quad x_{i_0}^{data'}(t_1), x_{i_0}^{data'}(t_2), \ldots, x_{i_0}^{data'}(t_N) \tag{5.7}$$

representing a second realization of the state variable x_{i_0}. The terms of this sequence are likely to be different from those of the first sequence 5.4, because of the stochastic nature of experimental data. Using the notations of the equation 5.3, this stochastic nature can be expressed in the form of the equation ($k \in \{1, 2, \ldots, N\}$):

$$x_{i_0}^{data}(t_k) = \mathbf{E}[x_{i_0}^{data}(t_k) \mid \mathbf{u}_{t_1,t_k}, x_{i_0}^{data}(t_{k-1}), x_{i_0}^{data}(t_{k-2}), \ldots, x_{i_0}^{data}(t_1)] + \epsilon_{x_{i_0}} \tag{5.8}$$

The number $\epsilon_{x_{i_0}}$ represents the process noise with respect to x_{i_0} and

$$\mathbf{E}[x_{i_0}^{data}(t_k) \mid \mathbf{u}_{t_1,t_k}, x_{i_0}^{data}(t_{k-1}), x_{i_0}^{data}(t_{k-2}), \ldots, x_{i_0}^{data}(t_1)]$$

the average of x_{i_0} at time t_k, given the objects $\mathbf{u}_{t_1,t_k}$, $x_{i_0}^{data}(t_{k-1})$, $x_{i_0}^{data}(t_{k-2})$, $\ldots$, $x_{i_0}^{data}(t_1)$. If the sequence 5.4 is assumed to be Markovian, then equation 5.8 can be simplified as:

$$x_{i_0}^{data}(t_k) = \mathbf{E}[x_{i_0}^{data}(t_k) \mid \mathbf{u}_{t_1,t_k}, x_{i_0}^{data}(t_{k-1})] + \epsilon_{x_{i_0}} \tag{5.9}$$

Intuitively, Markov property simply means that, given the present value of x_{i0}^{data}, the future value of it is independent of the past. This can be considered true in the case of the ASP, since all ASP models contain only first order differential equations.

Let $\mathcal{M}_{ASP}$ be a model of the activated sludge process. This model can be a black box (neural network, polynomial or ARMA) or mechanistically-based. Its identification (with respect to the state variable x_{i0}) is carried out on the basis of one realization of x_{i0} such as the series 5.4. In previous chapters, model identification has been viewed as an information transfer from a data series Υ_N to $\mathcal{M}_{ASP}$. But it can also be viewed as a *learning procedure*:

> The model $\mathcal{M}_{ASP}$ first "sees" a finite number of examples $x_{i0}^{data}(t_1)$, $x_{i0}^{data}(t_2)$, ..., $x_{i0}^{data}(t_N)$ about the real process dynamical behaviour (this is the "*teaching*" phase) and, then, based on this limited "experience", the model is supposed to exhibit an adequate amount of *intelligence* to infer the process future dynamics under all circumstances ("*production*" phase).

Mathematically, this procedure consists in minimizing the identification objective function ($\mathbf{p}$ being the parameter vector of $\mathcal{M}_{ASP}$):

$$J_{i0}(\mathbf{p}) = \sum_{k=1}^{N} \mid x_{i0}^{\mathcal{M}_{ASP}}(t_k) - x_{i0}^{data}(t_k) \mid^2 \tag{5.10}$$

Low values of $J_{i0}(\mathbf{p})$ do not always indicate a good model quality. They can also mean *data overfitting* by the identified model, a phenomenon that is not desirable in the case of the activated sludge process, an inherently stochastic system.

Data overfitting means that, in the course of the identification procedure (teaching phase), the model $\mathcal{M}_{ASP}$ learns the *process noise pattern*, in addition to the underlying process response function we want to model. Data overfitting is always associated with low values of the identification objective function $J_{i0}(\mathbf{p})$. In some cases, it may even bring this function down to zero by letting the model exactly fit all the N points of the data set Υ_N. This would usually happen when $\mathcal{M}_{ASP}$ is a neural network, ARMA, polynomial or mechanistic model with a number of parameters that is large enough compared to the size N of the data set Υ_N (for the polynomial type, for instance, take a polynomial function of degree $N-1$). When overfitting occurs, a low value of $J_{i0}(\mathbf{p})$ does not mean that the obtained model is a good approximation of the real process. On the contrary, a model that overfits the data is likely to produce bad predictions for the process future dynamics. These predictions will be worse if the process noise $\epsilon_{x_{i0}}$ is a strong one. If it is low, however, the data overfitting phenomenon may not be as critical. For example, in a well-controlled environment, such as a laboratory experiment, the process noise is usually very small and, as a result, data overfitting may not be too harmful. However, in the case of a real activated sludge wastewater treatment plant, the process noise is usually quite high and, consequently, data overfitting must be avoided.

Trying to avoid data overfitting and preventing ASP models from learning process noise may seem paradoxical. If models should not learn noise and shed light on the process uncertainty that is due to this noise, then what purpose should they be useful for? Do not we develop them to help us predict the uncertain and complex behaviour of the process? The following sub-sections examine these questions.

5.2.2 Why ASP Models Must be Prevented from Learning Noise?

An ASP model (of any type: neural network, ARMA, polynomial or mechanistic) is not meant to model noise. Noise is a random variable and, as such, it should be modeled by a probability distribution and not a set of algebraic or differential equations. In addition, because of the random aspect of noise, it is not possible to predict future values of it exactly. The best we can say about the next value of noise $\epsilon_{x_{i0}}$, for instance, is that it will fall in a given interval I with a certain probability $p_I = \mathbf{Pr}[\epsilon_{x_{i0}} \in I]$. But if $p_I \neq 1$, $\epsilon_{x_{i0}}$ may very well fall outside I, because $1 - p_I = \mathbf{Pr}[\epsilon_{x_{i0}} \notin I]$ is non-zero. Similarly, the realizations of $\epsilon_{x_{i0}}$ during the data collection time interval $[t_1, t_N]$ (see the time series in equation 5.4) may arise according to a certain pattern $\mathcal{P}$ that will never re-occur in the course of future process operation. If this is the case and overfitting is not prevented, the ASP model would then learn the noise pattern $\mathcal{P}$ during the teaching phase, as if it was one of the major process dynamical modes. But in doing so, the model has been just misled by noise: while the pattern $\mathcal{P}$ has arisen just fortuitously during $[t_1, t_N]$ and will not reappear again, the model has taken it "seriously" during the teaching phase and adapted its parameters to it. It will then keep reproducing it whenever the input perturbation conditions that prevailed during $[t_1, t_N]$ arise again. Result: poor model predictions of the real process response.

5.2.3 What Should ASP Models Learn from Data?

If process models must be prevented from learning noise, what should they then learn during the teaching phase? What information should they try to capture from the identification data set Υ_N? A qualitative and intuitive response to these questions has been given by Côté *et al.* (1995): "*It is important to recall that what we are looking for is the general tendency of the system rather than an overfitting of individual points. ...Again, the purpose of this work is not to match exactly with the experimental points, but to follow the variations.*" But how should we define and quantify a system "*general tendency*"? System "general tendency" and noise are always mixed up in the experimental data. How can we ensure that noise has been successfully filtered out during the teaching phase (model identification)? How can we provide a *principled* and *guaranteed* proof that the model is indeed learning the system general tendency and is not being fooled by noise during its identification? The first question is addressed in the last section of this Chapter, while the last two questions are the subject of next Chapters.

5.2.4 How Should Noise be Dealt With?

The purpose of ASP models is then to learn the process "general tendency" and separate it from noise. But noise is an integral part of the process behaviour. How should we deal with it, if ASP models are not supposed to take care of it? By reducing it and controlling it.

Noise is indeed an expression of process complexity. It cannot be eliminated completely. It can only be reduced and then controlled. To do so, research needs to be carried out in several different areas at the same time:

- **Choosing the right state variables:** As was explained previously, it is currently not possible to develop an ASP model that incorporates all variables influencing the ASP state. Therefore, the whole problem lies in selecting the smallest number of the most relevant process state variables. This requires a great deal of expertise and may necessitate the use of different models depending on the period of the year and the prevailing process operating conditions.

- **Integrating all existing process modelling technologies**: This point has been discussed in chapter 3.
- **developing process control strategies that help manage noise**: In this respect, stochastic control theory (Mybeck 1979; Ahmed, 1988; Borrie, 1992) can be a starting point for this research.

Investigation of these three points is outside the scope of this thesis. However, as was mentioned previously, noise cannot be eliminated and, therefore, it needs to be handled in any case. In what follows, the traditional approach to noise will be adopted: $\epsilon_{x_{i0}}$ is considered as a random variable with zero mean value. Its realizations can take positive and negative values and are dispersed around 0 according to certain fixed probability distribution function $F_{\epsilon_{x_{i0}}}$. This function characterizes the noise completely. The crux of the problem of uncertainty due to noise is that $F_{\epsilon_{x_{i0}}}$ is unknown and there is no way of determining it. Sometimes researchers may assume, for mathematical convenience, that the distribution function $F_{\epsilon_{x_{i0}}}$ is normal or uniform, for instance. But such assumptions are unreasonable and cannot be justified. Here is indeed the situation:

> On the one hand, the set $\mathcal{F}$ of probability distribution functions is extremely *large*. It contains <u>any</u> continuous real-valued function F that is increasing with $F(-\infty) = 0$ and $F(+\infty) = 1$ (Berger, 1993).
>
> On the other hand, the ASP *unknown* phenomena (the origin of noise, that is) change and perturb the process dynamics (about its "general tendency") in a way that current scientific knowledge is unable to predict or, even qualitatively, characterize.

So, if this is the case, how can a totally unknown "object" — the process noise — be, a priori, associated with one single and specific element of a large set $\mathcal{F}$? Any study that bases its results on assuming a specific probability distribution function, such as the normal or uniform distribution, for noise is a questionable one. It should not have more than an academic value.

In this research, $F_{\epsilon_{x_{i0}}}$ is considered fixed but unknown.

5.2.5 Model Cross Validation

Wastewater researchers and practicing engineers (WRPE) are usually aware of the data overfitting problem. That is why they do not always rely on the value of the identification objective function $J_{i_0}(\mathbf{p})$ as the sole criterion of model quality, neither they use a high degree polynomial model (a degree of $N-1$ would be enough), for instance, to bring $J_{i_0}(\mathbf{p})$ down to 0. They make use of one additional technique called *cross validation* to check overfitting and assess the ASP model quality. This technique consists in the following: the data set Υ_N is divided into two parts. The process model is identified on the basis of the first part — called identification sample — and the minimum value of the objective function is computed. Then the obtained model is tested on the second part — called validation sample — that the model has never "seen". If the model performs well on this sample, then it is retained. Otherwise, the model structure is adjusted and the cross validation technique repeated.

"*... There is no guarantee that cross validation will produce the optimal model. The smaller the validation sample and the higher the noise level, the more likely it is that cross-validation will fall short of this ideal*" (Smith, 1993). Cross validation relies exclusively on how the model performs on one single validation sample and, as a result, has at least two drawbacks (Smith, 1993):

- Withholding a validation sample reduces the number of examples (elements of the data set Υ_N are called examples) available to identify the model and thus *lowers* the model accuracy and reliability. This is very bothersome when we know that the whole sample size (identification sample + validation sample) is rarely large enough to even carry out the identification phase properly.

- The second drawback is more critical: the model's final parameter values and structure depend on which examples are in the validation sample. And the degree of variation between validation samples can be significant, especially in the case of a higher noise level. As a result, it is difficult to rule out the possibility that the validation sample is highly skewed in a way that distorts the final model. We may end up then with a model that performs well on the validation sample (that has a bias towards the identified model structure), but produces poor predictions for the process future dynamical behaviour. That is, we would have developed a useless model.

In a study by Cote et al. (1995), neural network models was used to improve the predictions of an ASP mechanistic model. These neural networks had 42 parameters *each* and were trained using a sample containing 140 examples. The authors have raised the issue of data overfitting and clearly stated that the goal is to model the "general tendency" of the system and avoid overfitting. They concluded that this goal has been achieved on the basis of the results of the model cross validation technique. Yet, according to studies on neural network ability to generalize and not overfit data, 140 examples are too few to train a 42-parameter neural network (indications are given in Haykin, 1994; pp. 179). Similar remarks can be made regarding the work of Capodaglio *et al.* (1992) where time series models containing up to 24 parameters have been identified with a number of about 240 examples. The danger with such studies is evident: developing highly complex black box models that pass the cross validation test just because of data overfitting, but would fail to produce good predictions in the course of the real plant operation.

This is why a principled and guaranteed method to evaluate the model performance in predicting the process future dynamical behaviour is needed. Subsequent Chapters aim at developing such a method.

5.2.6 Model Parsimony

Data overfitting occurs when the process model is highly complex compared to the size of the data set Υ_N to be used for model identification. Instances of such highly complex models are neural networks with a large number of hidden nodes, mechanistic or ARMA models with too many parameters or polynomial models with high degree. Having realized the danger of data overfitting by complex models, Box and Jenkins (1970), some of the pioneers of time series analysis, introduced the *principle of parsimony* referring to the use of "*the smallest possible number of parameters for adequate representation*". This principle has been described by the authors in a purely qualitative way: "*... Our objective, then, must be to obtain adequate but parsimonious models. Forecasting and control procedures could be seriously deficient if these models were either inadequate or unnecessarily prodigal in the use of parameters. Care and effort are needed in selecting the model. The process of selection is necessarily iterative, that is to say, it is a process of evolution, adaptation, or trial and error. ... The problem is to choose a suitable system of parsimonious parame-*

terization .... This is not a mathematical problem, but it is a question of finding out how the world tends to behave." (Box and Jenkins, 1970).

In this thesis, it is argued that the parsimony principle is just another statement of the principle of simplicity commonly attributed to William of Ockham (1290? - 1349?) and known as Occam's Razor (see Chapter 2). Box and Jenkins' foregoing statements define two key concepts: "*model adequacy*" and "*model parsimony*".

- **model adequacy**: for a model to be adequate, it must be complex enough so that it does not underfit the data. Data underfitting refers to the situation where the process model is too simple to learn the system "general tendency":

$$\begin{aligned} \textit{inadequate model} &\Longleftrightarrow \textit{data underfitting} \\ &\Longleftrightarrow \left(\begin{array}{c} \textit{model cannot learn even} \\ \textit{the system ``general tendency''} \end{array} \right) \\ &\Longleftrightarrow \textit{model too simple} \end{aligned}$$

- **model parsimony**: for a model to be parsimonious, it must be not too complex so that it does not overfit the data:

$$\begin{aligned} \textit{non-parsimonious model} &\Longleftrightarrow \textit{data overfitting} \\ &\Longleftrightarrow \left(\begin{array}{c} \textit{model learns both the system} \\ \textit{``general tendency'' and noise} \end{array} \right) \\ &\Longleftrightarrow \textit{model too complex} \end{aligned}$$

Both data underfitting and overfitting must be avoided. Consequently, an optimal process model structure (both adequate and parsimonious) needs to be determined. Box and Jenkins (see previous statement) asserted that the problem of determining such optimal model is not a mathematical one, but a "physical" one. In this thesis, it is argued that it is both:

- a ***physical*** one, because the general structure of the model must reflect all the available *a priori* knowledge about the physics of the system at hand, which is, in this thesis, the activated sludge process.

- a ***mathematical*** one, because the available *a priori* knowledge about ASP is not sufficient to build the perfect model of the process and, therefore, a mathematical tool to model and manage the process uncertainty needs to be developed.

This mathematical tool is developed in the next Chapter. Later on in Part IV of the thesis, a family of mechanistic models of the ASP is developed on the basis of a more realistic idea than the IAWPRC model's (Henze *et al.*, 1987) or Jeppsson (1996) model's. These models are then analyzed in the light of the mathematical framework defined in Chapter 6 and a method to select the optimal process model for a given size of the data set Υ_N is developed.

5.3 A Process Model is a Learning Machine

Consider the model differential equation that governs the dynamics of the ASP state variable x_{i_0}:

$$\dot{x}_{i_0} = f(t, \mathbf{x}, \mathbf{u}, \mathbf{p})$$

or

$$\frac{dx_{i_0}}{dt} = f(t, \mathbf{x}, \mathbf{u}, \mathbf{p}) \tag{5.11}$$

where t is the time, $\mathbf{x}$ is the process state vector, $\mathbf{u}$ is the process input vector, $\mathbf{p}$ is the parameter vector and f is a real-valued function. This equation represents one component of the vector differential equation:

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}, \mathbf{u}, \mathbf{p})$$

of an ASP model $\mathcal{M}_{ASP}$. However, the vectors $\mathbf{x}$, $\mathbf{u}$ and $\mathbf{p}$ in equation 5.11 do not necessarily contain all of their components. Normally, they should be denoted as $\mathbf{x}_{x_{i_0}}$, $\mathbf{u}_{x_{i_0}}$ and $\mathbf{p}_{x_{i_0}}$ and equation 5.11 should become:

$$\frac{dx_{i_0}}{dt} = f(t, \mathbf{x}_{x_{i_0}}, \mathbf{u}_{x_{i_0}}, \mathbf{p}_{x_{i_0}}) \tag{5.12}$$

in order to highlight the fact that $\mathbf{x}$, $\mathbf{u}$ and $\mathbf{p}$ contains only those state variables, input variables and parameters, respectively, that influence the dynamics of x_{i_0}. For the sake of simplicity, however, the notations of equation 5.11 will be retained, but the reader will keep in mind that it is the equation 5.12 that is meant.

Most ASP models, if not all, do not depend explicitly on time. In the model equations 4.23, for example, time t does not appear in the right-hand sides. A system such that is usually called an *autonomous system* (Slotine and Li, 1991). Intuitively, the autonomous property means that the state trajectory of the system is independent of the initial time. Mathematically, it means that t must be dropped from the right-hand side $f(t, \mathbf{x}, \mathbf{u}, \mathbf{p})$ of equation 5.11. Thus, the general model equation that governs x_{i0} becomes:

$$\frac{dx_{i0}}{dt} = f(\mathbf{x}, \mathbf{u}, \mathbf{p}) \tag{5.13}$$

All state variables (components of $\mathbf{x}$) are assumed to be directly and separately measurable (refer to Chapter 3 for this assumption).

As suggested by Henze *et al.* (1987), the Euler method can be used to numerically integrate equation 5.11: the time is discretized with a time step of Δt and then x_{i0} is computed at times

$$t_1 = \Delta t \ , \quad t_2 = 2\,\Delta t \ , \quad \ldots \ , \quad t_n = n\,\Delta t \ , \quad \ldots$$

using the following equation:

$$x_{i0}(t_n) = x_{i0}(t_{n-1}) + \Delta t\, f(\mathbf{x}(t_{n-1}), \mathbf{u}(t_{n-1}), \mathbf{p}) \tag{5.14}$$

Define $w^{\mathcal{M}_{ASP}}$ as the value of x_{i0} to be predicted by model $\mathcal{M}_{ASP}$, that is:

$$w^{\mathcal{M}_{ASP}} = x_{i0}(t_n)$$

Similarly, define the vector $\mathbf{v}$ as:

$$\mathbf{v} = [x_{i0}(t_{n-1}), \mathbf{x}(t_{n-1})^T, \mathbf{u}(t_{n-1})^T]^T \tag{5.15}$$

The superscript ^{T} means transposed vector. The number $w^{\mathcal{M}_{ASP}}$ takes values from a sub-set W of the real line $\Re$, and vector $\mathbf{v}$ from a multi-dimensional space V.

For the sake of simplicity, $w^{\mathcal{M}}$ will be used to designate $w^{\mathcal{M}_{ASP}}$.

Now introduce the real-valued function H defined as:

$$H(\mathbf{v}, \mathbf{p}) = x_{i0}(t_{n-1}) + \Delta t\, f(\mathbf{x}(t_{n-1}), \mathbf{u}(t_{n-1}), \mathbf{p}) \tag{5.16}$$

The expression of this function corresponds to that of the right-hand side of equation 5.14. The latter equation becomes then:

$$w^{\mathcal{M}} = H(\mathbf{v}, \mathbf{p}) \tag{5.17}$$

For a fixed parameter vector $\mathbf{p}$, $H(\,.\,,\mathbf{p})$ represents a mapping function from V to W:

$$\begin{array}{rccl} H(\,.\,,\mathbf{p}): & V & \longrightarrow & W \\ & \mathbf{v} & \longmapsto & w^{\mathcal{M}} = H(\mathbf{v},\mathbf{p}) \end{array} \tag{5.18}$$

The parameter vector $\mathbf{p}$ takes values from a multi-dimensional space denoted here as Γ. Define the set of functions $\mathcal{H}_{\mathcal{M}_{ASP}}$ of all mappings $H(\,.\,,\mathbf{p})$ with $\mathbf{p} \in \Gamma$:

$$\mathcal{H}_{\mathcal{M}_{ASP}} = \{H(\,.\,,\mathbf{p}) \mid \mathbf{p} \in \Gamma\} \tag{5.19}$$

This set will be indifferently denoted as $\mathcal{H}_{\mathcal{M}_{ASP}}$ or $\mathcal{H}_{\mathcal{M}}$.

Now assume that a sequence of instances of the couple $(\mathbf{v}, w)$:

$$\Upsilon_N: \quad (\mathbf{v}_1, w_1), (\mathbf{v}_2, w_2), \ldots, (\mathbf{v}_N, w_N)$$

can be obtained from the real process operation, and consider an algorithm $\mathcal{A}$ that receives the sequence Υ_N as input and produces a parameter vector $\mathbf{p}_{emp}$ corresponding to the function $H(\,.\,,\mathbf{p}_{emp}) \in \mathcal{H}_{\mathcal{M}}$ that best approximates the real process response. This algorithm corresponds, of course, to the model identification procedure which consists in minimizing an objective function:

$$R_{emp}(\mathbf{p}) = \frac{1}{N} \sum_{k=1}^{N} l(w_k, H(\mathbf{v}_k, \mathbf{p})) \tag{5.20}$$

The subscript $_{emp}$ means "empirical" and $l(w_k, H(\mathbf{v}_k, \mathbf{p}))$ represents a measure of the *loss* between the desired response w_k corresponding to the vector $\mathbf{v}_k$ and the model prediction represented by $H(\mathbf{v}_k, \mathbf{p})$. As was discussed previously, the quadratic loss function will be adopted in this thesis, i.e.:

$$l(w_k, H(\mathbf{v}_k, \mathbf{p})) = |w_k - H(\mathbf{v}_k, \mathbf{p})|^2 \tag{5.21}$$

In what follows, the function l will be called *quadratic loss* or simply *loss*, and the algorithm corresponding to the ASP model identification procedure will be denoted as $\mathcal{A}^l$.

A set of mapping functions equipped with an algorithm such as $\mathcal{A}^l$ is called a *learning machine* in the area of artificial intelligence and computational learning theory. *We have then shown above that the couple $(\mathcal{H}_{\mathcal{M}}, \mathcal{A}^l)$, composed of an ASP model and an identification procedure, is a learning machine.*

5.4 Fundamental Object of Interest in Model Identification

The objective of the learning machine $(\mathcal{H}_{\mathcal{M}}, \mathcal{A}^l)$ is to select that particular function:

$$H(\ .\ , \mathbf{p}_0) \in \mathcal{H}_{\mathcal{M}}$$

which *best* approximates the real *process response*. The exact meaning of the expression "best approximates" and that of "real process response" are defined below.

In section 5.2, it was explained that $(\mathcal{H}_{\mathcal{M}}, \mathcal{A}^l)$ must be prevented from learning the process noise in order to avoid overfitting. It was pointed out that the *purpose* of an ASP model is to approximate what was called the process "general tendency" and separate it from noise. To be able to evaluate the quality of such approximation, we need to define quantitatively what the process "general tendency" is. Here is a definition of it:

> Consider the variable w that was introduced in the previous section, and imagine a series of experiments similar to the ones discussed in section 5.1:
>
> > The vector $\mathbf{v} = [x_{i0}(t_{n-1}), \mathbf{x}(t_{n-1})^T, \mathbf{u}(t_{n-1})^T]^T$ is set at a fixed value, and then the variable w is measured very accurately. A numerical value of w is then obtained. Denote it as $w^{[1]}$. The same experiment is then repeated a large number of times: the vector $\mathbf{v}$ is set at exactly the same value and then w is measured very accurately. In that way, we obtain a long series of measurements:
> >
> > $$w^{[1]}\ ,\ w^{[2]}\ ,\ w^{[3]}\ ,\ \ldots$$
> >
> > of the same variable w, and for a given value of $\mathbf{v}$.
>
> These measurements are realizations of the variable w. They are dispersed around their expected average which is to be determined under the conditions that the vector $\mathbf{v}$ is given. Denote such average as:
>
> $$\mathbf{E}(w \mid \mathbf{v})$$

It can be represented solely as a function of the vector $\mathbf{v}$. If this function is denoted as g, then we have:

$$g(\mathbf{v}) = \mathbf{E}(w \mid \mathbf{v})$$

Mathematically, g is defined as:

$$g(\mathbf{v}) = \mathbf{E}(w \mid \mathbf{v}) = \int_W w\, P_{w|\mathbf{v}}(w|\mathbf{v})\, dw$$

where $P_{w|\mathbf{v}}(w|\mathbf{v})$ designates the probability density function (*pdf*) according to which w arises when the vector $\mathbf{v}$ is given. The function g:

$$\begin{array}{cccl} g: & V & \longrightarrow & W \\ & \mathbf{v} & \longmapsto & g(\mathbf{v}) = \mathbf{E}(w \mid \mathbf{v}) \end{array}$$

represents the *process "general tendency"* and is called *process response function*. It is the main object of interest in ASP model identification.

The purpose of an ASP model is then to *best* approximate the process "general tendency" g. Now we need to define the meaning of "*best approximation*". To do so, we will make use of the metric $\mathcal{D}$ defined in Chapter 4. It has been shown that this metric can be used to define the closeness of two given functions. If $H(\,.\,,\mathbf{p})$ is a function of the set $\mathcal{H}_{\mathcal{M}}$, then the distance between this function and g is measured by the number:

$$\mathcal{D}(H(\,.\,,\mathbf{p})\,,\,g) = \sqrt{\mathbf{E}(|H(\,.\,,\mathbf{p}) - g|^2)} = \sqrt{\int_V l(H(\mathbf{v},\mathbf{p}), g(\mathbf{v}))P_{\mathbf{v}}(\mathbf{v})\, d\mathbf{v}}$$

where $P_{\mathbf{v}}$ is the *pdf* of $\mathbf{v}$ and l is the quadratic loss. The smaller the number $\mathcal{D}(H(\,.\,,\mathbf{p})\,,\,g)$, the better the approximation of g by the model representation $H(\,.\,,\mathbf{p})$ corresponding to the parameter vector $\mathbf{p}$. Consequently:

"*better approximation of g by the model*" means "*smaller number $\mathcal{D}(H(\,.\,,\mathbf{p})\,,\,g)$*"

The objective of the ASP learning machine $\mathcal{LM} = (\mathcal{H}_{\mathcal{M}}, \mathcal{A}^l)$ is then to select, from the set $\mathcal{H}_{\mathcal{M}}$, that particular function $H(\,.\,,\mathbf{p}_0)$ that minimizes the distance $\mathcal{D}(H(\,.\,,\mathbf{p})\,,\,g)$ considered as a function of the parameter vector $\mathbf{p}$.

Chapter 6

Mathematical Framework for Uncertainty Modelling and Management

6.1 General Description

In a certain environment $\mathcal{E}$, a situation $\mathbf{v}$ arises randomly and a transformer $\mathcal{T}$ acts and assigns to this situation $\mathbf{v}$ a number w obtained as a result of the realization of a random trial. Formally, situation $\mathbf{v}$ represents a vector that takes values from an abstract space V called *instance space*. It is generated according to a *fixed* but *unknown* probability density function (*pdf*) $P_{\mathbf{v}}$ defined on V. The number w, which is dependent on $\mathbf{v}$, takes values from another space $W \subseteq \Re$ called *outcome space*. It is generated according to a conditional *pdf* $P_{w|\mathbf{v}}$ defined on W, also *fixed* but *unknown*. The mathematical object $(\mathbf{v}, w)$ arises then in the product space $Z = V \times W$ (called *sample space*) according to the joint *pdf* $P_{(\mathbf{v},w)} = P_{\mathbf{v}} P_{w|\mathbf{v}}$, which characterizes the probabilistic environment $\mathcal{E}$. In what follows, the couple $(\mathbf{v}, w)$ is denoted as z (to mean that it takes values from the sample space Z). Using this notation, the joint *pdf* $P_{(\mathbf{v},w)}$ is then denoted as P_z. The vector $\mathbf{v}$ will be indifferently called "situation" or "instance" and the number w "outcome" or "transformer's response".

If the behaviour of transformer $\mathcal{T}$ is governed by a process which is a dynamic one, this transformer would usually possess several different *operating modes*. To each mode would correspond a different *pdf* P_z and a different range of variation of $\mathbf{v}$ and

w. To illustrate what is meant by "operating mode" here, consider for instance the behaviour of an automotive engine: the operating conditions of such an engine are not the same when the car is climbing a hill and when it is driving along a highway. In the first case, the engine develops a very high torque and the speed is low, while in the second case, the same engine operates under opposite conditions: the speed is high but the torque is low. Another example that illustrates this concept of "operating mode" is a wastewater treatment plant using the activated sludge process: the operation of this plant can use little return of sludge and low solids in the aeration tank in order to achieve the objective of removing soluble substrate with relatively low oxygen supply. But this plant could also be operated with the purpose of aerobically destroying all of the organic solids in the waste, which can be done by returning all the sludge to the aeration tank. Thus, the same plant could operate under different operating conditions. In what follows, the operating mode of the transformer $\mathcal{T}$ will be denoted by $\mathcal{OM}$.

Associated with the environment $\mathcal{E} = (\mathcal{T}, \mathcal{OM}, z, P_z)$ is a *learning machine* $\mathcal{LM}$ whose objective is to understand the behaviour of the transformer $\mathcal{T}$. It receives a finite sequence Υ_N of N training examples:

$$\Upsilon_N : (\mathbf{v}_1, w_1), (\mathbf{v}_2, w_2), \ldots, (\mathbf{v}_N, w_N)$$

or, using the z-notation:

$$\Upsilon_N : z_1, z_2, \ldots, z_N$$

generated and measured in the probabilistic environment $\mathcal{E}$ as a result of one *realization* of this same environment. Based on these training examples, the learning machine $\mathcal{LM}$ selects a strategy that specifies the best approximation $w^{\mathcal{LM}}$ of the transformer's response for each instance $\mathbf{v}$. Once this strategy is selected, it will be used on *all* future situations $\mathbf{v}$ arising in the environment $\mathcal{E}$, in order to predict the transformer's responses. This strategy, which is mathematically a mapping function from V into W, is called a *decision rule* and is chosen from a fixed functional space $\mathcal{H}$ called *decision rule space*.

The goal of $\mathcal{LM}$ is then to select, from the space $\mathcal{H}$, that particular decision rule which best approximates the transformer's response. What is meant by the expression:

> "*decision rule which best approximates the transformer's response*"?

This question has already been addressed in the previous Chapter where it has been pointed out that "*best approximation of the transformer's response*" means "*closeness to the transformer's "general tendency"*". Closeness is understood in the sense of the metric $\mathcal{D}$. After receiving the sequence Υ_N of training examples, the learning machine $\mathcal{LM}$ selects that particular decision rule h_0 that minimizes $\mathcal{D}(h, g^{\mathcal{T}})$ on the space $\mathcal{H}$ (h designates an element of $\mathcal{H}$ and $g^{\mathcal{T}}$ the transformer's "general tendency"). Formally, this means finding the minimum of the function:

$$\begin{array}{rccl} \mathcal{D}(\,.\,, g^{\mathcal{T}}): & \mathcal{H} & \longrightarrow & \Re+ \\ & h & \longmapsto & \mathcal{D}(h, g^{\mathcal{T}}) \end{array}$$

and the decision rule h_0 at which this minimum is attained. To do so, $\mathcal{LM}$ implements an algorithm $\mathcal{A}$ whose ultimate goal is to find h_0 on the basis of the finite sequence Υ_N of training examples.

Note that the expression of $\mathcal{D}$ is the same as was given in previous chapter, that is:

$$\mathcal{D}(h, g^{\mathcal{T}}) = \sqrt{\mathbf{E}(\; l(h(\mathbf{v}), g^{\mathcal{T}}(\mathbf{v}))\;)} = \sqrt{\int_V l(h(\mathbf{v}), g^{\mathcal{T}}(\mathbf{v}))\, P_{\mathbf{v}}(\mathbf{v})\, d\mathbf{v}} \qquad (6.1)$$

where l is the quadratic loss. The learning machine's algorithm will be indifferently denoted as $\mathcal{A}$ or $\mathcal{A}^l$. The expression of $g^{\mathcal{T}}$ is also as defined in previous chapter:

$$g^{\mathcal{T}}(\mathbf{v}) = \mathbf{E}(w \mid \mathbf{v}) = \int_W w\, P_{w|\mathbf{v}}(w \mid \mathbf{v})\, dw \qquad (6.2)$$

Note that w is related to $g^{\mathcal{T}}(\mathbf{v})$ through the following relationship:

$$w = g^{\mathcal{T}}(\mathbf{v}) + \epsilon \qquad (6.3)$$

where ϵ is the noise associated with the probabilistic environment $\mathcal{E}$ (see equations 5.3 and 5.9 for a similar relationship). By the properties of conditional expectation, it follows from 6.3 that:

$$\mathbf{E}(\epsilon \mid \mathbf{v}) = 0 \qquad (6.4)$$

Remark: The decision rule space $\mathcal{H}$ is considered to be indexed by a subset of $\Re^n$ for some $n \geq 1$, that is, there exist an integer $n \geq 1$ and a subset $T \subseteq \Re^n$, such that the space $\mathcal{H}$ can be expressed as follows: $\mathcal{H} = \{h_{\mathbf{p}} \mid \mathbf{p} \in T\}$. This is the case in most applications, including the one that is developed in this thesis.

6.2 Overcoming the First Obstacle in Minimizing $\mathcal{D}(h, g^{\mathcal{T}})$

The objective of the learning machine $\mathcal{LM} = (\mathcal{H}, \mathcal{A}^l)$ is to minimize the distance $\mathcal{D}(h, g^{\mathcal{T}})$ over all the decision rule space $\mathcal{H}$. This distance involves two functions: h and $g^{\mathcal{T}}$. The function h is an element of the space $\mathcal{H}$ and, as such, it is well known to $\mathcal{LM}$: once the components of $\mathbf{v}$ are measured, the value of $h(\mathbf{v})$ is readily computable. The problem however is $g^{\mathcal{T}}$. Not only it is an unknown function and impossible to derive from first principles, but there is no operational way of getting even sample measurements or any empirical information about it. $g^{\mathcal{T}}$ is indeed buried in noise. What we can measure, with respect to the transformer's response, is the outcome w, and w contains in it both the value of $g^{\mathcal{T}}$ and noise, all mixed up.

So how should $\mathcal{LM}$ proceed to minimize $\mathcal{D}(h, g^{\mathcal{T}})$, when the only information it can get is in the form of noise-corrupted measurements of the outcome w and, of course, the instance $\mathbf{v}$? Theorem 6.1 will be of great help. Before stating it, we need the following definition:

Definition 6.1 (Expected Risk) *Let $\mathcal{E} = (\mathcal{T}, \mathcal{OM}, z, P_z)$ be a probabilistic environment and, associated with it, a learning machine $\mathcal{LM} = (\mathcal{H}, \mathcal{A}^l)$. Let $h \in \mathcal{H}$ be a decision rule. The expected risk $R(h)$ of h is defined as the expected value of the random variable:*

$$l(h(\mathbf{v}), w) = |h(\mathbf{v}) - w|^2$$

when the vector $z = (\mathbf{v}, w)$ is drawn at random in the sample space $Z = V \times W$ according to the pdf $P_z = P_{(\mathbf{v},w)}$ corresponding to environment $\mathcal{E}$. Formally, it is:

$$R(h) = \mathbf{E}(\ l(h(\mathbf{v}), w)\ ) = \int_{V \times W} l(h(\mathbf{v}), w)\, P_{(\mathbf{v},w)}(\mathbf{v}, w)\, d\mathbf{v}\, dw \tag{6.5}$$

Also, to simplify the notations, we need the following definition:

Definition 6.2 (Simplifying Notations) *Let $\mathcal{E} = (\mathcal{T}, \mathcal{OM}, z, P_z)$ be a probabilistic environment and, associated with it, a learning machine $\mathcal{LM} = (\mathcal{H}, \mathcal{A}^l)$. For* *every decision rule $h \in \mathcal{H}$, we define the real-valued function l_h on the sample space $Z = V \times W$ as follows:*

$$\forall (\mathbf{v}, w) \in V \times W, \quad l_h(\mathbf{v}, w) = l(h(\mathbf{v}), w) \tag{6.6}$$

Hence, using the z-notation, equations 6.6 and 6.5 become:

$$\forall z = (\mathbf{v}, w) \in Z, \quad l_h(z) = l(h(\mathbf{v}), w) \tag{6.7}$$

$$\forall h \in \mathcal{H}, \quad R(h) = \mathbf{E}(l_h(z)) = \int_Z l_h(z)\, P_z(z)\, dz \tag{6.8}$$

Theorem 6.1 (Transition $\mathcal{D}(h, g^{\mathcal{T}}) \longrightarrow R(h)$) *Let $\mathcal{E} = (\mathcal{T}, \mathcal{OM}, z, P_z)$ be a probabilistic environment and, associated with it, a learning machine $\mathcal{LM} = (\mathcal{H}, \mathcal{A}^l)$. Let $h_0 \in \mathcal{H}$ be a fixed decision rule. Then the function:*

$$h \longmapsto \mathcal{D}(h, g^{\mathcal{T}})$$

is minimal at h_0 if and only if the function:

$$h \longmapsto R(h)$$

is minimal at h_0.

Proof. Using equation 6.2, it can be shown that the equality:

$$R(h) = \int_{V \times W} [w - g^{\mathcal{T}}(\mathbf{v})]^2\, P_{(\mathbf{v},w)}(\mathbf{v}, w)\, d\mathbf{v}\, dw + [\mathcal{D}(h, g^{\mathcal{T}})]^2 \tag{6.9}$$

holds true for all $h \in \mathcal{H}$. Since the integral $\int_{V \times W} [w - g^{\mathcal{T}}(\mathbf{v})]^2\, P_{(\mathbf{v},w)}(\mathbf{v}, w)\, d\mathbf{v}\, dw$ is independent of h, it follows that $\mathcal{D}(h, g^{\mathcal{T}})$ is minimal if and only if $R(h)$ is minimal, and that both functions attain their minimum at the same function h_0.□

Theorem 6.1 is very important in simplifying the learning problem $\mathcal{LM}$ is faced with. What it means is that minimizing $\mathcal{D}(h, g^{\mathcal{T}})$ or, equivalently, the square of it

$[\mathcal{D}(h, g^{\mathcal{T}})]^2$ over $\mathcal{H}$ amounts to minimizing $R(h)$ over the decision rule space. Look at the expressions of these two functions $[\mathcal{D}(h, g^{\mathcal{T}})]^2$ and $R(h)$:

$$[\mathcal{D}(h, g^{\mathcal{T}})]^2 = \mathbf{E}(\ l(h(\mathbf{v}), g^{\mathcal{T}}(\mathbf{v}))\ ) \tag{6.10}$$

and

$$R(h) = \mathbf{E}(\ l(h(\mathbf{v}), w)\ ) \tag{6.11}$$

From these expressions, it can be seen that, in the course of minimizing $\mathcal{D}(h, g^{\mathcal{T}})$, theorem 6.1 allows us to replace the *unknown* and *non-measurable* noise-free value $g^{\mathcal{T}}(\mathbf{v})$ by the *measurable* noise-corrupted value w, without loosing information on that decision rule h_0 at which the minimum of $\mathcal{D}(h, g^{\mathcal{T}})$ is attained.

The following theorem will be helpful for the development of process uncertainty models such as inequality 4.25:

Theorem 6.2 (First Inequality) *Let $\mathcal{E} = (\mathcal{T}, \mathcal{OM}, z, P_z)$ be a probabilistic environment and, associated with it, a learning machine $\mathcal{LM} = (\mathcal{H}, \mathcal{A}^l)$. Then the inequality:*

$$[\mathcal{D}(h, g^{\mathcal{T}})]^2 \leq R(h) \tag{6.12}$$

holds true for any rule $h \in \mathcal{H}$.

Proof. This inequality is a direct consequence of equality 6.9.□

6.3 Second Obstacle: P_z is not known to $\mathcal{LM}$

Theorem 6.1 is still not enough for $\mathcal{LM}$ to proceed to the determination of the rule h_0 that minimizes $\mathcal{D}(h, g^{\mathcal{T}})$. This is because $R(h)$ is function of the *pdf* P_z: this *pdf* embodies all sources of uncertainty in the environment $\mathcal{E}$ and, as such, it is not known. The objective — and the power — of the framework developed here consists in avoiding any strong *a priori* assumption regarding the sources of uncertainty in $\mathcal{E}$. Consequently, in what follows, P_z is considered fixed but unknown.

Now, having taken this stand on P_z, we have to find a way of minimizing $R(h)$ on the basis of only a *finite* number N of training examples $z_1, z_2, \ldots, z_N$. How to do that? By introducing a principle called ***Inductive Principle of Empirical Risk Minimization*** ($\mathcal{IPERM}$). This principle has emerged in the mid-eighties as a result of an extensive research work by Vapnik (1982, 1995, 1998) and that of Vapnik and Chervonenkis (1968, 1981, 1991) in the area of mathematical statistics and its applications to computational machine learning theory.

6.4 Inductive Principle of Empirical Risk Minimization

Before we state the $\mathcal{IPERM}$, we need to define the meaning of *empirical risk* of a decision rule:

Definition 6.3 (Empirical Risk) *Let $\mathcal{E} = (\mathcal{T}, \mathcal{OM}, z, P_z)$ be a probabilistic environment and, associated with it, a learning machine $\mathcal{LM} = (\mathcal{H}, \mathcal{A}^l)$. Let $h \in \mathcal{H}$ be a decision rule and $\Upsilon_N = (z_1, z_2, \ldots, z_N)$ a finite sequence of N training examples generated and measured in the probabilistic environment $\mathcal{E}$ as a result of one realization of this same environment. The empirical risk $R_{emp}^{\Upsilon_N}(h)$ of h on the sequence Υ_N is defined as the arithmetic mean of the sequence of numbers:*

$$(l_h(z_i))_{i=1,2,\ldots,N}$$

that is:

$$R_{emp}^{\Upsilon_N}(h) = \frac{1}{N}\sum_{i=1}^{N} l_h(z_i) \tag{6.13}$$

Expected and empirical risks, $R(h)$ and $R_{emp}^{\Upsilon_N}(h)$, may seem to introduce new concepts in this framework, but they are not if we go back to the concepts of probability theory. To see that, <u>fix</u> a decision rule h in the space $\mathcal{H}$. Since z is a random variable (see the general description of the framework at section 5.1), the number $l_h(z)$ is then also a random variable. Denote it as ξ, that is:

$$\xi = l_h(z)$$

(Recall that h is fixed) From probability theory, we know that there are two measures of the *central tendency* of a random variable such as ξ:

- an *empirical measure*: given a series of realizations $\xi_1, \xi_2, \ldots, \xi_N$ of the variable ξ, this measure is constructed by computing the arithmetic average $(\sum_i \xi_i)/N$ of this series.

- a *mathematical measure*: this measure is expressed in terms of the *pdf* P_ξ of ξ, that is: $\int \xi P_\xi(\xi)\, d\xi$. It is called expected value.

In this framework, $R_{emp}^{\Upsilon_N}(h)$ represents the empirical measure of the central tendency of $\xi = l_h(z)$ and $R(h)$ represents the mathematical one. The former measure is *approximate* but *computable*, the latter is *exact* but *unknown*. Also, note that, under some conditions with respect to the dependency and heterogeneity of the realizations ξ_i, the empirical measure converges to the mathematical one when N is made infinitely large (White, 1984). This is known as the *Law of Large Numbers* in probability theory. Applying this law to the case of the expected and empirical risks, we get that $R_{emp}^{\Upsilon_N}(h)$ converges (in probability) to $R(h)$ as N is made infinitely large. That is:

$$R_{emp}^{\Upsilon_N}(h) \rightarrow R(h) \qquad as \;\; N \rightarrow \infty \tag{6.14}$$

The reader should note a very important fact here: *the convergence* 6.14 *is valid for a fixed decision rule h in the space $\mathcal{H}$.* This is called pointwise convergence, as opposed to another type of convergence (called uniform convergence) that is discussed briefly in the next sections. The term "pointwise" refers to the fact that the convergence 6.14 occurs only for fixed points (in mathematics, elements of any space can be called points) of $\mathcal{H}$ and not for all points of this space simultaneously.

Now, let us state the $\mathcal{IPERM}$. This principle consists in implementing the following two actions:

- **action 1**: *replace* the expected risk $R(h)$ by the empirical risk $R_{emp}^{\Upsilon_N}(h)$ computed on the basis of one training sequence Υ_N;

- **action 2**: *take* the decision rule $h_{emp}^{\Upsilon_N}$ at which $R_{emp}^{\Upsilon_N}(h)$ reaches its minimum as a good representation of the *best* rule h_0 that minimizes the expected risk $R(h)$.

Therefore, the implementation of the $\mathcal{IPERM}$ comes down to minimizing the empirical risk $R_{emp}^{\Upsilon_N}(h)$, instead of the expected one $R(h)$, over the space $\mathcal{H}$ and then choosing that decision rule $h_{emp}^{\Upsilon_N}$ at which the minimum of $R_{emp}^{\Upsilon_N}(h)$ is reached to describe the transformer's behaviour. Wastewater engineers have been using this procedure for process model identification for years. The reader may then wonder why we are developing a new mathematical framework, if all what we are going to do is to turn back to the traditional model identification procedure? What is the point?

This framework is not about inventing new procedures, but rationalizing existing ones and modelling the uncertainty that is associated with them. Wastewater engineers have been using the traditional identification procedure without being aware of the transitions:

$$\mathcal{D}(h, g^{\mathcal{T}}) \longrightarrow R(h) \longrightarrow R_{emp}^{\Upsilon_N}(h) \tag{6.15}$$

Their decision to rely on empirical risk minimization may be explained by the fact that mechanistic models are usually assumed to contain adequate *a priori* information about the real process and, as a result, very little information would be lost in the transition:

$$R(h) \longrightarrow R_{emp}^{\Upsilon_N}(h) \tag{6.16}$$

Now we know that this is not true: biological processes are very complex and all existing models represent just a simplified picture of the real process behaviour (see Jeppsson's statement in section 2.2). If the sequence Υ_N is a finite one, then there is definitely a loss of information in the transition 6.16, that has always been ignored by wastewater engineers. The aim of this framework is to rationalize and investigate the validity of this transition. First, we determine in what cases the replacement of $R(h)$ by $R_{emp}^{\Upsilon_N}(h)$ can be legitimatized and, second, evaluate the loss of information that occurs in the course of this replacement. To do so, we need to examine the applicability of the $\mathcal{IPERM}$, for which Vapnik's results will be of great help.

6.5 Applicability of the $\mathcal{IPERM}$

In the transition:

$$\mathcal{D}(h, g^{\mathcal{T}}) \longrightarrow R(h) \tag{6.17}$$

there is absolutely no information loss, by virtue of theorem 6.1. As a result, $R(h)$ can be considered as an exact measure of the performance of the decision rule h when this rule is selected by $\mathcal{LM}$ as an approximation of $g^{\mathcal{T}}$. The transition that is problematic is the second one:

$$R(h) \longrightarrow R_{emp}^{\Upsilon_N}(h)$$

$R_{emp}^{\Upsilon_N}(h)$ is indeed just an estimation of $R(h)$. Of course, one may argue that replacing $R(h)$ by $R_{emp}^{\Upsilon_N}(h)$, as suggested in action 1 of the $\mathcal{IPERM}$, can be legitimatized by the fact that, according to the Law of Large Numbers, $R_{emp}^{\Upsilon_N}(h)$ becomes a perfect estimation of $R(h)$ when the size N of the sequence Υ_N is made infinitely large. But, this fact cannot be used to justify action 2 of $\mathcal{IPERM}$. Here is indeed the problem:

> As was done above, denote the decision rules that minimize $R(h)$ and $R_{emp}^{\Upsilon_N}(h)$ as h_0 and $h_{emp}^{\Upsilon_N}$, respectively. This is equivalent to write that:
>
> $$R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N}) = \inf_{h \in \mathcal{H}} R_{emp}^{\Upsilon_N}(h) \tag{6.18}$$
>
> and
>
> $$R(h_0) = \inf_{h \in \mathcal{H}} R(h) \tag{6.19}$$
>
> Action 2 of the $\mathcal{IPERM}$ stipulates to take $h_{emp}^{\Upsilon_N}$ as a good representation of the best rule h_0. For this to be justified, we need to ensure that $h_{emp}^{\Upsilon_N}$ is very "close" to minimizing the expected risk $R(h)$ which is, as pointed out previously, an exact measure of the rule's performance (meaning the rule's closeness to $g^{\mathcal{T}}$ in the sense of $\mathcal{D}$). In more concrete terms, we need that the value $R(h_{emp}^{\Upsilon_N})$ of the expected risk at $h_{emp}^{\Upsilon_N}$ be close to the minimum one $R(h_0)$, for N sufficiently large. That is:
>
> $$R(h_{emp}^{\Upsilon_N}) \rightarrow R(h_0) \qquad as \quad N \rightarrow \infty \tag{6.20}$$
>
> (convergence is understood in probability)

It has been shown (Vapnik and Chervonenkis, 1991) that the pointwise convergence 6.14 does not guarantee the one that is really required for the purpose of the $\mathcal{IPERM}$, i.e., convergence 6.20. In other words, it is possible that convergence 6.14 be satisfied, but $R(h_{emp}^{\Upsilon_N})$ remains always far from $R(h_0)$ — even for large values of N —, meaning that $h_{emp}^{\Upsilon_N}$ would never constitute a good approximation to the transformer's behaviour. It is therefore important to verify whether the $\mathcal{IPERM}$ is applicable or not before using it in any learning problems.

Taking into consideration the foregoing comments, the following definition shall be adopted for the meaning of the applicability of the $\mathcal{IPERM}$:

Definition 6.4 (Applicability of the $\mathcal{IPERM}$) *Let $\mathcal{E} = (\mathcal{T}, \mathcal{OM}, z, P_z)$ be a probabilistic environment and, associated with it, a learning machine $\mathcal{LM} = (\mathcal{H}, \mathcal{A}^l)$. Let Υ_N be a finite sequence of N training examples from the environment $\mathcal{E}$ and let $h_{emp}^{\Upsilon_N}$ and h_0 be two decision rules that minimize the risks $R_{emp}^{\Upsilon_N}(h)$ and $R(h)$, respectively (refer to equations 6.18 and 6.19). The $\mathcal{IPERM}$ is said to be applicable to $(\mathcal{E}, \mathcal{LM})$ if, for any $\varepsilon > 0$, the following equality holds true:*

$$\lim_{N \to \infty} \mathbf{Pr} \left(\sup_{h \in \mathcal{H}} \delta[R(h), R_{emp}^{\Upsilon_N}(h)] > \varepsilon \right) = 0 \tag{6.21}$$

δ being a deviation measure defined on the real line.

Now that the applicability of $\mathcal{IPERM}$ has been defined, we need to develop a simple method of verifying it. In the foregoing discussion, it has been pointed out that the pointwise convergence 6.14 is not enough to guarantee the applicability of $\mathcal{IPERM}$. A more stringent condition regarding the empirical risk convergence needs to be imposed. Vapnik and Chervonenkis (1991) have shown that, for $\mathcal{IPERM}$ to be applicable, it is *necessary* and *sufficient* that the empirical risk $R_{emp}^{\Upsilon_N}(h)$ converges *uniformly* to the expected risk $R(h)$ over the whole space $\mathcal{H}$ (convergence is understood in probability). Mathematically, uniform convergence means that equation 6.21 holds true. Intuitively, it means that, as N is made infinitely large, the whole curve of $R_{emp}^{\Upsilon_N}(h)$ converges to that of $R(h)$ over the space $\mathcal{H}$. In this presentation, the theoretical part of such questions will not be detailed. Instead, the reader is referred to Vapnik's book "*Statistical Learning Theory*" (1998) for the details. In what follows, Vapnik's results are presented in a more practical fashion, allowing direct application to the case under study in this thesis. The mathematical rigor is, however, preserved throughout the whole presentation.

A criterion to verify the applicability of the $\mathcal{IPERM}$ is not the only thing that is needed here. We also want to know how much information is lost when $R(h)$ is replaced by $R_{emp}^{\Upsilon_N}(h)$. Here again, to evaluate this information loss, we need to define a measure of the deviation between $R(h)$ and $R_{emp}^{\Upsilon_N}(h)$. For this purpose, two deviation relative measures are introduced:

- *relative measure* δ_1 defined by:

$$\forall (a_1, a_2) \in \Re^2, \quad \delta_1[a_1, a_2] = \frac{a_1 - a_2}{\sqrt{a_1}} \tag{6.22}$$

- *relative measure* δ_2 defined by:

$$\forall (a_1, a_2) \in \Re^2, \quad \delta_2[a_1, a_2] = \frac{a_1 - a_2}{a_1} \tag{6.23}$$

Each one of these two measures will be associated with a different weak prior information about $(\mathcal{E}, \mathcal{LM})$.

Using these measures, the following theorem 6.3 defines sufficient conditions for the applicability of $\mathcal{IPERM}$ and helps evaluate the loss of information that occurs when $R(h)$ is replaced by $R_{emp}^{\Upsilon_N}(h)$:

Theorem 6.3 (Applicability of the $\mathcal{IPERM}$) *Let $\mathcal{E} = (\mathcal{T}, \mathcal{OM}, z, P_z)$ be a probabilistic environment and, associated with it, a learning machine $\mathcal{LM} = (\mathcal{H}, \mathcal{A}^l)$. Let Υ_N be a finite sequence of N training examples from the environment $\mathcal{E}$ and η a real number in the interval $]0, 1[$. Let δ be one of the deviation measures δ_1 or δ_2. If it is possible to establish some Weak Prior Information $\mathcal{WPI}$ about $(\mathcal{E}, \mathcal{LM})$ and construct a function $\mathcal{C}$ dependent on N, the whole set $\mathcal{H}$, $\mathcal{WPI}$ and the number η such that both statements 1 and 2 listed below hold true, then the $\mathcal{IPERM}$ is applicable to $(\mathcal{E}, \mathcal{LM})$. When such function:*

$$\mathcal{C} = \mathcal{C}(N, \mathcal{H}, \mathcal{WPI}, \eta)$$

exists, the $\mathcal{IPERM}$ is said to be δ-applicable to $(\mathcal{E}, \mathcal{LM})$ with the bound $\mathcal{C}(N, \mathcal{H}, \mathcal{WPI}, \eta)$.

- **Statement 1**: *for any $\eta \in]0, 1[$, the inequality:*

$$\sup_{h \in \mathcal{H}} \delta[R(h), R_{emp}^{\Upsilon_N}(h)] \leq \mathcal{C}(N, \mathcal{H}, \mathcal{WPI}, \eta)$$

  is satisfied with probability of at least $1 - \eta$.

- **Statement 2**: *when $\mathcal{H}, \eta$ and $\mathcal{WPI}$ are fixed, then:*

$$\lim_{N \to \infty} \mathcal{C}(N, \mathcal{H}, \mathcal{WPI}, \eta) = 0$$

Proof. Let $\varepsilon > 0$ and $\eta \in]0, 1[$ be two fixed numbers. From statement 2, we infer that:

$$\exists N_0 \in \aleph, \ \forall N > N_0, \ \mathcal{C}(N, \mathcal{H}, \mathcal{WPI}, \eta) < \varepsilon$$

Then, from statement 1, we get that for $N > N_0$, the inequality:

$$\sup_{h \in \mathcal{H}} \delta[R(h), R_{emp}^{\Upsilon_N}(h)] \leq \varepsilon$$

is satisfied with probability of at least $1 - \eta$. That is:

$$\mathbf{Pr}\left(\sup_{h \in \mathcal{H}} \delta[R(h), R_{emp}^{\Upsilon_N}(h)] > \varepsilon\right) < \eta$$

Thus, we have shown that, for any $\varepsilon > 0$:

$$\forall \eta \in]0, 1[, \ \exists N_0 \in \aleph, \ \forall N > N_0, \ \mathbf{Pr}\left(\sup_{h \in \mathcal{H}} \delta[R(h), R_{emp}^{\Upsilon_N}(h)] > \varepsilon\right) < \eta$$

which means, by definition, that:

$$\lim_{N \to \infty} \mathbf{Pr}\left(\sup_{h \in \mathcal{H}} \delta[R(h), R_{emp}^{\Upsilon_N}(h)] > \varepsilon\right) = 0 \qquad \square$$

Now recall that one of the objectives of this study is to develop an uncertainty model such as the one represented by inequality 4.25. The following theorem defines a way of developing such model:

Theorem 6.4 (Uncertainty Model) *Let $\mathcal{E} = (\mathcal{T}, \mathcal{OM}, z, P_z)$ be a probabilistic environment and, associated with it, a learning machine $\mathcal{LM} = (\mathcal{H}, \mathcal{A}^l)$. Let Υ_N be a finite sequence of N training examples from the environment $\mathcal{E}$ and η a real number in the interval $]0, 1[$. Let $\mathcal{WPI}$ be some weak prior information about $(\mathcal{E}, \mathcal{LM})$ and $h_{emp}^{\Upsilon_N}$ a decision rule at which the empirical risk $R_{emp}^{\Upsilon_N}(h)$ reaches its minimum.*

- *If the $\mathcal{IPERM}$ is δ_1-applicable to $(\mathcal{E}, \mathcal{LM})$ with the bound $C(N, \mathcal{H}, \mathcal{WPI}, \eta)$, then the inequality:*

$$[\mathcal{D}(h_{emp}^{\Upsilon_N}, g^{\mathcal{T}})]^2 \leq R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N}) + \frac{C^2(N, \mathcal{H}, \mathcal{WPI}, \eta)}{2}\left(1 + \sqrt{1 + \frac{4\, R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N})}{C^2(N, \mathcal{H}, \mathcal{WPI}, \eta)}}\right) \qquad (6.24)$$

  holds true with probability of at least $1 - \eta$.

- *If the $\mathcal{IPERM}$ is δ_2-applicable to $(\mathcal{E}, \mathcal{LM})$ with the bound $\mathcal{C}(N, \mathcal{H}, \eta, \mathcal{WPI})$, then the inequality:*

$$[\mathcal{D}(h_{emp}^{\Upsilon_N}, g^{\mathcal{T}})]^2 \leq \frac{R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N})}{(1 - \mathcal{C}(N, \mathcal{H}, \mathcal{WPI}, \eta))_+} \tag{6.25}$$

holds true with probability of at least $1 - \eta$, *where* $(a)_+ = \sup(a, 0)$.

Proof. If the $\mathcal{IPERM}$ is δ_1-applicable to $(\mathcal{E}, \mathcal{LM})$ with the bound $\mathcal{C}(N, \mathcal{H}, \eta, \mathcal{WPI})$, then, from theorem 6.3, it follows that (all inequalities hold with probability of at least $1 - \eta$):

$$\frac{R(h_{emp}^{\Upsilon_N}) - R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N})}{\sqrt{R(h_{emp}^{\Upsilon_N})}} \leq \mathcal{C}(N, \mathcal{H}, \mathcal{WPI}, \eta)$$

Hence:

$$\begin{aligned} R(h_{emp}^{\Upsilon_N}) &\leq R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N}) + \\ &\frac{\mathcal{C}^2(N, \mathcal{H}, \mathcal{WPI}, \eta)}{2}\left(1 + \sqrt{1 + \frac{4\, R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N})}{\mathcal{C}^2(N, \mathcal{H}, \mathcal{WPI}, \eta)}}\right) \end{aligned}$$

and then, from theorem 6.2, it follows that:

$$\begin{aligned} [\mathcal{D}(h_{emp}^{\Upsilon_N}, g^{\mathcal{T}})]^2 &\leq R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N}) + \\ &\frac{\mathcal{C}^2(N, \mathcal{H}, \mathcal{WPI}, \eta)}{2}\left(1 + \sqrt{1 + \frac{4\, R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N})}{\mathcal{C}^2(N, \mathcal{H}, \mathcal{WPI}, \eta)}}\right) \end{aligned}$$

Similarly, if the $\mathcal{IPERM}$ is δ_2-applicable to $(\mathcal{E}, \mathcal{LM})$ with the bound $\mathcal{C}(N, \mathcal{H}, \eta, \mathcal{WPI})$, then, from theorem 6.3, it follows that:

$$\frac{R(h_{emp}^{\Upsilon_N}) - R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N})}{R(h_{emp}^{\Upsilon_N})} \leq \mathcal{C}(N, \mathcal{H}, \mathcal{WPI}, \eta)$$

and then:

$$[\mathcal{D}(h_{emp}^{\Upsilon_N}, g^{\mathcal{T}})]^2 \leq R(h_{emp}^{\Upsilon_N}) \leq \frac{R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N})}{(1 - \mathcal{C}(N, \mathcal{H}, \mathcal{WPI}, \eta))_+} \qquad \square$$

The bound on the squared distance $[\mathcal{D}(h_{emp}^{\Upsilon_N}, g^{\mathcal{T}})]^2$, when it exists, is called *guaranteed deviation* between $h_{emp}^{\Upsilon_N}$ and $g^{\mathcal{T}}$, and denoted as φ or $\varphi(N, \mathcal{H}, R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N}), \mathcal{WPI}, \eta)$.

6.6 The Vapnik-Chervonenkis (VC) Dimension

One of the objects which the guaranteed deviation φ is dependent on is the whole set $\mathcal{H}$ of decision rules. Now we need to know exactly what characteristic of $\mathcal{H}$ affects φ and the uncertainty models 6.24 and 6.25. From our discussion of uncertainty model development in Chapter 3, it can be easily inferred that this characteristic is the complexity of $\mathcal{H}$. The objective of this section is to define a measure of this complexity. This measure is known as the *Vapnik-Chervonenkis dimension*, or simply VC dimension, named in honor of its originators, Vapnik and Chervonenkis (1968). The definition of this dimension is quite difficult to assimilate from the first reading. Because of this, an intuitive interpretation of VC dimension will be first given and, at the end of this section, a series of illustrative examples will be presented.

6.6.1 Intuitive Introduction

Consider the following concrete example:

- $V_1 = \Re$ and $W_1 = \Re$;
- $\mathcal{H} = \mathcal{H}_{line}$ is the set of all functions h from V into W such that:

$$\forall x \in V, \quad h(x) = p_1 x + p_2$$

  with $\mathbf{p} = (p_1, p_2) \in \Re^2$ is the parameter vector.

If we had to assign a number to the complexity of this set of functions, then intuitively the number two, corresponding to the number of parameters, would be the most suitable one. Consider now this second example:

- $V_2 = \Re$ and $W_2 = \Re$;
- $\mathcal{H} = \mathcal{H}_{sine}$ is the set of all functions h from V into W such that:

$$\forall x \in V, \quad h(x) = p_1 \sin(p_2 x)$$

  with $\mathbf{p} = (p_1, p_2) \in \Re^2$ is the parameter vector.

Since the number of parameters that define this set is also two, we may be tempted to again assign the number two to the complexity of this set. If we do so, it would mean that $\mathcal{H}_{line}$ and $\mathcal{H}_{sine}$ have the same degree of complexity, which is obviously not correct: the set $\mathcal{H}_{line}$ is a family of just straight lines, while $\mathcal{H}_{sine}$ is a complex family of curves that can take many different shapes. The "expressive power" of $\mathcal{H}_{sine}$ is indeed much higher than that of $\mathcal{H}_{line}$. As a result, it should be expected that the complexity of $\mathcal{H}_{sine}$ be much higher than that of $\mathcal{H}_{line}$, and that is what we get when we consider the VC dimension as a measure of the complexity of the decision rule space.

Intuitively, the VC dimension may be considered as equal to the maximum number of points that the curves representing the functions of the decision rule space can pass through simultaneously. Straight lines (functions defined by $h(x) = p_1 x + p_2$, space $\mathcal{H}_{line}$) can pass through any 2 points, but not any 3 points. Parabolas (functions defined by $h(x) = p_1 x^2 + p_2 x + p_3$, space $\mathcal{H}_{parab}$) can pass through any 3 points, but not any 4 points. Sine functions ($h(x) = p_1 \sin(p_2 x)$, space $\mathcal{H}_{sine}$) can pass through any number of points. Hence, if the VC dimension of a space $\mathcal{H}$ is denoted as $q(\mathcal{H})$, then:

$$q(\mathcal{H}_{line}) = 2$$

$$q(\mathcal{H}_{parab}) = 3$$

$$q(\mathcal{H}_{sine}) = \infty$$

The foregoing intuitive interpretation of VC dimension is approximate. A more precise definition of it is given in the next section.

6.6.2 Definitions

For every set I, the notation 2^I will designate the set of all subsets of I.

Definition 6.5 (VC Dimension of a Family of Sets) *Let $\boldsymbol{G}$ be some space ($\Re^n$ with $n > 0$ for example or any other space). Let $\mathcal{G}$ be a family of subsets of $\boldsymbol{G}$ (examples of $\mathcal{G}$ in the case of $\boldsymbol{G} = \Re^2$ are the family of all open (or closed) balls of $\Re^2$ or the family of all half planes of $\Re^2$) and I a finite subset of $\boldsymbol{G}$. Let $\Pi^{\mathcal{G}}(I)$ be the subset of 2^I defined as follows:*

$$\Pi^{\mathcal{G}}(I) = \{\Lambda \in 2^I | \exists F \in \mathcal{G}, \Lambda = F \cap I\}$$

The finite set I is said to be **shattered** *by the family of sets $\mathcal{G}$ if $\underline{\Pi^{\mathcal{G}}(I) = 2^I}$. The largest integer q such that some finite subset $I \subset \boldsymbol{G}$ of* **size** *q is shattered by $\mathcal{G}$ is* *called the* Vapnik-Chervonenkis dimension *(VC dimension) of the family $\mathcal{G}$. It is denoted by $q = q(\mathcal{G})$. If such integer q does not exist, then the VC dimension of $\mathcal{G}$ is said to be infinite.*

Definition 6.6 (VC Dimension of a Family of Functions) *Let $\mathcal{F}$ be a family of real-valued functions on some space $\boldsymbol{G}$ and I a finite subset of $\boldsymbol{G}$. For every function $f \in \mathcal{F}$, define the subset pos(f) of the space $\boldsymbol{G}$ as follows:*

$$pos(f) = \{a \in \boldsymbol{G} |\ f(a) > 0\}$$

Then define the family $pos(\mathcal{F})$ of subsets of $\boldsymbol{G}$ as follows:

$$pos(\mathcal{F}) = \{pos(f) |\ f \in \mathcal{F}\}$$

The finite set I is said to be shattered by the family of real-valued functions $\mathcal{F}$, if it is shattered by the family of subsets $pos(\mathcal{F})$. The Vapnik-Chervonenkis dimension *(VC dimension) $q(\mathcal{F})$ of the family $\mathcal{F}$ of real-valued functions is, by definition, equal to the* Vapnik-Chervonenkis dimension *of the family of subsets $pos(\mathcal{F})$:*

$$q(\mathcal{F}) = q(pos(\mathcal{F}))$$

The VC dimension is then a purely combinatorial concept that has, *a priori*, no connection with the geometric notion of dimension. In most situations, it is difficult to evaluate the VC dimension by analytic means. Usually, all that it is possible is to determine a bound on the VC dimension, that is, establish an inequality of the form: $q(\mathcal{F}) \leq q_0$ ($q_0 \in \aleph$). Also in some cases the VC dimension is simply approximated by the free parameters of the family $\mathcal{F}$. The following theorem shows how to determine it in some particular cases. It also establishes a link with the geometric notion of dimension.

Theorem 6.5 (VC Dimension and Vector Space) *Let $\mathcal{F}$ be a family of real-valued functions on some space $\boldsymbol{G}$. Fix any function f_0 from $\boldsymbol{G}$ into $\Re$ and let $\mathcal{F}_0$ be the new family of functions defined by $\mathcal{F}_0 = f_0 + \mathcal{F} = \{f_0 + f |\ f \in \mathcal{F}\}$. If $\mathcal{F}$ is an m-dimensional real vector space, then the VC dimension $q(\mathcal{F}_0)$ of $\mathcal{F}_0$ is equal to m:*

$$q(\mathcal{F}_0) = m$$

Proof. Refer to Wenocur and Dudley (1981) for the proof of this theorem □.

6.6.3 Examples

- **Example 1:** Consider the family of functions $h_{\mathbf{p}}$ defined from the space $\mathbf{G} = \Re^n$ ($n \in \aleph^\circ$) into $\{0, 1\}$ by:

$$\forall \mathbf{x} = (x_1, x_2, \ldots, x_n) \in \Re^n, \qquad h_{\mathbf{p}}(\mathbf{x}) = \psi(\sum_{i=1}^{n} p_i x_i)$$

where $\mathbf{p} = (p_1, p_2, \ldots, p_n, \theta) \in \Re^{n+1}$ is the parameter vector and ψ is defined by (real threshold θ):

$$\psi(a) = \begin{cases} 1 & \text{if } a \geq \theta \\ 0 & \text{if } a < \theta \end{cases}$$

This family of functions is known as the *perceptron* and is used in pattern recognition. Its VC dimension is equal to $n + 1$ (Anthony and Biggs, 1992).

- **Example 2:** Consider the family of real-valued functions $h_{\mathbf{p}}$ defined on some space $\mathbf{G}$ by:

$$\forall \mathbf{x} \in \mathbf{G}, \qquad h_{\mathbf{p}}(\mathbf{x}) = \sum_{i=1}^{n} p_i \psi_i(\mathbf{x})$$

where $\mathbf{p} = (p_1, p_2, \ldots, p_n) \in \Re^n$ is the parameter vector and $\psi_1, \psi_2, \ldots, \psi_n$ is a sequence of n linearly independent real-valued functions. The VC dimension of this family of functions is equal to n (Vapnik, 1982). Note that the determination of this VC dimension results directly from theorem 6.5.

- **Example 3:** Consider the family of functions $h_{\mathbf{p}}$ defined on $\mathbf{G} = \Re^2$ by:

$$\forall (x, y) \in \Re^2, \qquad h_{\mathbf{p}}(x, y) = (y - poly_n(x, \mathbf{p}))^2$$

where $\mathbf{p} = (p_0, p_1, p_2, \ldots, p_n) \in \Re^{n+1}$ is the parameter vector and $poly_m(x, \mathbf{p})$ is a polynomial function of degree n defined by:

$$\forall x \in \Re, \qquad poly_n(x, \mathbf{p}) = p_0 + p_1 x + p_2 x^2 + \ldots + p_n x^n$$

The VC dimension of this family of functions $h_{\mathbf{p}}$ is at most $2n + 2$ (Vapnik, 1995)

- **Example 4:** Consider the family of functions $h_{\mathbf{p}}$ defined on $\mathbf{G} = \Re$ by:

$$\forall x \in \Re, \quad h_{\mathbf{p}}(x) = p_1 \sin(p_2 x)$$

where $\mathbf{p} = (p_1, p_2) \in \Re^2$ is the parameter vector. The VC dimension of this family of functions is infinite (Vapnik, 1995).

From these examples, it can be seen that, generally speaking, the VC dimension of a family of functions is not always related to the number of parameters. It can be larger (example 4), equal (examples 1 and 2) or smaller (see Vapnik, 1995 where new types of learning machines were constructed) than the number of parameters.

6.7 VC dimension and applicability of the $\mathcal{IPERM}$

In section 6.5, the concept of applicability of $\mathcal{IPERM}$ and that of guaranteed deviation between the decision rule $h_{emp}^{\Upsilon_N}$ that minimizes the empirical risk and the transformer's response function $g^{\mathcal{T}}$ were introduced. However, no methodology has been developed to determine the expression of the function $\mathcal{C} = \mathcal{C}(N, \mathcal{H}, \mathcal{WPI}, \eta)$ (see theorems 6.3 and 6.4), which is the key function in implementing those concepts. In this section, some fundamental results with respect to the determination of such function are presented. These results make use of the VC dimension concept defined in the previous section and they are due to Vapnik (1998). Applications of these results to model identification and quality evaluation will be discussed in next Chapters.

Before stating these results, we need to define a new space $l_{\mathcal{H}}$ and five different conditions.

Definition 6.7 (Space $l_{\mathcal{H}}$) *Let $\mathcal{E} = (\mathcal{T}, \mathcal{OM}, z, P_z)$ be a probabilistic environment and, associated with it, a learning machine $\mathcal{LM} = (\mathcal{H}, \mathcal{A}^l)$. For every decision rule $h \in \mathcal{H}$ and a real number $\beta \in \Re^+$, we define the real-valued functions $l_{h,\beta}$ on the sample space $Z = V \times W$ as follows:*

$$\forall z \in Z, \quad l_{h,\beta}(z) = l_h(z) - \beta$$

The functional space of all functions $l_{h,\beta}$ will be denoted by $l_{\mathcal{H}}$:

$$l_{\mathcal{H}} = \{l_{h,\beta} | \ (h, \beta) \in \mathcal{H} \times \Re^+\}$$

Now let us define the following conditions ***C.1***, ***C'.1***, ***C.2***, ***C.3*** and ***C'.3***:

C.1 Weak Prior Information (1):
There exists a positive number $M \in]0, +\infty[$ *such that:*

$$\sup_{h \in \mathcal{H}, z \in Z} l_h(z) = M$$

C'.1 Weak Prior Information (2):
There exist a pair $(s, \tau) \in \Re^2$ *with* $s > 2$ *and* $\tau < +\infty$ *such that:*

$$\sup_{h \in \mathcal{H}} \frac{\mathbf{E}^{1/s}([l_h(z)]^s)}{R(h)} < \tau$$

C.2 VC Dimension:
The VC dimension $q = q(l_{\mathcal{H}})$ *of the functional space* $l_{\mathcal{H}}$ *is finite.*

C.3 i.i.d. condition:
The training examples:

$$z_1, z_2, \ldots, z_N$$

of the sequence Υ_N *are independent and identically distributed (i.i.d.).*

C'.3 Weaker i.i.d. condition:
The real-valued random variables:

$$l_h(z_1); \; l_h(z_2); \; \ldots; \; l_h(z_N)$$

obtained by computing the values of l_h *at each one of the training examples* z_i *of the sequence* Υ_N, *are independent and identically distributed (i.i.d.) for any* $h \in \mathcal{H}$.

Theorem 6.6 ($\mathcal{IPERM}$ applicability and VC (1)) *Let* $\mathcal{E} = (\mathcal{T}, \mathcal{OM}, z, P_z)$ *be a probabilistic environment and, associated with it, a learning machine* $\mathcal{LM} = (\mathcal{H}, \mathcal{A}^l)$. *Let* Υ_N *be a finite sequence of* N *training examples from the environment* $\mathcal{E}$ *and* η *a real number in the interval* $]0, 1[$. *If the conditions* ***C.1***, ***C.2*** *and* ***C.3*** *are satisfied, then the* $\mathcal{IPERM}$ *is* δ_1*-applicable to* $(\mathcal{E}, \mathcal{LM})$ *with the bound:*

$$\mathcal{C} = \sqrt{M \zeta} \tag{6.26}$$

where:

- *The number* ζ *is:*

$$\zeta = 4\frac{\left[q\left(\ln\left(\frac{2N}{q}\right)+1\right)-\ln\left(\frac{\eta}{4}\right)\right]}{N}$$

- q *is the VC dimension* $q(l_{\mathcal{H}})$ *of the space* $l_{\mathcal{H}}$.

Proof. Vapnik (1998) showed that, for any $\varepsilon > 0$, the following inequality holds true:

$$\mathbf{Pr}\left(\sup_{h\in\mathcal{H}} \delta_1[R(h), R_{emp}^{\Upsilon_N}(h)] > \varepsilon\right) < 4\ \exp\left[\left(\frac{q\left(\ln\left(\frac{2N}{q}\right)+1\right)}{N} - \frac{\varepsilon^2}{4M}\right)N\right] \quad (6.27)$$

when conditions **C.1**, **C.2** and **C.3** are satisfied (Vapnik (1998), see inequalities 5.24 and 5.12 at pages 197 and 192 respectively). Set the right hand side of the above inequality equal to η. Then the expression of ε is:

$$\varepsilon = \sqrt{M\zeta}$$

and, therefore, from Vapnik's inequality, it follows that the inequality:

$$\sup_{h\in\mathcal{H}} \delta_1[R(h), R_{emp}^{\Upsilon_N}(h)] < \sqrt{M\zeta}$$

holds true with probability of at least $1-\eta$.□

Theorem 6.7 ($\mathcal{IPERM}$ applicability and VC (2)) *Let* $\mathcal{E} = (\mathcal{T}, \mathcal{OM}, z, P_z)$ *be a probabilistic environment and, associated with it, a learning machine* $\mathcal{LM} = (\mathcal{H}, \mathcal{A}^l)$. *Let* Υ_N *be a finite sequence of* N *training examples from the environment* $\mathcal{E}$ *and* η *a real number in the interval* $]0,1[$. *If the conditions* ***C'.1***, ***C.2*** *and* ***C.3*** *are satisfied, then the* $\mathcal{IPERM}$ *is* δ_2*-applicable to* $(\mathcal{E}, \mathcal{LM})$ *with the bound:*

$$\mathcal{C} = \gamma(s)\,\tau\sqrt{\zeta} \quad (6.28)$$

where:

- $\gamma(s) = \sqrt[s]{\frac{1}{2}\left(\frac{s-1}{s-2}\right)^{s-1}}$
- *The number* ζ *is:*

$$\zeta = 4\frac{\left[q\left(\ln\left(\frac{2N}{q}\right)+1\right)-\ln\left(\frac{\eta}{4}\right)\right]}{N}$$

- q *is the VC dimension* $q(l_{\mathcal{H}})$ *of the space* $l_{\mathcal{H}}$.

Proof. Vapnik (1998) showed that, for any $\varepsilon > 0$, the following inequality holds true:

$$\mathbf{Pr}\left(\sup_{h\in\mathcal{H}} \delta_2[R(h), R_{emp}^{\Upsilon_N}(h)] > \gamma(s)\,\tau\,\varepsilon\right) < 4\,\exp\left[\left(\frac{q\left(\ln\left(\frac{2N}{q}\right)+1\right)}{N} - \frac{\varepsilon^2}{4}\right)N\right] \quad (6.29)$$

when conditions **C'.1**, **C.2** and **C.3** are satisfied (Vapnik (1998), see inequalities 5.43 and 5.12 at pages 210 and 192 respectively). Set the right hand side of the above inequality equal to η. Then the expression of ε is:

$$\varepsilon = \sqrt{\zeta}$$

and, therefore, the inequality:

$$\sup_{h\in\mathcal{H}} \delta_2[R(h), R_{emp}^{\Upsilon_N}(h)] < \gamma(s)\,\tau\sqrt{\zeta}$$

holds true with probability of at least $1-\eta$.□

Note that $\mathcal{WPI}$ is represented by the number M in theorem 6.6 and by the numbers s and τ in theorem 6.7.

The following theorem uses a weaker *i.i.d.* condition (***C'.3***):

Theorem 6.8 (Using condition *C'.3*) *If the third condition* ***C.3*** *in the two previous theorems* 6.6 *and* 6.7 *is replaced by the condition* ***C'.3*** *and the two other conditions,* ***C.1*** *and* ***C.2*** *for theorem* 6.6 *and* ***C'.1*** *and* ***C.2*** *for theorem* 6.7, *are kept unchanged, then the* $\mathcal{IPERM}$ *is still applicable to* $(\mathcal{E}, \mathcal{LM})$ *with respect to the same deviation measures* δ_1 *and* δ_2 *and with the same bounds* 6.26 *and* 6.27, *respectively.*

Proof. To prove inequalities 6.27 and 6.29, Vapnik (1982,1998) made use of the weaker i.i.d. condition only. As a result, these inequalities remain true if condition **C.3** is replaced by condition **C'.3**. Consequently, the foregoing proofs of theorems 6.7 and 6.6 are still valid with condition **C'.3**.□

Using theorems 6.6, 6.7, 6.8 and 6.4, it is now possible to develop uncertainty models for $(\mathcal{E}, \mathcal{LM})$ with a guaranteed deviation φ that is readily computable:

Theorem 6.9 (Uncertainty Model and VC) *Let $\mathcal{E} = (\mathcal{T}, \mathcal{OM}, z, P_z)$ be a probabilistic environment and, associated with it, a learning machine $\mathcal{LM} = (\mathcal{H}, \mathcal{A}^l)$. Let Υ_N be a finite sequence of N training examples from the environment $\mathcal{E}$ and η a real number in the interval $]0, 1[$. Let $h_{emp}^{\Upsilon_N}$ be a decision rule at which the empirical risk $R_{emp}^{\Upsilon_N}(h)$ reaches its minimum.*

- *If the conditions **C.1**, **C.2** and **C'.3** are satisfied, then the inequality:*

$$[\mathcal{D}(h_{emp}^{\Upsilon_N}, g^{\mathcal{T}})]^2 \leq R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N}) + \frac{M\zeta}{2}\left(1 + \sqrt{1 + \frac{4\,R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N})}{M\zeta}}\right) \qquad (6.30)$$

  holds true with probability of at least $1 - \eta$.

- *If the conditions **C'.1**, **C.2** and **C'.3** are satisfied, then the inequality:*

$$[\mathcal{D}(h_{emp}^{\Upsilon_N}, g^{\mathcal{T}})]^2 \leq \frac{R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N})}{(1 - \gamma(s)\,\tau\sqrt{\zeta})_+} \qquad (6.31)$$

  holds true with probability of at least $1 - \eta$.

  ⋆ $(a)_+ = \sup(a, 0)$ *for any number* $a \in \Re$;

  ⋆ $\gamma(s) = \sqrt[s]{\frac{1}{2}\left(\frac{s-1}{s-2}\right)^{s-1}}$

  ⋆ *The number ζ is:*

$$\zeta = 4\frac{\left[q\left(\ln\left(\frac{2N}{q}\right) + 1\right) - \ln\left(\frac{\eta}{4}\right)\right]}{N} \qquad (6.32)$$

  ⋆ *q is the VC dimension $q(l_{\mathcal{H}})$ of the space $l_{\mathcal{H}}$.*

Proof. This theorem is a direct consequence of theorems 6.8 and 6.4. □

Theorem 6.9 is the most important result of this Chapter. It establishes two uncertainty models, $\mathcal{UM}_1$ and $\mathcal{UM}_2$, for $(\mathcal{E}, \mathcal{LM})$. The first one, $\mathcal{UM}_1$, is based on the weak prior information $\mathcal{WPI}(1)$ and is defined by inequality 6.30. The right-hand side of this inequality represents the guaranteed deviation φ_1 between $h_{emp}^{\Upsilon_N}$ and $g^{\mathcal{T}}$, developed on the basis of $\mathcal{WPI}(1)$. Using this function φ_1, the uncertainty model $\mathcal{UM}_1$ can be re-written as follows:

$$\mathcal{UM}_1 : \qquad [\mathcal{D}(h_{emp}^{\Upsilon_N}, g^{\mathcal{T}})]^2 \leq \varphi_1(N, \mathcal{H}, R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N}), \mathcal{WPI}(1), \eta) \qquad (6.33)$$

with:

$$\varphi_1(N, \mathcal{H}, R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N}), \mathcal{WPI}(1), \eta) = R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N}) + \frac{M\zeta}{2}\left(1 + \sqrt{1 + \frac{4\, R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N})}{M\zeta}}\right) \quad (6.34)$$

The second model, $\mathcal{UM}_2$, is based on the weak prior information $\mathcal{WPI}(2)$ and is defined by inequality 6.31. Denoting the right-hand side of this inequality as φ_2 (guaranteed deviation developed on the basis of $\mathcal{WPI}(2)$), the uncertainty model $\mathcal{UM}_2$ can be re-written as:

$$\mathcal{UM}_2 : \qquad [\mathcal{D}(h_{emp}^{\Upsilon_N}, g^{\mathcal{T}})]^2 \leq \varphi_2(N, \mathcal{H}, R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N}), \mathcal{WPI}(2), \eta) \quad (6.35)$$

with:

$$\varphi_2(N, \mathcal{H}, R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N}), \mathcal{WPI}(2), \eta) = \frac{R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N})}{(1 - \gamma(s)\,\tau\sqrt{\zeta})_+} \quad (6.36)$$

Part IV

UNCERTAINTY MODELLING AND MANAGEMENT IN THE ASP

Chapter 7

Modelling the Uncertainty in the Activated Sludge Process

7.1 Introduction

The purpose of this Chapter is to investigate the uncertainty that underlies the behaviour of a simple activated sludge wastewater treatment plant composed of a completely mixed reactor and a secondary clarifier. This investigation is carried out under the conditions that the plant's behaviour is approximated by the following mechanistic model $\mathcal{M}$:

$$\mathcal{M}\begin{cases} \dot{S} = \frac{(Q_{in}S_{in}+Q_rS_r)}{V} - \frac{(Q_{in}+Q_r)}{V}S - \frac{kS}{K_S+S}X \\ \\ \dot{X} = \frac{(Q_{in}X_{in}+Q_rX_r)}{V} - \frac{(Q_{in}+Q_r)}{V}X + \frac{\mu_H S}{K_S+S}X - bX \end{cases} \tag{7.1}$$

The variables and parameters of this model are defined below:

- S is the substrate concentration in the bio-reactor and effluent (process first state variable).
- X is the microorganisms concentration in the bio-reactor and effluent (process second state variable).
- Q is the flow rate.
- the subscript $_{in}$ means influent.
- the subscript $_r$ means recycle.

- V is the bio-reactor volume.
- μ_H, b, k, K_S are the model parameters.

The equations of model $\mathcal{M}$ can be re-arranged differently: define the input vector $\mathbf{u}$ as:

$$\mathbf{u} = \left(\frac{Q_{in}}{V}, \frac{Q_r}{V}, S_{in}, S_r, X_{in}, X_r\right) \tag{7.2}$$

and the unit vectors $\vec{e_i}$ of the space $\Re^6$ whose ith component is 1 and all other components are 0. Then model $\mathcal{M}$ can be re-written as:

$$\mathcal{M} \begin{cases} \dot{S} &= (u_S.u_Q^T) - (\vec{a}.u_Q^T)S + r_S(S, X, \mathbf{p}_S) \\ \dot{X} &= (u_X.u_Q^T) - (\vec{a}.u_Q^T)X + r_X(S, X, \mathbf{p}_X) \end{cases} \tag{7.3}$$

where:

- The superscript ^{T} means transposed vector.
- $u_Q = [\mathbf{u}.\vec{e_1}^T, \mathbf{u}.\vec{e_2}^T] = [\frac{Q_{in}}{V}, \frac{Q_r}{V}]$.
- $u_S = [\mathbf{u}.\vec{e_3}^T, \mathbf{u}.\vec{e_4}^T] = [S_{in}, S_r]$.
- $u_X = [\mathbf{u}.\vec{e_5}^T, \mathbf{u}.\vec{e_6}^T] = [X_{in}, X_r]$.
- $\vec{a} = [1, 1]$.
- $r_S(S, X, \mathbf{p}_S)$ is the rate of biodegradation of S. Its expression is defined as:

$$r_S(S, X, \mathbf{p}_S) = -\frac{kS}{K_S + S}X$$

- $r_X(S, X, \mathbf{p}_X)$ is the rate of growth of X defined as:

$$r_X(S, X, \mathbf{p}_X) = \frac{\mu_H S}{K_S + S}X - bX$$

- $\mathbf{p}_S$ and $\mathbf{p}_X$ represent parameter vectors defined by:

$$\mathbf{p}_S = (k, K_S)$$

$$\mathbf{p}_X = (\mu_H, b, K_S)$$

In ASP mathematical modelling, one often hears assertions such as "the more data we have, the better it is for model identification". While this is a fact, such assertion implies that the only way of getting the best approximation of the plant's behaviour would be by supplying an infinite amount of data to the model identification procedure. Since data sets are always finite and usually of a small size, poor predictions by the identified model are often attributed to the lack of adequate amount of data. In addition, if the process model is too complex, lack of model identifiability is also blamed. If, on the other hand, the model is too simple the explanation of poor model predictions is usually straightforward: the simple model structure does not account for all the process dynamical modes.

Up until now, there is no general methodology to account for the effect of each of these parameters — lack of data, model complexity, model simplicity — on the quality of an ASP model prediction. We do not have answers to the questions of *"when and what causes what"*. In this Chapter and next ones, it will be shown how the mathematical framework of Chapter 6 can be utilized to address such crucial questions. More specifically, in this Chapter, uncertainty models for the ASP will be developed. Based on these models and other ones developed in Chapter 9, a study of uncertainty management in the ASP will be carried out in Chapters 8 and 10.

The purpose of next section is to define the mathematical objects:

$$\mathcal{E}\ ,\ \mathcal{T}\ ,\ \mathbf{v}\ ,\ w\ , P_{\mathbf{v}}\ ,\ P_{w|\mathbf{v}}$$

(see previous Chapter) for a wastewater treatment plant. The subscript $_{asp}$ will be used occasionally to indicate that these objects pertain to an activated sludge process plant.

7.2 Environment $\mathcal{E}_{asp}$ for a Wastewater Treatment Plant

The probabilistic environment $\mathcal{E}_{asp}$ for a wastewater treatment plant can be an urban area, a city, a small community or a watershed. The transformer $\mathcal{T}_{asp}$ is the wastewater treatment plant itself, which is located within the environment $\mathcal{E}_{asp}$. This plant uses an activated sludge process (this study is limited to the ASP only) to treat the wastewater generated in $\mathcal{E}_{asp}$. The situation $\mathbf{v}$ encompasses the inputs to the plant and the state variables of the ASP. It takes all its values in a space V. The probability density function $P_\mathbf{v}$ is a characteristic of the nature and amount of uncertainty associated with the environment $\mathcal{E}_{asp}$. Two environments $\mathcal{E}_{asp_1}$ and $\mathcal{E}_{asp_2}$ with similar features (population, people's customs, types of industries, climate, plant configuration, ...) would have almost the same probability density function. The outcome w is the future value of one state variable of the treatment process. In this study, w can be either the substrate concentration S or the microorganisms concentration X. In what follows, the former concentration is selected: $w = S(t)$. This variable takes values from some subspace W of $\Re$. The conditional probability density function, $P_{w|\mathbf{v}}$, of the outcome w given the instance $\mathbf{v}$ is a characteristic of the plant $\mathcal{T}_{asp}$. Two plants $\mathcal{T}_{asp_1}$ and $\mathcal{T}_{asp_1}$ with similar design, history, operating mode and control strategy would have almost the same conditional probability density function.

7.3 Learning Machine for the Environment $\mathcal{E}_{asp}$

As was suggested by Henze *et al.* (1987), the model's first differential equation:

$$\dot{S} = (u_S.u_Q^T) - (\vec{a}.u_Q^T)S + r_S(S, X, \mathbf{p}_S) \tag{7.4}$$

can be integrated using the Euler method: the time is discretized with a time step of Δt and then the values of the substrate concentration S are computed at times

$$t_1 = \Delta t,\ t_2 = 2\,\Delta t,\ \ldots, t_n = n\,\Delta t,\ \ldots$$

using the following equation:

$$\begin{aligned} S(t_n) &= S(t_{n-1}) + \\ &\Delta t\,[(u_S(t_{n-1}).u_Q^T(t_{n-1})) - (\vec{a}.u_Q^T(t_{n-1}))\,S(t_{n-1}) + r_S(S(t_{n-1}), X(t_{n-1}), \mathbf{p}_S)] \end{aligned}$$

For the sake of notation simplification, the subscript $_n$ is used to designate a variable value at time instant $t_n = n\,\Delta t$. Using this convention, the previous equation becomes:

$$S_n = S_{n-1} + \Delta t\,[(u_{S_{n-1}}.u_{Q_{n-1}}^T) - (\vec{a}.u_{Q_{n-1}}^T)\,S_{n-1} + r_S(S_{n-1}, X_{n-1}, \mathbf{p}_S)] \tag{7.5}$$

The situation vector $\mathbf{v}$ consists of the values S_{n-1} and X_{n-1} of the substrate and microorganisms concentrations, and of the relevant components of the input vector $\mathbf{u}_{n-1}$ at time instant $(n-1)\,\Delta t$, i.e., the components of $u_{Q_{n-1}}$

$$\frac{Q_{in_{n-1}}}{V}, \frac{Q_{r_{n-1}}}{V}$$

and those of $u_{S_{n-1}}$

$$S_{in_{n-1}}, S_{r_{n-1}}$$

Formally, $\mathbf{v}$ is:

$$\mathbf{v} = [S_{n-1}, X_{n-1}, \frac{Q_{in_{n-1}}}{V}, \frac{Q_{r_{n-1}}}{V}, S_{in_{n-1}}, S_{r_{n-1}}] \qquad (7.6)$$

In most real-world situations, the variations of X over time are relatively small. In this study, it will be considered that these small variations of $X(t)$ are known. If not, then $X(t)$ will be assumed close to X_0, i.e., $X(t) \simeq X_0$, where X_0 (and S_0) are the initial concentration(s) which, in turn, correspond to the steady state conditions of the process. This assumption needs to be investigated further in connection with the process observability study (see Chapter 3, section on research needs for the ASP), which is beyond the scope of this work.

Thus, if the approximation $X(t) \approx X_0$ is assumed, $\mathbf{v}$ would become:

$$\mathbf{v} = [S_{n-1}, \frac{Q_{in_{n-1}}}{V}, \frac{Q_{r_{n-1}}}{V}, S_{in_{n-1}}, S_{r_{n-1}}] \qquad (7.7)$$

Otherwise, $\mathbf{v}$ will be defined by equation 7.6.

Since the outcome w is equal to the concentration S_n to be predicted, equation 7.5 can be written as:

$$w = H(\mathbf{v}, \mathbf{p}_S)$$

where the function H is defined by:

$$H(\mathbf{v}, \mathbf{p}_S) = S_{n-1} + \Delta t\,[(u_{S_{n-1}}.u^T_{Q_{n-1}}) - (\vec{a}.u^T_{Q_{n-1}})\,S_{n-1} + r_S(S_{n-1}, X_{n-1}, \mathbf{p}_S)]$$

The parameter vector $\mathbf{p}_S = (k, K_S)$ takes values from a certain space

$$\Gamma = \Gamma_k \times \Gamma_{K_S} \subset \Re^2$$

Based on the extensive experience of researchers and practitioners in the area of wastewater engineering, it has been established that k and K_S take only positive values for any operating mode $\mathcal{OM}$ of the plant $\mathcal{T}_{aps}$, and vary within a pretty narrow range of $\Re^{+}$. Therefore, the parameter space Γ is only a small subset of $\Re^{+2}$. Also, given the fact that the outcome w and all components of the vector $\mathbf{v}$ represent physical variables (a concentration, a flow rate or a volume) and, therefore, are necessarily bounded, the outcome space W and the instance space V are subsets of $\Re^{+}$ and $\Re^{+6}$ respectively. Note that the components of the parameter vector $\mathbf{p}_S$ appear only in one term in the expression of $H(\mathbf{v}, \mathbf{p}_S)$: the rate $r_S(S_{n-1}, X_{n-1}, \mathbf{p}_S)$ of substrate biodegradation.

Now define the functional space $\mathcal{H}^{\mathcal{M}}$ by:

$$\mathcal{H}^{\mathcal{M}} = \{H(\,.\,, \mathbf{p}_S) : \ V \rightarrow W \mid \mathbf{p}_S \in \Gamma\}$$

The functions $H(\,.\,, \mathbf{p}_S)$ of this space compute the value of the solution $S^{\mathcal{M}}_{\mathbf{p}_S}$ to the differential equation 7.4 at time $t = n\,\Delta t$:

$$\forall \mathbf{v} \in V, \quad H(\mathbf{v}, \mathbf{p}_S) = S^{\mathcal{M}}_{\mathbf{p}_S}(n\,\Delta t)$$

$S^{\mathcal{M}}_{\mathbf{p}_S}(n\,\Delta t)$ represents the model prediction for S at $t = n\,\Delta t$.

Now, denote the process model identification algorithm as $\mathcal{A}^l$, where l is the quadratic loss function. Then the couple $\mathcal{LM}_{asp} = (\mathcal{H}^{\mathcal{M}}, \mathcal{A}^l)$ represents the learning machine to be associated with the environment $\mathcal{E}_{asp}$, and the space $\mathcal{H}^{\mathcal{M}}$ the decision rule space.

If the components of the instance vector $\mathbf{v}$ are denoted as $v_1, v_2, v_3, v_4, v_5, v_6$ and those of the parameter vector $\mathbf{p}_S$ as p_1, p_2, then the expression of $H(\mathbf{v}, \mathbf{p}_S)$ becomes:

$$H(\mathbf{v}, \mathbf{p}_S) = v_1 + \Delta t \left(\begin{bmatrix} v_5 \\ v_6 \end{bmatrix} [v_3 \ v_4] - \begin{bmatrix} 1 \\ 1 \end{bmatrix} [v_3 \ v_4] \ v_1 - \frac{p_1 \ v_1}{p_2 + v_1} v_2 \right) \tag{7.8}$$

7.4 VC Dimension of the Space $l_{\mathcal{H}^{\mathcal{M}}}$

The space $l_{\mathcal{H}^{\mathcal{M}}}$ has been defined in Chapter 6. Its VC dimension $q(l_{\mathcal{H}^{\mathcal{M}}})$ is needed for the calculation of the guaranteed deviation φ and for uncertainty model development. As pointed out in Chapter 6, evaluating VC dimensions by analytical means is usually difficult and, in most situations, all that is possible to determine is a bound on the value of this dimension. This section is about finding a bound on $q(l_{\mathcal{H}^{\mathcal{M}}})$, using rigorous mathematical proofs.

Theorem 7.1 *The VC dimension $q = q(l_{\mathcal{H}^{\mathcal{M}}})$ of the functional space $l_{\mathcal{H}^{\mathcal{M}}}$ associated with the decision rule space $\mathcal{H}^{\mathcal{M}}$ is at most 8.*

Proof. The VC dimension $q(l_{\mathcal{H}^{\mathcal{M}}})$ of $l_{\mathcal{H}^{\mathcal{M}}}$ is by definition equal to $q(pos(l_{\mathcal{H}^{\mathcal{M}}}))$. For every $(h, \beta) \in \mathcal{H}^{\mathcal{M}} \times \Re^{+}$ and $(\mathbf{v}, w) \in V \times W$, the following inequalities are equivalent:

$$l_{h,\beta}(\mathbf{v}, w) > 0$$

$$l_h(\mathbf{v}, w) - \beta > 0$$

$$l(w, h(\mathbf{v})) - \beta > 0$$

Considering the parameter vector $\mathbf{p}_S \in \Gamma$ that corresponds to the decision rule h and then replacing h by $H(\mathbf{v}, \mathbf{p}_S)$, it is easy to see that, for every $(\mathbf{p}_S, \beta) \in \Gamma \times \Re^{+}$ and $(\mathbf{v}, w) \in V \times W$, all the following inequalities are also equivalent:

$$l(w, h(\mathbf{v})) - \beta > 0$$

$$l(w, H(\mathbf{v}, \mathbf{p}_S)) - \beta > 0$$

$$(w - H(\mathbf{v}, \mathbf{p}_S))^2 - \beta > 0$$

$$\left(w - v_1 - \Delta t.\left(\begin{bmatrix} v_5 \\ v_6 \end{bmatrix} [v_3\ v_4] - \begin{bmatrix} 1 \\ 1 \end{bmatrix} [v_3\ v_4]\ v_1 - \frac{p_1\ v_1}{p_2 + v_1} v_2\right)\right)^2 - \beta > 0$$

$$\left(w - v_1 - \Delta t \begin{bmatrix} v_5 \\ v_6 \end{bmatrix} [v_3\ v_4] + \Delta t \begin{bmatrix} 1 \\ 1 \end{bmatrix} [v_3\ v_4]\ v_1 + \Delta t \frac{p_1\ v_1}{p_2 + v_1} v_2\right)^2 - \beta > 0$$

Define the function $\chi(\mathbf{v}, w)$ by:

$$\chi(\mathbf{v}, w) = w - v_1 - \Delta t \begin{bmatrix} v_5 \\ v_6 \end{bmatrix} [v_3\ v_4] + \Delta t \begin{bmatrix} 1 \\ 1 \end{bmatrix} [v_3\ v_4]\ v_1$$

Then the following inequalities are all equivalent to the previous ones:

$$\left(\chi(\mathbf{v}, w) + \Delta t \frac{p_1\ v_1}{p_2 + v_1} v_2\right)^2 - \beta > 0$$

$$((p_2 + v_1)\ \chi(\mathbf{v}, w) + \Delta t\ p_1\ v_1\ v_2)^2 - \beta(p_2 + v_1)^2 > 0$$

$$\begin{aligned}&\chi^2(\mathbf{v}, w)\ v_1^2 + (p_2)^2\ \chi^2(\mathbf{v}, w) + 2\ p_2\ \chi^2(\mathbf{v}, w)\ v_1 + \\ &\quad 2\ p_2\ p_1\ \Delta t\ \chi(\mathbf{v}, w)\ v_1\ v_2 + 2\ p_1\ \Delta t\ \chi(\mathbf{v}, w)\ (v_1)^2\ v_2 + \\ &\qquad (p_1)^2\ (\Delta t)^2\ (v_1)^2\ (v_2)^2 - \beta\ (p_2)^2 - 2\ \beta\ p_2\ v_1 - \beta\ (v_1)^2 > 0\end{aligned}$$

From this last inequality, it can be seen that the space $pos(l_{\mathcal{H}^{\mathcal{M}}})$ is a subset of the space $pos(f_0 + \mathcal{F})$, where:

- f_0 is the real-valued function $(\mathbf{v}, w) \mapsto \chi^2(\mathbf{v}, w)v_1^2$;
- $\mathcal{F}$ is the 8-dimensional vector space generated by the 8 following real-valued functions:

$$\begin{array}{ll} f_1 : (\mathbf{v}, w) \mapsto \chi^2(\mathbf{v}, w) & f_2 : (\mathbf{v}, w) \mapsto \chi^2(\mathbf{v}, w)\ v_1 \\ f_3 : (\mathbf{v}, w) \mapsto \chi(\mathbf{v}, w)\ v_1\ v_2 & f_4 : (\mathbf{v}, w) \mapsto \chi(\mathbf{v}, w)(v_1)^2\ v_2 \\ f_5 : (\mathbf{v}, w) \mapsto (v_1)^2(v_2)^2 & f_6 : (\mathbf{v}, w) \mapsto 1 \\ f_7 : (\mathbf{v}, w) \mapsto v_1 & f_8 : (\mathbf{v}, w) \mapsto (v_1)^2 \end{array}$$

Therefore, from theorem 6.5, it can be seen that:

$$q(l_{\mathcal{H}^{\mathcal{M}}}) = q(pos(l_{\mathcal{H}^{\mathcal{M}}})) \leq q(pos(f_0 + \mathcal{F})) = q(f_0 + \mathcal{F})) = 8$$

Thus: $q(l_{\mathcal{H}^{\mathcal{M}}}) \leq 8$.□

Theorem 7.2 *If the operating mode $\mathcal{OM}$ of the plant $\mathcal{T}_{asp}$ is such that the effluent soluble substrate concentration $v_1 = S_{n-1}$ is negligible in comparison with the values taken by the parameter $K_S = p_2$, that is: $S_{n-1} \ll K_S$, then the VC dimension $q = q(l_{\mathcal{H}^{\mathcal{M}}})$ of the functional space $l_{\mathcal{H}^{\mathcal{M}}}$ is at most 3.*

Proof. When the plant operating mode $\mathcal{OM}$ is such that $v_1 \ll p_2$, the form of the functions $l_{h,\beta}$ $(h = H(\ .\ , \mathbf{p}_S) \in \mathcal{H}^{\mathcal{M}}$ and $\beta \in \Re^+)$ of the space $l_{\mathcal{H}^{\mathcal{M}}}$ becomes as follows:

$$\begin{aligned}
\forall(\mathbf{v},w)\in V\times W: l_{h,\beta}(\mathbf{v},w) &= l(w,H(\mathbf{v},\mathbf{p}_S))-\beta \\
&= \left(\chi(\mathbf{v},w)+\Delta t\,\frac{p_1\,v_1}{p_2}\,v_2\right)^2-\beta \\
l_{h,\beta}(\mathbf{v},w) &= \chi^2(\mathbf{v},w)+2\left(\frac{p_1}{p_2}\right)\Delta t\;\chi(\mathbf{v},w)\;v_1\;v_2+ \\
&\left(\frac{p_1}{p_2}\right)^2(\Delta t)^2\;(v_1)^2\;(v_2)^2-\beta
\end{aligned}$$

Therefore, the space $l_{\mathcal{H}\cdot\mathcal{M}}$ is a subset of the functional space $f_0+\mathcal{F}$, where:

- f_0 is the real-valued function $(\mathbf{v},w)\mapsto\chi^2(\mathbf{v},w)$;
- $\mathcal{F}$ is the 3-dimensional vector space generated by the 3 following real-valued functions:

$$\begin{aligned}
f_1 &: (\mathbf{v},w)\mapsto 2\;\Delta t\;\chi(\mathbf{v},w)\,v_1\,v_2 \\
f_2 &: (\mathbf{v},w)\mapsto (\Delta t)^2\;(v_1)^2\;(v_2)^2 \\
f_3 &: (\mathbf{v},w)\mapsto -1
\end{aligned}$$

Hence, from theorem 6.5, we infer that:

$$q(l_{\mathcal{H}\cdot\mathcal{M}})\leq q(f_0+\mathcal{F})=3$$

The theorem is proved.□

7.5 Conditions of Applicability of the $\mathcal{IPERM}$

Theorem 6.9 is a powerful one. It states that, if a set of conditions ***C.1***, ***C'.1***, ***C.2***, ***C.3***, ***C'.3*** on the couple $(\mathcal{E}_{asp},\mathcal{LM}_{asp})$ are satisfied, then it is possible to develop uncertainty models of the type of inequality 4.25 for the activated sludge process. Formally, this means:

$$\textbf{\textit{C.1}},\ \textbf{\textit{C'.1}},\ \textbf{\textit{C.2}},\ \textbf{\textit{C.3}},\ \textbf{\textit{C'.3}} \Rightarrow \text{uncertainty models are possible to develop} \qquad (7.9)$$

All conditions of implication 7.9 are quite weak. However, it is important, from an engineering point of view, to understand the mathematics of these conditions, analyze their content and explain why they are reasonably satisfied in the case of the ASP. If physicists are primarily concerned with the mathematical validity of the logical implication 7.9 itself, engineers need to investigate the "edges" of this implication:

1. how the conditions relate to reality and how to assess them; and

2. how to make use of the uncertainty models and improve ASP mathematical modelling procedures.

This section is about investigating the first point. It will be shown, in particular, that the foregoing conditions are reasonably satisfied in the case of $(\mathcal{E}_{asp}, \mathcal{LM}_{asp})$. To do so, we will make use of qualitative reasoning based on good engineering judgment and heuristic arguments with respect to the real behaviour of $\mathcal{T}_{asp}$ being approximated by the model $\mathcal{M}$. Simple examples to illustrate this reasoning will be given. Existing mathematical proofs and classical results of mathematical statistics will also be utilized whenever it is possible.

7.5.1 Condition *C.1*: $\mathcal{W}$eak $\mathcal{P}$rior $\mathcal{I}$nformation (1)

The main feature that makes Vapnik's framework interesting and useful in the area of wastewater engineering — or complex systems engineering in general — lies in its ambition to estimate the process response function $g^{\mathcal{T}_{asp}}$ without having to assume any prior information about the probability density function $P_z = P_{(\mathbf{v},w)}$, which is the embodiment of all sources of uncertainty in $\mathcal{E}_{asp}$. To carry out the estimation of $g^{\mathcal{T}_{asp}}$, this framework utilizes the $\mathcal{IPERM}$ which tries to establish a reliable conclusion about the value of the expected risk $R(h)$, based on the knowledge of the empirical one $R_{emp}^{\Upsilon_N}(h)$ only. Unfortunately, the inductive procedure implemented by this principle cannot be used for *any* arbitrary probability density function P_z. Vapnik (1982) has indeed demonstrated that, in the particular case of "*large deviations*" of the real-valued random variable:

$$l_h(z) = l(w, h(\mathbf{v}))$$

with $h \in \mathcal{H}^{\mathcal{M}}$, it is impossible to obtain a guaranteed empirical estimator of the expected risk $R(h)$. In the very simplest terms, "large deviations" means here that some "unusual" large values of the foregoing variable:

$$l_h(z) = l(w, h(\mathbf{v})) = [S^{data}(t_n) - S_{\mathrm{PS}}^{\mathcal{M}}(t_n)]^2$$

occur in the environment $\mathcal{E}_{asp}$ with a small probability, but their contribution into the mathematical expectation $R(h)$ of $l_h(z)$ is, however, considerable. If P_z has such property, then the $\mathcal{IPERM}$ will not be applicable and, as a result, the following question arises: what characterizes the probability density functions for which the

$\mathcal{IPERM}$ can be implemented? The objective here is to find a sufficient condition $\mathcal{SC}$ on P_z that guarantees the applicability of this inductive principle. In order to insure the usefulness of the latter principle in solving real-world problems, this condition $\mathcal{SC}$ should not be too *restrictive*. In other words, $\mathcal{SC}$ has to be made as weak as possible, so that the set $\mathcal{P}_{SC}$ of probability density functions on Z that satisfy $\mathcal{SC}$ be as large as possible. In that way, the $\mathcal{IPERM}$ will apply to a wide spectrum of density functions and, as a result, its usefulness will augment.

The first condition $\mathcal{SC}$ that has been widely used in computational learning theory is expressed as follows:

$$\textbf{\textit{C.1}}: \qquad \exists M \in \Re_{+}^{\circ}, \sup_{h \in \mathcal{H}^{\mathcal{M}}, z \in Z} l_h(z) = M$$

or, by using a more explicit form:

$$\textbf{\textit{C.1}}: \qquad \exists M \in \Re_{+}^{\circ}, \sup_{h \in \mathcal{H}^{\mathcal{M}}, w \in W, \mathbf{v} \in V} (w - h(\mathbf{v}))^2 = M$$

The condition ***C.1*** means that the random variable:

$$l_h(z) = (w - h(\mathbf{v}))^2$$

which expresses the deviation between the *real outcome* $w = S(t_n)$ that arises in $\mathcal{E}_{asp}$ as a response to the instance $\mathbf{v}$ and the *prediction* $h(\mathbf{v}) = S_{\text{ps}}^{\mathcal{M}}(t_n)$ of the model $\mathcal{M}$ is **bounded**. The main advantage of this type of condition is its simplicity. Several inductive procedures, corresponding to different expressions of the guaranteed deviation φ, have been developed on the basis of the condition ***C.1***. Vapnik (1982, 1995, 1998), for example, used the concept of VC dimension to develop this inductive procedure, while Haussler (1992) and Barlett *et al.* (1994) used the concepts of pseudo dimension and fat-shattering dimension, respectively. In this thesis, only Vapnik's results, which are believed to lead to the best uncertainty models (Theorems 6.6, 6.8 and 6.9) will be considered.

There is no doubt that the condition ***C.1*** is satisfied in the case of an activated sludge wastewater treatment plant $\mathcal{T}_{asp}$ being approximated by the learning machine $\mathcal{LM}_{asp} = (\mathcal{H}^{\mathcal{M}}, \mathcal{A}^l)$. Indeed, not only the model parameters are bounded (this point will be discussed extensively in sub-sections 7.5.3 "*Condition **C.2***" and 7.5.4 "*Conditions **C.3** and **C'.3***"), but also the instance $\mathbf{v}$ and the outcome w, since they represent measurements of physical variables of the plant $\mathcal{T}_{asp}$. The problem, however, with the use of ***C.1*** as a base to develop the inductive procedure of $\mathcal{IPERM}$ is that there is no general method to estimate the value of the bound M. Because of that, it is suggested here to develop an empirical procedure to carry out this estimation. One such procedure can be as follows:

1. select typical values for the two components k and K_S of the parameter vector $\mathbf{p}_S$. Plant operators usually have an approximate idea about the values of these parameters. Some typical values can also be found in Metcalf and Eddy (1991), Table 8-7. These values can also be taken as equal to those that are obtained from a preliminary model identification. Designate the selected values as k_0 and K_{S0}, and the parameter vector as $\mathbf{p}_{S0}$

2. Using the identification data set

$$\Upsilon_N : (\mathbf{v}_1, w_1), (\mathbf{v}_2, w_2), \cdots, (\mathbf{v}_N, w_N)$$

   compute the losses:

$$(w_1 - H(\mathbf{v}_1, \mathbf{p}_{S0}))^2, (w_2 - H(\mathbf{v}_2, \mathbf{p}_{S0}))^2, \cdots, (w_N - H(\mathbf{v}_N, \mathbf{p}_{S0}))^2$$

3. Compute the maximum

$$M_0 = \max_{i \in \{1,2,\ldots,N\}} (w_i - H(\mathbf{v}_i, \mathbf{p}_{S0}))^2$$

4. Take M as equal to κM_0, where κ is number greater than one.

Another procedure of estimating M can be based on the knowledge of the empirical risk value $R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N})$. The latter represents indeed an average of the quadratic loss

$$l_{h_{emp}^{\Upsilon_N}}(z) = (w - h_{emp}^{\Upsilon_N}(\mathbf{v}))^2$$

that is computed on the basis of the data set Υ_N. This average is certainly less than the bound M. Thus, M can be taken as equal to $\kappa R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N})$, with $\kappa > 1$. It is this expression that will be adopted for M in the rest of this work. The number κ is considered to vary between 1 and 100:

$$M = \kappa R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N}) \qquad \text{with} \quad \kappa \in]1, 100] \tag{7.10}$$

Expression 7.10 will be used to investigate the guaranteed deviation given in inequality 6.30. In the next Chapter, it will be shown that certain rates of model improvement, as N increases, are relatively independent of the value of the prior information M or, equivalently, κ.

7.5.2 Condition *C'.1*: $\mathcal{W}$eak $\mathcal{P}$rior $\mathcal{I}$nformation (2)

As pointed out in previous sub-section, condition ***C.1*** is fully satisfied for $\mathcal{T}_{asp}$ when the learning machine is implemented by a mechanistic model such as $\mathcal{M}$. However, this is not guaranteed, if $\mathcal{LM}$ is implemented by a black box model such as a neural network, for example. Anyone who has dealt with neural network training knows indeed that the weights can take positive or negative values that are arbitrarily large, during the training procedure. As a result, it is unreasonable to assume the existence of a bound M on the quadratic loss $l_h(z)$, when h is implemented by a neural network. Taking into account this problem, Vapnik (1998) developed another inductive procedure based on an even *weaker* prior information. He expressed the latter by the following condition:

$$\textbf{\textit{C'.1}}: \qquad \exists\,(s,\tau) \in]2,+\infty[\times\Re, \quad \sup_{h\in\mathcal{H}} \frac{\mathbf{E}^{1/s}([l_h(z)]^s)}{R(h)} < \tau$$

Note that $R(h) = \mathbf{E}(l_h(z))$ and, as a result, this condition can also be written as:

$$\textbf{\textit{C'.1}}: \qquad \exists\,(s,\tau) \in]2,+\infty[\times\Re, \quad \sup_{h\in\mathcal{H}} \frac{\mathbf{E}^{1/s}([l_h(z)]^s)}{\mathbf{E}(l_h(z))} < \tau$$

This condition is weaker than ***C.1*** because it does not require that the random variable $l_h(z)$ be bounded; indeed, it is possible to have cases where ***C'.1*** is satisfied while ***C.1*** is not. As a result, the inductive procedure developed on the basis of ***C'.1*** would apply to a wider spectrum of probability density functions and learning machines.

Definition 7.1 *Let $\underline{p}$ and $\underline{q}$ be two logic (boolean) propositions.*

- *if $\underline{p}$ implies $\underline{q}$ and $\underline{q}$ implies $\underline{p}$, then the propositions $\underline{p}$ and $\underline{q}$ are said to be equivalent;*
- *if $\underline{p}$ implies $\underline{q}$ and $\underline{q}$ does not imply $\underline{p}$, then the propositions $\underline{q}$ is said to be weaker than $\underline{p}$;*

With condition ***C'.1***, the probability density function of $l_h(z)$ need not be concentrated on a certain specific interval $[0, M]$, as was the case with condition ***C.1***. On the contrary, this density function can spread over the whole set $\Re^+$ of positive numbers, which means that $l_h(z)$ would be allowed to take ***any*** large value in $\Re^+$ with a certain probability. As these large values should not occur in the form of "large deviations" (see the beginning of previous sub-section), some restrictions must, however, be imposed on the ***tail*** of the density function of $l_h(z)$, and *that is what the condition* ***C'.1*** *is all about*: describing the types of distribution tails that are acceptable for the application of the $\mathcal{IPERM}$. To appreciate the essence of the restrictions described by ***C'.1***, consider the following simple example:

Example 7.1. Fix a decision rule h in $\mathcal{H}^{\mathcal{M}}$ and denote the random variable $l_h(z)$ by ξ. Consider the following two different cases where ξ is governed by:

- **First case:** the *Gaussian* density function,

$$f_\xi^G(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad x \in [0, +\infty[$$

- **Second case:** the *Cauchy* density function,

$$f_\xi^C(x) = \frac{1}{\pi} \frac{1}{1+x^2}, \qquad x \in [0, +\infty[$$

The graphs of the two functions f_ξ^G and f_ξ^C resemble each other remarkably well: they are both bell-shaped. But the way the tail of the first function approaches the horizontal axis is completely different from that of the second one. The tail of f_ξ^G converges towards the foregoing axis so rapidly that even when we multiply its values $f_\xi^G(x)$ (for x approaching the infinity) by the large numbers x^s, the integral:

$$\int_0^{+\infty} x^s f_\xi^G(x)\, dx$$

remains ***finite*** for any positive real number s and, *a fortiori*, for s greater than 2. As a result, the ratio:

$$\frac{\left(\int_0^{+\infty} x^s f_\xi^G(x)\, dx\right)^{1/s}}{\int_0^{+\infty} x f_\xi^G(x)\, dx} = \frac{\mathrm{E}^{1/s}(\xi^s)}{\mathrm{E}(\xi)}$$

will also be finite and, therefore, the condition *C'.1* is readily satisfied[1] for the density function f_ξ^G. This function is said to have a "***light tail***". On the other hand, the tail of f_ξ^C approaches the axis

[1] *Remark:* If the ratio:

$$\frac{\mathrm{E}^{1/s}([l_h(z)]^s)}{\mathrm{E}(l_h(z))}$$

is finite for a fixed $s > 2$ and all $h \in \mathcal{H}^{\mathcal{M}}$, then all what is needed for its supremum to be bounded is that the space $\mathcal{H}^{\mathcal{M}}$ be a compact set, which is always the case in the application considered here.

so slowly that even the expectation:

$$\mathbf{E}(\xi) = \int_0^{+\infty} x f_\xi^C(x)\,dx$$

of ξ is infinite. The other integrals:

$$\mathbf{E}(\xi^s) = \int_0^{+\infty} x^s f_\xi^C(x)\,dx$$

are also infinite for the same reason, and reach $+\infty$ even faster than the expectation $\mathbf{E}(\xi)$. Consequently, the ratio:

$$\frac{\mathbf{E}^{1/s}(\xi^s)}{\mathbf{E}(\xi)}$$

cannot be bounded for any $s > 2$ and, as a result, the condition ***C'.1*** can never be satisfied for the density f_ξ^C. This function gives too much weighting to large values of x.

The condition ***C'.1*** holds true for probability density functions with "light tails", i.e., tails that approach the horizontal axis fast enough, so that there exist an $s_0 > 2$ for which the s_0-th moment $\mathbf{E}([l_h(z)]^{s_0})$ of $l_h(z)$ is a finite number. The lighter the density function tail (i.e. the faster this tail approaches the axis), the greater can be the number s_0. In concrete terms, this amounts to saying that when large values of $l_h(z)$ occur very rarely in $\mathcal{E}_{asp}$, then ***C'.1*** can be satisfied with a high number $s_0 > 2$. In the case of the wastewater treatment plant $\mathcal{T}_{asp}$, it has been pointed out in the previous sub-section that values of $l_h(z)$ that are greater than a certain threshold M do not occur at all. This is because condition ***C.1*** holds true for this plant $\mathcal{T}_{asp}$ and, therefore, the density function tail of $l_h(z)$ is not only "light", but it is actually null beyond the threshold M. Consequently, it is legitimate to consider that ***C'.1*** is satisfied with a high number $s_0 > 2$ for the case of $(\mathcal{E}_{asp}, \mathcal{LM}_{asp})$. Accordingly, the value

$$\gamma(s_0) = \sqrt[s_0]{\frac{1}{2}\left(\frac{s_0 - 1}{s_0 - 2}\right)^{s_0 - 1}}$$

of the function γ that appears in the guaranteed deviation φ (right-hand side) of inequality 6.31 (theorem 6.9) will be almost equal to 1, since $\gamma(s)$ approaches the unit as soon as s becomes greater than 3.

As far as the bound τ is concerned, Vapnik (1982) has reported that it varies between the narrow limits of 1.35 and 2.45 for all parametric models of probability density functions that are conventionally used to describe the uncertainty underlying error functions such as $w - h(\mathbf{v})$. One of the main advantages of τ over the bound M of condition ***C.1*** is that τ is dimensionless and, therefore, would not change very much when the process model $\mathcal{M}$ (that is, the decision rule space $\mathcal{H}^{\mathcal{M}}$) is changed. The bound τ characterizes more the type of the density function P_z, while the value of M depends on both the density function P_z and the process model $\mathcal{M}$ that is used to approximate the behaviour of $\mathcal{T}_{asp}$. Moreover, the fact that τ is dimensionless helps make its value independent of the parameters of the density function P_z (see Vapnik's calculations, 1982, pp. 33).

Since the bound τ varies within a narrow range, one can choose to use the average value of this range, which is about 2, as an approximation of τ to compute the guaranteed deviation φ between $g^{\mathcal{T}_{asp}}$ and $h_{emp}^{\Upsilon_N}$ (as it is expressed by right-hand side of inequality 6.31). However, in the case of $(\mathcal{E}_{asp}, \mathcal{LM}_{asp})$, it is argued that τ is rather closer to the lower limit of the range of variation and, as a result, it is more appropriate to select τ from the lower part of the range [1.35, 2]. The rationale that underlies this assertion is explained below.

Qualitatively, the number τ is a sort of "gross measure" of the amount of uncertainty underlying the occurrence of the variable $l_h(z)$ in the environment $\mathcal{E}_{asp}$. The higher this uncertainty, the larger the value of τ. The uncertainty is considered here to be characterized by the degree of dispersion of the random variable around its expected value.

To illustrate this dependence between τ and the uncertainty on $l_h(z)$, consider the simple case where $s = 2$. The ratio that appears in the condition ***C'.1*** becomes:

$$\begin{aligned}\frac{\sqrt{\mathbf{E}([l_h(z)]^2)}}{\mathbf{E}(l_h(z))} &= \frac{\sqrt{\mathbf{Var}(l_h(z)) + \mathbf{E}^2(l_h(z))}}{\sqrt{\mathbf{E}^2(l_h(z))}} \\ &= \sqrt{\frac{\mathbf{Var}(l_h(z))}{\mathbf{E}^2(l_h(z))} + 1} \\ &= \sqrt{\mathbf{CV}^2(l_h(z)) + 1}\end{aligned}$$

and, therefore:

$$\tau = \sup_{h \in \mathcal{H}} \frac{\sqrt{\mathbf{E}([l_h(z)]^2)}}{\mathbf{E}(l_h(z))} = \sup_{h \in \mathcal{H}} \sqrt{\mathbf{CV}^2(l_h(z)) + 1}$$

$\mathbf{Var}(l_h(z))$ being the variance of $l_h(z)$ and $\mathbf{CV}(l_h(z))$ its coefficient of variation. These equations show that τ is directly related to $\mathbf{CV}$, which is a measure of spread of $l_h(z)$. As a result, the bound τ becomes also a characteristic of the dispersion and, hence, the degree of uncertainty that underlies the random variable $l_h(z)$.

Consequently, in order to be able to decide about an estimation of τ for the case of $(\mathcal{E}_{asp}, \mathcal{LM}_{asp})$, we shall first carry out a qualitative assessment of the degree of dispersion of the variable:

$$l_h(z) = (w - h(\mathbf{v}))^2 = (w - H(\mathbf{v}; \mathbf{p}_S))^2$$

about its expected value, when $h = H(\ .\ , \mathbf{p}_S)$ is a decision rule from the space $\mathcal{H}^{\mathcal{M}}$.

The model $\mathcal{M}$, which has served as a base for the definition of $\mathcal{H}^{\mathcal{M}}$, has a remarkable property that no one of the other *black box* models possess: it is a mechanistically-based model and, as a result, a great deal of knowledge about the fundamental mechanisms that govern the dynamics of the treatment plant $\mathcal{T}_{asp}$ has already been incorporated into the model equations at the time of their development. This knowledge preexists in $\mathcal{M}$ even before the learning phase (or the identification procedure) starts. It is embodied in the celebrated *Monod* equation:

$$\mu(S) = Y \frac{k\,S}{K_S + S}$$

(Y is the yield coefficient, $k\,Y = \mu_H$) that describes the specific growth rate $\mu(S)$ of the microorganisms as a function of the limiting substrate concentration S (Monod, 1942). This knowledge is also expressed by the orderly fashion according to which the components of the instance vector $\mathbf{v}$ and the parameter vector $\mathbf{p}_S$ appear in the model equations, reflecting the reactor physical flow regime and the biochemical mechanisms of biodegradation. Part of this prior knowledge is also the fact that the parameter space $\Gamma = \Gamma_k \times \Gamma_{K_S}$ within which the search of the best set of parameters will be carried out, is a small subset of $R_+^{\circ 2}$, as all model parameters are concentrated around their typical values. In contrast to this mechanistic model $\mathcal{M}$, black box models that have been used to describe the behaviour of wastewater treatment plants (Novotny and Capodaglio, 1992; Hiraoka *et al.*, 1990; Côté *et al.*, 1995) are just arbitrary families of functions that contain no prior information about these plants. The functions that are implemented by these black box models act on the components of $\mathbf{v}$ in a way that has nothing to do with the real process governing the dynamics of the plant under study. The parameters that define these functions can take, a priori, any value in $]-\infty, +\infty[$. In the case of a neural network model, for instance, there is no theoretical or empirical rule and/or information that would impose limitations on the range of variation of the network weights, when they are applied to a wastewater treatment system.

As a result of this comparison "*mechanistic models versus black box models*", one can readily assert that the variability of the difference:

$$w - h(\mathbf{v})$$

between the real outcome w and the prediction $h(\mathbf{v})$ produced by a decision rule h is very much lower when $h \in \mathcal{H}^{\mathcal{M}}$ than when h is a function representing a black box model, such as a neural network or an ARMA time series model. As a confirmation of this fact, note indeed that the condition ***C.1***, which states that the loss:

$$l_h(z) = (w - h(\mathbf{v}))^2$$

must be bounded, is fully satisfied when the decision rule space is $\mathcal{H}^{\mathcal{M}}$, while in the case of a black box model, it would always be possible to find a set of weights or parameters such that the foregoing loss is arbitrarily large. The black box model would start giving reasonable predictions of the plant behaviour only after it has consumed a great deal of input/output training data during the learning phase, especially when the number of parameters of this model is high. In consequence, it is fairly legitimate to conclude that the degree of dispersion of the random variable $l_h(z)$ will be rather low, when it is the mechanistic model $\mathcal{M}$ that is used to generate the decision rule space and, as a result, the bound τ will be much closer to the lower limit of its range of variation. Therefore, for the computation of the guaranteed deviation φ corresponding to the case of $(\mathcal{E}_{asp}, \mathcal{LM}_{asp})$, τ will be assumed to take values from the interval $[1.35, 2]$

Note that the foregoing rationale could be used for any other mechanistic model $\mathcal{M}'$ of the activated sludge process. Consequently, it is suggested to use the following estimates of s and τ whenever a mechanistic model is being used to approximate the activated sludge process behaviour:

- $\tau \in [1.35, 2]$;
- s is high enough so that the value of:

$$\gamma(s) = \sqrt[s]{\frac{1}{2}\left(\frac{s-1}{s-2}\right)^{s-1}}$$

is equal to 1.

7.5.3 Condition *C.2*: VC dimension

In virtue of the results of theorem 6.5, it is easy to see that the VC dimension of any space whose elements are functions that depend on a *finite* set of parameters in a rational way (rational functions of parameters) is finite.

Proof Outline:
Indeed, it suffices to notice that the inequality:

$$\frac{a_1 f_1(\vec{x}) + a_2 f_2(\vec{x}) + \cdots + a_p f_p(\vec{x})}{a'_1 f'_1(\vec{x}) + a'_2 f'_2(\vec{x}) + \cdots + a'_{p'} f'_{p'}(\vec{x})} > 0$$

where a_i and a'_i are the parameters and f_i and f'_j are real-valued functions on a certain fixed space (p and p' are two elements of $\aleph^o$), is equivalent to the inequality:

$$(a_1 f_1(\vec{x}) + a_2 f_2(\vec{x}) + \cdots + a_p f_p(\vec{x}))(a'_1 f'_1(\vec{x}) + a'_2 f'_2(\vec{x}) + \cdots + a'_{p'} f'_{p'}(\vec{x})) > 0$$

or:

$$\sum_{i=1}^{p} \sum_{j=1}^{p'} (a_i a'_j)\; f_i(\vec{x})\, f'_j(\vec{x}) > 0$$

If the product $a_i a'_j$ is denoted by λ_{ij} and the function $f_i(\vec{x})\, f'_j(\vec{x})$ by $\varrho_{ij}(\vec{x})$, then the previous inequality becomes:

$$\sum_{i=1}^{p} \sum_{j=1}^{p'} \lambda_{ij}\; \varrho_{ij}(\vec{x}) > 0$$

The set $\mathcal{F}$ of all functions $\vec{x} \longmapsto \sum_{i=1}^{p} \sum_{j=1}^{p'} \lambda_{ij}\; \varrho_{ij}(\vec{x})$, with λ_{ij} taking any value in $\Re$ is a pp'-dimensional vector space. Therefore, the VC dimension of the space of rational functions is finite and does not exceed pp'.

Consequently, the issue here is not to show that the VC dimension is finite (it is already the case here), but to find its exact value. However, computing VC dimensions is, in most cases, an extremely difficult mathematical problem and, because of that, we usually have to content to just find upper bounds on these dimensions. But if these bounds are too high, then the value of the guaranteed deviation φ (right-hand sides of inequalities 6.30 and 6.31) computed on the basis of these bounds, would be high too and, therefore, useless from a process management point of view — although mathematically correct. Also, when a high bound is used to compute the value of φ, a lot more training examples would be required to reduce φ to a reasonable value, *not because the nature of the environment $\mathcal{E}_{asp}$ and its transformer $\mathcal{T}_{asp}$ demands it, but just because it was not possible to establish a mathematical proof for smaller bounds on the VC dimension in question.*

Therefore, it is important to endeavor to determine the lowest possible bound, if not the exact value, of this dimension. In face of the absence of general and systematic methods of computing VC dimensions, it is suggested here to use good engineering judgment and heuristic arguments — combined with existing mathematical proofs — to estimate the VC dimension of the functional space $l_{\mathcal{H},\mathcal{M}}$.

In theorem 7.2, it has been shown that the VC dimension $q(l_{\mathcal{H},\mathcal{M}})$ of this space is equal or less than 3, when the $\mathcal{OM}$ of $\mathcal{T}_{asp}$ is such that $S_{n-1} \ll K_S$. The latter condition is satisfied in most real-world situations for the activated sludge process (see for instance Metcalf and Eddy,1991, pp. 393; Grady and Lim, 1980, pp 323, 374, 442, 451). To reach the conclusion $q(l_{\mathcal{H},\mathcal{M}}) \leq 3$ in theorem 7.2, the space $l_{\mathcal{H},\mathcal{M}}$ was compared to another space denoted $f_0 + \mathcal{F}$ and defined as (same notations as those used in the proof of theorem 7.2):

$$f_0 + \mathcal{F} = \{f_0 + \lambda_1 f_1 + \lambda_2 f_2 + \lambda_3 f_3 \mid (\lambda_1, \lambda_2, \lambda_3) \in \Re^3\}$$

It was shown that $l_{\mathcal{H},\mathcal{M}} \subseteq f_0 + \mathcal{F}$ and, from this relation, the inequality $q(l_{\mathcal{H},\mathcal{M}}) \leq q(f_0 + \mathcal{F})$ was inferred, which has finally led to $q(l_{\mathcal{H},\mathcal{M}}) \leq 3$, because $q(f_0 + \mathcal{F}) = 3$ in virtue of theorem 6.5. The space $l_{\mathcal{H},\mathcal{M}}$ is, however, a rather small subset of $f_0 + \mathcal{F}$. Indeed, the three coefficients:

$$\lambda_1 = \frac{p_1}{p_2} = \frac{k}{K_S}$$
$$\lambda_2 = \left(\frac{p_1}{p_2}\right)^2 = \left(\frac{k}{K_S}\right)^2$$
$$\lambda_3 = \beta$$

do not take any value in $\Re$. They all vary within limited ranges:

- **The first coefficient** λ_1: Based on the extensive experience of researchers and practitioners in the area of wastewater engineering, it has been established that both parameters k and K_S vary within narrow ranges, $[k^-, k^+]$ and $[K_S^-, K_S^+]$ respectively, of the set of strictly positive numbers $\Re_+^\circ$. Researchers have even reported typical values for these parameters corresponding to wastewaters with various origins (Benefield and Randall, 1980; Metcalf and Eddy, 1991). As a result, the ratio $k/K_S = p_1/p_2 = \lambda_1$ would also vary within a narrow range $[\lambda_1^-, \lambda_1^+]$, where $\lambda_1^- = k^-/K_S^+$ and $\lambda_1^+ = k^+/K_S^-$.

- **The second coefficient** λ_2: This coefficient is totally dependent on the first one ($\lambda_2 = \lambda_1^2$), which reduces even more the *"expressive power"* of the whole set $l_{\mathcal{H}\mathcal{M}}$.

- **The third coefficient** λ_3: This coefficient $\lambda_3 = \beta$ takes only positive numbers, by definition of the space $l_{\mathcal{H}\mathcal{M}}$.

Taking into account these restrictions, one can readily anticipate that the VC dimension of $l_{\mathcal{H}\mathcal{M}}$, a small subset of $f_0 + \mathcal{F}$, would be strictly less than that of the whole super-set $f_0 + \mathcal{F}$, which is $q(f_0 + \mathcal{F}) = 3$. In other words, the dimension $q(l_{\mathcal{H}\mathcal{M}})$ is equal or less than 2. To corroborate and clarify this heuristic reasoning, consider the following simple example:

Example 7.2. Let ψ_1, ψ_2 and ψ_3 be three real-valued functions on the plane $\Re^2$. The VC dimension of the functional vector-space:

$$\mathcal{F}_{ex} = <\psi_1, \psi_2, \psi_3> = \{a\,\psi_1 + b\,\psi_2 + c\,\psi_3 \mid (a, b, c) \in \Re^3\}$$

is equal to 3, in virtue of theorem 6.5. Now consider the subset $\mathcal{F}_{ex_0}$ of $\mathcal{F}_{ex}$ which contains all functions $a\,\psi_1 + b\,\psi_2 + c\,\psi_3$ of $\mathcal{F}_{ex}$, **except** that the parameter a is restricted to taking only strictly positive real values:

$$\mathcal{F}_{ex_0} = \{a\,\psi_1 + b\,\psi_2 + c\,\psi_3 \mid (a, b, c) \in \Re_+^\circ \times \Re \times \Re\}$$

Then it can be shown that the VC dimension $q(\mathcal{F}_{ex_0})$ of this subset is reduced to 2. The fact of imposing a restriction, "$a > 0$", on just one parameter has, therefore, caused the VC dimension to drop from 3 down to 2.

A particular case: Consider the case where:

$$\forall (x, y) \in \Re^2 : \psi_1(x, y) = y; \quad \psi_2(x, y) = x; \quad \psi_3(x, y) = 1$$

Then the space $pos(\mathcal{F}_{ex})$ is the set of all half-planes of $\Re^2$. Its VC dimension is 3. The subset $\mathcal{F}_{ex_0}$ contains all half-planes of $\Re^2$ delimited by a straight line with a negative slope. Its VC dimension is 2.

Theorem 7.3 *Let ψ_1, ψ_2 and ψ_3 be three real-valued functions on a fixed space $\mathcal{G}$ and $\mathcal{F}_{ex}$ the functional space:*

$$\mathcal{F}_{ex} = \{a\,\psi_1 + b\,\psi_2 + c\,\psi_3 \mid (a, b, c) \in \Re^3\}$$

If just one of the parameters, a, b or c, is restricted to taking only strictly positive real values, then the VC dimension of $\mathcal{F}_{ex}$ drops from 3 down to 2.

In the light of the results of this example, and the foregoing discussion with regard to the limited range of variation of the coefficients λ_1, λ_2 and λ_3, it is then reasonable to assume that:

$$q(l_{\mathcal{H}^{\mathcal{M}}}) \leq 2 \tag{7.11}$$

It is this upper bound, 2 that is, that will be used in what follows to compute the value of the guaranteed deviation φ for $(\mathcal{E}_{asp}, \mathcal{LM}_{asp})$.

7.5.4 *C.3* and *C'.3*: the *i.i.d.* and the weaker *i.i.d* conditions

The amount of information $\mathcal{I}(\Upsilon_N)$ that is contained in a set of training examples:

$$\Upsilon_N : (\mathbf{v}_1, w_1), (\mathbf{v}_2, w_2), \cdots, (\mathbf{v}_N, w_N)$$

is characterized by the size N of the set and the degree of statistical dependence among the examples $(\mathbf{v}_i, w_i)$. For a given size N, the information $\mathcal{I}(\Upsilon_N)$ is maximal when the examples are *i.i.d.* (independent, identically distributed): the independence among the examples guarantees that every additional example in the sequence Υ_N carries a new information about the real system. Most inductive procedures that have been developed for the $\mathcal{IPERM}$ are based on this *i.i.d.* condition (Vapnik, 1982, 1998; Haussler, 1992). Unfortunately, the examples that will be used for training the learning machine $\mathcal{LM}_{asp} = (\mathcal{H}^{\mathcal{M}}, \mathcal{A}^l)$ do not satisfy this condition for the reasons explained below:

The instance vector $\mathbf{v}$ has been defined in equation 7.6 as:

$$\mathbf{v} = [S_{n-1}, X_{n-1}, \frac{Q_{in_{n-1}}}{V}, \frac{Q_{r_{n-1}}}{V}, S_{in_{n-1}}, S_{r_{n-1}}]$$

and the outcome as: $w = S_n$, where the subscript $_n$ indicates that the variables in question correspond to time instant $t_n = n\,\Delta t$. To emphasize the fact that the training is carried out specifically for time t_n, denote $\mathbf{v}$ and w by $\mathbf{v}_{t_n}$ and w_{t_n}, respectively. In the next section, it will be explained that it is possible to use the following examples:

$$(\mathbf{v}_{t_1}, w_{t_1}), (\mathbf{v}_{t_2}, w_{t_2}), \cdots, (\mathbf{v}_{t_N}, w_{t_N})$$

to train the machine $\mathcal{LM}_{asp}$. From the definition of $\mathbf{v}$ and w, it is easy to see that the components of these training examples are related in the following way:

$$\begin{aligned}
w_{t_1} &= (\mathbf{v}_{t_2})_1 \\
w_{t_2} &= (\mathbf{v}_{t_3})_1 \\
&\vdots \\
w_{t_i} &= (\mathbf{v}_{t_{i+1}})_1 \\
&\vdots \\
w_{t_{N-1}} &= (\mathbf{v}_{t_N})_1
\end{aligned}$$

where $(\mathbf{v}_{t_i})_1$ designates the first component of the vector $\mathbf{v}_{t_i}$, and, as a result, all the examples $(\mathbf{v}_{t_i}, w_{t_i})$ are dependent on each other. Therefore, the implementation of the $\mathcal{IPERM}$ for $(\mathcal{E}_{asp}, \mathcal{LM}_{asp})$ cannot be rigorously based on the *i.i.d.* condition ***C.3*** as it is defined in Chapter 6. However, this is not a problem because all proofs of theorems of the $\mathcal{IPERM}$ are actually based on a weaker condition ***C'.3*** that can be expressed as follows:

the real-valued random variables:

$$l_h(\mathbf{v}_{t_1}, w_{t_1});\ l_h(\mathbf{v}_{t_2}, w_{t_2});\ \cdots;\ l_h(\mathbf{v}_{t_N}, w_{t_N})$$

or, using the z-notation:

$$l_h(z_{t_1});\ l_h(z_{t_2});\ \cdots;\ l_h(z_{t_N})$$

are i.i.d. for any $h \in \mathcal{H}^{\mathcal{M}}$.

This condition is weaker than ***C.3*** because it does not require that the random variables $z_{t_1}, z_{t_2}, \cdots, z_{t_N}$ be independent. It is indeed possible to have $z_{t_1}, z_{t_2}, \cdots, z_{t_N}$ dependent, while $l_h(z_{t_1});\ l_h(z_{t_2});\ \cdots;\ l_h(z_{t_N})$ are independent. To illustrate this fact, consider the following simple example that shows that the logical implication:

$$(f(\mathcal{X}_1) \text{ and } f(\mathcal{X}_2) \text{ are independent}) \Rightarrow (\mathcal{X}_1 \text{ and } \mathcal{X}_2 \text{ are independent})$$

f being a continuous function and $\mathcal{X}_1$ and $\mathcal{X}_2$ are two random variables, is false, although the converse statement of it is always true.

Example 7.3. Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be two real-valued random variables jointly distributed with the following probability density function:

$$f_{\mathcal{X}_1,\mathcal{X}_2}(x_1, x_2) = \begin{cases} \frac{(1+x_1 x_2)}{4} & \text{if } |x_1| < 1 \text{ and } |x_2| < 1 \\ 0 & \text{otherwise} \end{cases}$$

The marginal density functions can be easily determined from the joint one:

$$f_{\mathcal{X}_1}(x_1) = \begin{cases} 1/2 & \text{if } |x_1| < 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad f_{\mathcal{X}_2}(x_2) = \begin{cases} 1/2 & \text{if } |x_2| < 1 \\ 0 & \text{otherwise} \end{cases}$$

Obviously $f_{\mathcal{X}_1,\mathcal{X}_2}(x_1, x_2) \neq f_{\mathcal{X}_1}(x_1)\, f_{\mathcal{X}_2}(x_2)$ and, as a result, $\mathcal{X}_1$ and $\mathcal{X}_2$ are dependent. Now consider the function $f : \Re \longrightarrow \Re,\ x \longmapsto x^2$ and let's show that the random variables $f(\mathcal{X}_1)$ and $f(\mathcal{X}_2)$ are independent. To do so, designate by $\mathbf{Pr}[\overline{event}]$ the probability of the event "$\overline{event}$". For $x_1 \in [0, 1]$ and $x_2 \in [0, 1]$, we find:

$$\begin{aligned} \mathbf{Pr}[f(\mathcal{X}_1) < x_1, f(\mathcal{X}_2) < x_2] &= \mathbf{Pr}[{\mathcal{X}_1}^2 < x_1, {\mathcal{X}_2}^2 < x_2] \\ &= \mathbf{Pr}[-\sqrt{x_1} < \mathcal{X}_1 < \sqrt{x_1}, -\sqrt{x_2} < \mathcal{X}_2 < \sqrt{x_2}] \\ &= \int_{-\sqrt{x_1}}^{\sqrt{x_1}} \int_{-\sqrt{x_2}}^{\sqrt{x_2}} \frac{1 + a\,b}{4}\, da\, db \\ &= \sqrt{x_1}\,\sqrt{x_2} \\ &= \mathbf{Pr}[{\mathcal{X}_1}^2 < x_1]\, \mathbf{Pr}[{\mathcal{X}_2}^2 < x_2] \\ &= \mathbf{Pr}[f(\mathcal{X}_1) < x_1]\, \mathbf{Pr}[f(\mathcal{X}_1) < x_2] \end{aligned}$$

Thus, $f(\mathcal{X}_1)$ and $f(\mathcal{X}_2)$ are independent random variables.

Therefore, despite the mutual dependence of the training examples $z_{t_1}, z_{t_2}, \cdots, z_{t_N}$, it is *mathematically* possible that condition ***C'.3*** be satisfied in the case of the plant $\mathcal{T}_{asp}$ and its environment $\mathcal{E}_{asp}$.

The question that arises now is: how reasonable is it to consider that ***C'.3*** is *actually* satisfied in the case of $(\mathcal{E}_{asp}, \mathcal{LM}_{asp})$? This question is examined below.

From equation 6.3, it follows that the loss $l_h(z_{t_i})$ corresponding to an example z_{t_i} and a decision rule $h = H(\ .\ , \mathbf{p}_S)$ can be expressed as:

$$\begin{aligned} l_h(z_{t_i}) &= (w_{t_i} - H(\mathbf{v}_{t_i}, \mathbf{p}_S))^2 \\ &= (g^{\mathcal{T}_{asp}}(\mathbf{v}_{t_i}) + \epsilon_{t_i} - H(\mathbf{v}_{t_i}, \mathbf{p}_S))^2 \\ &= ([g^{\mathcal{T}_{asp}}(\mathbf{v}_{t_i}) - H(\mathbf{v}_{t_i}, \mathbf{p}_S)] + \epsilon_{t_i})^2 \end{aligned}$$

As was discussed previously, the parameter space $\Gamma \ni \mathbf{p}_S$ that generates the decision rule space:

$$\mathcal{H}^{\mathcal{M}} = \{H(\, . \, , \mathbf{p}_S) : \; V \rightarrow W \mid \mathbf{p}_S \in \Gamma\}$$

is a small subset of $R_+^{\circ 2}$, as both parameters k and K_S vary within limited ranges and are concentrated around their typical values. These typical values were obtained as a result of the extensive experience of researchers and practitioners in the area of wastewater engineering. Furthermore, in the process of optimizing model parameters, some computerized wastewater treatment plant modelling software packages make it possible to experiment with *manual* adjustment of these parameters before launching any automated optimizing procedure. For instance, the GPS-X (Hydromantis, 1995) uses an advanced graphical user interface and greatly facilitates interactive simulations that help gain more insight into the optimal values of model parameters, before even starting to use the built-in optimizer tool. These pre-simulations help make the parameter space Γ even smaller, concentrated around some heuristically pre-determined optimal values of the parameters to be identified. Adding to that the fact that the model equation $H(\mathbf{v}_{t_i}, \mathbf{p}_S)$ carries in it a substantial amount of prior information about the real plant (see previous discussion about condition ***C'.1***), it follows that the difference $g^{\mathcal{T}_{asp}}(\mathbf{v}_{t_i}) - H(\mathbf{v}_{t_i}, \mathbf{p}_S)$ may approach 0, if $g^{\mathcal{T}_{asp}}$ happens to being an element of the set $\mathcal{H}^{\mathcal{M}}$. If not, then this difference will almost approach some constant Cte from which all the dependences on the different components of vector $\mathbf{v}$ are removed. Expressed with the jargon used in the area of time-series analysis, the operation of subtracting $H(\mathbf{v}_{t_i}, \mathbf{p}_S)$ from w_{t_i} consists in fact in *removing existing seasonal variation and trend* from the plant's responses. At the end of this operation, we will be left with a series ν_{t_i} such that:

$$\nu_{t_i} = w_{t_i} - H(\mathbf{v}_{t_i}, \mathbf{p}_S) = Cte + \epsilon_{t_i}$$

that is entirely oscillatory about the constant Cte, which could be 0. Note, however, that, contrary to what is usually done in time series analysis, the set of functions used here to remove the elements attributable to seasonality and trend is not an arbitrary one; it is generated by the mechanistic model $\mathcal{M}$ into which prior information about the plant's behaviour has been incorporated at the time of its development.

Hence, the condition ***C'.3*** comes down to assuming that:

$$\boldsymbol{C'.3}_{eq} : \qquad (Cte + \epsilon_{t_1})^2, (Cte + \epsilon_{t_2})^2, \cdots, (Cte + \epsilon_{t_N})^2 \;\text{ are independent}$$

or, equivalently:

$$C'.3_{eq}: \qquad (\epsilon_{t_1})^2, (\epsilon_{t_2})^2, \cdots, (\epsilon_{t_N})^2 \text{ are independent}$$

In real-world situations, the random variables ϵ_{t_i} (noise), are not necessarily independent. However, the condition $C'.3_{eq}$ is weaker than the mutual independence of ϵ_{t_i} (refer to example 3). In other words, assuming that $C'.3_{eq}$ is satisfied does not prevent the noise ϵ_{t_i} from being dependent. The weakness of condition $C'.3_{eq}$ means that it is likely to be satisfied for a wider spectrum of real systems, including ASP systems. Thus, it is not very restrictive to assume in this study that $(\epsilon_{t_i})^2$ and, as a result $l_h(z_{t_i})$, are independent so that the $\mathcal{IPERM}$ can be implemented for $(\mathcal{E}_{asp}, \mathcal{LM}_{asp})$.

But what if some dependencies happened to exist among the terms of the sequence $l_h(z_{t_1})$, $l_h(z_{t_2})$, $\cdots$, $l_h(z_{t_N})$? Would that mean that the whole theoretical framework of the $\mathcal{IPERM}$ has to be rejected or would it still hold?

In the foregoing discussion, the concept of dependence of a random sequence was viewed as a "*crisp*" one: a sequence is either independent or dependent; there is nothing in between. This approach is not appropriate because sequence dependence does not have the same nature for all random sequences. In some cases, for instance, the internal dependencies among successive terms of a sequence can be so strong that any N-th term, for N arbitrarily large, would still be affected by the value of the first term of the sequence in question. In other cases, however, the terms would decorrelate so rapidly with the time shift that the effect of any term disappears just a few steps away from it. Here are two examples that illustrate these facts:

Example 7.4.(adapted from White, 1984) Let x_{t_1}, x_{t_2}, x_{t_3}, ...be an *i.i.d* sequence whose terms are uniformly distributed on $[0, 1]$, and let y be an $\mathcal{N}(0, 1)$ random variable, independent of all the random variables x_{t_1}, x_{t_2}, x_{t_3}, ... Define the sequence $a_{t_i} = x_{t_i} + y$, for all $i \in \{1, 2, 3, \ldots\}$. Then a_{t_1}, a_{t_2}, a_{t_3}, ...is an example of sequence where the internal dependencies among successive terms is very strong to the extent that, no matter how far into the future we take an observation on a_{t_N}, the initial value a_{t_1} still determines to some extent what a_{t_N} will be. This is because of the common component y to all terms of a_t, which results in the fact that a_{t_N} is correlated to a_{t_1} for any value of N. Note that the sequence a_{t_i} is stationary, meaning that their statistical properties do not change with the time shift.

Example 7.5.(adapted from White, 1984) Now consider the sequence a_{t_i} defined by the recursive equation

$$a_{t_i} = \rho\, a_{t_{i-1}} + \iota_{t_i} \quad \text{for all} \quad i \in \{1, 2, 3, \ldots\}$$

where $|\rho| < 1$ and ι_{t_i} is an *i.i.d.* Gaussian sequence with mean 0 and variance 1. In other words, the sequence $a_{t_1}, a_{t_2}, a_{t_3}, \ldots$ is what is called a Gaussian $AR(1)$ process ("AR" stands for *Auto-Regressive*). In the case of this sequence, the correlation among the terms a_{t_i} decreases exponentially with the time shift and, as a result, the effect of any term a_{t_i} disappears just a few time steps away from it.

To quantify the degree of internal dependence among successive terms of a generic sequence $a_{t_1}, a_{t_2}, \ldots, a_{t_n}, \cdots$, several measures have been developed in mathematical statistics. One of them is called the *strong mixing* or *α-mixing* coefficients and is defined as follows:

$$\forall m \in \aleph^\circ, \quad \alpha(m) = \sup_{i \in \aleph^\circ} \left(\sup_{A \in \mathcal{A}_1^i, B \in \mathcal{B}_{i+m}^\infty} |\mathbf{Pr}(A \cap B) - \mathbf{Pr}(A)\,\mathbf{Pr}(B)| \right)$$

where $\mathcal{A}_1^i$ is the σ-field generated by the random variables $a_{t_1}, a_{t_2}, \ldots, a_{t_i}$ and $\mathcal{B}_{i+m}^\infty$ is the σ-field generated by the random variables $a_{t_{i+m}}, a_{t_{i+m+1}}, \ldots$, and $\mathbf{Pr}[\overline{event}]$ the probability of the event "$\overline{event}$" (for more information on this definition, refer to Doukhan (1994), Bradley (1986) or White (1984)). Intuitively, the number $\alpha(m)$ measures how much dependence exists between those events that generate two terms of the sequence $a_{t_1}, a_{t_2}, \ldots$, separated by m time periods. Here are some examples that illustrate more the significance of the α-mixing coefficient:

Example 7.6. Given a sequence $a_{t_1}, a_{t_2}, \ldots, a_{t_n}, \cdots$, if $\alpha(m) = 0$ for some $m \in \aleph^\circ$, then terms that are m periods apart in this sequence are independent.

Example 7.7. If the sequence $a_{t_1}, a_{t_2}, \ldots, a_{t_n}, \cdots$ is m_0-independent, that is a_i is independent of a_{i-j} for all $j > m_0$, then $\alpha(m) = 0$ for all $m > m_0$.

Example 7.8. If $a_{t_1}, a_{t_2}, \ldots, a_{t_n}, \cdots$ is an independent sequence, then $\alpha(m) = 0$ for all $m \in \aleph^\circ$.

Example 7.9. Let $a_{t_1}, a_{t_2}, \ldots, a_{t_n}, \cdots$ be a random sequence such that $a_{t_i} = \rho\, a_{t_{i-1}} + \iota_i$ for all $i \in \aleph^\circ$, where $|\rho| < 1$ and ι_i is an *i.i.d.* Gaussian sequence with mean 0 and variance 1. In other words, $a_{t_1}, a_{t_2}, \ldots, a_{t_n}, \cdots$ is a Gaussian $AR(1)$ process. Then $\alpha(m)$ dwindles to 0 as $m \to 0$

(Other examples can be found in Doukhan (1994)). From these examples, it can be seen that the α-mixing coefficient $\alpha(m)$ is a characteristic of the nature of internal dependence among the terms of a random sequence. In particular, if $\alpha(m)$ converges to 0, then it becomes a fact that:

F. 1 *The faster the convergence of $\alpha(m)$ to 0, the weaker the dependence among the terms of the sequence $a_{t_1}, a_{t_2}, \ldots, a_{t_n}, \cdots$, the closer is this sequence to be an independent one.*

Another fact that will be useful in the next discussion is as follows:

F. 2 *The weaker the dependence among the terms of the sequence $a_{t_1}, a_{t_2}, \ldots, a_{t_n}, \ldots$, the higher the information content of finite sub-sequences $a_{i_0+1}, a_{i_0+2}, \cdots, a_{i_0+N}$ (i_0 and N being two elements of $\aleph^\circ$).*

Sequences for which $\alpha(m)$ converges to 0 are called *α-mixing sequences.*

Now that a measure of the internal dependence of a sequence is defined, it is possible to examine judiciously the aforementioned question on the validity of the $\mathcal{IPERM}$ in the dependent case: does the $\mathcal{IPERM}$ still hold true if a certain amount of dependency exists among the terms of the sequence $l_h(z_{t_1}), l_h(z_{t_2}), \cdots, l_h(z_{t_N})$?

The first fact to point out in this respect is that the *i.i.d.* conditions ***C.3*** and ***C'.3*** are not absolutely necessary for the $\mathcal{IPERM}$ to be applicable to $(\mathcal{E}_{asp}, \mathcal{LM}_{asp})$. This principle holds indeed in cases where the sequences $(z_{t_i})_{i=1,N}$ and $(l_h(z_{t_i}))_{i=1,N}$ are dependent ones. White and Wooldridge (1990) and White (1990), for instance, have derived an inductive procedure that is valid for a wide range of dependency types among the terms of the sequence $(z_i)_{i=1,N}$ and, *a fortiori*, of the sequence $(l_h(z_i))_{i=1,N}$. To this procedure corresponds a guaranteed deviation φ_{dep} similar to the ones obtained in theorem 6.9. The expression of φ_{dep} is dependent on the nature of internal dependence of the sequence $(z_i)_{i=1,N}$, in addition to all the variables on which the guaranteed deviations φ of theorem 6.9 are dependent, that is: $N, \mathcal{H}$, $R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N})$, $\mathcal{WPI}$ and η. If we limit this discussion to sequences that are α-mixing, then the main characteristic of the nature of sequence dependence is the rate of convergence

$\mathcal{RC}_\alpha$ of the α-mixing coefficient $\alpha(m)$ to 0. Hence, φ_{dep} would be a function of the variables listed in the following equation:

$$\varphi_{dep} = \varphi_{dep}(N, \mathcal{H}, R^{\Upsilon_N}_{emp}(h^{\Upsilon_N}_{emp}), \mathcal{WPI}, \eta, \mathcal{RC}_\alpha)$$

The following examples illustrate what is meant by rate of convergence $\mathcal{RC}_\alpha$ of $\alpha(m)$ to 0:

Example 7.10. When $\alpha(m)$ has an expression of the form:

$$\forall m \in \aleph^\circ, \quad \alpha(m) = \alpha_0\, m^{-a} \tag{7.12}$$

with $\alpha_0 > 0$ and $a > 0$, that is a power function of m, the rate of convergence to 0 is quite slow. $\mathcal{RC}_\alpha$ can be characterized in this case by the triplet: $(power, \alpha_0, a)$.

Example 7.11. When $\alpha(m)$ has an expression of the form:

$$\forall m \in \aleph^\circ, \quad \alpha(m) = \alpha_0\, a^{m} \tag{7.13}$$

with $\alpha_0 > 0$ and $0 < a < 1$, that is an exponential function of m, the rate of convergence to 0 is fast. $\mathcal{RC}_\alpha$ can be characterized here by the triplet: $(exponential, \alpha_0, a)$. Note that ARMA($\tilde{p}$,$\tilde{q}$) random sequences with finite $\tilde{p}$ and $\tilde{q}$ (for definition of ARMA($\tilde{p}$,$\tilde{q}$) sequences, see Box and Jenkins (1970)) exhibit a degree of dependence that is comparable to $\mathcal{RC}_\alpha = (exponential, \alpha_0, a)$ (White and Wooldridge (1990).

Example 7.12. When $\alpha(m)$ has an expression of the form:

$$\forall m \in \aleph^\circ, \quad \alpha(m) = \alpha_0\, a^{c^m} \tag{7.14}$$

with $\alpha_0 > 0$ and $0 < a < 1$ (exponential of the exponential), the rate of convergence to 0 is extremely fast. $\mathcal{RC}_\alpha$ can be characterized here by the triplet: $(exponential\ of\ exponential, \alpha_0, a)$. A random sequence with such degree of dependence is actually very close to be an independent sequence (White and Wooldridge (1990).

Example 7.13. When $\alpha(m)$ is such that:

$$\forall m \in \aleph^\circ, \quad \alpha(m) = 0 \tag{7.15}$$

the random sequence is statistically independent. Equation 7.15 expresses the fastest possible rate of convergence to 0.

How does $\mathcal{RC}_\alpha$ affect the value of φ_{dep}? To address this question, assume that it is possible to generate in $\mathcal{E}_{asp}$ two α-mixing sequences:

$$\Upsilon : \quad z_{t_1}, z_{t_2}, \cdots, z_{t_n}, \cdots$$

and

$$\Upsilon' : \quad z'_{t_1}, z'_{t_2}, \cdots, z'_{t_n}, \cdots$$

that describe the same environment $\mathcal{E}_{asp}$, but have different internal dependencies characterized by two different α-mixing coefficient α and α', respectively. Assume that α(m) converges to 0 faster than α'(m) and denote this assumption by:

$$\mathcal{RC}_{\alpha} > \mathcal{RC}_{\alpha'}$$

Mathematically, this means that:

$$\lim_{m \to 0} \frac{\alpha(m)}{\alpha'(m)} = 0$$

or, equivalently:

$$\alpha(m) = o(\alpha'(m))$$

From *fact* **F. 1**, it follows then that the dependence among the terms of Υ would be weaker than that among the terms of Υ'. Therefore, from *fact* **F. 2**, an N-subsequence Υ_N of Υ would contain more information about the environment $\mathcal{E}_{asp}$ than another N-subsequence Υ'_N of Υ'. Consequently, it should be expected that, for a **fixed** integer N, the decision rule $h^{\Upsilon_N}_{emp}$ determined on the basis of N examples from Υ be closer to the plant response function $g^{\mathcal{T}_{asp}}$ than the decision rule $h^{\Upsilon'_N}_{emp}$ determined on the basis of N examples from Υ'. In terms of the guaranteed deviation, this should translate into the following inequality:

$$\varphi_{dep}(N, \mathcal{H}, R^{\Upsilon_N}_{emp}(h^{\Upsilon_N}_{emp}), \mathcal{WPI}, \eta, \mathcal{RC}_{\alpha}) < \varphi_{dep}(N, \mathcal{H}, R^{\Upsilon'_N}_{emp}(h^{\Upsilon'_N}_{emp}), \mathcal{WPI}, \eta, \mathcal{RC}_{\alpha'})$$

In particular, if the sequence Υ is such that: $\forall m \in \aleph^{\circ}$, $\alpha(m) = 0$, that is if Υ is independent, then it would obviously satisfy the relation:

$$\mathcal{RC}_{\alpha} > \mathcal{RC}_{\alpha'}$$

and, as a result:

$$\varphi(N, \mathcal{H}, R^{\Upsilon_N}_{emp}(h^{\Upsilon_N}_{emp}), \mathcal{WPI}, \eta) < \varphi_{dep}(N, \mathcal{H}, R^{\Upsilon'_N}_{emp}(h^{\Upsilon'_N}_{emp}), \mathcal{WPI}, \eta, \mathcal{RC}_{\alpha'}) \tag{7.16}$$

where $\varphi(N, \mathcal{H}, R^{\Upsilon_N}_{emp}(h^{\Upsilon_N}_{emp}), \mathcal{WPI}, \eta) = \varphi_{dep}(N, \mathcal{H}, R^{\Upsilon'_N}_{emp}(h^{\Upsilon'_N}_{emp}), \mathcal{WPI}, \eta, \mathcal{RC}_{\alpha=0})$ is the guaranteed deviation corresponding to the independent case. Therefore, to achieve a certain *fixed* level of accuracy (meaning a certain value of the deviation $\mathcal{D}(h, g^{\mathcal{T}_{asp}})$) in the estimation of the plant response function $g^{\mathcal{T}_{asp}}$, the number of examples N_{dep}

required in the dependent case would be larger than the number N required in the independent case. For instance, to transform inequality 7.16 into equality, we need to use a larger number of examples N_{dep} in the expression of φ_{dep}:

$$\varphi(N, \mathcal{H}, R_{emp}^{\Upsilon_N}(h_{emp}^{\Upsilon_N}), \mathcal{WPI}, \eta) = \varphi_{dep}(N_{dep}, \mathcal{H}, R_{emp}^{\Upsilon'_{N_{dep}}}(h_{emp}^{\Upsilon'_{N_{dep}}}), \mathcal{WPI}, \eta, \mathcal{RC}_{\alpha'})$$

with $N_{dep} > N$. Thus, the main consequence that would result from the sequence:

$$l_h(z_1), l_h(z_2), \cdots, l_h(z_n), \cdots$$

$(h \in \mathcal{H})$ being dependent, instead of independent as stated in condition ***C'.3***, is a *larger* size of the required training sample. Based on this result, it is now possible to re-state the condition ***C'.3*** in a milder form:

- ***C".3 (Degree of dependence allowed):***

  The random sequence:

  $$l_h(z_1);\ l_h(z_2);\ \ldots;\ l_h(z_N)$$

  obtained by computing the values of l_h, $h \in \mathcal{H}^{\mathcal{M}}$, at each one of the training examples z_i of the sequence Υ_N, is independent, identically distributed (i.i.d.) for all $h \in \mathcal{H}$. If this is not the case, the foregoing sequence is, at least, stationary and the dependence among successive terms of this sequence is weak *to the extent that the guaranteed deviations φ of theorem 6.9 (right-hand sides of inequalities 6.30 and 6.31) is a reasonably valid bound on the deviation squared $[\mathcal{D}(h_{emp}^{\Upsilon_N}, g^{\mathcal{T}_{asp}})]^2$.*

In what follows, this condition will be assumed to be satisfied and, as discussed previously, this assumption is quite weak.

7.5.5 Last Condition: Ergodicity

As pointed out at the beginning of previous section, training of the machine

$$\mathcal{LM}_{asp} = (\mathcal{H}^{\mathcal{M}}, \mathcal{A}^l)$$

associated with the environment $\mathcal{E}_{asp}$ is carried out for a specific time t_n. This time is arbitrary, but fixed. The examples $(\mathbf{v}_1, w_1), (\mathbf{v}_2, w_2), \cdots, (\mathbf{v}_N, w_N)$ to be used for machine training should therefore correspond to a series of realizations of the real plant $\mathcal{T}_{asp}$ and its environment $\mathcal{E}_{asp}$ at time t_n. In practice, this is not possible, because the instance vector $\mathbf{v}$ and the outcome w are measured only **once** at any time instant t. And what is obtained from these measurements is actually a *time series*:

$$(\mathbf{v}_{t_1}, w_{t_1}), (\mathbf{v}_{t_2}, w_{t_2}), \cdots, (\mathbf{v}_{t_n}, w_{t_n}), \cdots$$

or using the z-notation:

$$z_{t_1};\ z_{t_2};\ \cdots;\ z_{t_n}; \cdots$$

whose terms represent the couples instance/outcome at successive time instants t_1, $t_2, \cdots, t_n, \cdots$. This time series is constructed from the plant input/output data:

$$(\mathbf{u}^{data}(t_0), S^{data}(t_0)), (\mathbf{u}^{data}(t_1), S^{data}(t_1)), \cdots, (\mathbf{u}^{data}(t_n), S^{data}(t_n)), \cdots$$

that are made available for process identification. It corresponds to *one realization* of the environment $\mathcal{E}_{asp}$ in time. This realization would usually — if not always — be the only one that is available for investigating the plant's behaviour. Fortunately, according to the classical theorems of mathematical statistics, this single realization is enough to achieve adequate machine training, if the random sequence:

$$l_h(z_{t_1}), l_h(z_{t_2}), \cdots, l_h(z_{t_n}), \cdots \tag{7.17}$$

where $h \in \mathcal{H}^{\mathcal{M}}$, is an *ergodic* one. The property of ergodicity is just another characterization of the memory of the random sequence in question, i.e., the degree of dependence among its successive terms. It occurs when these terms decorrelate sufficiently rapidly with the time shift so that the time-average:

$$\frac{1}{N}\sum_{i=1}^{N} l_h(z_{t_i})$$

computed on the basis of an infinitely large sample size N, coincides with the so-called *ensemble* average:

$$\mathbf{E}(\xi) = \int x\, P_\xi(x)\, dx$$

where ξ represents the random variable $l_h(z_{t_n})$, corresponding to the n-th term of the time series 7.17. The notion of ergodicity is usually defined for stationary random sequences only, so that the ensemble average $\mathbf{E}(\xi)$ keeps the same value for all time instants t_n.

Note that the value of $\mathbf{E}(\xi)$ is exactly the same as that of the expected risk $R(h)$ and, as a result, when the sequence 7.17 is ergodic, the following equality:

$$R(h) = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} l_h(z_{t_i}) \tag{7.18}$$

is valid. As a result, the empirical risk (which is an empirical estimation of $R(h)$) can be estimated on the basis of the following equation:

$$R_{emp}^{\Upsilon_N}(h) = \frac{1}{N} \sum_{i=1}^{N} l_h(z_{t_i})$$

For each $h = H(\,.\, , \mathbf{p}_S) \in \mathcal{H}^{\mathcal{M}}$, the summation

$$\begin{aligned} \sum_{i=1}^{N} l_h(z_{t_i}) &= \sum_{i=1}^{N} l(h(\mathbf{v}_{t_i}), w_{t_i}) \\ &= \sum_{i=1}^{N} |h(\mathbf{v}_{t_i}) - w_{t_i}|^2 \\ &= \sum_{i=1}^{N} |H(\mathbf{v}_{t_i}, \mathbf{p}_S) - w_{t_i}|^2 \\ &= \sum_{i=1}^{N} |S_{\mathbf{p}_S}^{\mathcal{M}}(t_i) - S^{data}(t_i)|^2 \\ &= J_S(\mathbf{p}_S) \end{aligned} \tag{7.19}$$

represents exactly the so-called objective function, $J_S(\mathbf{p}_S)$, that is used in the classical procedure of process identification. Consequently, $R_{emp}^{\Upsilon_N}(h)$ is obtained by dividing the objective function value $J_S(\mathbf{p}_S)$ by N, that is:

$$R_{emp}^{\Upsilon_N}(h) = \frac{J_S(\mathbf{p}_S)}{N} \tag{7.20}$$

It is then legitimate, if the ergodicity condition is satisfied, to use the terms of the finite sub-sequence:

$$z_{t_1};\ z_{t_2};\ \cdots;\ z_{t_N}$$

as the training examples for the machine $\mathcal{LM}_{asp}$.

The question that arises now is the following: how reasonable is it to assume that the sequence 7.17 is ergodic? In the previous section, it has been shown that it is reasonable to consider that the terms of the sequence $l_h(z_{t_i})$ are *i.i.d.* (condition ***C'.3***). In addition, it was argued that internal dependencies are allowed, as long as they are weak and the sequence in question is stationary (condition ***C".3***). It is of course understood that weak dependencies implies at least the α-mixing property, that is:

$$\lim_{m \to \infty} \alpha(m) = 0$$

Both conditions, ***C'.3*** and ***C".3***, are stronger than the condition of ergodicity. It has indeed been proven in mathematical statistics that the following logical implications hold true for any random sequence (White, 1984):

$$(i.i.d.) \Rightarrow (\text{stationarity and } \alpha\text{-mixing}) \Rightarrow (\text{ergodicity})$$

Therefore, in view of the arguments of the previous section, the ergodicity condition is fully satisfied for the sequence:

$$l_h(z_{t_1}), l_h(z_{t_2}), \cdots, l_h(z_{t_n}), \cdots$$

and, as a result, the learning machine $\mathcal{LM}_{asp}$ or, equivalently, the model $\mathcal{M}$ can be trained using the classical model identification procedure. Also, for simulation purposes, the expected risk $R(h)$ can be estimated with the summation $\frac{1}{N'}\sum_{i=1}^{N'} l_h(z_{t_i})$ (see equation 7.18), where N' is a very large integer. That is, for $h = H(\,..\,\mathbf{p}_S) \in \mathcal{H}^{\mathcal{M}}$:

$$\begin{aligned} R(h) &\approx \frac{1}{N'}\sum_{i=1}^{N'} l_h(z_{t_i}) \\ &= \frac{1}{N'}\sum_{i=1}^{N'} |H(\mathbf{v}_{t_i}, \mathbf{p}_S) - w_{t_i}|^2 \\ &= \frac{1}{N'}\sum_{i=1}^{N'} |S^{\mathcal{M}}_{\mathbf{p}_S}(t_i) - S^{data}(t_i)|^2 \end{aligned} \quad (7.21)$$

Similarly, the squared deviation $[\mathcal{D}(h, g^{\mathcal{T}_{asp}})]^2$ can be estimated using the following summation (see equation 6.1):

$$\begin{aligned} [\mathcal{D}(h, g^{\mathcal{T}_{asp}})]^2 &\approx \frac{1}{N'}\sum_{i=1}^{N'} l(h(\mathbf{v}_{t_i}), g^{\mathcal{T}_{asp}}(\mathbf{v}_{t_i})) \\ &= \frac{1}{N'}\sum_{i=1}^{N'} |h(\mathbf{v}_{t_i}) - g^{\mathcal{T}_{asp}}(\mathbf{v}_{t_i})|^2 \\ &= \frac{1}{N'}\sum_{i=1}^{N'} |H(\mathbf{v}_{t_i}, \mathbf{p}_S) - g^{\mathcal{T}_{asp}}(\mathbf{v}_{t_i})|^2 \\ &= \frac{1}{N'}\sum_{i=1}^{N'} |S^{\mathcal{M}}_{\mathbf{p}_S}(t_i) - g^{\mathcal{T}_{asp}}(t_i)|^2 \end{aligned} \quad (7.22)$$

where $g^{\mathcal{T}_{asp}}(t_i)$ is defined as:

$$\begin{aligned} g^{\mathcal{T}_{asp}}(t_i) &= g^{\mathcal{T}_{asp}}(\mathbf{v}_{t_i}) \\ &= \mathbf{E}(w_{t_i} \mid \mathbf{v}_{t_i}) \\ &= \mathbf{E}(S^{data}(t_i) \mid \mathbf{v}_{t_i}) \\ &= \mathbf{E}(S^{data}(t_i) \mid S^{data}(t_{i-1}), u_{Q(t_{i-1})}, u_{S(t_{i-1})}) \end{aligned} \quad (7.23)$$

(see the definitions of u_Q and u_S at the beginning of this Chapter and refer to equation 7.7 for the expression of $\mathbf{v}_{t_i}$). Note that, in real-world situations, the plant's response (or "general tendency") function $g^{\mathcal{T}_{asp}}$ is totally unknown. However, in a simulated environment, this function can be assumed to be known and, therefore, compared to the model prediction $S^{\mathcal{M}}_{\mathbf{p}_S}(t_i)$. In both equations 7.21 and 7.22, N' represents a very large integer.

Notation Changes

In view of equations 7.20, 7.21 and 7.22, notations regarding model prediction function, the deviation $\mathcal{D}$, the empirical and expected risks will be changed using more familiar objects, such as $S^{\mathcal{M}}_{\mathbf{p}_S}$, $\mathbf{p}_S$ and $J_S(\mathbf{p}_S)$. In what follows,

- the model prediction function $S^{\mathcal{M}}_{\mathbf{p}_S}$ (that is the solution to the first differential equation of model $\mathcal{M}$ (equations 7.1)) will be simply denoted as $S_{\mathbf{p}_S}$.
- the empirical risk $R^{\Upsilon_N}_{emp}(h)$ will be denoted as $R^{\Upsilon_N}_{emp}(\mathbf{p}_S)$:

$$R^{\Upsilon_N}_{emp}(\mathbf{p}_S) = \frac{J_S(\mathbf{p}_S)}{N} \quad (7.24)$$

- the expected risk $R(h)$ will be denoted as $R(\mathbf{p}_S)$:

$$R(\mathbf{p}_S) = \frac{1}{N'} \sum_{i=1}^{N'} |S_{\mathbf{p}_S}(t_i) - S^{data}(t_i)|^2 \quad (7.25)$$

  with N' being a very large integer.

- the deviation $\mathcal{D}(h, g^{\mathcal{T}_{asp}})$ will be denoted as $\mathcal{D}(S_{\mathbf{p}_S}, g^{\mathcal{T}_{asp}})$:

$$\mathcal{D}(S_{\mathbf{p}_S}, g^{\mathcal{T}_{asp}}) = \frac{1}{N'} \sum_{i=1}^{N'} |S_{\mathbf{p}_S}(t_i) - g^{\mathcal{T}_{asp}}(t_i)|^2 \quad (7.26)$$

  with N' being a very large integer.

- the parameter vector $\mathbf{p}_S$ corresponding to the decision rule $h_{emp}^{\Upsilon_N}$ (see equation 6.18 for definition of $h_{emp}^{\Upsilon_N}$) will be denoted as $\mathbf{p}_S^{\Upsilon_N}$. In other words:

$$h_{emp}^{\Upsilon_N} = H(\ .\ , \mathbf{p}_S^{\Upsilon_N}) \tag{7.27}$$

Note also that at any time instant t_i:

$$h_{emp}^{\Upsilon_N}(\mathbf{v}_{t_i}) = H(\mathbf{v}_{t_i}, \mathbf{p}_S^{\Upsilon_N}) = S_{\mathbf{p}_S^{\Upsilon_N}}(t_i) \tag{7.28}$$

- the parameter vector $\mathbf{p}_S$ corresponding to the decision rule h_0 (see equation 6.19 for definition of h_0) will be denoted as $\mathbf{p}_S^0$. In other words:

$$h_0 = H(\ .\ .\mathbf{p}_S^0) \tag{7.29}$$

Similarly, at any time instant t_i:

$$h_0(\mathbf{v}_{t_i}) = H(\mathbf{v}_{t_i}, \mathbf{p}_S^0) = S_{\mathbf{p}_S^0}(t_i) \tag{7.30}$$

In the foregoing equations, $\mathbf{p}_S$ represents the parameter vector that defines the decision rule h, that is: $h = H(\ .\ , \mathbf{p}_S)$.

7.5.6 Discussion Summary

The main points of the discussion of the $\mathcal{IPERM}$ conditions (sections 7.5.1 to 7.5.4) are recapitulated below:

- **Condition *C.1***: $\mathcal{W}$eak $\mathcal{P}$rior $\mathcal{I}$nformation (1)

$$M = \kappa\, R_{emp}^{\Upsilon_N}(\mathbf{p}_S^{\Upsilon_N}) = \kappa\left(\frac{J_S(\mathbf{p}_S^{\Upsilon_N})}{N}\right) \qquad \text{with} \quad \kappa \in]1, 100] \tag{7.31}$$

- **Condition *C'.1***: $\mathcal{W}$eak $\mathcal{P}$rior $\mathcal{I}$nformation (2)

$$\tau \in [1.35, 2] \tag{7.32}$$

and

$$s \text{ is high enough such that}: \quad \gamma(s) = \sqrt[s]{\frac{1}{2}\left(\frac{s-1}{s-2}\right)^{s-1}} = 1 \tag{7.33}$$

- **Condition *C.2***: VC Dimension

$$q(l_{\mathcal{H}^{\mathcal{M}}}) \leq 2 \tag{7.34}$$

- **Conditions *C.3*, *C'.3* and *C".3*:** Degree of dependence allowed

  The internal dependence among the terms of the sequence:

$$|S_{\mathrm{p}_S}(t_1)-S^{data}(t_1)|^2,\ |S_{\mathrm{p}_S}(t_2)-S^{data}(t_i)|^2,\ \cdots,\ |S_{\mathrm{p}_S}(t_n)-S^{data}(t_n)|^2,\ \cdots \tag{7.35}$$

  is weak to the extent that the time series 7.35 is close to be *i.i.d.*.

7.6 Expressions of the Guaranteed Deviation and Uncertainty Models for the ASP

Now that all conditions of applicability of the $\mathcal{IPERM}$ have been assessed, it is possible to establish numerical expressions for the guaranteed deviation between the real plant's response function, $g^{\mathcal{T}_{asp}}$, and the empirical model estimation $S_{\mathrm{p}_S^{\Upsilon_N}}$. Once these numerical expressions are established, uncertainty models for $(\mathcal{E}_{asp}, \mathcal{LM}_{asp})$ (that is: ASP plant $\mathcal{T}_{asp}$ being approximated by model $\mathcal{M}$) can be obtained using inequalities 6.30 and 6.31.

Two guaranteed deviation functions have been defined in theorem 6.9: φ_1 based on the weak prior information $\mathcal{WPI}(1)$ and φ_2 based on the weak prior information $\mathcal{WPI}(2)$.

Guaranteed Deviation φ_1

The formal expression of φ_1 is given by equation 6.34 which is as follows:

$$\varphi_1 = R_{emp}^{\Upsilon_N}(\mathrm{p}_S^{\Upsilon_N}) + \frac{M\zeta}{2}\left(1+\sqrt{1+\frac{4\,R_{emp}^{\Upsilon_N}(\mathrm{p}_S^{\Upsilon_N})}{M\zeta}}\right) \tag{7.36}$$

Using equation 7.31, this expression can be transformed into:

$$\begin{aligned}
\varphi_1 &= R_{emp}^{\Upsilon_N}(\mathrm{p}_S^{\Upsilon_N}) + \frac{\kappa\, R_{emp}^{\Upsilon_N}(\mathrm{p}_S^{\Upsilon_N})\,\zeta}{2}\left(1+\sqrt{1+\frac{4}{\kappa\zeta}}\right)\\
&= R_{emp}^{\Upsilon_N}(\mathrm{p}_S^{\Upsilon_N})\left(1+\frac{\kappa\zeta}{2}\left(1+\sqrt{1+\frac{4}{\kappa\zeta}}\right)\right)\\
&= \left(\frac{J_S(\mathrm{p}_S^{\Upsilon_N})}{N}\right)\left(1+\frac{\kappa\zeta}{2}\left(1+\sqrt{1+\frac{4}{\kappa\zeta}}\right)\right)
\end{aligned} \tag{7.37}$$

Now, the expression of ζ is given by equation 6.32 which is:

$$\zeta = 4\frac{\left[q\left(\ln\left(\frac{2N}{q}\right)+1\right) - \ln\left(\frac{\eta}{4}\right)\right]}{N} \tag{7.38}$$

Let $1-\eta$ be set to 0.95 for instance, i.e., $\eta = 0.05$. Then, since $q(l_{\mathcal{H},\mathcal{M}}) \leq 2$ (inequality 7.34), the value of q can be replaced by 2 in equation 7.38:

$$\begin{aligned} \zeta &= 4\frac{\left[2\left(\ln\left(\frac{2N}{2}\right)+1\right) - \ln\left(\frac{0.05}{4}\right)\right]}{N} \\ &= 4\left(\frac{2\ \ln N + 6.4}{N}\right) = \zeta(N) \end{aligned} \tag{7.39}$$

Hence, the guaranteed deviation function φ_1 can be computed using the equation:

$$\varphi_1 = \left(\frac{J_S(\mathbf{p}_S^{\Upsilon_N})}{N}\right)\left(1 + \frac{\kappa\zeta(N)}{2}\left(1 + \sqrt{1 + \frac{4}{\kappa\zeta(N)}}\right)\right) \tag{7.40}$$

with $\zeta(N)$ given by equation 7.39.

Guaranteed Deviation φ_2

The formal expression of φ_2 is given by equation 6.36 which is as follows:

$$\varphi_2 = \frac{R_{emp}^{\Upsilon_N}(\mathbf{p}_S^{\Upsilon_N})}{(1 - \gamma(s)\,\tau\sqrt{\zeta})_+} \tag{7.41}$$

From equation 7.33, it follows that:

$$\varphi_2 = \frac{(J_S(\mathbf{p}_S^{\Upsilon_N})/N)}{(1 - \tau\sqrt{\zeta(N)})_+} \tag{7.42}$$

with $\zeta(N)$ given by equation 7.39.

Uncertainty Model $\mathcal{UM}_1$

This model, $\mathcal{UM}_1$, is formally defined by inequality 6.30. In the case of the ASP plant $\mathcal{T}_{asp}$ being approximated by the process model $\mathcal{M}$, the uncertainty model $\mathcal{UM}_1$ can be developed using the expression 7.40 of the guaranteed deviation function φ_1:

ASP Uncertainty Model 1

$$\mathcal{UM}_1: \quad [\mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_N}}, g^{\mathcal{T}_{asp}})]^2 \leq \left(\frac{J_S(\mathbf{p}_S^{\Upsilon_N})}{N}\right)\left(1 + \frac{\kappa\zeta(N)}{2}\left(1 + \sqrt{1 + \frac{4}{\kappa\zeta(N)}}\right)\right) \tag{7.43}$$

with:

$$\zeta(N) = 4\left(\frac{2\ \ln N + 6.4}{N}\right) \tag{7.44}$$

Uncertainty Model $\mathcal{UM}_2$

Similarly, the second uncertainty model, $\mathcal{UM}_2$, which is formally defined by inequality 6.31 can be developed for the ASP using the expression 7.42 of the guaranteed deviation function φ_2:

ASP Uncertainty Model 2

$$\mathcal{UM}_2: \quad [\mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_N}}, g^{\mathcal{T}_{asp}})]^2 \leq \frac{(J_S(\mathbf{p}_S^{\Upsilon_N})/N)}{\left(1 - \tau\sqrt{\zeta(N)}\right)_+} \tag{7.45}$$

with:

$$\zeta(N) = 4\left(\frac{2\ \ln N + 6.4}{N}\right) \tag{7.46}$$

As a reminder, here is a description of the objects used in the foregoing uncertainty models:

- N is the size of the data set Υ_N used for process model identification.
- $\mathbf{p}_S^{\Upsilon_N}$ is the parameter vector that minimizes the process model identification objective function $J_S(\mathbf{p}_S)$, computed on the basis of the finite data set Υ_N of size N.
- $S_{\mathbf{p}_S^{\Upsilon_N}}$ is the solution to the first differential equation of model $\mathcal{M}$, corresponding to the "best" parameter vector $\mathbf{p}_S^{\Upsilon_N}$. It represents the model prediction function of the substrate concentration.
- $g^{\mathcal{T}_{asp}}$ represents the true ASP plant's response function.

- $\mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_N}}, g^{\mathcal{T}_{asp}})$ represents the deviation between $S_{\mathbf{p}_S^{\Upsilon_N}}$ and $g^{\mathcal{T}_{asp}}$.

- $J_S(\mathbf{p}_S^{\Upsilon_N})$ is the value of the process model identification objective function J_S at the parameter vector $\mathbf{p}_S^{\Upsilon_N}$. It represents the minimum of J_S on the whole parameter space.

- κ is a number to be selected from the interval $]1, 100]$. It represents a prior information about the ASP plant behaviour being approximated by process model $\mathcal{M}$. Formally, it is defined as:

$$\kappa = \frac{\sup_{\mathbf{p}_S, t_n} |S_{\mathbf{p}_S}(t_n) - S^{data}(t_n)|^2}{\left(\frac{J_S(\mathbf{p}_S^{\Upsilon_N})}{N}\right)}$$

  It is dimensionless.

- τ is a number to be selected from the interval $[1.35, 2]$. It represents a prior information of another type about the ASP plant behaviour being approximated by process model $\mathcal{M}$. Formally, it is defined as:

$$\tau = \sup_{\mathbf{p}_S} \frac{\left(\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}|S_{\mathbf{p}_S}(t_i) - S^{data}(t_i)|^{2s}\right)\right)^{1/s}}{\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}|S_{\mathbf{p}_S}(t_i) - S^{data}(t_i)|^{2}\right)}$$

  where s is any number greater than 3. The number τ is also dimensionless.

Chapter 8

Uncertainty Management in the Activated Sludge Process (Part 1)

This Chapter and Chapter 10 are devoted to the subject of uncertainty management in the activated sludge process. This Chapter concentrates on investigating the effect of the size of the data set used for ASP model identification on the quality of the identified model. The process model is fixed (it is the model $\mathcal{M}$ defined by equations 7.1) and all investigation is based on the ASP uncertainty models $\mathcal{UM}_1$ and $\mathcal{UM}_2$ developed in previous Chapter:

$$\mathcal{UM}_1 : \quad [\mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_N}}, g^{\mathcal{T}_{asp}})]^2 \leq \left(\frac{J_S(\mathbf{p}_S^{\Upsilon_N})}{N}\right)\left(1 + \frac{\kappa\zeta(N)}{2}\left(1 + \sqrt{1 + \frac{4}{\kappa\zeta(N)}}\right)\right) \tag{8.1}$$

$$\mathcal{UM}_2 : \quad [\mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_N}}, g^{\mathcal{T}_{asp}})]^2 \leq \frac{(J_S(\mathbf{p}_S^{\Upsilon_N})/N)}{\left(1 - \tau\sqrt{\zeta(N)}\right)_+} \tag{8.2}$$

where:

$$\zeta(N) = 4\left(\frac{2\,\ln N + 6.4}{N}\right) \tag{8.3}$$

In Chapter 10, the study of ASP uncertainty management is continued by investigating the effect of process model structure complexity on the quality of the identified model.

8.1 Comparing the Uncertainty Models $\mathcal{UM}_1$ and $\mathcal{UM}_2$

In previous Chapters, it has been shown that both uncertainty models $\mathcal{UM}_1$ and $\mathcal{UM}_2$ are valid for the activated sludge process being approximated by the process model $\mathcal{M}$. Before we start utilizing these models for uncertainty control, we need to know first how they compare to each other: when do they coincide? when do they differ and why?

8.1.1 $\mathcal{UM}_1$ and $\mathcal{UM}_2$ Coincide Asymptotically ($N \to \infty$)

In this section, it will be shown that $\mathcal{UM}_1$ and $\mathcal{UM}_2$ are actually the same when the size N of the data set Υ_N is very large, that is when $N \to \infty$. To do so, we will make use of the following classical result of calculus:

$$\forall a \in \Re, \quad (1+x)^a = 1 + a\,x \qquad \text{when} \quad x \to 0 \tag{8.4}$$

The following numerical calculations carried out for $a = 0.5$ are presented to convince the reader of this simple equality:

$$\text{for } x = 0.1 \quad (1+x)^{0.5} = 1.0488 \;\text{ and }\; 1 + 0.5\,x = 1.0500$$

$$\text{for } x = 0.08 \quad (1+x)^{0.5} = 1.0392 \;\text{ and }\; 1 + 0.5\,x = 1.0400$$

$$\text{for } x = 0.05 \quad (1+x)^{0.5} = 1.0246 \;\text{ and }\; 1 + 0.5\,x = 1.0250$$

$$\text{for } x = 0.01 \quad (1+x)^{0.5} = 1.0050 \;\text{ and }\; 1 + 0.5\,x = 1.0050$$

The right-hand side of inequality 8.1, denoted as φ_1 (guaranteed deviation), can be transformed as follows:

$$\begin{aligned} \varphi_1 &= \left(\frac{J_S(\mathbf{p}_S^{\Upsilon_N})}{N}\right)\left(1 + \frac{\kappa\,\zeta(N)}{2}\left(1 + \sqrt{1 + \frac{4}{\kappa\,\zeta(N)}}\right)\right) \\ &= \left(\frac{J_S(\mathbf{p}_S^{\Upsilon_N})}{N}\right)\left(1 + \frac{\kappa\,\zeta(N)}{2}\left(1 + \sqrt{\frac{4}{\kappa\,\zeta(N)}}\sqrt{\frac{\kappa\,\zeta(N)}{4} + 1}\right)\right) \end{aligned} \tag{8.5}$$

From equation 8.3, it easy to see that $\zeta(N)$ becomes very close to 0, as N increases, i.e., $\lim_{N \to \infty} \zeta(N) = 0$. Therefore, using equation 8.4, it follows that for very large values of N:

$$\begin{aligned}
\varphi_1 &= \left(\frac{J_S(\mathbf{p}_S^{\Upsilon_N})}{N}\right)\left(1+\frac{\kappa\zeta(N)}{2}\left(1+\sqrt{\frac{4}{\kappa\zeta(N)}}\sqrt{\frac{\kappa\zeta(N)}{4}+1}\right)\right)\\
&= \left(\frac{J_S(\mathbf{p}_S^{\Upsilon_N})}{N}\right)\left[1+\frac{\kappa\zeta(N)}{2}\left(1+\sqrt{\frac{4}{\kappa\zeta(N)}}\left(\frac{\kappa\zeta(N)}{8}+1\right)\right)\right]\\
&= \left(\frac{J_S(\mathbf{p}_S^{\Upsilon_N})}{N}\right)\left[1+\frac{\kappa\zeta(N)}{2}\left(1+\sqrt{\frac{4}{\kappa\zeta(N)}}\frac{\kappa\zeta(N)}{8}+\sqrt{\frac{4}{\kappa\zeta(N)}}\right)\right]\\
&= \left(\frac{J_S(\mathbf{p}_S^{\Upsilon_N})}{N}\right)\left[1+\frac{\kappa\zeta(N)}{2}\left(1+\frac{\sqrt{\kappa\zeta(N)}}{4}+\frac{2}{\sqrt{\kappa\zeta(N)}}\right)\right]\\
&= \left(\frac{J_S(\mathbf{p}_S^{\Upsilon_N})}{N}\right)\left[1+\frac{\kappa\zeta(N)}{2}+\frac{\kappa\zeta(N)\sqrt{\kappa\zeta(N)}}{8}+\sqrt{\kappa\zeta(N)}\right]
\end{aligned}$$

Since $\zeta(N)$ approaches 0 as N becomes very large, it follows that:

$$\frac{\kappa\zeta(N)}{2}+\frac{\kappa\zeta(N)\sqrt{\kappa\zeta(N)}}{8}=o\left(\sqrt{\kappa\zeta(N)}\right)$$

Therefore:

$$\begin{aligned}
\varphi_1 &= \left(\frac{J_S(\mathbf{p}_S^{\Upsilon_N})}{N}\right)\left(1+\sqrt{\kappa\zeta(N)}+o\left(\sqrt{\kappa\zeta(N)}\right)\right)\\
&\approx \left(\frac{J_S(\mathbf{p}_S^{\Upsilon_N})}{N}\right)\left(1+\sqrt{\kappa\zeta(N)}\right)
\end{aligned}$$

Hence, when $N\to\infty$, φ_1 becomes:

$$\varphi_1=\left(\frac{J_S(\mathbf{p}_S^{\Upsilon_N})}{N}\right)\left(1+\sqrt{\kappa}\sqrt{\zeta(N)}\right) \tag{8.6}$$

Now, let us examine the right-hand side of inequality 8.2, which will be denoted as φ_2. When $N\to\infty$, φ_2 becomes:

$$\begin{aligned}
\varphi_2 &= \frac{(J_S(\mathbf{p}_S^{\Upsilon_N})/N)}{(1-\tau\sqrt{\zeta(N)})_+}\\
&= \frac{(J_S(\mathbf{p}_S^{\Upsilon_N})/N)}{\left(1-\tau\sqrt{\zeta(N)}\right)}\\
&= \left(\frac{J_S(\mathbf{p}_S^{\Upsilon_N})}{N}\right)\left(1-\tau\sqrt{\zeta(N)}\right)^{-1}\\
&= \left(\frac{J_S\left(\mathbf{p}_S^{\Upsilon_N}\right)}{N}\right)\left(1+\tau\sqrt{\zeta(N)}\right)
\end{aligned}$$

Hence, when $N \to \infty$, φ_2 becomes:

$$\varphi_2 = \left(\frac{J_S\left(\mathbf{p}_S^{\Upsilon_N}\right)}{N}\right)\left(1 + \tau\sqrt{\zeta(N)}\right) \tag{8.7}$$

Consequently, when $N \to \infty$, the two uncertainty models $\mathcal{UM}_1$ and $\mathcal{UM}_2$ become:

$$\mathcal{UM}_{1\{N\to\infty\}}: \qquad [\mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_N}}, g^{\mathcal{T}_{asp}})]^2 \leq \left(\frac{J_S(\mathbf{p}_S^{\Upsilon_N})}{N}\right)\left(1 + \sqrt{\kappa}\sqrt{\zeta(N)}\right) \tag{8.8}$$

$$\mathcal{UM}_{2\{N\to\infty\}}: \qquad [\mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_N}}, g^{\mathcal{T}_{asp}})]^2 \leq \left(\frac{J_S\left(\mathbf{p}_S^{\Upsilon_N}\right)}{N}\right)\left(1 + \tau\sqrt{\zeta(N)}\right) \tag{8.9}$$

These two inequalities show clearly that $\mathcal{UM}_1$ and $\mathcal{UM}_2$ coincide asymptotically, and suggest that the two prior information numbers κ and τ are related according to the equation:

$$\sqrt{\kappa} = \tau$$

or

$$\kappa = \tau^2 \tag{8.10}$$

Consequently, since τ takes values from the narrow range $[1.35, 2]$ (see previous Chapter), it follows that κ must actually be selected from a relatively small range $[1.35^2, 2^2] = [1.82, 4]$, instead of the *a priori* large interval $]1, 100]$ suggested in the previous Chapter:

$$\kappa \in [1.82, 4] \quad \text{instead of:} \quad \kappa \in]1, 100] \tag{8.11}$$

8.1.2 How do $\mathcal{UM}_1$ and $\mathcal{UM}_2$ Compare When N is Small?

Here, the more realistic case of data sets of small size is examined. The following numerical example will show that model $\mathcal{UM}_2$ is more conservative than model $\mathcal{UM}_1$.

Example 7.1. How many data points N do we need in order to guarantee that the squared deviation $[\mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_N}}, g^{\mathcal{T}_{asp}})]^2$ is less than 5 times the minimum of the averaged value of the identification objective function $J_S\left(\mathbf{p}_S^{\Upsilon_N}\right)/N$? This question can be re-expressed in a more direct way:

$$\textit{find } N \textit{ such that: } [\mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_N}}, g^{\mathcal{T}_{asp}})]^2 \leq 5 \times \left(\frac{J_S\left(\mathbf{p}_S^{\Upsilon_N}\right)}{N}\right) \tag{8.12}$$

Let us use models $\mathcal{UM}_1$ and $\mathcal{UM}_2$ successively to address this question.

Model $\mathcal{UM}_1$

This model will be used here with $\kappa = \kappa_{avg} = 2.91$. From inequalities 8.1 and 8.12. it follows that:

$$\frac{\kappa\,\zeta(N)}{2}\left(1+\sqrt{1+\frac{4}{\kappa\,\zeta(N)}}\right)=4$$

or

$$\frac{2.91\,\zeta(N)}{2}\left(1+\sqrt{1+\frac{4}{2.91\,\zeta(N)}}\right)=4$$

or again

$$1.45\,\zeta(N)\left(1+\sqrt{1+\frac{1}{0.73\,\zeta(N)}}\right)=4$$

Solving this equation for $\zeta(N)$ leads to:

$$\zeta(N) = 1.105$$

that is:

$$4\left(\frac{2\ \ln N + 6.4}{N}\right) = 1.105$$

Solving again this equation leads to $N = 52$.

Model $\mathcal{UM}_2$

This model will be used here with $\tau = \tau_{avg} = 1.67$. From inequalities 8.2 and 8.12. it follows that:

$$\frac{1}{\left(1-\tau\sqrt{\zeta(N)}\right)}=5$$

or

$$\frac{1}{\left(1-1.67\sqrt{\zeta(N)}\right)}=5$$

Solving this equation for $\zeta(N)$ gives:

$$\zeta(N) = \left(\frac{4}{5\times 1.67}\right)^2$$

that is:

$$4\left(\frac{2\ \ln N + 6.4}{N}\right) = 0.23$$

Solving this equation for N gives the solution $N = 311$.

This example has showed that models $\mathcal{UM}_1$ and $\mathcal{UM}_2$ require minimum

$$N = 52 \quad \text{and} \quad N = 311$$

respectively, in order to guarantee that the squared deviation $[\mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_N}}, g^{\mathcal{T}_{asp}})]^2$ be less than

$$5 \times \left(\frac{J_S\left(\mathbf{p}_S^{\Upsilon_N}\right)}{N} \right)$$

In general, model $\mathcal{UM}_2$ requires a much larger amount of data than $\mathcal{UM}_1$, to guarantee the same degree of process model accuracy. Why is this difference between two models that are supposed to account for the same uncertainty? The answer has to do with the *prior information* $\mathcal{WPI}$ on which each uncertainty model is based.

In the previous Chapter, it was mentioned that the prior information $\mathcal{WPI}(1)$ on which model $\mathcal{UM}_1$ is based, i.e.,

$$\mathcal{WPI}(1): \qquad \exists \kappa \in \Re^{\circ}_{+}, \quad \frac{\sup_{\mathbf{p}_S, t_n} |S_{\mathbf{p}_S}(t_n) - S^{data}(t_n)|^2}{\left(\frac{J_S(\mathbf{p}_S^{\Upsilon_N})}{N}\right)} = \kappa$$

is stronger than the prior information $\mathcal{WPI}(2)$ on which model $\mathcal{UM}_2$ is based:

$$\mathcal{WPI}(2): \qquad \exists \tau \in \Re^{\circ}_{+}, \ \sup_{\mathbf{p}_S} \frac{\left(\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}|S_{\mathbf{p}_S}(t_i) - S^{data}(t_i)|^{2s}\right)\right)^{1/s}}{\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}|S_{\mathbf{p}_S}(t_i) - S^{data}(t_i)|^{2}\right)} = \tau$$

(with $s > 3$). With the stronger prior information $\mathcal{WPI}(1)$ as a base, uncertainty model $\mathcal{UM}_1$ assumes that a high amount of information pre-exists within the process model $\mathcal{M}$. Because of that, $\mathcal{UM}_1$ considers that $\mathcal{M}$ does not need too much data for its identification. It is possible to formalize this qualitative explanation by making use of the concepts of *system information content, transfer and conversion*, which were introduced earlier in this thesis:

Let $\mathcal{I}$ denote a universal measure of system information content. The total information content of the real ASP is then $\mathcal{I}(ASP)$. From an information-based point of view, the objective of process modelling is to develop a mathematical system $\mathcal{S}$ such that: $\mathcal{I}(ASP) = \mathcal{I}(\mathcal{S})$. The traditional approach of developing such a system $\mathcal{S}$ consists in writing a mechanistic model $\mathcal{M}$ for the ASP and identifying it on the basis of a finite data set Υ_N. Since the identified model does not produce perfect predictions of the real process behaviour, an uncertainty model $\mathcal{UM}$ is developed on the basis of some weak prior information $\mathcal{WPI}$ about the process. Given the objects $\mathcal{M}$, Υ_N and $\mathcal{WPI}$ — which can be considered as the components of the system $\mathcal{S}$ —, this uncertainty model determines the information gap between the process model prediction and the real process behaviour. Or, conversely, given a certain degree of process model accuracy, $\mathcal{UM}$ determines the minimum size of Υ_N that is needed to achieve that accuracy. In terms of "system information content", this means that the purpose of $\mathcal{UM}$ is to determine the size N of Υ_N such that the information content of

$$\mathcal{S} = (\mathcal{M}, \Upsilon_N, \mathcal{WPI}) \tag{8.13}$$

becomes equal to some <u>fixed</u> fraction υ_0 ($\upsilon_0 \in]0, 1[$) of $\mathcal{I}(ASP)$, that is:

$$\mathcal{I}(\mathcal{S}) = \mathcal{I}(\mathcal{M}) + \mathcal{I}(\Upsilon_N) + \mathcal{I}(\mathcal{WPI}) = \upsilon_0 \mathcal{I}(ASP) \tag{8.14}$$

The latter equation can be solved for $\mathcal{I}(\Upsilon_N)$:

$$\mathcal{I}(\Upsilon_N) = \upsilon_0 \mathcal{I}(ASP) - \mathcal{I}(\mathcal{M}) - \mathcal{I}(\mathcal{WPI}) \tag{8.15}$$

In previous Chapters, it has been suggested to take $\mathcal{I}(\Upsilon_N)$ as equal to ςN, where the coefficient $\varsigma \in [0, 1]$ accounts for the data set statistical dependence:

$$\mathcal{I}(\Upsilon_N) = \varsigma N \tag{8.16}$$

From equations 8.15 and 8.16, it follows that:

$$N = \frac{\upsilon_0 \mathcal{I}(ASP) - \mathcal{I}(\mathcal{M}) - \mathcal{I}(\mathcal{WPI})}{\varsigma} \tag{8.17}$$

Now consider the two foregoing types of prior information, $\mathcal{WPI}(1)$ and $\mathcal{WPI}(2)$, about the ASP. Since $\mathcal{WPI}(1)$ is stronger than $\mathcal{WPI}(2)$, it follows that $\mathcal{I}(\mathcal{WPI}(1))$ is higher than $\mathcal{I}(\mathcal{WPI}(2))$:

$$\mathcal{I}(\mathcal{WPI}(1)) > \mathcal{I}(\mathcal{WPI}(2)) \tag{8.18}$$

Since the numbers υ_0, $\mathcal{I}(ASP)$, $\mathcal{I}(\mathcal{M})$ and ς are fixed, it can then be easily inferred from 8.17 and 8.18 that the size N_1 obtained from $\mathcal{UM}_1$ is less than the size N_2 obtained from $\mathcal{UM}_2$:

$$N_1 < N_2$$

This result shows then that the stronger the prior information about the ASP, the smaller the data set size required for process model identification. In the extreme case where the prior information characterizes *completely* the real plant response function, no data will be needed to calibrate the model. Conversely, if the prior information is extremely weak, then the uncertainty model would compute a very large size for the identification data set.

Now that the two uncertainty models, $\mathcal{UM}_1$ and $\mathcal{UM}_2$, of the ASP are compared, which one should we use for uncertainty management in the activated sludge process? This question is addressed below:

The main bothersome point about using uncertainty models, $\mathcal{UM}_1$ and $\mathcal{UM}_2$, is the prior information κ for $\mathcal{UM}_1$ and τ for $\mathcal{UM}_2$. There is indeed no general methodology to determine the values of these numbers. And, with respect to the absence of such methodology, model $\mathcal{UM}_2$ has a significant advantage: its prior information τ has been reported to vary within a limited range (Vapnik, 1982). In the previous Chapter, it was argued indeed that τ for the ASP takes values from the small interval $[1.35, 2]$. As a result, it may be reasonable to use model $\mathcal{UM}_2$ for uncertainty management in the ASP with a value of τ equal to:

$$\tau = \tau_{avg} = \frac{1.35 + 2}{2} = 1.67$$

However, from the numerical computations of example 7.1 and the foregoing discussion, it follows that $\mathcal{UM}_2$ requires a very large number of data points for the process model identification, which represents a disadvantage of $\mathcal{UM}_2$.

On the other hand, model $\mathcal{UM}_1$ is much less prodigal in terms of identification data point requirements, but it is based on a prior information κ that can *a priori* take any finite value greater than one, which is obviously very bothersome. However, if it is agreed that the relationship 8.10, which states that κ equals the square of τ and which was induced from asymptotic comparison of $\mathcal{UM}_1$ and $\mathcal{UM}_2$, is a valid one, then the prior information κ will be much less problematic for the use of $\mathcal{UM}_1$: equation 8.10 implies indeed that κ must take values from the relatively small interval $[1.82, 4]$. In addition, it will be shown in the next sections that certain rates of process model improvement, as N increases, are independent of κ. Because of all that, model $\mathcal{UM}_1$ will be selected and used in what follows for ASP uncertainty management.

8.2 Does an Infinite Data Set Guarantee a Perfect Estimation of the Process Response?

As pointed out earlier in this thesis, the ultimate goal of ASP modelling procedures is to determine the real process response function, which was denoted as $g^{\mathcal{T}_{asp}}$. Among all modelling procedures, mechanistic modelling aims at deriving $g^{\mathcal{T}_{asp}}$ from first principles. Though mechanistic models that have been developed during the last three decades are quite complex, "*they are still greatly simplifying the representation of many species of organisms*" (Jeppsson, 1996). Because of this, it has become a fact that the perfect mechanistic model of the ASP does not exist yet. To remedy this situation and improve mechanistic models as much as possible, wastewater researchers and practicing engineers resort to exploiting another source of information: "*empirical data*". And the question that arises is:

Can this new information source of empirical data
compensate for our limited knowledge (8.19)
of the fundamental mechanisms of the process?

This is a fundamental question. Addressing it and establishing a definite answer for it will provide us with directions as to what type of research must be carried on in the area of ASP modelling:

- *if the answer to question 8.19 is yes*, then there will be no need to carry out more research on investigating and modelling *all* ASP mechanisms, which is clearly very costly, time consuming and usually not conclusive. Rather, the real bottleneck in ASP modelling would become the ***amount of empirical data*** available for identifying the mechanistic model $\mathcal{M}$. It will then be more appropriate to concentrate on collecting as much data as possible and improving the mathematical and computational aspects of process model identification procedures, since it is guaranteed that more empirical data will compensate for all those unknown *hard-to-model* ASP mechanisms.

- *if, however, the answer to question 8.19 is no*, then there will be no way of escaping the difficult task of investigating and modelling in detail ASP mechanisms. The bottleneck in ASP modelling will be, in this case, both the ***amount of empirical data*** and the ***quality of the mechanistic model***.

Now that the importance of question 8.19 is explained, we can start addressing it. To do so, we will assume that:

1. the mechanistic model $\mathcal{M}$ is defined by equations 7.1. It is a simple model which does not account for all process mechanisms.

2. an infinite amount of empirical data is available for the identification of model $\mathcal{M}$. This means that "data" is not a limiting factor in the course of the determination of $g^{\mathcal{T}_{asp}}$ through model identification.

3. an infinite computing power is available to carry out the model identification procedure. This is to guarantee that the identification algorithm $\mathcal{A}^l$, whose purpose is to minimize the objective function $J_S(\mathbf{p}_S)$, will end up finding the exact *global minimum* of this function. In other words, local minima problems are completely resolved by the available infinite computing power.

With these three conditions defined, question 8.19 can be re-expressed in an equivalent way:

Does the process model identification procedure,
carried out under the foregoing conditions
1. (a simple process model is fixed), 2. (infinite data set) (8.20)
and 3. (infinite computing power),
guarantee the exact determination
of the process response function $g^{\mathcal{T}_{asp}}$?

The following theorem addresses the latter question.

Theorem 8.1 *Let Υ_N be a sequence of N data points of the ASP plant $\mathcal{T}_{asp}$. Let $\mathbf{p}_S^{\Upsilon_N}$ and $\mathbf{p}_S^0$ be two parameter vectors that globally minimize the identification objective function*

$$J_S(\mathbf{p}_S) = \sum_{i=1}^{N} |S_{\mathbf{p}_S}(t_i) - S^{data}(t_i)|^2$$

and the expected risk

$$R(\mathbf{p}_S) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} |S_{\mathbf{p}_S}(t_i) - S^{data}(t_i)|^2$$

respectively. Then the random sequence $\mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_N}}, g^{\mathcal{T}_{asp}})$ converges in probability to its minimum value $\mathcal{D}(S_{\mathbf{p}_S^0}, g^{\mathcal{T}_{asp}})$, as N is made infinitely large.

Proof. Using equation 6.2 and the notations of the mathematical framework of Chapter 6, it can be shown that the equality:

$$R(h) = \int_{V\times W} [w - g^{\mathcal{T}_{asp}}(\mathbf{v})]^2 \, P_{(\mathbf{v},w)}(\mathbf{v}, w) \, d\mathbf{v} \, dw + [\mathcal{D}(h, g^{\mathcal{T}_{asp}})]^2 \tag{8.21}$$

holds true for all decision rules h. In particular, for $h_{emp}^{\Upsilon_N}$, we get:

$$R(h_{emp}^{\Upsilon_N}) = \int_{V\times W} [w - g^{\mathcal{T}_{asp}}(\mathbf{v})]^2 \, P_{(\mathbf{v},w)}(\mathbf{v}, w) \, d\mathbf{v} \, dw + [\mathcal{D}(h_{emp}^{\Upsilon_N}, g^{\mathcal{T}_{asp}})]^2$$

Or, equivalently:

$$[\mathcal{D}(h_{emp}^{\Upsilon_N}, g^{\mathcal{T}_{asp}})]^2 = R(h_{emp}^{\Upsilon_N}) - \int_{V\times W} [w - g^{\mathcal{T}_{asp}}(\mathbf{v})]^2 \, P_{(\mathbf{v},w)}(\mathbf{v}, w) \, d\mathbf{v} \, dw$$

Because the $\mathcal{IPERM}$ is applicable to $(\mathcal{E}_{asp}, \mathcal{LM}_{asp})$, the following convergence holds true in probability:

$$R(h_{emp}^{\Upsilon_N}) \to R(h_0) \qquad as \quad N \to \infty$$

Hence, as $N \to \infty$, $[\mathcal{D}(h_{emp}^{\Upsilon_N}, g^{\mathcal{T}_{asp}})]^2$ converges in probability to the following number:

$$R(h_0) - \int_{V\times W} [w - g^{\mathcal{T}_{asp}}(\mathbf{v})]^2 \, P_{(\mathbf{v},w)}(\mathbf{v}, w) \, d\mathbf{v} \, dw$$

which is, by virtue of equation 8.21, equal to $[\mathcal{D}(h_0, g^{\mathcal{T}_{asp}})]^2$. Therefore, $\mathcal{D}(h_{emp}^{\Upsilon_N}, g^{\mathcal{T}_{asp}})$ converges in probability to $\mathcal{D}(h_0, g^{\mathcal{T}_{asp}})$ or, by adjusting the notations, $\mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_N}}, g^{\mathcal{T}_{asp}})$ converges in probability to $\mathcal{D}(S_{\mathbf{p}_S^0}, g^{\mathcal{T}_{asp}})$, as N is made infinitely large.□

Recall from previous Chapters that:

$S_{\mathbf{p}_S^{\Upsilon_N}}$: is a time function that computes, for each time instant t, the prediction $S_{\mathbf{p}_S^{\Upsilon_N}}(t)$ by model $\mathcal{M}$ for the substrate concentration, when model $\mathcal{M}$ is identified on the basis of the finite data set Υ_N.

$S_{\mathbf{p}_S^0}$: is also a time function that computes, for each time instant t, the *best* possible prediction $S_{\mathbf{p}_S^0}(t)$ that model $\mathcal{M}$ can produce for the substrate concentration. This prediction is obtained when model $\mathcal{M}$ is identified on the basis of an infinite data set.

Now, what theorem 8.1 states is that, as N increases, the deviation between *(1)* the identified process model prediction function $S_{\mathbf{p}_S^{\Upsilon_N}}$ and *(2)* the real process response function $g^{\mathcal{T}_{asp}}$ shrinks continuously until it reaches its minimum, which is

$\mathcal{D}(S_{\mathrm{p}_S^0}, g^{\mathcal{T}_{asp}})$. In other words, as N increases, the identified process model keeps improving more and more until it reaches the best possible prediction function $S_{\mathrm{p}_S^0}$ of the real process behaviour. However, this "*best function*" $S_{\mathrm{p}_S^0}$ does not equal the real process response function $g^{\mathcal{T}_{asp}}$, because there is a nonzero deviation $\mathcal{D}(S_{\mathrm{p}_S^0}, g^{\mathcal{T}_{asp}})$ between these two functions. This means that, though the identified process model improves as N increases, it will ***never*** reach the real process response function $g^{\mathcal{T}_{asp}}$ we are looking for, even for infinitely large values of N! $S_{\mathrm{p}_S^0}$ will be the best prediction function that one can reach (note that $S_{\mathrm{p}_S^0}$ can be reached in theory only). As a result, even under the ideal conditions of infinite data points and computing power, there will *always* be a *residual gap* with a magnitude of $\mathcal{D}(S_{\mathrm{p}_S^0}, g^{\mathcal{T}_{asp}})$ between the identified process model that will be used in real-world situations for the management of the plant $\mathcal{T}_{asp}$ and the "*reality*" as it is expressed by $g^{\mathcal{T}_{asp}}$.

Consequently, the answer to questions 8.19 and 8.20 is no: empirical data cannot compensate for our limited knowledge of process mechanisms, even if an infinite amount of data and computing power are made available to the model identification procedure. But, if this is the case, how could some researchers be able to achieve very satisfactory results with a black-box model (such as a neural network model, for example) trained on the basis of a relatively large data set? A black-box model contains absolutely no information about the process before training (see discussion of this fact in a previous Chapter), yet process modellers could use just empirical data to transfer a great deal of information about the ASP into this black-box model and, thus, produce a satisfactory ASP model. So, we should, in fact, conclude that it is possible to compensate for the lack of information about process mechanisms by using empirical data. How to explain this discrepancy?

The negative answer to questions 8.19 and 8.20 was established for a fixed mechanistic model $\mathcal{M}$. But, a black-box model such a neural network is not one fixed model: it is actually a series of models where the number of nodes — and therefore the model parameters — can be increased or decreased during the training, until the optimal model structure is attained. Therefore, black-box model training procedures exploit not only the information contained in the data set, but also the possibility of *adjusting* the model structure in order to better approximate the real process response function. And what makes black-box models such as neural networks successful in doing so is the fact that neural networks are *universal approximators*, which means that their structure can always be adjusted to uniformly approximate practically *any* continuous function to any desired degree of accuracy (see Haykin, 1994, pp. 181 and 189).

In the case of the mechanistic model $\mathcal{M}$, however, model improvement relies exclusively on the information content of the data set. The use of model structure adjustment to enhance the approximation of the real process behaviour cannot be carried out with just one single model. That is why, in the next Chapter, an innovative modelling approach will be devised for the ASP in order to develop a series of mechanistic models for this process. This will provide us with that advantage of model structure "adjustability" in the area of mechanistic modelling too and will contribute to the development of a process model which has the advantages of both mechanistic and black-box models, an *all-advantage* model that is.

Now that the foregoing discrepancy has been explained, it should be noted that complex (mechanistic or black-box) model structures are likely to achieve a better approximation of $g^{\mathcal{T}_{asp}}$, but they would require a large number of data points for their identification. Since data are usually scarce, there is a need to develop a methodology to tune the model structure complexity to the small number of data points available for the identification procedure. This methodology will be presented in Chapter 10.

In the remainder of this Chapter, we will carry on investigating the uncertainty that underlies the ASP plant behaviour, being approximated by mechanistic model $\mathcal{M}$.

8.3 Introducing a New Concept: Process Model Maximal and Marginal Improvements

It has been shown in the previous section that, even under ideal conditions of infinite data points and computing power, process model identification does not guarantee the exact determination of the real process response function $g^{\mathcal{T}_{asp}}$. However, it is a fact that the identified process model *improves* as the amount of data N supplied to the identification procedure increases. To manage the process model quality, we need to develop a measure of *this model improvement*. Some of the questions that arise in this respect and need to be addressed are: what is the maximal improvement we can achieve for model $\mathcal{M}$, and how to quantify it? How much does $\mathcal{M}$ improve if the number of identification data points increases by one point, for instance? How does the model improvement rate vary with N? Is it constant or does it vary with N? What practical conclusions can be drawn from answers to these questions and what recommendations can be made to wastewater treatment plant managers and consultants with respect to the number of points to be used for model identification?

These questions are quite challenging and have never been investigated in the area of wastewater engineering, despite their crucial importance. In this section, they are addressed in the light of the uncertainty models developed in the previous Chapter.

8.3.1 Measures of Process Model Improvement

Process Model Improvement as N Increases from N_1 to N_2

In what follows, the process model improvement when the number of identification data points increases from N_1 to N_2 will be denoted as $\mathcal{IP}_{\mathcal{M}}(N_1, N_2)$. The purpose of this section is to develop a measure of $\mathcal{IP}_{\mathcal{M}}(N_1, N_2)$.

A measure must satisfy two conditions:

Measure Condition n° 1: it characterizes the magnitude of the "concept" it is quantifying: to each concept magnitude corresponds one unique measure.

Measure Condition n° 2: it increases (or decreases) as the magnitude of the "concept" it is quantifying increases (or decreases).

Now, to develop a measure of the concept

"process model improvement"

we need to think about what characterizes an identified model being better than another identified model. One natural way of quantifying it would be:

$$\mathcal{IP}_{\mathcal{M}}(N_1, N_2) = \parallel \mathbf{p}_S^{\Upsilon_{N_1}} - \mathbf{p}_S^{\Upsilon_{N_2}} \parallel \tag{8.22}$$

that is, the euclidian distance between the parameter vectors $\mathbf{p}_S^{\Upsilon_{N_1}}$ and $\mathbf{p}_S^{\Upsilon_{N_2}}$. This definition is, however, not appropriate, because the difference between parameter vectors does not characterize the amount by which an identified model has improved: a large difference $\| \mathbf{p}_S^{\Upsilon_{N_1}} - \mathbf{p}_S^{\Upsilon_{N_2}} \|$ does not mean necessarily that the identified process model $S_{\mathbf{p}_S^{\Upsilon_N}}$ has improved a lot as N increased from N_1 to N_2. Consequently, the expression 8.22 does not satisfy measure condition n° 2 and, as a result, the difference $\| \mathbf{p}_S^{\Upsilon_{N_1}} - \mathbf{p}_S^{\Upsilon_{N_2}} \|$ cannot be adopted as a measure of $\mathcal{IP}_{\mathcal{M}}(N_1, N_2)$.

Another way of quantifying $\mathcal{IP}_{\mathcal{M}}(N_1, N_2)$ would be:

$$\mathcal{IP}_{\mathcal{M}}(N_1, N_2) = \mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_{N_1}}}, S_{\mathbf{p}_S^{\Upsilon_{N_2}}}) \tag{8.23}$$

that is, the functional deviation $\mathcal{D}$ between the two identified models $S_{\mathbf{p}_S^{\Upsilon_{N_1}}}$ and $S_{\mathbf{p}_S^{\Upsilon_{N_2}}}$. This definition is not appropriate either, because of the lack of process model identifiability pointed out earlier in this thesis: it is indeed possible that the identification objective function $J_S(\mathbf{p}_S)$ corresponding to N_2, for instance, attains its minimum at two *different* parameter vectors $(\mathbf{p}_S^{\Upsilon_{N_2}})_1$ and $(\mathbf{p}_S^{\Upsilon_{N_2}})_2$. When this is the case, the equation 8.23 will assign two different numbers, $\mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_{N_1}}}, S_{(\mathbf{p}_S^{\Upsilon_{N_2}})_1})$ and $\mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_{N_1}}}, S_{(\mathbf{p}_S^{\Upsilon_{N_2}})_2})$, to the same concept $\mathcal{IP}_{\mathcal{M}}(N_1, N_2)$. This is in contradiction with measure condition n° 1 and, as a result, the deviation $\mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_{N_1}}}, S_{\mathbf{p}_S^{\Upsilon_{N_2}}})$ cannot be adopted as a measure of $\mathcal{IP}_{\mathcal{M}}(N_1, N_2)$.

A third possible way of quantifying $\mathcal{IP}_{\mathcal{M}}(N_1, N_2)$ would consist in utilizing the deviation between the identified process model and the real process response function, i.e.:

$$\mathcal{IP}_{\mathcal{M}}(N_1, N_2) = \mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_{N_1}}}, g^{\mathcal{T}_{asp}}) - \mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_{N_2}}}, g^{\mathcal{T}_{asp}}) \tag{8.24}$$

According to theorem 8.1, the deviation $\mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_N}}, g^{\mathcal{T}_{asp}})$ decreases more and more, as N increases, until it reaches its minimum value $\mathcal{D}(S_{\mathbf{p}_S^0}, g^{\mathcal{T}_{asp}})$. Hence, the larger the deviation between N_1 and N_2, i.e., the higher the amount by which the identified model $S_{\mathbf{p}_S^{\Upsilon_N}}$ has improved, the bigger the difference $\mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_{N_1}}}, g^{\mathcal{T}_{asp}}) - \mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_{N_2}}}, g^{\mathcal{T}_{asp}})$. Thus, the expression 8.24 satisfies the measure condition n° 2. However, because of the lack of process model identifiability, the parameter vectors $\mathbf{p}_S^{\Upsilon_{N_1}}$ and $\mathbf{p}_S^{\Upsilon_{N_2}}$ are not necessarily unique and, as a result, the difference $\mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_{N_1}}}, g^{\mathcal{T}_{asp}}) - \mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_{N_2}}}, g^{\mathcal{T}_{asp}})$ is not unique for every fixed couple (N_1, N_2). Measure condition n° 1 is therefore not satisfied for the expression 8.24 and, consequently, the latter cannot be adopted as a measure of $\mathcal{IP}_{\mathcal{M}}(N_1, N_2)$. Nevertheless, uncertainty model $\mathcal{UM}_1$ (equation 8.1) suggest to replace the deviations $\mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_{N_i}}}, g^{\mathcal{T}_{asp}})$ in equation 8.24 by the upper bounds on the squares of these deviations, i.e., the guaranteed deviation:

$$\varphi_1 = \varphi_1(N_i) = \left(\frac{J_S(\mathbf{p}_S^{\Upsilon_{N_i}})}{N_i}\right)\left(1 + \frac{\kappa\zeta(N_i)}{2}\left(1 + \sqrt{1 + \frac{4}{\kappa\zeta(N_i)}}\right)\right) \tag{8.25}$$

Using this idea, the process model improvement can then be quantified in the following way:

$$\mathcal{IP}_{\mathcal{M}}(N_1, N_2) = \varphi_1(N_1) - \varphi_1(N_2) \tag{8.26}$$

Note that the expression of $\varphi_1(N_i)$ is the product of two terms: the empirical risk

$$R_{emp}^{\Upsilon_{N_i}}(\mathbf{p}_S^{\Upsilon_{N_i}}) = \frac{J_S(\mathbf{p}_S^{\Upsilon_{N_i}})}{N_i} \tag{8.27}$$

and the mathematical function

$$\phi(N_i) = 1 + \frac{\kappa\zeta(N_i)}{2}\left(1 + \sqrt{1 + \frac{4}{\kappa\zeta(N_i)}}\right) \tag{8.28}$$

that is:

$$\varphi_1(N_i) = R_{emp}^{\Upsilon_{N_i}}(\mathbf{p}_S^{\Upsilon_{N_i}}) \times \phi(N_i) \tag{8.29}$$

Using this decomposition (equation 8.29), equation 8.26 can then be re-written as:

$$\mathcal{IP}_{\mathcal{M}}(N_1, N_2) = R_{emp}^{\Upsilon_{N_1}}(\mathbf{p}_S^{\Upsilon_{N_1}}) \times \phi(N_1) - R_{emp}^{\Upsilon_{N_2}}(\mathbf{p}_S^{\Upsilon_{N_2}}) \times \phi(N_2) \tag{8.30}$$

Before discussing the measure conditions for expression 8.30, we need to examine an important point regarding the vector $\mathbf{p}_S^{\Upsilon_N}$: *global* versus *local* minimum.

By definition, $\mathbf{p}_S^{\Upsilon_N}$ represent the parameter vector at which the identification objective function or, equivalently, the empirical risk attains its global minimum. In practice, there is no general method or algorithm that guarantees the determination of the exact value of the global minimum of $R_{emp}^{\Upsilon_N}(\mathbf{p}_S)$. The best we can get in reality is a local minimum, which we will denote here as $(\mathbf{p}_S^{\Upsilon_N})_{loc}$. Therefore, the identified process model that will be used in real ASP control is in fact $S_{(\mathbf{p}_S^{\Upsilon_N})_{loc}}$, and not $S_{\mathbf{p}_S^{\Upsilon_N}}$. This situation is, of course, not desirable, because $S_{\mathbf{p}_S^{\Upsilon_N}}$, if it could be obtained, would be a better model. But, unfortunately, it is not accessible in practice and we have to be content to use a less effective model which is $S_{(\mathbf{p}_S^{\Upsilon_N})_{loc}}$. Can the uncertainty model $\mathcal{UM}_1$ (inequality 8.1) be used with the parameter vector $(\mathbf{p}_S^{\Upsilon_N})_{loc}$ instead of $\mathbf{p}_S^{\Upsilon_N}$? Yes, and this is one of the strengths of the general methodology developed in this thesis for uncertainty control. It suffices indeed to replace $\mathbf{p}_S^{\Upsilon_N}$ by $(\mathbf{p}_S^{\Upsilon_N})_{loc}$ in the inequality 8.1. We then obtain uncertainty model $\mathcal{UM}_1'$:

$$[\mathcal{D}(S_{(\mathbf{p}_S^{\Upsilon_N})_{loc}}, g^{\mathcal{T}_{asp}})]^2 \leq \left(\frac{J_S((\mathbf{p}_S^{\Upsilon_N})_{loc})}{N}\right)\left(1+\frac{\kappa\zeta(N)}{2}\left(1+\sqrt{1+\frac{4}{\kappa\zeta(N)}}\right)\right) \qquad (8.31)$$

Accordingly, equation 8.30 can be adjusted by replacing the empirical risk global minimums by the local minimums:

$$\mathcal{IP}_{\mathcal{M}}(N_1, N_2) = R_{emp}^{\Upsilon_{N_1}}((\mathbf{p}_S^{\Upsilon_{N_1}})_{loc}) \times \phi(N_1) - R_{emp}^{\Upsilon_{N_2}}((\mathbf{p}_S^{\Upsilon_{N_2}})_{loc}) \times \phi(N_2) \qquad (8.32)$$

Now, we need to check whether or not the two measure conditions are satisfied for the expression 8.32.

Note first that the measure condition n° 1 is well satisfied for the original expression 8.30: indeed, for every fixed data sequence Υ_N, the value of $\phi(N)$ and that of the empirical risk global minimum $R_{emp}^{\Upsilon_N}(\mathbf{p}_S^{\Upsilon_N})$ are unique (as for $R_{emp}^{\Upsilon_N}(\mathbf{p}_S^{\Upsilon_N})$, it is a mathematical property of function global minimum). However, in expression 8.32, the local minimum $R_{emp}^{\Upsilon_N}((\mathbf{p}_S^{\Upsilon_N})_{loc})$ is of course not necessarily unique, because a function can have several local minimums. However, it is conjectured that, given an optimization algorithm and a certain amount of computing power, $R_{emp}^{\Upsilon_{N_1}}((\mathbf{p}_S^{\Upsilon_N})_{loc})$ changes very little with N. And, what will affect the value of $R_{emp}^{\Upsilon_N}((\mathbf{p}_S^{\Upsilon_N})_{loc})$ significantly is the complexity or, equivalently, the VC dimension of the process model used to approximate the ASP plant dynamical behaviour. This point will be investigated in subsequent Chapters of this thesis. In this Chapter, however, the ASP model has been fixed and, because of that, $R_{emp}^{\Upsilon_N}((\mathbf{p}_S^{\Upsilon_N})_{loc})$ will be considered to be practically constant:

$$R_{emp}^{\Upsilon_N}((\mathbf{p}_S^{\Upsilon_N})_{loc}) \approx c \qquad (8.33)$$

with the number $c \in \Re_+^\circ$ depends primarily on the degree of complexity (VC dimension, that is) of the process model. Equation 8.32 becomes then:

$$\mathcal{IP}_{\mathcal{M}}(N_1, N_2) = c \times (\phi(N_1) - \phi(N_2)) \qquad (8.34)$$

Measure conditions n° 1 and n° 2 are both satisfied for expression 8.34 (the proof is straightforward). As a result, $c \times (\phi(N_1) - \phi(N_2))$ can be adopted as a measure of the process model improvement $\mathcal{IP}_{\mathcal{M}}(N_1, N_2)$.

Process Model Maximal Improvement

Process model maximal improvement, denoted here as $\mathcal{IP}_{\mathcal{M}}^{max}$, is defined as the value of $\mathcal{IP}_{\mathcal{M}}(N_1, N_2)$ when N_1 approaches its minimum (which is 2) and N_2 its maximum (which is infinity). That is:

$$\begin{aligned}\mathcal{IP}_{\mathcal{M}}^{max} &= \mathcal{IP}_{\mathcal{M}}(2, \infty) \\ &= c \times (\phi(2) - \phi(\infty)) \\ \mathcal{IP}_{\mathcal{M}}^{max} &= c \times 7.78\,\kappa \left(1 + \sqrt{1 + \frac{1}{3.89\,\kappa}}\right) \end{aligned} \tag{8.35}$$

Knowing that $\kappa \in [1.82, 4]$ and, therefore, $3.89 \times \kappa \gg 1$, it is possible to apply the approximation 8.4 to the square root of equation 8.35:

$$\begin{aligned}\mathcal{IP}_{\mathcal{M}}^{max} &= c \times 7.78\,\kappa \left(1 + \sqrt{1 + \frac{1}{3.89\,\kappa}}\right) \\ &= c \times 7.78\,\kappa \left(1 + 1 + \frac{1}{7.78\,\kappa}\right) \\ &= c \times (15.56\,\kappa + 1) \\ \mathcal{IP}_{\mathcal{M}}^{max} &\approx 15.56\,c\,\kappa \end{aligned} \tag{8.36}$$

From this equation, it follows that $\mathcal{IP}_{\mathcal{M}}^{max}$ increases — linearly — with the prior information κ. Does this make sense? Yes: if a low value is selected for κ, it means that we are pretending to possess a strong prior information about the ASP dynamical behaviour, being approximated by model $\mathcal{M}$. In other words, a low value of κ is as if we are assuming that a high amount of information is available to the process modeller before even the beginning of the identification procedure. Because of this, model $\mathcal{M}$ would need to draw very little information from the plant empirical data (during its identification), to reach its highest degree of accuracy. Hence, the maximum amount by which $\mathcal{M}$ would improve, as N approaches infinity, would not be very significant. That is: a low value of κ implies a low value of $\mathcal{IP}_{\mathcal{M}}^{max}$ and vice-versa.

Process Model Marginal Improvement

"*The more data we have, the better it is for model identification*". The following discussion is about quantifying this statement, in the marginal sense: how better does model $\mathcal{M}$ become, when the size of the identification data set is increased by one point, for instance? Consider the following two cases:

- ***case A***: $N_0 = 10$ points are used for model identification.
- ***case B***: $N_0 = 100$ points are used for model identification.

Does an increase of N_0 by one unit lead to exactly the same model improvement in both cases A and B? If yes, that is if the model $\mathcal{M}$ improves uniformly as N_0 increases, then what is the value of the model improvement constant rate? If no, then how does the model improvement rate vary with N_0?

To address these questions, consider the model improvement when the number of identification data points is increased from N_0 to certain number N. From our previous discussion, this improvement is expressed as:

$$\mathcal{IP}_{\mathcal{M}}(N_0, N) = c \times (\phi(N_0) - \phi(N)) \tag{8.37}$$

To obtain the amount of model improvement *per one identification data point*, we divide equation 8.37 by $N_0 - N$:

$$\frac{\mathcal{IP}_{\mathcal{M}}(N_0, N)}{N_0 - N} = c \times \frac{\phi(N_0) - \phi(N)}{N_0 - N} \tag{8.38}$$

The ratio

$$\frac{\mathcal{IP}_{\mathcal{M}}(N_0, N)}{N_0 - N}$$

represents the *model improvement average rate between* N_0 and N. To compute the *model improvement rate at* N_0, denoted as $\varpi(N_0)$, we take the limit of the expressions of equation 8.38, as N approaches N_0:

$$\varpi(N_0) = \lim_{N \to N_0} \frac{\mathcal{IP}_{\mathcal{M}}(N_0, N)}{N_0 - N} = c \times \lim_{N \to N_0} \frac{\phi(N_0) - \phi(N)}{N_0 - N} \tag{8.39}$$

By definition, the limit

$$\lim_{N \to N_0} \frac{\phi(N_0) - \phi(N)}{N_0 - N}$$

represents the derivative

$$\frac{\partial \phi}{\partial N}(N_0)$$

of the function ϕ with respect to N at N_0. Therefore, equation 8.39 can be re-written as:

$$\varpi(N_0) = c \times \frac{\partial \phi}{\partial N}(N_0) \tag{8.40}$$

Note that the derivative of ϕ with respect to N is a negative function. Since we are not interested in the derivative sign, the following expression will be adopted for $\varpi(N_0)$:

$$\varpi(N_0) = c \times \left| \frac{\partial \phi}{\partial N}(N_0) \right| = -c \times \frac{\partial \phi}{\partial N}(N_0) \tag{8.41}$$

Equation 8.41 defines the expression of the process model improvement rate $\varpi(N_0)$ at N_0, that is, the process model improvement when N_0 increases by *one* point. This rate will also be called *process model marginal improvement.*

The expression of $\varpi(N_0)$ can be developed explicitly by differentiating equations 8.28 (for ϕ) and 8.3 (for ζ):

$$\varpi(N_0) = c \times \left(\frac{\phi(N_0) - 1}{\zeta(N_0)} - \frac{\kappa}{2\,\phi(N_0) - \kappa\,\zeta(N_0) - 2} \right) \times \left(\frac{\zeta(N_0)}{N_0} - \frac{8}{{N_0}^2} \right) \tag{8.42}$$

Now that we have developed a measure of the process model marginal improvement, answers to the questions posed at the beginning of this sub-section (cases A and B) become straightforward:

- When the size N of the identification data set is increased by one point, model $\mathcal{M}$ improves by an amount of $\varpi(N)$.
- An increase of N by one unit does not lead to the same model improvement in cases A ($N = 10$) and B ($N = 100$), because the values of $\varpi(10)$ and $\varpi(100)$ are different. Consequently, model $\mathcal{M}$ does not improve uniformly as N increases.
- Note that, because $\varpi(100) < \varpi(10)$ (compute the derivative of $\varpi(N)$ with respect and verify that it is always negative), model $\mathcal{M}$ improves much less when one point is added to 100 points than to 10 points.

8.3.2 Recapitulation - Relationships among Model Improvement Measures

Three measures of process model improvement were defined:

- Process model improvement as N increases from N_1 to N_2:

$$\mathcal{IP}_{\mathcal{M}}(N_1, N_2) = c \times (\phi(N_1) - \phi(N_2)) \tag{8.43}$$

- Process model maximal improvement:

$$\mathcal{IP}_{\mathcal{M}}^{max} = 15.56\, c\, \kappa \tag{8.44}$$

- Process model marginal improvement:

$$\varpi(N) = c \times \left(\frac{\phi(N) - 1}{\varsigma(N)} - \frac{\kappa}{2\,\phi(N) - \kappa\,\varsigma(N) - 2} \right) \times \left(\frac{\varsigma(N)}{N} - \frac{8}{N^2} \right) \tag{8.45}$$

where:

$$\phi(N) = 1 + \frac{\kappa\,\varsigma(N)}{2} \left(1 + \sqrt{1 + \frac{4}{\kappa\,\varsigma(N)}} \right) \tag{8.46}$$

$$c \approx R_{emp}^{\Upsilon_N} \left((\mathrm{p}_S^{\Upsilon_N})_{loc} \right) \tag{8.47}$$

$$\varsigma(N) = 4 \left(\frac{2\, \ln N + 6.4}{N} \right) \tag{8.48}$$

Note that the unit of all model improvement measures (equations 8.43, 8.44 and 8.45) is the square of the unit of the substrate concentration S. For example, if S is expressed in mg/l as COD, then model improvement measures will be expressed in $(mg/l)^2$ as COD. To define new measures with the same unit as the substrate concentration unit, it suffices to take the square root of the right-hand side of equations 8.43, 8.44 and 8.45.

The following relationships can be easily verified:

$$\mathcal{IP}_{\mathcal{M}}(N_1, N_2) = \int_{N_1}^{N_2} \varpi(N)\, dN \tag{8.49}$$

$$\mathcal{IP}_{\mathcal{M}}^{max} = \int_{2}^{+\infty} \varpi(N)\, dN \tag{8.50}$$

8.4 Practical Recommendations for Wastewater Treatment Plant Managers and Consultants

The ultimate goal of uncertainty management is to reduce the deviation $\mathcal{D}(S_{\mathrm{p}_S^{\Upsilon_N}}, g^{\mathcal{T}_{asp}})$ as much as possible. Ideally, we would like to be able to bring this deviation down to zero, that is: $\mathcal{D}(S_{\mathrm{p}_S^{\Upsilon_N}}, g^{\mathcal{T}_{asp}}) = 0$ in order to achieve: $S_{\mathrm{p}_S^{\Upsilon_N}} = g^{\mathcal{T}_{asp}}$. However, it was shown above that, with model $\mathcal{M}$, it is impossible to reduce $\mathcal{D}(S_{\mathbf{p}_S^{\Upsilon_N}}, g^{\mathcal{T}_{asp}})$ to zero through the use of the process model identification procedure, even under the ideal conditions of infinite data points and computing power. So the question that arises is then: what is the best process model improvement we can achieve, and at what cost?

To address such a question, we will make use of the uncertainty model (see inequality 8.31):

$$\mathcal{UM}_1' : \qquad [\mathcal{D}(S_{(\mathbf{p}_S^{\Upsilon_N})_{loc}}, g^{\mathcal{T}_{asp}})]^2 \leq R_{emp}^{\Upsilon_N}((\mathbf{p}_S^{\Upsilon_N})_{loc}) \times \phi(N) \tag{8.51}$$

where $R_{emp}^{\Upsilon_N}((\mathbf{p}_S^{\Upsilon_N})_{loc})$ is the value of the empirical risk at a parameter vector $(\mathbf{p}_S^{\Upsilon_N})_{loc}$ at which a certain local minimum is attained and the function ϕ is defined by equation 8.28. The idea here is to utilize the fact that the right-hand side of inequality 8.51, i.e., the guaranteed deviation $\varphi_1(N)$, is known and can be easily computed. Thus, by reducing it, we will reduce the deviation $\mathcal{D}(S_{(\mathbf{p}_S^{\Upsilon_N})_{loc}}, g^{\mathcal{T}_{asp}})$, which is the measure of uncertainty in ASP. As was mentioned previously, the first term of $\varphi_1(N)$, that is $R_{emp}^{\Upsilon_N}((\mathbf{p}_S^{\Upsilon_N})_{loc})$, changes very little with the identification data set size N and, because of that, it was considered to be almost constant. The second term, however, is the one that governs most of the variations of the guaranteed deviation $\varphi_1(N)$. So, to reduce the latter function (which would result in reducing the deviation $\mathcal{D}(S_{(\mathbf{p}_S^{\Upsilon_N})_{loc}}, g^{\mathcal{T}_{asp}})$), we need to act on $\phi(N)$ and minimize it. The minimum of the latter function is 1 and, as a result, the best we can achieve, according to uncertainty model $\mathcal{UM}_1'$, is a guarantee that the deviation $\mathcal{D}(S_{(\mathbf{p}_S^{\Upsilon_N})_{loc}}, g^{\mathcal{T}_{asp}})$ is lower than the empirical risk local minimum value:

$$R_{emp}^{\Upsilon_N}((\mathbf{p}_S^{\Upsilon_N})_{loc}) = \frac{J_S((\mathbf{p}_S^{\Upsilon_N})_{loc})}{N} \tag{8.52}$$

that is:

$$\mathcal{D}(S_{(\mathbf{p}_S^{\Upsilon_N})_{loc}}, g^{\mathcal{T}_{asp}}) \leq R_{emp}^{\Upsilon_N}((\mathbf{p}_S^{\Upsilon_N})_{loc}) \tag{8.53}$$

Now, to reduce $R_{emp}^{\Upsilon_N}((\mathbf{p}_S^{\Upsilon_N})_{loc})$, we will need to improve the process model structure. This point will be investigated in subsequent Chapters of this thesis. In the rest of this Chapter, where the process model structure is fixed, we will concentrate on the effect of the data set size N on the quality of identified model $\mathcal{M}$. We will do this through the investigation of the mathematical function $\phi(N)$ and the use of process model improvement measures defined in previous section.

8.4.1 A Fundamental Result

The purpose of this section is to prove a fundamental result with respect to the practice of ASP model identification. This result is stated below:

In the course of identifying the ASP model $\mathcal{M}$ with respect to the substrate concentration S, 80% of the model maximum improvement occurs at a number of data points of about $N_{80\%} \approx 15$ to 18. To achieve the other 20%, N has to be increased from the relatively small number $N_{80\%}$ to infinity. This result is practically independent of the prior information $\mathcal{WPI}$ as it is expressed by the value of the number κ.

Here is the proof of this result:

As pointed out in previous section, the process model improvement that occurs when the number of data points increases from its minimum 2 to a certain value N is equal to:

$$\mathcal{IP}_{\mathcal{M}}(2, N) = c \times (\phi(2) - \phi(N)) \tag{8.54}$$

while the maximum improvement is equal to:

$$\mathcal{IP}_{\mathcal{M}}^{max} = c \times (\phi(2) - \phi(\infty)) = 15.57\, c\, \kappa \tag{8.55}$$

So the question to be addressed here is: what should be the value of N in order to achieve:

$$\mathcal{IP}_{\mathcal{M}}(2, N) = 0.8 \times \mathcal{IP}_{\mathcal{M}}^{max} \; ? \tag{8.56}$$

This equation can be re-written as:

$$\frac{\mathcal{IP}_{\mathcal{M}}(2, N)}{\mathcal{IP}_{\mathcal{M}}^{max}} = 0.8$$

or, equivalently:

$$\frac{c \times (\phi(2) - \phi(N))}{15.57\, c\, \kappa} = 0.8$$

$$\frac{\phi(2) - \phi(N)}{15.57\, \kappa} = 0.8 \tag{8.57}$$

Hence the foregoing question comes down to solving the equation 8.57. The function f defined by the expression in the left-hand side of this equation, i.e.:

$$\begin{array}{rccl} f: & \{2,3,4,\cdots\} & \rightarrow & [0,1[\\ & N & \mapsto & f(N) = \frac{\phi(2)-\phi(N)}{15.57\,\kappa} \end{array} \tag{8.58}$$

is continuous and strictly increasing from 0 to 1. Note indeed that $f(N)$ is also equal to:

$$f(N) = \frac{\phi(2) - \phi(N)}{\phi(2) - \phi(\infty)}$$

which then makes the previous statement about the variations of $f(N)$ easier to prove. Therefore, f is an injective function and, as a result, the equation 8.57 in N has at most one solution. Denote this solution, when it exists, as $N_{80\%}$. To determine $N_{80\%}$, we will first limit the search in the subset of integers $\{2,3,\cdots,16\}$. If we can find one number in this subset that satisfies equation 8.57, then this number is equal to the one we are looking for, i.e., $N_{80\%}$, because of the injectivity of the function $f(N)$.

For $N \in \{2,3,\cdots,16\}$ and $\kappa \in [1.82, 4]$, the following inequality can be proved in a straightforward manner:

$$1.82 \times 2.98 \leq \kappa\,\zeta(N) \leq 4 \times 15.57$$

that is:

$$5.42 \leq \kappa\,\zeta(N) \leq 62.28$$

or, equivalently:

$$0.06 \leq \frac{4}{\kappa\,\zeta(N)} \leq 0.74$$

The ratio $4/\kappa\,\zeta(N)$ is therefore small enough to justify the use of equation 8.4 for the square root:

$$\sqrt{1 + \frac{4}{\kappa\,\zeta(N)}}$$

(Note indeed that:

$$(1 + 0.74)^{0.5} = 1.320 \quad \text{and} \quad 1 + 0.5 \times 0.74 = 1.370$$

and similarly,

$$(1 + 0.06)^{0.5} = 1.029 \quad \text{and} \quad 1 + 0.5 \times 0.06 = 1.030)$$

Hence:

$$\sqrt{1+\frac{4}{\kappa\zeta(N)}} \approx 1+\frac{2}{\kappa\zeta(N)}$$

Consequently, the expression of $\phi(N)$ can be simplified as follows:

$$\begin{aligned}\forall N \in \{2,\cdots,16\},\ \forall\kappa \in [1.82,4],\quad \phi(N) &= 1+\frac{\kappa\zeta(N)}{2}\left(1+\sqrt{1+\frac{4}{\kappa\zeta(N)}}\right)\\ &\approx 1+\frac{\kappa\zeta(N)}{2}\left(1+1+\frac{2}{\kappa\zeta(N)}\right)\\ \phi(N) &\approx 2+\kappa\zeta(N) \qquad (8.59)\end{aligned}$$

Thus, for $N \in \{2,3,\cdots,16\}$ and $\kappa \in [1.82,4]$, the expression of the function $f(N)$ becomes:

$$f(N) = \frac{\phi(2)-\phi(N)}{15.57\,\kappa} \approx \frac{(2+\kappa\zeta(2))-(2+\kappa\zeta(N))}{15.57\,\kappa} = \frac{\zeta(2)-\zeta(N)}{15.57} \qquad (8.60)$$

i.e., an expression that is independent of κ. Therefore, equation 8.57 is also independent of κ for $N \in \{2,3,\cdots,10\}$, and can re-written as:

$$\frac{\zeta(2)-\zeta(N)}{15.57} = 0.8 \qquad (8.61)$$

or:

$$\zeta(N) = \zeta(2) - 0.8 \times 15.57$$

$$\zeta(N) = 3.12 \qquad (8.62)$$

Solving this equation in N gives:

$$N = N_{80\%} \approx 15 \text{ data points} \qquad (8.63)$$

The foregoing approximations were used mainly to show why the amount of data that is required to achieve 80% of the model maximum improvement is practically independent of the prior information κ. It is however possible to solve equation 8.57 rigorously for each fixed value of κ, using a simple spreadsheet. The results are presented in table 7.1

It is therefore clear that $N_{80\%}$ is practically independent of κ and has a value of about 15 to 18.

Table 8.1: The number $N_{80\%}$ is practically independent of κ

Value of κ	Value of $N_{80\%}$
1.82	18
2.00	18
2.30	17
2.50	17
2.70	17
2.91	17
3.50	17
4.00	17
5.00	16
10.00	16
15.00	16
30.00	16
40.00	16
50.00	16
100.00	16

8.4.2 Marginal Cost of Process Model Improvement

The concepts of "*marginal cost*" and "*marginal revenue*" are used extensively in economics. They constitute very powerful decision tools for determining, among others, the profit-maximizing level of a firm output (see the articles of Ordover and Machlup in the McGraw-Hill Encyclopedia of Economics, Greenwald (1994)). The intent of this section is to develop similar concepts that can be used in WWT process control strategy design.

Motivation

One of the major duties of ASP engineers is to develop a strategy to control the dynamics of the activated sludge WWT plant, so that the latter meets the regulatory effluent quality standards at all (or most of) times. To do so, they can make use of two distinct approaches: *structural control* and *operational control*. The former approach consists in acting on the wastewater treatment (WWT) plant units by physically adding or taking off one (or several) unit(s) or by changing the physical characteristics of existing ones. The latter approach consists in acting on the process control variables to influence the process dynamics and make them consistent with regulatory effluent standards. In order to develop the most cost-effective control strategy for the ASP using (one or both of) these two approaches, process engineers need to know the cost and performance of each physical or mathematical tool they have at their disposal for process control strategy development. And one of these tools is of course the process model which represents a fundamental component of every operational control strategy. The purpose of this section is to develop a procedure to help process engineers determine the cost Ct_{mrg} of improving *an-already-identified* model by one unit (unit of Ct_{mrg}: \$/unit of the modeled process variable). The cost Ct_{mrg} is called *marginal cost of process model improvement.* Here is an example of circumstances where the computation of this cost would be very useful:

> *A process model of an activated sludge WWT plant is developed and identified. Denote this model $\mathcal{M}$. Based on $\mathcal{M}$, an operational control strategy is developed for the plant. After this control strategy is implemented and used correctly for a certain period of time, the plant manager realizes that his facility still violates the effluent quality standards (note that the manager could reach this conclusion just by carrying out a series of computer simulations of his plant). Because of this, he asks his consultant engineer to review and improve the process control strategy. In the course of reviewing the latter strategy, the consultant shows that several different options are technically feasible and each one of them has the potential of improving this strategy. A partial list of these options could be:*
>
> 1. *improving the accuracy of the existing identified process model $\mathcal{M}$ by adding more points to the identification data set.*
> 2. *developing a new process model structure which may possibly show better performance.*
> 3. *change nothing model wise, but start using some chemicals at a certain stage of the treatment process.*
> 4. *adding a flow equalization unit to the plant and consider eliminating the current operational problems for a longer period of time, say, ten*

years.

The consultant has established that all these options are technically viable, but to decide which one to select, he has to carry out a cost-benefit analysis of all these options. To do so, he needs to know how much money is required to achieve a certain "amount" of benefit for each option. In particular, for option 1, he needs to know the cost of improving the existing identified model accuracy by, say, one unit, i.e., 1 mg/l as COD. That is the marginal Cost Ct_{mrg}.

Developing the Procedure

Now that the motivation of the marginal cost of model improvement is explained, consider that model $\mathcal{M}$ (see equations 7.1) has been identified, with respect to the substrate concentration S, using a data set of size N and a certain optimization algorithm $\mathcal{A}$. Denote the identified model as $S_{(\mathbf{p}_S^{\Upsilon_N})_{loc}}$, where $(\mathbf{p}_S^{\Upsilon_N})_{loc}$ is a parameter vector at which the identification objective function attains a local minimum. Note that we said "*a* parameter vector" instead of "*the* parameter vector". This is because of the lack of process model identifiability pointed out earlier in this thesis. In section 8.3.1, it was pointed out that, when the identification data set size N is increased by one point, the process model improves by an amount of $\varpi(N)$ in $(mg/l)^2$ as COD, or $\sqrt{\varpi(N)}$ in mg/l as COD, where $\varpi(N)$ is given by equation 8.45. Now, denote the cost of obtaining one extra data point as Ct_{pt}. The latter cost includes the costs of measuring all the variables that are needed to carry out the model identification, that is:

- the influent and recycle flowrates Q_{in} and Q_r
- the substrate concentration in the influent S_{in}
- the substrate concentration in the recycle S_r
- the substrate concentration in the effluent S
- possibly the micro-organisms concentration X

Using $\sqrt{\varpi(N)}$ and Ct_{pt}, the model marginal improvement in mg/l as COD per dollar can be obtained from the equation:

$$\omega = \frac{\sqrt{\varpi(N)}}{Ct_{pt}} \tag{8.64}$$

and the marginal cost of model improvement in dollars per mg/l as COD can be inferred as the inverse of ω:

$$Ct_{mrg} = \frac{1}{\omega} = \frac{Ct_{pt}}{\sqrt{\varpi(N)}} \tag{8.65}$$

The computation of Ct_{mrg} is detailed in the following procedure developed for the particular case of model $\mathcal{M}$:

1. *Compute the empirical risk local minimum* $R_{emp}^{\Upsilon_N}\left((\mathbf{p}_S^{\Upsilon_N})_{loc}\right)$ *on the basis of* N *data points using any optimization algorithm* $\mathcal{A}$*:*

$$c = R_{emp}^{\Upsilon_N}\left((\mathbf{p}_S^{\Upsilon_N})_{loc}\right) \tag{8.66}$$

2. *Compute the average of the prior information* κ *as:*

$$\bar{\kappa} = \frac{1.82 + 4}{2} = 2.91 \tag{8.67}$$

3. *Compute* $\zeta(N)$ *and* $\phi(N)$ *using the equations:*

$$\zeta(N) = 4\left(\frac{2\ \ln N + 6.4}{N}\right) \tag{8.68}$$

and

$$\phi(N) = 1 + \frac{\bar{\kappa}\,\zeta(N)}{2}\left(1 + \sqrt{1 + \frac{4}{\bar{\kappa}\,\zeta(N)}}\right) \tag{8.69}$$

4. *Compute* $\varpi(N)$ *and then* $\sqrt{\varpi(N)}$ *using the equation:*

$$\varpi(N) = c \times \left(\frac{\phi(N) - 1}{\zeta(N)} - \frac{\bar{\kappa}}{2\,\phi(N) - \bar{\kappa}\,\zeta(N) - 2}\right) \times \left(\frac{\zeta(N)}{N} - \frac{8}{N^2}\right) \tag{8.70}$$

5. *Compute the total cost* Ct_{pt} *of measuring all the variables* Q_{in}*,* Q_r*,* S_{in}*,* S_r*,* S *and* X *once (in dollars).*

6. *Compute the marginal cost of model improvement* Ct_{mrg} *(in dollars per* mg/l *as* COD*) using the equation:*

$$Ct_{mrg} = \frac{Ct_{pt}}{\sqrt{\varpi(N)}} \tag{8.71}$$

Illustrative Example

Case A: Consider that model $\mathcal{M}$ is identified, with respect to the substrate concentration S, on the basis of $N = 10$ data points and let us compute the marginal cost of improvement of $\mathcal{M}$ using the foregoing procedure.

Assume that the value of the empirical risk local minimum was found to be:

$$R_{emp}^{\Upsilon_{10}}\left((\mathbf{p}_S^{\Upsilon_{10}})_{loc}\right) = 3.2\ (mg/l)^2\ as\ COD$$

The values of $\phi(10)$ and $\varsigma(10)$ can be computed as follows:

$$\varsigma(10) = 4\left(\frac{2\ \ln 10 + 6.4}{10}\right) = 4.40$$

and

$$\phi(10) = 1 + \frac{2.91 \times \varsigma(10)}{2}\left(1 + \sqrt{1 + \frac{4}{2.91 \times \varsigma(10)}}\right) = 14.74$$

Then $\varpi(10)$ is computed in $(mg/l)^2\ as\ COD$:

$$\varpi(10) = 3.2 \times \left(\frac{\phi(10) - 1}{\varsigma(10)} - \frac{2.91}{2\,\phi(10) - 2.91\,\varsigma(10) - 2}\right) \times \left(\frac{\varsigma(10)}{10} - \frac{8}{10^2}\right) = 3.37$$

and

$$\sqrt{\varpi(10)} = 1.83\ mg/l\ as\ COD$$

Assume that the cost Ct_{pt} of obtaining one data point is:

$$Ct = \$120$$

Then the marginal cost of model improvement is:

$$Ct_{mrg} = \frac{120}{1.83} = \$65/mg/l\ as\ COD$$

Case B: In this case, the model is considered to be identified on the basis of $N = 100$ data points and the marginal cost of model improvement can be computed in the same way:

The values of $\phi(10)$ and $\zeta(10)$ can be computed as follows:

$$\zeta(100) = 4\left(\frac{2\ \ln 100 + 6.4}{100}\right) = 0.62$$

and

$$\phi(100) = 1 + \frac{2.91 \times \zeta(100)}{2}\left(1 + \sqrt{1 + \frac{4}{2.91 \times \zeta(100)}}\right) = 3.52$$

Then $\varpi(10)$ is computed in $(mg/l)^2$ *as COD*:

$$\varpi(100) = 3.2 \times \left(\frac{\phi(100) - 1}{\zeta(100)} - \frac{2.91}{2\,\phi(100) - 2.91\,\zeta(100) - 2}\right) \times \left(\frac{\zeta(100)}{100} - \frac{8}{100^2}\right) = 0.05$$

and

$$\sqrt{\varpi(100)} = 0.23\ mg/l\ as\ COD$$

Then the marginal cost of model improvement is:

$$Ct_{mrg} = \frac{120}{0.23} = \$522/mg/l\ as\ COD$$

Thus, to improve the accuracy (or, equivalently, to reduce the uncertainty) of an identified model $S_{(\mathbf{P}_S^{\Upsilon_N})_{loc}}$ by one extra mg/l as COD, we need to spend minimum \$65 (which is relatively cheap) if the model has originally been identified on the basis of $N = 10$ data points. Therefore, if the process engineer is not satisfied with a plant control strategy (PCS) that is based on $S_{(\mathbf{P}_S^{\Upsilon_{10}})_{loc}}$ and that requires a more accurate process model than $S_{(\mathbf{P}_S^{\Upsilon_{10}})_{loc}}$ to be effective, then he may possibly consider ordering a few more data points to re-conduct the process model identification procedure (of course, he would do that only if he establishes that "adding more points" is the most economical option among all other alternatives that are technically viable for improving the PCS). However, if he is not satisfied with the PCS in the second case where he used a hundred data points to identify the process model $S_{(\mathbf{P}_S^{\Upsilon_{100}})_{loc}}$, then he may have to consider changing the whole process model structure, utilize another modelling technology or change completely the PCS design. In other words, if, after using 100 data points for model identification, the model accuracy is still not adequate for developing an effective control strategy, then it is a wise decision to just get rid of $\mathcal{M}$ and think about something else. Indeed, improving the identified process model accuracy at $N = 100$ becomes a very costly alternative: \$522/mg/l as COD.

This simple example shows clearly how powerful are the concept of "process model marginal improvement" and that of "process model improvement marginal cost", as they constitute excellent decision tools in designing cost-effective control strategies for WWT plants. The author believes that they may find several other uses in WWT process modelling and control, as was the case of the concepts of "marginal cost" and "marginal revenue" in economics. However, since WWT process control strategy design is outside the scope of this work, no more examples of the use of these concepts are given in this thesis.

8.4.3 Solving the Practical Case n° 1 of Chapter 1

In Chapter 1 (Introduction), it was pointed out that one of the major applications of the general methodology developed in this thesis is the determination of the prediction accuracy of an identified ASP model and the degree of confidence associated with this accuracy. A scenario for such application was called "*Practical Case n° 1*" and formulated as follows (Chapter 1):

- **Practical Case n° 1**: *Consider the case where the manager of an activated sludge wastewater treatment plant has formulated a request to his consulting engineer, asking him to develop a model for the plant. Also the manager specified that he is able to supply as much data as the consultant may need and that he wants to obtain the most accurate model with the lowest risk possible. Because this model is going to be part of a comprehensive control system using the most recent control technologies and because the manager is aware of the difficulty in modelling the complete dynamics of the process, he asked the consultant to provide him with the* ***prediction accuracy*** *and the* ***degree of confidence*** *associated with this accuracy. The manager needs this information (accuracy and confidence) to compare the performance of the model of his consultant to other empirical models that have been developed for his plant by his own employees.*

Solving this case can be carried out using the following *general* procedure:

Procedure (P1):

1. *Select the state variable that will be concerned with the process model identification. Denote this state variable as* x_{i_0}, *for instance.*

2. *Find out, from the model differential equation that governs the dynamics of* x_{i_0} *(i.e., equation of* $\dot{x}_{i_0}$*), what are the model parameters that affect directly the*

latter state variable. Denote the vector whose components are these parameters as $\mathbf{p}_{x_{i0}}$. *Count the components of this vector and denote the obtained number as* $n_{\mathbf{p}_{x_{i0}}}$.

3. *Set a value to the integration time step* Δt. *Determine the VC dimension* q *of the space of functions defined by:*

$$\begin{pmatrix} x_{i0}(n\,\Delta t) \\ x_{i0}((n-1)\,\Delta t) \\ \mathbf{u}^T((n-1)\,\Delta t) \end{pmatrix} \mapsto (x_{i0}(n\,\Delta t) - sol_{\mathbf{p}_{x_{i0}}}(n\,\Delta t))^2 - \beta$$

and parameterized by the vector $\mathbf{p}_{x_{i0}}$ *and the positive number* β. *The function* $sol_{\mathbf{p}_{x_{i0}}}$ *represents the solution to the model differential equation governing* $\dot{x}_{i0}$ *and the vector* $\mathbf{u}$ *the model input vector. The integer* n *is arbitrary but fixed.*

If it is not possible to determine this VC dimension analytically, then take the number $n_{\mathbf{p}_{x_{i0}}}$ *as an approximation of it:*

$$q \approx n_{\mathbf{p}_{x_{i0}}}$$

4. *Using the identification data set* Υ_N, *containing* N *data points, and any computer optimization algorithm* $\mathcal{A}$, *minimize the value of the objective function:*

$$J_{x_{i0}}(\mathbf{p}_{x_{i0}}) = \sum_{j=1}^{N} |sol_{\mathbf{p}_{x_{i0}}}(t_j) - x_{i0}^{data}(t_j)|^2$$

where t_j *represents the time instants at which the data* x_{i0}^{data} *have been collected. Since computer algorithms do not guarantee the exact determination of the global minimum, the obtained minimum is to be considered as a local one. Denote it* $J_{x_{i0}}\left((\mathbf{p}_{x_{i0}}^{\Upsilon_N})_{loc}\right)$, *where* $(\mathbf{p}_{x_{i0}}^{\Upsilon_N})_{loc}$ *is a parameter vector at which the objective function local minimum is attained.*

5. *Compute the average value of the objective function by dividing it by* N, *and denote the obtained number as* $R_{emp}^{\Upsilon_N}\left((\mathbf{p}_{x_{i0}}^{\Upsilon_N})_{loc}\right)$ *(empirical risk):*

$$R_{emp}^{\Upsilon_N}\left((\mathbf{p}_{x_{i0}}^{\Upsilon_N})_{loc}\right) = \frac{J_{x_{i0}}\left((\mathbf{p}_{x_{i0}}^{\Upsilon_N})_{loc}\right)}{N}$$

6. *Select, from the interval* $]0, 1[$, *a value for the degree of confidence,* $1-\eta$, *with which the model prediction accuracy is to be determined. Usually,* $1-\eta$ *is set to 0.9 or 0.95 (i.e.,* $\eta = 10\%$ *or 5%), but it can take any value between 0 and 1 (0 and 1 are excluded).*

7. *Compute the value of* $\zeta(N)$ *using the equation:*

$$\zeta = 4\frac{\left[q\left(\ln\left(\frac{2N}{q}\right)+1\right)-\ln\left(\frac{\eta}{4}\right)\right]}{N}$$

8. *Compute the average of the prior information* κ *as:*

$$\bar{\kappa} = \frac{1.82+4}{2} = 2.91 \tag{8.72}$$

9. *Compute the guaranteed deviation* φ_1 *using the equation:*

$$\varphi_1 = R_{emp}^{\Upsilon_N}\left((\mathrm{p}_{x_{i0}}^{\Upsilon_N})_{loc}\right)\left(1+\frac{\bar{\kappa}\zeta(N)}{2}\left(1+\sqrt{1+\frac{4}{\bar{\kappa}\zeta(N)}}\right)\right)$$

10. *Conclude that the exact value of the weighted root mean square deviation*

$$\mathcal{D}(sol_{(\mathrm{p}_{x_{i0}}^{\Upsilon_N})_{loc}}, g^{\mathcal{T}_{asp}})$$

between the identified process model prediction function $sol_{(\mathrm{p}_{x_{i0}}^{\Upsilon_N})_{loc}}$ *and the real process response function* $g^{\mathcal{T}_{asp}}$ *is guaranteed to be less than* $\sqrt{\varphi_1(N)}$, *with a confidence of* $1-\eta$:

$$\mathcal{D}(sol_{(\mathrm{p}_{x_{i0}}^{\Upsilon_N})_{loc}}, g^{\mathcal{T}_{asp}}) \leq \sqrt{\varphi_1(N)} \quad \text{with a confidence of: } 1-\eta \tag{8.73}$$

In concrete and simplified terms, inequality 8.73 means that, with a confidence of $1-\eta$, *the deviation between the model prediction* $sol_{(\mathrm{p}_{x_{i0}}^{\Upsilon_N})_{loc}}(t)$ *and the actual process response* $g^{\mathcal{T}_{asp}}(t)$ *for the state variable* x_{i0} *at any time instant* t, *is, in average, less than the number* $\sqrt{\varphi_1(N)}$.

Chapter 9

Modelling the Activated Sludge Process: a New Approach

9.1 Motivation and Purpose

As pointed out earlier in this thesis, the ultimate goal of the ASP modelling task is to determine the real process response function $g^{\mathcal{T}_{asp}}$. With respect to the substrate concentration, this response function has been defined by equation 7.23 which is reported below:

$$g^{\mathcal{T}_{asp}}(t_i) = \mathbf{E}(\ S^{data}(t_i) \mid S^{data}(t_{i-1}), u_{Q(t_{i-1})}, u_{S(t_{i-1})}\ ) \tag{9.1}$$

The general procedure to determine $g^{\mathcal{T}_{asp}}$ consists of two steps:

- *Process model development:* first we develop a process model $\mathcal{M}$. Based on this model, a family $\mathcal{H}^{\mathcal{M}}$ of time-functions, $S_{\mathbf{p}_S}$, parameterized by the model free parameter vector $\mathbf{p}_S$ is generated:

$$\mathcal{H}^{\mathcal{M}} = \{S_{\mathbf{p}_S} : t \mapsto S_{\mathbf{p}_S}(t) \mid \mathbf{p}_S \in \Gamma\} \tag{9.2}$$

  The set Γ denotes the model parameter space which is a subset of the space $\Re^n$, where $n = n_{\mathbf{p}_S}$ is the number of parameters in the vector $\mathbf{p}_S$.

- *Process model identification:* then, within this family $\mathcal{H}^{\mathcal{M}}$, we search for the closest possible function to $g^{\mathcal{T}_{asp}}$. Let us denote this closest function as $(S_{\mathbf{p}_S})_{cl}$. The search for $(S_{\mathbf{p}_S})_{cl}$ is generally carried out on the basis of two items:

Table 9.1: Process model identification can lead to three different functions

	Infinite Computing Power and Infinite N	**Infinite Computing Power and Finite N**	**Finite Computing Power and Finite N**
$(S_{\mathrm{P}_S})_{cl}$	$S_{\mathrm{P}_S^0}$	$S_{\mathrm{P}_S^{\Upsilon_N}}$	$S_{(\mathrm{P}_S^{\Upsilon_N})_{loc}}$

1. a set of empirical data points Υ_N of size N obtained from the real process operation.
2. a computer algorithm $\mathcal{A}$ that implements the search procedure using the foregoing data set Υ_N.

In the previous Chapter, it was pointed out that, depending on the size N of Υ_N and the computing power that is used to implement algorithm $\mathcal{A}$, the search for $(S_{\mathrm{P}_S})_{cl}$ can possibly lead to three different functions: $S_{\mathrm{P}_S^0}$, $S_{\mathrm{P}_S^{\Upsilon_N}}$ and $S_{(\mathrm{P}_S^{\Upsilon_N})_{loc}}$. Table 9.1 explains the conditions under which each one of these functions is obtained.

In particular, it has been shown in section 8.2 that, even under ideal conditions of infinite data set and computing power, the function $(S_{\mathrm{P}_S})_{cl}$ does not match the real response function $g^{\mathcal{T}_{asp}}$, when the process model is fixed. There is indeed a nonzero deviation:

$$gap = \mathcal{D}(S_{\mathrm{P}_S^0}, g^{\mathcal{T}_{asp}}) = \mathcal{D}(\mathcal{H}^{\mathcal{M}}, g^{\mathcal{T}_{asp}}) \tag{9.3}$$

between the best possible approximation $S_{\mathrm{P}_S^0}$ and $g^{\mathcal{T}_{asp}}$. This deviation is larger when the selected model contains a low amount of information about the real process biochemical mechanisms. For instance, models such as the straight line or a simple neural network could very well be used to approximate the ASP behaviour by making them fit the identification data. However, these models would generate very poor predictions even under ideal conditions of an infinite data set and computing power, because the *gap* defined by equation 9.3 would be huge.

It is true that the function $S_{\mathrm{P}_S^0}$ has no more than a theoretical significance. It is indeed totally inaccessible in practice, and the function that will be used in reality for process control is actually $S_{(\mathrm{P}_S^{\Upsilon_N})_{loc}}$. Because of that, we may wonder why we are considering in this discussion the performance of a function such as $S_{\mathrm{P}_S^0}$ that has no clear practical usefulness, and not concentrating on the one on the basis of which process control strategies will be developed. The reason is simple: *if a model cannot perform well at the theoretical level, i.e., under ideal conditions, then it would have no chance to do better under less favourable conditions which always prevail in reality.* Therefore, we need first to make sure that the selected process model $\mathcal{M}$ has the potential of minimizing the residual gap defined by equation 9.3 down to its lowest possible value, which is 0. That is:

$$\mathcal{D}(\mathcal{H}^{\mathcal{M}}, g^{\mathcal{T}_{asp}}) = 0 \tag{9.4}$$

The property expressed by the latter equation will guarantee us that every effort made to improve the model accuracy — by adding more data points or increasing the computing power — will indeed contribute to achieve better approximation of $g^{\mathcal{T}_{asp}}$, and *not* $S_{\mathbf{P}^0_S}$.

In this Chapter we concentrate on constructing process models that have the potential of minimizing the gap expressed by equation 9.3 and, therefore, satisfy property 9.4. The idea to be utilized here consists in making the family $\mathcal{H}^{\mathcal{M}}$ large enough so that it may possibly include $g^{\mathcal{T}_{asp}}$ or, at least, minimize the distance between $\mathcal{H}^{\mathcal{M}}$ and the latter function (note indeed that if $g^{\mathcal{T}_{asp}} \in \mathcal{H}^{\mathcal{M}}$, then automatically $gap = \mathcal{D}(\mathcal{H}^{\mathcal{M}}, g^{\mathcal{T}_{asp}}) = 0$). To implement this idea, we can make use of two different types of models:

1. *Use large black-box models*: As was discussed in previous Chapters, black-box models such as neural networks and polynomial models are universal approximators. Because of this, they have indeed the potential of minimizing the gap $\mathcal{D}(\mathcal{H}^{\mathcal{M}}, g^{\mathcal{T}_{asp}})$ down to 0.

2. *Use detailed mechanistic models that describe ASP mechanisms as they take place in reality*: These models are not universal approximators, but they are considered to have the potential of minimizing the gap $\mathcal{D}(\mathcal{H}^{\mathcal{M}}, g^{\mathcal{T}_{asp}})$ because they contain a great deal of information about the process behaviour. They are developed specifically for the ASP and are supposed to achieve good approximation of one specific function: $g^{\mathcal{T}_{asp}}$.

In this study, it is proposed to make use of the second type of models, i.e., detailed mechanistic models. A priori, such models would lead to a low value of the gap

$\mathcal{D}(\mathcal{H}^{\mathcal{M}}, g^{\mathcal{T}_{asp}})$ without making the family $\mathcal{H}^{\mathcal{M}}$ unnecessarily too large (which is of course not the case of black-box models). As a result, detailed mechanistic models would require much less data points and computing power for identification than their black-box counterparts. But, what has prevented researchers from developing them systematically is a problem that is still misunderstood within the wastewater community. Up until now, wastewater researchers have attributed their concern about developing detailed mechanistic models to the problem of *"lack of parameter identifiability of complex models"*, while the true problem is in fact what we called earlier in this thesis the *data overfitting phenomenon*: *"a too complex model can fit almost anything"*. However, as will be explained in the next Chapter, this issue of data overfitting can be totally resolved, thanks to the general methodology of uncertainty management developed in previous Chapters. As a result, there is no more obstacles that stand in front of developing detailed mechanistic models of the ASP and, because of this, we have taken the liberty in this study to consider complex ASP mechanistic models.

Note however that the current Chapter is not about developing and adding again another model to the already-large pool of existing mechanistic models of the ASP (see Jeppsson, 1996, for a literature review of existing mechanistic models). Rather, the objective here is to devise a new approach of dealing with the *almost-infinite* degree of complexity of the ASP behaviour, through mechanistic modelling procedures. The idea is to construct an *infinite series $\mathcal{NS}$ of nested ASP mechanistic models* of increasing complexity, with the following goal in mind: as the complexity of the process model is allowed to increase within this nested series $\mathcal{NS}$ of models, the process mechanisms are described better and, as a result, the residual gap of equation 9.3 converges to zero, letting then the ideal process model $S_{\mathrm{p}^0_S}$ approach more and more the process response function $g^{\mathcal{T}_{asp}}$. Formally, this means that the series $\mathcal{NS}$ of nested mechanistic models has to be such that:

$$gap \longrightarrow 0, \quad \text{as the complexity of } \mathcal{M} \text{ increases within the series } \mathcal{NS} \tag{9.5}$$

or, equivalently:

$$S_{\mathrm{p}^0_S} \longrightarrow g^{\mathcal{T}_{asp}}, \quad \text{as the complexity of } \mathcal{M} \text{ increases within the series } \mathcal{NS} \tag{9.6}$$

Now, since the ideal process model $S_{\mathrm{P}_S^0}$ corresponds to an infinite number of identification data points ($N = \infty$) and infinite computing power, statement 9.6 becomes:

$$(S_{\mathrm{P}S})_{cl} \rightarrow g^{\mathcal{T}_{asp}}, \quad \text{as} \begin{cases} \bullet \text{ the complexity of } \mathcal{M} \text{ increases within the series } \mathcal{NS} \\ \bullet\ N \rightarrow \infty \\ \bullet \text{ computing power increases} \end{cases} \tag{9.7}$$

The function $(S_{\mathrm{P}S})_{cl}$ was defined at the beginning of this Chapter as the closest possible function of the family $\mathcal{H}^{\mathcal{M}}$ to $g^{\mathcal{T}_{asp}}$. In real-world situations, it is actually equal to $S_{(\mathrm{P}_S^{\Upsilon_N})_{loc}}$ and, as a result, statement 9.7 can re-written as:

$$S_{(\mathrm{P}_S^{\Upsilon_N})_{loc}} \rightarrow g^{\mathcal{T}_{asp}}, \quad \text{as} \begin{cases} \bullet \text{ the complexity of } \mathcal{M} \text{ increases within the series } \mathcal{NS} \\ \bullet\ N \rightarrow \infty \\ \bullet \text{ computing power increases} \end{cases} \tag{9.8}$$

Statement 9.8 is of significant importance at both the theoretical and practical levels:

- *At the theoretical level*, it demonstrates that it is possible to determine the unknown process response function $g^{\mathcal{T}_{asp}}$. Also, it defines exactly what it takes to guarantee a good approximation of this function $g^{\mathcal{T}_{asp}}$: (1) a series of process models within which the complexity of $\mathcal{M}$ can be increased, (2) a large number of data points and (3) a high computing power.

- *At the practical level*, it suggests what should be done to manage the uncertainty in the activated sludge process. It shows indeed that the uncertainty management variables are the model complexity, the number N and the computing power. All what remains to do practically is then to develop an uncertainty management strategy using these variables, which will be discussed in the next Chapter.

To conceive the idea of developing a nested series of ASP mechanistic models, the author was inspired, to a great extent, by neural networks and what has made them so successful in some areas of systems science and engineering. As pointed out earlier in this thesis, neural networks constitute indeed a nested series of black-box models that have the property of universal approximators. This property means that neural network structure complexity can always be adjusted to uniformly approximate practically any continuous function to any desired degree of accuracy. Now, given this powerful property, what remains to be solved before neural networks can be adopted as process models is how to carry out neural network structure adjustment in order to match the dynamics of the studied process. The only procedure that exists up until now to resolve this adjustment problem is what we called neural network training, calibration or identification. When neural network training is taken into account, the property of "universal approximators" becomes equivalent, in some cases, to the fact that the condition 9.8 is satisfied for practically any continuous function — including $g^{\mathcal{T}_{asp}}$. Therefore, what we want to do in the current Chapter is to construct a series $\mathcal{NS}$ of nested mechanistic models that plays for the set $\mathcal{F}_{ASP}$ of all ASP response functions the role of neural networks for the set $\mathcal{F}_{CF}$ of practically all continuous functions:

$$\begin{aligned} \textit{Neural networks} &= \textit{Universal approximators for the functions of } \mathcal{F}_{CF} \\ \mathcal{NS} &= \textit{Universal approximators for the functions of } \mathcal{F}_{ASP} \end{aligned}$$

The nested series $\mathcal{NS}$ of ASP mechanistic models is developed in next section.

9.2 Developing a Nested Series of Mechanistic Models

9.2.1 Fuzziness and the Bi-Substrate Hypothesis

Wastewater treatment using the activated sludge process involves two main components:

- *microorganisms*, i.e., those living species that biodegrade the organic waste and derive energy and food from it.
- *organic waste*, i.e., those contaminants that have to be eliminated from the wastewater before its disposal into the natural environment. Organic waste is called substrate.

Note that, in the real activated sludge bio-reactor, the component "microorganisms" includes a great variety of biological species. Each one of these species has a different behaviour and different way of interacting with the substrate. Similarly, the component "organic waste", which is generated by human activities, can include almost any type of organic material.

Most of the first ASP models developed in the early seventies used two state variables to describe the activated sludge process: microorganisms concentration and substrate concentration. In 1980, Dold *et al.* introduced the bi-substrate hypothesis (BS hypothesis) to explain a precipitous drop in oxygen consumption rate that takes place at the instant of feed termination (see Dold *et al.*, 1980, for details). The BS hypothesis consists in dividing the organic waste to be biodegraded into two fractions:

- the readily biodegradable fraction (RB) which is utilized by microorganisms at a very rapid rate.
- the slowly biodegradable fraction (SB) which requires storage and enzymatic breakdown (i.e. *hydrolysis*) prior to transfer through the cell wall and utilization by microorganisms.

It is the author's opinion that the discovery of the foregoing BS hypothesis constitutes the major achievement in the area of ASP mechanistic modelling during the last two decades. Henze *et al.* (1987) pointed out in their report on Activated Sludge Model n° 1 that "*the degradation of slowly biodegradable organic matter is very important to realistic modelling of activated sludge systems because it is primarily responsible for the attainment of realistic space-time and real time dependent electron acceptor profiles*". Yet, the BS hypothesis is just a simplification of the real biochemical behaviour of the process because of the reasons explained below.

The BS hypothesis is based on the traditional scientific thought that requires that the logical value of a proposition is either 1 ("*true*") or 0 ("*false*"). In the case of the BS hypothesis, this traditional thought implies that a chemical compound in the organic waste is either "readily biodegradable" or "slowly biodegradable"; there is nothing in between. This way of thinking uses crisp logic which is based on the classical mathematics of set theory: given an object x and a set S, we have either $x \in S$ or $x \notin S$. If we define a membership function f_S for the set S, then this function would have two values: $f_S(x) = 1$ if $x \in S$, and $f_S(x) = 0$ if $x \notin S$, i.e., f_S is a binary function.

However, the world is not crisp. It is fuzzy in the sense that the logical value of a proposition is not necessarily binary, but it can take other values between 0 and 1 in order to express the degree of truthfulness of that proposition. The concepts of *fuzzy logic* and *membership functions of fuzzy sets* were first introduced by Zadeh (1965) to characterize the uncertainty that underlies many complex systems. Fuzzy logic is now being used extensively and successfully in modelling and control of various engineering systems. In the case of the ASP, fuzzy logic technology can also be implemented to resolve the "lexical" uncertainty — *slow* versus *fast* biodegradation — that underlies the bi-substrate hypothesis in modelling the activated sludge process. This would amount to introducing two membership functions f_{slow} and f_{fast} that characterize the ambiguous meaning of "slow" and "fast" biodegradability of a chemical compound in the organic waste. By doing so, we can imagine an improvement of ASP mechanistic modelling procedures by combining them with the fuzzy logic modelling technology. Nevertheless, in this study, fuzzy logic concepts will not be utilized. Rather, we will continue expanding the mechanistic modelling methodology, by introducing a more realistic hypothesis with respect to substrate biodegradability. This hypothesis will be designated as *the multi-substrate hypothesis* and will be defined as follows:

The organic waste is divided into a series of substrate groups:

$$\mathcal{SG}_1, \mathcal{SG}_2, \ldots, \mathcal{SG}_l, \ldots \tag{9.9}$$

Each group $\mathcal{SG}_l$ of this series is characterized by a fixed degree of substrate biodegradability and contains all chemical compounds that biodegrade according to this fixed degree. As the index l of this series increases, the biodegradability degree of the compounds in the group $\mathcal{SG}_l$ decreases in such a way that $\mathcal{SG}_1$ contains extremely rapidly biodegradable compounds and $\mathcal{SG}_\infty$ contains inert organic compounds, i.e.:

$\mathcal{SG}_1$ *contains extremely rapidly biodegradable compounds*
$\mathcal{SG}_2$ *contains very rapidly biodegradable compounds*
$\mathcal{SG}_3$ *contains rapidly biodegradable compounds*
$\vdots$
$\mathcal{SG}_i$
$\vdots$
$\mathcal{SG}_{k-1}$ *contains slowly biodegradable compounds*
$\mathcal{SG}_k$ *contains very slowly biodegradable compounds*
$\mathcal{SG}_{k+1}$ *contains extremely slowly biodegradable compounds*

⋮

$\mathcal{SG}_{\infty}$ *contains inert organic compounds*

How do researchers view this multi-substrate hypothesis? They support it. Here are indeed the opinions of two different researchers with regard to this hypothesis :

- Referring to the bi-substrate hypothesis, Dold and Marais (1986) pointed out that "*the two biodegradable COD fractions* [the SB and RB fractions] *are utilized for synthesis of new cell mass. The rates of utilization of the two, however, differ greatly, by approximately an order of magnitude. Even though within each group there may be a range of rates, the differences within each group are relatively small compared with the differences between the two groups*". In this excerpt, Dold and Marais (1986) acknowledged that the activated sludge process behaves in reality according to the multi-substrate hypothesis. At the same time, they conjectured that the bi-substrate partition has a primary importance in describing the process behaviour, while the multi-substrate one is secondary. Based on this presumption, a model that takes into account the bi-substrate partition *only* was developed — it is the IAWPRC model n° 1 — and, as a result, the wastewater community ended up with *one fixed model* of the highly complex activated sludge process. If, for any reason, ASP engineers want a model prediction accuracy that takes into account more than just the primary phenomena, they would not be able to obtain it, because the IAWPRC model n° 1 was not developed for that purpose. This situation resembles the one where a group of university students, for example, are given one single approximate equation:

$$\sin x = x \tag{9.10}$$

for the function sine and are asked to use it in carrying out all their physics projects. If, for any reason, the students need higher accuracy in estimating the sine function, they would get stuck with equation 9.10. They would not know how to improve this equation and benefit from any additional computing power they can obtain from university to carry out their projects. Equation 9.10 is indeed "inflexible" and does not show how *secondary terms* can be used to improve the estimation of the sine function. But, on the other hand, if the students were given the complete equation, i.e.

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots + (-1)^k \frac{x^{2k+1}}{(2k+1)!} + \cdots \tag{9.11}$$

they would be in a much better position to carry out their projects in an optimal fashion: given the computing power they have, the nature of project they are dealing with and the degree of accuracy required by that project, the students will be able to decide what secondary terms to keep in equation 9.11 and what to drop.

This situation is of course imaginary, but it reflects exactly what is needed in the case of the activated sludge process. Indeed, the ultimate goal of process modelling is to develop a cost-effective control strategy for the process. The effectiveness of this strategy depends to a great extent on the accuracy of the process model. Because of this, one of ASP engineers' main concern is to possess the most accurate model of the ASP. But since the ASP perfect model does not exist, it is crucial for engineers to have a handy methodology to improve the prediction accuracy of a given process model, as more computing power and real data become available. This would not be possible with the IAWPRC model n° 1. Refer indeed to the previous Chapter where it was shown that, even under ideal conditions of infinite data and computing power, a fixed model does not lead to a perfect estimation of the real process response function. Consequently, there is a need to develop a more general process model that shows how primary *as well as* secondary phenomena affect the process behaviour.

- Jeppsson (1996): "*The organic matter that is received by a WWT plant includes all kinds of different molecular structures. Different organisms deal with different substrates in different time scales, which makes it probable that an entire set of biodegradation processes with time constants ranging* [in almost a continuous fashion] *from fast to slow biodegradability is at work here*". Thus, Jeppsson also acknowledges the multi-substrate hypothesis. But, to deal with it, he chose to simplify the modelling procedure: "*Since there is no apparent upper limit to the number of substrates which would really need to be included, the opposite solution is suggested ..., that is, we model only one type of organic biodegradable substrate*" (Jeppsson, 1996). Here, the author would like to put forth a famous saying by Albert Einstein: "*Everything should be made as simple as possible — but not simpler*". Because of the high degree of complexity of the activated sludge process, the author believes that Jeppsson's suggestion is "simpler", and the one implemented in this work is just "as simple as possible".

The next section shows how the multi-substrate hypothesis can be put into practice in order to develop a nested series of ASP mechanistic models.

9.2.2 Equations of the Nested Series of Models

Using the idea of Jeppsson (1996), define the following process state variable X_{COD} by:

$$X_{COD} = S_I + S_S + (X_I + X_P) + X_S \tag{9.12}$$

where the terms of the right-hand side are as defined in Henze *et al.* (1987):

- S_I: soluble inert organic matter;
- S_S: readily biodegradable substrate;
- X_I and X_P: particulate inert organic matter;
- X_S: slowly biodegradable substrate;

Note that the variable X_{COD} measures the total pollution strength of the wastewater.

Designate the concentration of organic matter (mg/l as COD, for example) in the substrate group $\mathcal{SG}_l$ as X_{Sl}. Then X_{COD} can also be expressed as:

$$X_{COD} = X_{S1} + X_{S2} + X_{S3} + \cdots + X_{Sl} + \cdots \tag{9.13}$$

Now, the question that arises here is: how do microorganisms deal with this series of substrates X_{S1}, X_{S2}, X_{S3}, $\cdots$, X_{Sl}, $\cdots$?

Do microorganisms act upon
these substrate groups
independently and synthesize (***mechanism 1***)
cells directly from them,

OR

do they first hydrolyze the constituents
of substrate groups $\mathcal{SG}_l$, $l \geq 2$, sequentially,
release the hydrolyzed matter
to the bulk liquid as soluble **(mechanism 2)**
compounds belonging to the first group $\mathcal{SG}_1$,
and then synthesize cells from
the constituents of the latter group only?

This is still an open question, as "*the exact mechanism cannot yet be stated explicitly*" (Dold *et al.*, 1980). However, Dold *et al.* (1980) pointed out that "*from extensive simulation study, it appears that the two substrates* [SB and RB fractions] *are acted on independently by the same active mass*". Consequently, in what follows, only mechanism 1 will be considered. Based on this mechanism, two nested series of mechanistic models will be developed using the Monod equation for the first series, and the Tiessier equation for the second one.

Using the Monod Equation

Heterotrophic Biodegradation

As indicated previously, in the case of the bi-substrate hypothesis, Dold *et al.* (1980) reported that, "*from extensive simulation study, it appears that the two substrates* [SB and RB fractions] *are acted on independently by the same active mass and that the sludge synthesized is the sum of these metabolic reactions*". In what follows, we will adopt a similar mechanism for the case of the multi-substrate hypothesis, i.e.,

Microorganisms act upon substrate groups $\mathcal{SG}_l$, $1 \leq l \leq \infty$,
independently and synthesize cells directly from them, (9.14)
and
the sludge synthesized is the sum of all these metabolic reactions.

Let us express the above statement in a formal way. Denote the concentration of active heterotrophic biomass as $X_{B,H}$. This biomass acts on each substrate group

$\mathcal{SG}_l$, $1 \leq l \leq \infty$, to produce more biomass $(X_{B,H})_l$:

$$X_{Sl} \longrightarrow (X_{B,H})_l \tag{9.15}$$

The first part of statement 9.14 is equivalent to the fact that the biochemical transformation 9.15 involves synthesis only if $l = 1$, and hydrolysis and synthesis if $l \geq 2$. Designate the rate of this transformation as $r_{X_{Sl}}$ and note that, by definition, the following equality holds true:

$$r_{X_{Sl}} = \frac{dX_{Sl}}{dt} = -\frac{1}{Y_H}\frac{d(X_{B,H})_l}{dt} = -\frac{1}{Y_H} r_{(X_{B,H})_l} \tag{9.16}$$

where t is the time, Y_H is the heterotrophic yield and $r_{(X_{B,H}),l}$ is the growth rate of the heterotrophic biomass on the constituents of substrate group $\mathcal{SG}_l$. Then, the second part of statement 9.14 says that we can write that the total growth rate $r_{X_{B,H}}$ of $X_{B,H}$ in the bioreactor as equal to:

$$r_{X_{B,H}} = \frac{dX_{B,H}}{dt} = r_{(X_{B,H})_1} + r_{(X_{B,H})_2} + r_{(X_{B,H})_3} + \cdots + r_{(X_{B,H})_l} + \cdots \tag{9.17}$$

Now, we need to find an explicit expression of the total growth rate $r_{X_{B,H}}$ given by equation 9.17. To do so, we will develop an expression for each partial growth rate $r_{(X_{B,H})_l}$.

The Monod equation was originally developed for soluble nutrients and, as a result, it is suitable to describe the biodegradation rate of the constituents of group $\mathcal{SG}_1$ which is composed of *only* soluble matter. Hence, the expression of the growth rate $r_{(X_{B,H})_1}$ is:

$$r_{(X_{B,H})_1} = \frac{\hat{\mu}_{H_1} X_{S1}}{K'_{S1} + X_{S1}} X_{B,H} \tag{9.18}$$

where $\hat{\mu}_H$ and K'_S are two parameters, usually known as the maximum specific growth rate and half saturation constant, respectively. As for the other substrate groups $\mathcal{SG}_l$, $l \geq 2$, the biochemical transformation 9.15 involves hydrolysis and synthesis. To describe the rate of it, we will use here the expression derived by Dold *et al.* (1980) for the case of the bi-substrate hypothesis, i.e.,

$$r_{(X_{B,H})_l} = \frac{\hat{\mu}_{H_l} X_{Sl}/X_{B,H}}{K'_{Sl} + X_{Sl}/X_{B,H}} X_{B,H} = \frac{\hat{\mu}_{H_l} X_{Sl}}{K'_{Sl} X_{B,H} + X_{Sl}} X_{B,H} \tag{9.19}$$

where $\hat{\mu}_{H_l}$ and K'_{Sl} are two parameters. Using equations 9.17, 9.18 and 9.19, the expression of the total growth rate $r_{X_{B,H}}$ of $X_{B,H}$ can be derived as follows:

$$r_{X_{B,H}} = \frac{\hat{\mu}_{H_1} X_{S1}}{K'_{S1} + X_{S1}} X_{B,H} + \sum_{l=2}^{\infty} \frac{\hat{\mu}_{H_l} X_{Sl}}{K'_{Sl} X_{B,H} + X_{Sl}} X_{B,H} \tag{9.20}$$

This rate model is detailed and take into account both primary and secondary biochemical phenomena that affect ASP behaviour, but the problem with it is that it involves a quasi-infinite number of state variables $X_{S1}, X_{S2}, X_{S3}, \ldots$, which poses at least two difficulties at the practical level:

1. *measurability and observability*: how to measure and what to measure?
2. *curse of dimensionality*: the computing power required for large state spaces is enormous.

So what to do? We suggest to make use of the *concepts of information content, transfer and conversion* introduced earlier in this thesis. Here is the idea: the following series of process state variables:

$$X_{S1}, X_{S2}, X_{S3}, \cdots, X_{Sl}, \cdots \tag{9.21}$$

contains a certain amount of information about the process. If we can transform this series into another one that contains the same amount of information and whose terms can be incorporated into the rate model 9.20 as *model parameters*, then the foregoing problems 1. and 2. with the quasi-infinite process state space would be resolved. Why? Because we now possess a powerful mathematical framework to deal with model parameter identification, model structure complexity selection and process uncertainty. Here is how the transformation of the series 9.21 can be carried out:

Define the fractions

$$\Omega_l = \frac{X_{Sl}}{X_{COD}} \tag{9.22}$$

of X_{Sl} in X_{COD}, and replace X_{Sl} by $\Omega_l X_{COD}$ in equation 9.20. We obtain the following equation:

$$r_{X_{B,H}} = \frac{\hat{\mu}_{H_1}\Omega_1 X_{COD}}{K'_{S1} + \Omega_1 X_{COD}} X_{B,H} + \sum_{l=2}^{\infty} \frac{\hat{\mu}_{H_l}\Omega_l X_{COD}}{K'_{Sl} X_{B,H} + \Omega_l X_{COD}} X_{B,H} \tag{9.23}$$

Now divide the numerator and denominator of each term by Ω_l:

$$r_{X_{B,H}} = \frac{\hat{\mu}_{H_1} X_{COD}}{(K'_{S1}/\Omega_1) + X_{COD}} X_{B,H} + \sum_{l=2}^{\infty} \frac{\hat{\mu}_{H_l} X_{COD}}{(K'_{Sl}/\Omega_l) X_{B,H} + X_{COD}} X_{B,H} \tag{9.24}$$

then denote the ratios K'_{Sl}/Ω_l as K_{Sl}, and replace them in previous equation:

$$\boxed{r_{X_{B,H}} = \frac{\hat{\mu}_{H_1} X_{COD}}{K_{S1} + X_{COD}} X_{B,H} + \sum_{l=2}^{\infty} \frac{\hat{\mu}_{H_l} X_{COD}}{K_{Sl} X_{B,H} + X_{COD}} X_{B,H}} \tag{9.25}$$

Nitrification

Denote the concentration of active autotrophic biomass as $X_{B,A}$, the concentration of nitrogen in $NH_4^+ + NH_3$ as S_{NH} and the concentration of nitrogen in $NO_3^- + NO_2^-$ as S_{NO}. The biomass $X_{B,A}$ is responsible for the transformation:

$$S_{NH} \longrightarrow S_{NO} \tag{9.26}$$

The rate of this transformation is proportional to the growth rate $r_{X_{B,A}}$ which will be expressed according to the Monod Equation:

$$r_{X_{B,A}} = \frac{\hat{\mu}_A S_{NH}}{K_{NH} + S_{NH}} X_{B,A} \tag{9.27}$$

The ammonia/ammonium is generated in the ASP bioreactor according to the transformation:

$$X'_{ND} \longrightarrow S_{NH} \tag{9.28}$$

where X'_{ND} designates the total concentration of entrapped organic nitrogen. This conversion will be called here *nitrogen mineralization*. It occurs in parallel to the transformations 9.15 and, because of this, we will divide X'_{ND} into a series of subgroups $(X'_{ND})_1$, $(X'_{ND})_2$, $(X'_{ND})_3$, ..., $(X'_{ND})_l$, ... — similarly to what we did for the carbonaceous matter. Each $(X'_{ND})_l$ represents the concentration of the organic nitrogen entrapped in the constituents of the substrate group $\mathcal{SG}_l$. The transformation 9.28 can then be viewed as a lumped form of the following conversions:

$$(X'_{ND})_l \longrightarrow (S_{NH})_l \tag{9.29}$$

Since nitrogen in $(X'_{ND})_1$ is all in soluble form, the transformation 9.29 involves nitrogen ammonification only for $l = 1$; its rate $r_{(S_{NH})_1}$ will be expressed by (see Dold and Marais, 1986):

$$r_{(S_{NH})_1} = k'_a \, (X'_{ND})_1 \, X_{B,A} \tag{9.30}$$

where k'_a is a model parameter. As for $l \geq 2$, the biochemical transformation involves both hydrolysis and ammonification. It is associated with transformation 9.15 and, because of this, its rate $r_{S_{NH}}$ is considered to be described by the equation (see Henze *et al.*, 1987):

$$r_{(S_{NH})_l} = r_{(X_{B,H})_l} \left(\frac{(X'_{ND})_l}{X_{Sl}} \right) = \frac{\hat{\mu}_{H_l} \, (X'_{ND})_l}{K'_{Sl} X_{B,H} + X_{Sl}} X_{B,H} \tag{9.31}$$

Similarly to what we did for the carbonaceous matter, define the fractions

$$\Omega'_l = \frac{(X'_{ND})_l}{X'_{ND}} \tag{9.32}$$

of $(X'_{ND})_l$ in X'_{ND}. Note here that for both series $(\Omega_l)_{l\geq 1}$ and $(\Omega'_l)_{l\geq 1}$, we have:

$$\Omega_1 + \Omega_2 + \Omega_3 + \cdots + \Omega_l + \cdots = 1 \tag{9.33}$$

and

$$\Omega'_1 + \Omega'_2 + \Omega'_3 + \cdots + \Omega'_l + \cdots = 1 \tag{9.34}$$

Equations 9.30 and 9.31 can then be re-written as:

$$r_{(S_{NH})_1} = (k'_a\, \Omega'_1)\, X'_{ND}\, X_{B,H} \tag{9.35}$$

and for $l \geq 2$

$$r_{(S_{NH})_l} = \frac{\hat{\mu}_{H_l}\, \Omega'_l X'_{ND}}{K'_{Sl} X_{B,H} + \Omega_l X_{COD}} X_{B,H} = \frac{(\hat{\mu}_{H_l}\, (\Omega'_l/\Omega_l)) X'_{ND}}{(K'_{Sl}/\Omega_l) X_{B,H} + X_{COD}} X_{B,H} \tag{9.36}$$

Denote the products $k'_a\, \Omega'_1$ and $\hat{\mu}_{H_l}\, (\Omega'_l/\Omega_l)$ as k_a and k_l, respectively. Then previous equations can be simplified as follows:

$$r_{(S_{NH})_1} = k_a\, X'_{ND}\, X_{B,H} \tag{9.37}$$

and for $l \geq 2$

$$r_{(S_{NH})_l} = \frac{k_l X'_{ND}}{K_{Sl} X_{B,H} + X_{COD}} X_{B,H} \tag{9.38}$$

Now, using the rule that the total production rate of nitrogen in the form of ammonia/ammonium is equal to the sum of partial production rates of the transformations 9.29, we obtain the rate $r_{S_{NH}}$ of the conversion 9.28:

$$\boxed{r_{S_{NH}} = k_a\, X'_{ND}\, X_{B,H} + \sum_{l=2}^{\infty} \frac{k_l X'_{ND}}{K_{Sl} X_{B,H} + X_{COD}} X_{B,H}} \tag{9.39}$$

k_a, k_l and K_{Sl} are model parameters.

Biomass Decay

The rates $(r_{X_{B,H}})_{dec}$ and $(r_{X_{B,A}})_{dec}$ of heterotrophic and autotrophic biomass decay are described by first order equations (Henze *et al.*, 1987):

$$(r_{X_{B,H}})_{dec} = b_H\, X_{B,H} \tag{9.40}$$

$$(r_{X_{B,A}})_{dec} = b_A\, X_{B,A} \tag{9.41}$$

where b_H and b_A are model parameters.

Using the Tiessier Equation

Heterotrophic Biodegradation

The Tiessier equation (Tiessier, 1936) states that:

$$\mu = a(1 - \exp(-k'S))$$

where μ is the specific growth rate, S is the substrate concentration and a and k' are constants. In terms of this model, equation 9.17 (which is a formal expression of the statement 9.14) can be written as:

$$r_{X_{B,H}} = \sum_{l=1}^{\infty} a_l[1 - \exp(-k_l' X_{Sl})]X_{B,H} = X_{B,H} \sum_{l=1}^{\infty}[a_l - a_l \exp(-k_l' X_{Sl})] \tag{9.42}$$

or, by writing the summation term in a compact form:

$$r_{X_{B,H}} = X_{B,H} \sum_{l=1}^{\infty} c_l \exp(-k_l' X_{Sl}) \tag{9.43}$$

where c_l is real number that takes both positive and negative values and k_l' is a postive constant. Using the coefficients Ω_l defined previously, the latter equation can be transformed into:

$$r_{X_{B,H}} = X_{B,H} \sum_{l=1}^{\infty} c_l \exp(-k_l X_{COD}) \tag{9.44}$$

where $k_l = k_l'\Omega_l$.

$$r_{X_{B,H}} = X_{B,H} \sum_{l=1}^{\infty} c_l \exp(-k_l X_{COD}) \tag{9.45}$$

Note that in the case of small variations of the concentration $X_{B,H}$, the expression of the rate $r_{X_{B,H}}$ can be simplified as follows:

$$r_{X_{B,H}} = \sum_{l=1}^{\infty} c_l \exp(-k_l X_{COD}) \tag{9.46}$$

Remark on the exponential term: in the case of mechanism 1, the term $\exp(-k_l X_{COD})$ arose in the rate expression $r_{X_{B,H}}$ because of the Tiessier equation. A qualitative/approximate reasoning shows that this same term would also be arise in the case of mechanism 2. Here is the idea of this reasoning:

Mechanism 2 states that microorganisms first hydrolyze the constituents of substrate groups $\mathcal{SG}_l$, $l \geq 2$ sequentially, release the hydrolyzed matter to the bulk liquid as soluble compounds belonging to the first group $\mathcal{SG}_1$, and then synthesize cells from the constituents of the latter group only. The substrate biodegradation process can then be modeled as a collection of consecutive reactions acting in the following way:

$$\cdots \; X_{Sl} \longrightarrow X_{Sl-1} \longrightarrow \cdots \longrightarrow X_{S2} \longrightarrow X_{S1} \longrightarrow X_{B,H} \qquad (9.47)$$

Developing the exact kinetic equations for this mechanism is a challenging mathematical problem. Note indeed that the *linear* assumption that we implemented in the case of mechanism 1 (the total sludge synthesized is the sum of partial sludge productions by individual biochemical reactions) is not valid for this mechanism. In what follows, we will make use of the following qualitative reasoning to develop these rate equations:

Exact Rate Equations for Two Consecutive reactions:

Consider the following consecutive reactions:

$$A \xrightarrow{k_1} B \xrightarrow{k_2} C$$

when they take place in a batch reactor, with initial concentrations C_{B_0} and C_{C_0} both equal to 0. Assume that both reactions are first order reactions with rate constants k_1 and k_2. Then it is possible to show rigorously that the concentration C_B of component B varies with time t according to the following equation:

$$C_B = \frac{k_1 \, C_{A_0}}{k_1 - k_2}(\exp(-k_1 \, t) - \exp(-k_2 \, t))$$

or, equivalently:

$$C_B = c_1' \, \exp(-k_1 \, t) + c_2' \, \exp(-k_2 \, t) \qquad (9.48)$$

with c_1', c_2', k_1 and k_2 are constants.

This equation suggests that the kinetic model structure for a mechanism of consecutive first order reactions involves exponential functions rather than rational functions such as Monod equation.

Generalization to the Process Reactions 9.47:

As Henze *et al.* (1987) suggested in their report on Activated Sludge Model n° 1 (pages 27, 28), we will consider here that ASP reactions may be described by first order kinetics. Based on this assumption, we will generalize equation 9.48 to ASP reactions 9.47 in the following way:

$$X_{S1} = c_1' \, \exp(-k_1 \, t) + c_2' \, \exp(-k_2 \, t) + c_3' \, \exp(-k_3 \, t) + \cdots + c_l' \, \exp(-k_l \, t) + \cdots \qquad (9.49)$$

Since the ASP bioreactor is a continuous flow reactor and not a batch reactor, it is meaningless to express the substrate concentration as a direct function of time. Rather, X_{S1} is governed by the variations of the concentration X_{COD} in the bioreactor. Consequently, X_{S1} will be expressed as:

$$X_{S1} = c'_1 \exp(-k_1 X_{COD}) + c'_2 \exp(-k_2 X_{COD}) + \cdots + c'_l \exp(-k_l X_{COD}) + \cdots \quad (9.50)$$

Assuming that the heterotrophic growth rate is also described by first order kinetics, we get:

$$\begin{aligned} r_{X_{B.H}} &= kX_{S1} \\ &= (c_1 \exp(-k_1 X_{COD}) + c_2 \exp(-k_2 X_{COD}) + \cdots + c_l \exp(-k_l X_{COD}) + \cdots) \end{aligned}$$

or, equivalently

$$r_{X_{B.H}} = \sum_{l=1}^{\infty} c_l \exp(-k_l X_{COD}) \quad (9.51)$$

where the constants c_l are the products of c'_l and k. Note that the constants k_l take positive values only, while c_l can be positive or negative.

Nitrification

Here again, first order kinetics will be assumed. Using the same notations as for "*mechanism 1*", the autotrophic growth rate $r_{X_{B.A}}$ is expressed as:

$$r_{X_{B.A}} = k_{NH} S_{NH} \quad (9.52)$$

Nitrogen mineralization occurs in parallel to the transformations 9.47. To develop the rate equations for this process, we will use a similar rationale as the one implemented for the case of the Monod equation (see last part of sub-section "Nitrification" for the Monod equation), except that it is the Tiessier model that is used here for the mineralization reactions. Hence, the nitrogen mineralization rate can be expressed as follows:

$$r_{S_{NH}} = X_{B.H} \sum_{l=1}^{L} c_{NHl} \exp(-k_{NHl} X'_{ND}) \quad (9.53)$$

Biomass Decay

Biomass decay will be described using the same equations as in mechanism 1:

$$(r_{X_{B.H}})_{dec} = b_H \, X_{B.H} \tag{9.54}$$

$$(r_{X_{B.A}})_{dec} = b_A \, X_{B.A} \tag{9.55}$$

where b_H and b_A are model parameters.

9.2.3 Recapitulation

In the case of the Monod model, equations 9.25, 9.27, 9.39, 9.40 and 9.41 define a nested series $\mathcal{NS}_1$ of mechanistic models of the ASP. These models are described for a fixed integer L using the traditional matrix format in table 9.2. If the model corresponding to the integer L is denoted as $\mathcal{M}_L^1$, then we can write the property that justifies the adjective "nested" as:

$$\mathcal{M}_1^1 \subset \mathcal{M}_2^1 \subset \mathcal{M}_3^1 \subset \cdots \subset \mathcal{M}_L^1 \subset \cdots \subset \mathcal{M}_\infty^1 \tag{9.56}$$

Similarly, in the case of the Tiessier model, equations 9.50, 9.51, 9.52, 9.53 and 9.54 define another nested series $\mathcal{NS}_2$ of mechanistic models of the ASP. These models are described for a fixed integer L in table 9.3 using the matrix format. If the model corresponding to the integer L is denoted as $\mathcal{M}_L^2$, the nested property can be expressed as:

$$\mathcal{M}_1^2 \subset \mathcal{M}_2^2 \subset \mathcal{M}_3^2 \subset \cdots \subset \mathcal{M}_L^2 \subset \cdots \subset \mathcal{M}_\infty^2 \tag{9.57}$$

9.3 Uncertainty Model for $\mathcal{NS}_1$ and $\mathcal{NS}_2$

Now that we have developed nested series of process models for the ASP, we need to investigate the uncertainty that is associated with these models when they are used to approximate ASP dynamics. To do so, we will concentrate in the rest of this thesis on the heterotrophic biodegradation and, more specifically, on the process model prediction of the state variable X_{COD}. Note that this variable X_{COD} is similar to the one used by Jeppsson (1996) in his reduced-order process models.

Table 9.2: Matrix Formulation of Nested Series $\mathcal{NS}_1$ of Mechanistic Models $\mathcal{M}_L^1$

Component $i \rightarrow$	1	2	3	4	5	6	Process Rate ρ_j
Process $j \downarrow$	X_{COD}	$X_{B,H}$	$X_{B,A}$	S_{NO}	S_{NH}	X'_{ND}	$[ML^{-3}T^{-1}]$
1. Heterotrophic biomass growth	$-\frac{1}{Y_H}$	1			$-i_{XB}$		$\frac{\hat{\mu}_{H_1} X_{COD}}{K_{S1}+X_{COD}} X_{B,H} + \sum_{l=2}^{L} \frac{\hat{\mu}_{H_l} X_{COD}}{K_{Sl} X_{B,H}+X_{COD}} X_{B,H}$
2. Autotrophic biomass growth			1	$\frac{1}{Y_A}$	$-i_{XB} - \frac{1}{Y_A}$		$\frac{\hat{\mu}_A S_{NH}}{K_{NH}+S_{NH}} X_{B,A}$
3. Decay of heterotrophic biomass	1	-1				i_{XB}	$b_H X_{B,H}$
4. Decay of autotrophic biomass	1		-1			i_{XB}	$b_A X_{B,A}$
5. Nitrogen mineralization					1	-1	$k_a X'_{ND} X_{B,H} + \sum_{l=2}^{L} \frac{k_l X'_{ND}}{K_{Sl} X_{B,H}+X_{COD}} X_{B,H}$
Observed Conversion Rates $[ML^{-3}T^{-1}]$	$r_i = \sum_j \nu_{ij} \rho_j$						

Table 9.3: Matrix Formulation of Nested Series $\mathcal{NS}_2$ of Mechanistic Models $\mathcal{M}_L^2$

Component $i \rightarrow$	1	2	3	4	5	6	Process Rate ρ_j
Process $j \downarrow$	X_{COD}	$X_{B.H}$	$X_{B.A}$	S_{NO}	S_{NH}	X'_{ND}	$[ML^{-3}T^{-1}]$
1. Heterotrophic biomass growth	$-\frac{1}{Y_H}$	1			$-i_{XB}$		$\sum_{l=1}^{\infty} c_l \exp(-k_l X_{COD})\, X_{B.H}$
2. Autotrophic biomass growth			1	$\frac{1}{Y_A}$	$-i_{XB} - \frac{1}{Y_A}$		$k_{NH} S_{NH}$
3. Decay of heterotrophic biomass	1	-1				i_{XB}	$b_H X_{B.H}$
4. Decay of autotrophic biomass	1		-1			i_{XB}	$b_A X_{B.A}$
5. Nitrogen mineralization					1	-1	$\sum_{l=1}^{L} c_{NHl} \exp(-k_{NHl} X'_{ND})\, X_{B.H}$
Observed Conversion Rates $[ML^{-3}T^{-1}]$	$r_i = \sum_j \nu_{ij} \rho_j$						

From table 9.2, it can be seen that, using a model $\mathcal{M}_L^1$ (L is a fixed integer) for mechanism 1, the dynamics of X_{COD} are governed by the differential equation:

$$\frac{dX_{COD}}{dt} = (u_{COD}.u_Q^T) - (\vec{a}.u_Q^T)\, X_{COD} - \frac{1}{Y_H}\frac{\hat{\mu}_{H_1} X_{COD}}{K_{S1} + X_{COD}} X_{B,H} - \frac{1}{Y_H}\sum_{l=2}^{L}\frac{\hat{\mu}_{H_l} X_{COD}}{K_{Sl}X_{B,H} + X_{COD}} X_{B,H} + b_H\, X_{B,H} + b_A\, X_{B,A} \quad (9.58)$$

where:

- $u_Q = [\frac{Q_{in}}{V}, \frac{Q_r}{V}]$.
- V is the bio-reactor volume.
- $u_{COD} = [X_{CODin}, X_{CODr}]$.
- $\vec{a} = [1, 1]$.
- L is an integer.
- The superscript ^{T} means transposed vector.
- the subscript $_{in}$ means influent.
- the subscript $_r$ means recycle.

and the other mathematical objects are as defined in previous sections of this Chapter. In order to simplify the notations, the variable X_{COD} will be denoted simply as X. Equation 9.58 can then be re-written as:

$$\frac{dX}{dt} = (u_{COD}.u_Q^T) - (\vec{a}.u_Q^T)\, X - \frac{1}{Y_H}\frac{\hat{\mu}_{H_1} X}{K_{S1} + X} X_{B,H} - \frac{1}{Y_H}\sum_{l=2}^{L}\frac{\hat{\mu}_{H_l} X}{K_{Sl}X_{B,H} + X} X_{B,H} + b_H\, X_{B,H} + b_A\, X_{B,A} \quad (9.59)$$

The latter equation involves $2L + 2$ parameters:

$$\mathbf{p}_L = (\hat{\mu}_{H_1}, K_{S1}, \hat{\mu}_{H_2}, K_{S2}, \hat{\mu}_{H_3}, K_{S3}, \cdots, b_H, b_A) \quad (9.60)$$

and the objective of the model identification procedure is to determine the values of these parameters by minimizing the objective function:

$$J(\mathbf{p}_L) = \sum_{i=1}^{N} |X_{\mathbf{p}_L}^{\mathcal{M}_L}(t_i) - X^{data}(t_i)|^2 \quad (9.61)$$

where:

- $X^{data}(t_1)$, $X^{data}(t_2)$, $X^{data}(t_3)$, ..., $X^{data}(t_N)$ are the empirical data obtained from the real operation of the bioreactor. Denote the set of these data as Υ_N.
- $X^{\mathcal{M}_L^1}_{\mathbf{p}_L}(t)$ represents the prediction function of variable X by process model $\mathcal{M}_L^1$. Mathematically, it is the solution of the differential equation 9.59.

Equation 9.59 shows that there is interaction between the dynamics of X and that of $X_{B,H}$ and $X_{B,A}$. The study of this interaction and its effect on the mathematical framework developed in Chapter 5 is outside the scope of this thesis. In the remaining of this study, it is assumed that the time variations of $X_{B,H}$ and $X_{B,A}$ are known either by measuring them directly, or by carrying out an observability study and developing an *observer* for $X_{B,H}$ and $X_{B,A}$ (see Fossard and Normand-Cyrot, 1993, for an example of observer development for a biochemical process using Monod equation).

When the identification objective function $J(\mathbf{p}_L)$ is minimized over the model parameter space, it is usually not possible to guarantee that the obtained minimum is a global one. Because of this, we will consider in what follows that the minimization algorithm leads to a local minimum of $J(\mathbf{p}_L)$ reached at a certain parameter vector denoted as $(\mathbf{p}_L^{\Upsilon_N})_{loc}$. Process control strategy design will then be based on the model prediction function $X^{\mathcal{M}_L}_{(\mathbf{p}_L^{\Upsilon_N})_{loc}}$. If we designate the real process response function, with respect to X, as $g^{\mathcal{T}_{asp}}$, then using the results of previous Chapters we can write that the squared deviation between the latter function and $X^{\mathcal{M}_L}_{(\mathbf{p}_L^{\Upsilon_N})_{loc}}$ is bounded as follows:

$$[\mathcal{D}(X^{\mathcal{M}_L^1}_{(\mathbf{p}_L^{\Upsilon_N})_{loc}}, g^{\mathcal{T}_{asp}})]^2 \leq \left(\frac{J_S((\mathbf{p}_L^{\Upsilon_N})_{loc})}{N}\right)\left(1+\frac{\kappa_L\,\zeta(N,q_{1\,est})}{2}\left(1+\sqrt{1+\frac{4}{\kappa_L\,\zeta(N,q_{1\,est})}}\right)\right) \tag{9.62}$$

where:

$$\zeta(N,q_{1\,est}) = 4\frac{\left[q_{1\,est}\left(\ln\left(\frac{2N}{q_{1\,est}}\right)+1\right)-\ln\left(\frac{\eta}{4}\right)\right]}{N} \tag{9.63}$$

- $q_{1\,est}$ is an estimated value of the VC dimension of the space $l_{\mathcal{H}^{\mathcal{M}_L^1}}$.
- κ_L is a number that represents a prior information about the ASP behaviour being approximated by model $\mathcal{M}_L^1$. It is assumed that it takes values from the

interval [1.82,4].

- The number η belongs to]0,1[. When it is set to a certain value from this interval, then inequality 9.62 holds true with a probability of at least $1-\eta$.

Similarly, using a model $\mathcal{M}_L^2$ (L is a fixed integer) for mechanism 2 and table 9.3, the differential equation that governs the dynamics of X can be written as:

$$\begin{aligned}\frac{dX}{dt} &= (u_{COD}.u_Q^T) - (\vec{a}.u_Q^T)\, X - \\ &\frac{1}{Y_H}\sum_{l=1}^{L} c_l \exp(-k_l\, X)\, X_{B,H} + b_H\, X_{B,H} + b_A\, X_{B,A}\end{aligned} \tag{9.64}$$

Using the same notations as before, the uncertainty equation for process model $\mathcal{M}_L^2$ can be written as:

$$[\mathcal{D}(X^{\mathcal{M}_L^2}_{(\mathbf{p}_L^{\Upsilon_N})_{loc}}, g^{\mathcal{T}_{asp}})]^2 \leq \left(\frac{J_S((\mathbf{p}_L^{\Upsilon_N})_{loc})}{N}\right)\left(1+\frac{\kappa_L\,\varsigma(N,q_{2est})}{2}\left(1+\sqrt{1+\frac{4}{\kappa_L\,\varsigma(N,q_{2est})}}\right)\right) \tag{9.65}$$

where:

$$\varsigma(N,q_{2est}) = 4\frac{\left[q_{2est}\left(\ln\left(\frac{2N}{q_{2est}}\right)+1\right)-\ln\left(\frac{\eta}{4}\right)\right]}{N} \tag{9.66}$$

- $\mathbf{p}_L$ is the parameter vector:

$$\mathbf{p}_L = (c_1, k_1, c_2, k_2, \cdots, c_L, k_L, b_H, b_A)$$

- q_{2est} is an estimated value of the VC dimension of the space $l_{\mathcal{H}^{\mathcal{M}_L^2}}$.

- κ_L is a number that represents a prior information about the ASP behaviour being approximated by model $\mathcal{M}_L^2$. It is assumed that it takes values from the interval [1.82,4].

- The number η belongs to]0,1[. When it is set to a certain value from this interval, then inequality 9.65 holds true with a probability of at least $1-\eta$.

Chapter 10

Uncertainty Management in the Activated Sludge Process (Part 2)

10.1 The Issue of Data Scarcity

The effect of the number N of identification data points on the quality of process model prediction was investigated in Chapter 8: the process model was fixed and the number N was used as an uncertainty control variable. In most cases, however, N is usually fixed and small and, as such, it cannot be manipulated in order to control the uncertainty that underlies ASP behaviour. In this Chapter, another uncertainty control variable — *the process model structure complexity* — is introduced and a new strategy to reduce the process uncertainty using this variable is developed. An intuitive explanation of the author's idea is presented below.

Because of its biological aspect, the activated sludge process "*has a complexity unparalleled in the chemical industry*" (Jeppsson, 1996). As a result, to describe its dynamical behaviour fully, we would need highly complex models that involve a quasi-infinite number of state variables and parameters. Examples of such models are the nested series $\mathcal{NS}$ of mechanistic models $\mathcal{M}_L$ developed in previous Chapter. Although such models are detailed and aim to take into account a large number of biochemical phenomena, they are useless until the values of the parameters they involve are correctly identified. The model parameter identification procedure that is traditionally used by wastewater researchers comes down to minimizing an objective function such that:

$$J(\mathbf{p}_L) = \sum_{i=1}^{N} |X_{\mathbf{p}_L}^{\mathcal{M}_L}(t_i) - X^{data}(t_i)|^2 \quad (10.1)$$

over the model parameter space Γ (all mathematical objects in equation 10.1 are as defined in previous Chapter). The parameter vector $\mathbf{p}_L^{\Upsilon_N}$ at which the minimum of this function is attained and the identified model $X_{\mathbf{p}_L^{\Upsilon_N}}^{\mathcal{M}_L}$ will be used as a basis for process control strategy design.

Now imagine that the size N of the identification data set Υ_N is small, and the expression of the model prediction function $X_{\mathbf{p}_L}^{\mathcal{M}_L}$ involves a large number of parameters, i.e., the integer L is large compared to N. Then, in this case, it is guaranteed to find a parameter vector $\mathbf{p}_L^{\Upsilon_N}$ in the space Γ that minimizes the function $J(\mathbf{p}_L)$ down to a very low value, if not to zero. This fact is true not only for mechanistic models $\mathcal{M}_L$, but for neural network or polynomial models as well. Let us illustrate this fact by considering the latter type of models, for instance, which we will denote as $X_{\mathbf{p}_L}^{\mathcal{P}_L}$:

$$X_{\mathbf{p}_L}^{\mathcal{P}_L}(t) = p_0 + p_1 t + p_2 t^2 + \cdots + p_L t^L \quad (10.2)$$

i.e., a polynomial function of degree L with $\mathbf{p}_L = (p_0, p_1, \cdots, p_L)$ the parameter vector. In order to satisfy the condition "*L large compared to N*", assume that $L = N - 1$, for example, and consider the following system of linear equations:

$$\left\{ \begin{array}{lcl} p_0 + t_1 p_1 + t_1^2 p_2 + \cdots + t_1^L p_L & = & X^{data}(t_1) \\ p_0 + t_2 p_1 + t_2^2 p_2 + \cdots + t_2^L p_L & = & X^{data}(t_2) \\ & \vdots & \\ p_0 + t_N p_1 + t_N^2 p_2 + \cdots + t_N^L p_L & = & X^{data}(t_N) \end{array} \right. \quad (10.3)$$

The unknowns in this system are the model parameters p_0, p_1, p_2, $\ldots$, p_L and the coefficients are $a_{i,j} = t_i^j$, i.e., the $N \times N$ system matrix is:

$$A = (a_{i,j}) = \begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^L \\ 1 & t_2 & t_2^2 & \cdots & t_2^L \\ & & & \vdots & \\ 1 & t_N & t_N^2 & \cdots & t_N^L \end{pmatrix} \quad (10.4)$$

Such type of matrix is usually called Vandermonde matrix. Its determinant is never null because $t_1 < t_2 < \cdots < t_N$. As a result, it is guaranteed to find one unique set of parameters p_1, p_2, $\ldots$, p_L such that all equalities of system 10.3 hold true simultaneously. If we designate such set of parameters as $\mathbf{p}_L^{\Upsilon_N}$, then all the following equalities hold true simultaneously:

$$\begin{aligned} X^{\mathcal{P}_L}_{\mathbf{p}_L^{\Upsilon_N}}(t_1) &= X^{data}(t_1) \\ X^{\mathcal{P}_L}_{\mathbf{p}_L^{\Upsilon_N}}(t_2) &= X^{data}(t_2) \\ &\vdots \\ X^{\mathcal{P}_L}_{\mathbf{p}_L^{\Upsilon_N}}(t_N) &= X^{data}(t_N) \end{aligned} \quad (10.5)$$

or, equivalently:

$$\begin{aligned} |X^{\mathcal{P}_L}_{\mathbf{p}_L^{\Upsilon_N}}(t_1) - X^{data}(t_1)|^2 &= 0 \\ |X^{\mathcal{P}_L}_{\mathbf{p}_L^{\Upsilon_N}}(t_2) - X^{data}(t_2)|^2 &= 0 \\ &\vdots \\ |X^{\mathcal{P}_L}_{\mathbf{p}_L^{\Upsilon_N}}(t_N) - X^{data}(t_N)|^2 &= 0 \end{aligned} \quad (10.6)$$

By adding the latter equalities up, we obtain:

$$J(\mathbf{p}_L^{\Upsilon_N}) = \sum_{i=1}^{N} |X^{\mathcal{P}_L}_{\mathbf{p}_L^{\Upsilon_N}}(t_i) - X^{data}(t_i)|^2 = 0 \quad (10.7)$$

Hence, the vector $\mathbf{p}_L^{\Upsilon_N}$ minimizes the identification objective function down to its lowest value 0. In addition, it is unique. Does that mean that the identified model $X^{\mathcal{P}_L}_{\mathbf{p}_L^{\Upsilon_N}}$ is the best process model? Of course *not*. All what the latter model has done indeed is that it has used its high "expressive power" (due to large number of parameters compared to the number of data points) to overfit the identification data points and bring the value of $J(\mathbf{p}_L)$ down to 0. And, as was explained earlier in this thesis, data overfitting is not beneficial at all for complex system modelling. It must always be avoided in the course of model identification, whether we are using neural networks, polynomial models or mechanistic ones.

Thus, a complex process model cannot be correctly identified when the data set size is small. So what to do? Many researchers have settled on the other extreme of the continuum of model structure complexity: they chose to adopt reduced-order models, i.e., mechanistic models that involve a small number of parameters to describe the ASP behaviour (Jeppsson, 1996). Of course, overfitting does not occur with such simple models. There is, however, a risk that the selected reduced-order model be too simple to correctly approximate the dynamical variations of the real process response function $g^{\mathcal{T}_{asp}}$. If this happens, we say that the model underfits the data, which we must also avoid in the course of process identification (see Chapter 5).

Hence, the current situation in the area of ASP modelling is that too complex models cause overfitting, and too simple ones cause underfitting. In this Chapter, we develop an innovative methodology that helps determine, for a fixed data set of small size N, the optimal model structure complexity that minimizes both overfitting and underfitting and, therefore, reduces the amount of uncertainty underlying the process behaviour. This methodology is based on a principle called the "*Inductive Principle of Structural Risk Minimization*" ($\mathcal{IPSRM}$).

10.2 Inductive Principle of Structural Risk Minimization

$\mathcal{IPSRM}$ was first introduced by Vapnik, and the reader is referred to the references Vapnik (1982, 1995, 1998) for theoretical details of this principle. In what follows, $\mathcal{IPSRM}$ is explained briefly and implemented at the practical level for the case of the activated sludge process.

The $\mathcal{IPSRM}$ is based on a simple rationale explained below.

In the last section of previous Chapter, it was concluded that the squared deviation $[\mathcal{D}(X^{\mathcal{M}_L}_{(\mathbf{p}_L^{\Upsilon_N})_{loc}}, g^{\mathcal{T}_{asp}})]^2$ between:

- the identified model $X^{\mathcal{M}_L}_{(\mathbf{p}_L^{\Upsilon_N})_{loc}}$ on the basis of which a process control strategy will be developed

and

- the real process response function $g^{\mathcal{T}_{asp}}$

is bounded as follows:

$$[\mathcal{D}(X^{\mathcal{M}_L}_{(\mathbf{p}_L^{\Upsilon_N})_{loc}}, g^{\mathcal{T}_{asp}})]^2 \leq \left(\frac{J_S((\mathbf{p}_L^{\Upsilon_N})_{loc})}{N}\right)\left(1 + \frac{\kappa_L\,\varsigma(N, q_{est})}{2}\left(1 + \sqrt{1 + \frac{4}{\kappa_L\,\varsigma(N, q_{est})}}\right)\right) \tag{10.8}$$

where:

$$\varsigma(N, q_{est}) = 4\frac{\left[q_{est}\left(\ln\left(\frac{2N}{q_{est}}\right) + 1\right) - \ln\left(\frac{\eta}{4}\right)\right]}{N} \tag{10.9}$$

Denote the first term of right-hand side of inequality 10.8 as $c(q_{est})$ and second term as $\phi(N, q_{est})$:

$$c(q_{est}) = \frac{J_S((\mathbf{p}_L^{\Upsilon_N})_{loc})}{N} \tag{10.10}$$

$$\phi(N, q_{est}) = 1 + \frac{\kappa_L\, \zeta(N, q_{est})}{2}\left(1 + \sqrt{1 + \frac{4}{\kappa_L\, \zeta(N, q_{est})}}\right) \tag{10.11}$$

Then, inequality 10.8 can be re-written as:

$$[\mathcal{D}(X^{\mathcal{M}_L}_{(\mathbf{p}_L^{\Upsilon_N})_{loc}}, g^{\mathcal{T}_{asp}})]^2 \leq c(q_{est}) \times \phi(N, q_{est}) \tag{10.12}$$

The number η is of course assumed to be fixed in the interval $]0, 1[$; for example $\eta = 0.05$, i.e., $1 - \eta = 0.95$.

Assume now that the number N is set at a fixed value N_0 that is small. Vapnik (1995) considered that N_0 is small if the ratio

$$\frac{N_0}{q_{est}}$$

is small, say, less than 20. In this study, we will adopt the same definition of "N_0 small". Inequality 10.13 becomes then:

$$[\mathcal{D}(X^{\mathcal{M}_L}_{(\mathbf{p}_L^{\Upsilon_{N_0}})_{loc}}, g^{\mathcal{T}_{asp}})]^2 \leq c(q_{est}) \times \phi(N_0, q_{est}) \tag{10.13}$$

As the complexity q_{est} of process model $\mathcal{M}_L$ increases, the intensity of the underfitting phenomenon and the empirical risk

$$c(q_{est}) = \frac{J((\mathbf{p}_L^{\Upsilon_{N_0}})_{loc})}{N_0} = R_{emp}^{\Upsilon_{N_0}}((\mathbf{p}_L^{\Upsilon_{N_0}})_{loc}) \tag{10.14}$$

decrease. At the same time, the intensity of the overfitting phenomenon and the mathematical function $\phi(N_0, q_{est})$ increase. This suggests that the product:

$$\varphi_1(N_0, q_{est}) = c(q_{est}) \times \phi(N_0, q_{est}) \tag{10.15}$$

which represents the guaranteed deviation, passes through a minimum value

$$\varphi_1(N_0, (q_{est})_{min})$$

at which underfitting, overfitting and the uncertainty value $\mathcal{D}(X^{\mathcal{M}_L}_{(\mathbf{p}_L^{\Upsilon_{N_0}})_{loc}}, g^{\mathcal{T}_{asp}})$ are minimal. The purpose of $\mathcal{IPSRM}$ is to determine the value of $(q_{est})_{min}$ at which this minimum is attained and that sets the optimal model structure complexity. Once $(q_{est})_{min}$ and its corresponding integer L_{min} are determined, the identified model:

$$X^{\mathcal{M}_{L_{min}}}_{(\mathbf{p}_{L_{min}}^{\Upsilon_{N_0}})_{loc}}$$

can be used as a basis for process control strategy design, because it minimizes the guaranteed deviation $\varphi_1(N_0, q_{est})$ and, consequently, the process uncertainty.

The $\mathcal{IPSRM}$ will be used below to solve the practical case n° 2 formulated in Chapter 1 of this thesis.

10.3 Solving the Practical Case n° 2 of Chapter 1

In Chapter 1 (Introduction), it was pointed out that one of the major applications of uncertainty management techniques developed in this thesis is the determination of the optimal model structure for a fixed (and usually small) number of identification data points. A scenario for such application was called "*Practical Case n° 2*" and formulated as follows (see Chapter 1):

- Practical Case n° 2: *Consider the case where the manager of the wastewater treatment plant cannot supply an unlimited amount of data. He responds that actually only a sequence of about $N_0 = 10$ input-output empirical data are available. Still, the consultant is asked to develop the best possible* **model structure** *based on this limited amount of data. The manager is aware of the problem of developing accurate models with very few data, especially for a highly complex and uncertain system such as the ASP. He knows that the risk associated with this model is likely to be high, but he is expecting the consultant to extract maximum information from the small amount of data supplied to him, so that the lowest possible risk under such circumstances can be achieved.*

Solving this case can be carried out using the procedure outlined below. Note that this procedure can be implemented for either $\mathcal{NS}_1$ (mechanism 1) and $\mathcal{NS}_2$ (mechanism 2).

Procedure (P2):

1. *Using the N_0 empirical data points, minimize the objective function:*

$$J(\mathbf{p}_1) = \sum_{i=1}^{N_0} |X_{\mathbf{p}_1}^{\mathcal{M}_1}(t_i) - X^{data}(t_i)|^2 \tag{10.16}$$

where the time function $X_{\mathbf{p}_1}^{\mathcal{M}_1}$ represents the solution to the differential equation 9.59 for mechanism 1 or 9.64 for mechanism 2, with $L = 1$. We get a local minimum $J((\mathbf{p}_1^{\Upsilon_{N_0}})_{loc})$.

2. *Repeat step 1 for the objective functions $J(\mathbf{p}_2)$, $J(\mathbf{p}_3)$, $J(\mathbf{p}_4)$, ..., corresponding to the integers $L = 2, 3, 4, \ldots$. We obtain the local minimums $J((\mathbf{p}_2^{\Upsilon_{N_0}})_{loc})$, $J((\mathbf{p}_3^{\Upsilon_{N_0}})_{loc})$, $J((\mathbf{p}_4^{\Upsilon_{N_0}})_{loc})$, ..., with:*

$$\ldots < J((\mathbf{p}_4^{\Upsilon_{N_0}})_{loc}) < J((\mathbf{p}_3^{\Upsilon_{N_0}})_{loc}) < J((\mathbf{p}_2^{\Upsilon_{N_0}})_{loc}) < J((\mathbf{p}_1^{\Upsilon_{N_0}})_{loc})$$

3. *For any given L, estimate the VC dimension $q_{est} = q_{est\,L}$ of the space $l_{\mathcal{H}^{\mathcal{M}_L}}$. This estimation can be based on the number of model parameters and the range over which these parameters are allowed to vary.*

4. *Compute the empirical risks $c(1)$, $c(2)$, $c(3)$, $c(4)$, ... by dividing all the local minimums $J((\mathbf{p}_1^{\Upsilon_{N_0}})_{loc})$, $J((\mathbf{p}_2^{\Upsilon_{N_0}})_{loc})$, $J((\mathbf{p}_3^{\Upsilon_{N_0}})_{loc})$, $J((\mathbf{p}_4^{\Upsilon_{N_0}})_{loc})$, ... by N_0:*

$$c(q_{est}) = \frac{J((\mathbf{p}_L^{\Upsilon_{N_0}})_{loc})}{N_0}$$

5. *Select, from interval $]0, 1[$, a value for the degree of confidence, $1-\eta$, with which the model prediction accuracy is to be determined. Usually, $1-\eta$ is set to 0.9 or 0.95 (i.e., $\eta = 10\%$ or 5%), although it can take any value between 0 and 1 (0 and 1 are excluded).*

6. *Compute the values of $\zeta(N_0, 1)$, $\zeta(N_0, 2)$, $\zeta(N_0, 3)$, $\zeta(N_0, 4)$, ... using the equation:*

$$\zeta(N_0, q_{est}) = 4\frac{\left[q_{est}\left(\ln\left(\frac{2N_0}{q_{est}}\right) + 1\right) - \ln\left(\frac{\eta}{4}\right)\right]}{N_0}$$

7. *Set κ_L at its average value:*

$$\bar{\kappa} = \frac{1.82 + 4}{2} = 2.91 \tag{10.17}$$

8. *Compute the values $\varphi_l(N_0, 1)$, $\varphi_l(N_0, 2)$, $\varphi_l(N_0, 3)$, $\varphi_l(N_0, 4)$, ... of the guaranteed deviation φ_l using the equation:*

$$\varphi_l(N_0, q_{est}) = c(q_{est}) \left(1 + \frac{\bar{\kappa}\zeta(N_0, q_{est})}{2} \left(1 + \sqrt{1 + \frac{4}{\bar{\kappa}\zeta(N_0, q_{est})}}\right)\right)$$

If N_0 is very small compared to q_{est}, i.e. $N_0/q_{est} < 10$ for example, then multiply all foregoing values of $\varphi_l(N_0, q_{est})$ by a correcting function $Cor(N_0, q_{est})$ such that (Vapnik, 1998; pp.524):

$$Cor(N_0, q_{est}) = 1 + \frac{\frac{q_{est}}{N_0} \ln N_0}{2\left(1 - \frac{q_{est}}{N_0}\right)}$$

for instance.

9. *Plot the values $\varphi_l(N_0, 1)$, $\varphi_l(N_0, 2)$, $\varphi_l(N_0, 3)$, $\varphi_l(N_0, 4)$, ... of the guaranteed deviation φ_l as a function of the q_{est}. We obtain a curve representing the function:*

$$q_{est} \mapsto \varphi_l(N_0, q_{est})$$

10. *Determine the minimum value of the curve plotted in step 9 and the integer $(q_{est})_{min}$ at which this minimum is attained. Find out the integer $L = L_{min}$ corresponding to the VC dimension $(q_{est})_{min}$. The deviation between the real process response function and the identified model prediction function is minimal for $L = L_{min}$ and, as a result, the optimal model structure for $N = N_0$ is $\mathcal{M}_{L_{min}}$.*

11. *Use the model $\mathcal{M}_{L_{min}}$ as a basis for process control strategy design.*

10.4 Illustrative Simulations

The objective of these simulations is twofold:

- illustrate the implementation of the $\mathcal{IPSRM}$ and that of the foregoing procedure.
- confirm the theory of the $\mathcal{IPSRM}$.

10.4.1 Simulation Methodology

The true process response function $g^{\mathcal{T}_{asp}}$, with respect to the state variable X, is considered to be known over a period of 24 hours. The N_0 data points are generated by adding a noise to the values of $g^{\mathcal{T}_{asp}}(t)$, about every two hours.

In reality, all that process engineers would know are the N_0 data points. The process response function is always unknown, and the objective of process modelling, identification and validation is to determine it. In this simulation study, however, we need to assume that $g^{\mathcal{T}_{asp}}$ is known in order to be able to verify the theory: does the $\mathcal{IPSRM}$ lead to the best approximation of the true process response function or not?

Once the data points are generated, we start the model identification procedure, which consists in minimizing the objective function:

$$J(\mathbf{p}_L) = \sum_{i=1}^{N_0} |X_{\mathbf{p}_L}^{\mathcal{M}_L}(t_i) - X^{data}(t_i)|^2$$

for models $\mathcal{M}_L$ ($L = 1, 2, 3, \ldots$) of a nested series $\mathcal{NS}$ ($\mathcal{NS}_1$ for mechanism 1 and $\mathcal{NS}_2$ for mechanism 2). The software Matlab of The MathWorks, Inc is the platform on which all function minimization tasks were carried out. At least 4 algorithms were tried:

- Simplex search method (using Matlab built-in functions)
- Quasi-Newton method (using Matlab built-in functions)
- Global optimization by multilevel coordinate search (MCS), developed by W. Huyer and A. Neumaier (http://solon.cma.univie.ac.at/~neum/software/mcs/)
- Genetic algorithms, implementing differential evolution, developed R. Storn (http://www.icsi.berkeley.edu/~storn/)

In addition, manual parameter adjustment and analytical investigation of the effects of model parameters on the variations of the objective functions $J(\mathbf{p}_L)$ were also carried out. The goal of all this computational endeavor was, of course, to overcome the local minima problem. Three computers were devoted to this task:

- a desktop, Pentium 166MHz, 16 MB RAM.

- a desktop, 486 DX2 - 66, 8MB RAM.
- a laptop, Pentium 166MHz, 160MB RAM.

Once the minimums of the functions $J(\mathbf{p}_L)$ are found and the VC dimensions

$$q_{est} = q_{est\,L}$$

estimated, computation of the guaranteed deviation values becomes straightforward.

In this simulation work, the variable X was taken equal to:

$$X = X_{COD} = S_S + X_S \tag{10.18}$$

The other variables (S_I, X_I, X_P), which represent substrates that are non-biodegradable in the absolute sense, are ignored.

Simulated Activated Sludge System

The activated sludge system used in this study to simulate the real ASP consists of one completely mixed aerobic bioreactor followed by a secondary clarifier. The operational variables of this system are presented in table 10.1 and the influent wastewater characteristics (using the state variables and notations of the IAWPRC model n° 1) in table 10.2. The dynamics of the simulated activated sludge system are generated using one model $\mathcal{M}_{L_0}$ of the nested series $\mathcal{NS}$, corresponding to a fixed integer L_0 selected arbitrarily. The closest description of these dynamics by the IAWPRC model n° 1 corresponds to the set of parameter values presented in table 10.3. Comparison of the simulated system dynamics to the IAWPRC model n° 1 was carried out using the DOS-based software SSSP, developed by Bidstrup and Grady (1987).

10.4.2 Simulation Results for the Case of the Monod Equation

The key task in implementing procedure (P2) is the minimization of the objective function:

$$J(\mathbf{p}_L) = \sum_{i=1}^{N} |X_{\mathbf{p}_L}^{\mathcal{M}_L}(t_i) - X^{data}(t_i)|^2 \tag{10.19}$$

Table 10.1: Operational variables of the simulated WWT plant

Variables	**Values**
Average flow rate	1000 m^3/day
Recycle flow rate	500 m^3/day
Reactor volume	200 m^3
Sludge age	10 $days$
Dissolved oxygen concentration	2 gO_2/m^3

Table 10.2: Steady-state influent Characteristics

Variables	**Values**
Heterotrophic organisms, $X_{B,H}$	0.0 $gCOD/m^3$
Autotrophic organisms, $X_{B,A}$	0.0 $gCOD/m^3$
Particulate inert organics, X_I	35.0 $gCOD/m^3$
Slowly biodegradable substrate, X_S	150.0 $gCOD/m^3$
Readily biodegradable substrate, S_S	115.0 $gCOD/m^3$
Soluble ammonia N, S_{NH}	25.0 gN/m^3
Soluble Nitrate/Nitrite N, S_{NO}	0.0 gN/m^3
Soluble biodegradable organic nitrogen N, S_{ND}	6.5 gN/m^3
Particulate biodegradable organic nitrogen N, X_{ND}	8.5 gN/m^3
Oxygen, S_O	0.0 gO_2/m^3
Alkalinity, S_{ALK}	5.0 $mole/m^3$

Table 10.3: Closest IAWPRC model parameters for the simulated system

Variables	**Values**
$\hat{\mu}_H$	4.0 day^{-1}
K_S	10.0 $gCOD/m^3$
K_{OH}	0.1 gO_2/m^3
Y_H	0.67 $gCOD/gCOD$
b_H	0.62 day^{-1}
η_g	0.8
K_{NO}	0.2 gN/m^3
k_h	2.2 day^{-1}
K_X	0.15 $gCOD/gCOD$
η_h	0.4
k_a	0.16 $m^3/gCOD/day$
f_P	0.08
i_{XB}	0.086
i_{XP}	0.06
$\hat{\mu}_A$	0.65 day^{-1}
K_{NH}	1.0 gN/m^3
K_{OA}	0.4 gO_2/m^3
Y_A	0.24 $gCOD/gN$
b_A	0.12 day^{-1}

for $L = 1, 2, 3, \cdots$, with $X_{\mathbf{p}_L}^{\mathcal{M}_L}$ the solution to the differential equation:

$$\frac{dX}{dt} = (u_{COD}.u_Q^T) - (\vec{a}.u_Q^T)\,X - \frac{1}{Y_H}\frac{\hat{\mu}_{H_1}X}{K_{S1}+X}X_{B,H} - \frac{1}{Y_H}\sum_{l=2}^{L}\frac{\hat{\mu}_{H_l}X}{K_{Sl}X_{B,H}+X}X_{B,H} + b_H\,X_{B,H} + b_A\,X_{B,A} \qquad (10.20)$$

In the course of minimizing the function $J(\mathbf{p}_L)$, the kinetic term:

$$\frac{1}{Y_H}\frac{\hat{\mu}_{H_1}X}{K_{S1}+X}X_{B,H} + \frac{1}{Y_H}\sum_{l=2}^{L}\frac{\hat{\mu}_{H_l}X}{K_{Sl}X_{B,H}+X}X_{B,H} \qquad (10.21)$$

which is a combination of Monod-like equations, has showed a great deal of numerical instabilities. No reasonable value of the function minimum could be attained for $L > 2$, despite the large number of simulations that were carried out using all four function minimization algorithms. Because of this, the claims of Dold *et al.* (1980, 1986) with respect to mechanism 1 — using the Monod equation — for the bi-substrate hypothesis could not be confirmed for the multi-substrate hypothesis. The nested series using the Monod equation was then abandoned in this thesis, although a fundamental (experimental and/or theoretical) proof to refute or confirm Mechanism 1 and the Monod equation is still to be investigated and developed. The task of establishing such proof is, however, beyond the scope of this work.

10.4.3 Simulation Results for the Case of the Tiessier Equation

Numerical instability was less of a problem in the case of the model equations using the Tiessier model (which can also be considered to describe the Mechanism 2 — see the previous Chapter). The results are presented below and confirm the theory of $\mathcal{IPSRM}$.

The true process response function $g^{\mathcal{T}_{asp}}$ and the data points used in this simulation are presented in Figure 10.1. The function $g^{\mathcal{T}_{asp}}$ is generated as the solution to the differential equation:

$$\frac{dX}{dt} = (u_{COD}.u_Q^T) - (\vec{a}.u_Q^T)\,X - \frac{1}{Y_H}\sum_{l=1}^{L} c_l \exp(-k_l\,X) + b_H\,X_{B,H} + b_A\,X_{B,A} \qquad (10.22)$$

where $L = 3$, $Y_H = 0.67\ gCOD/gCOD$, $c_1 = 3600\ gCOD/m^3/day$, $k_1 = 0.001\ [gCOD/m^3]^{-1}$, $c_2 = -1000\ gCOD/m^3/day$, $k_2 = 0.005\ [gCOD/m^3]^{-1}$, $c_3 = -3300\ gCOD/m^3/day$, $k_3 = 0.01\ [gCOD/m^3]^{-1}$, $b_H = 0.65\ day^{-1}$, $b_A = 0.2\ day^{-1}$, and the influent characteristics are as defined in section "Simulation Methodology".

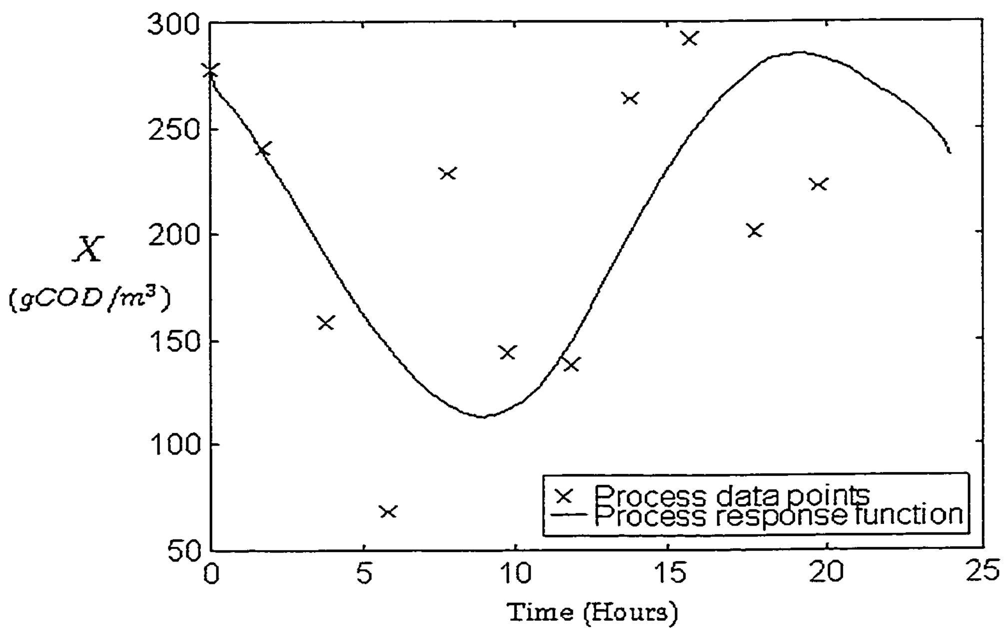

Figure 10.1: The true process response function $g^{T_{asp}}$ and the data points used for process identification

The process data points are generated by adding a noise ϵ_X to the values of the process response function $g^{T_{asp}}$, about every two hours. The noise ϵ_X is uniformly distributed within the interval $[-d, d]$, where d is 50% of the average of X over one day. The number of points generated is $N_0 = 11$.

The results of process model identification (steps 1 and 2 of procedure (P2)) for $L = 1$, $L = 2$, $L = 3$ and $L = 4$ are presented in Figures 10.2, 10.3, 10.4 and 10.5 respectively. The complexity of a model $\mathcal{M}_L$ is measured by its VC dimension $q_{est} = q_{est_L}$ (which is VC dimension of the space $l_{\mathcal{H},\mathcal{M}_L}$). Because the parameters b_H and b_A vary within limited ranges and the parameters k_l are all positive, the VC dimension q_{est_L} will be taken equal to L: $q_{est_L} = L$ (step 3).

Visual inspection of the Figures 10.2 to 10.5 shows that process model of complexity 2 provides the best approximation to the true process response function $g^{T_{asp}}$. Does the theory of $\mathcal{IPSRM}$ lead to the same result? The analysis below shows that the answer is positive and, therefore, the theory is in accordance with the simulation results. Following the instructions of steps 4 to 8 of procedure (P2), we can compute the values of the guaranteed deviation $\varphi_1(N_0, q_{est})$ for model complexities 1, 2, 3 and 4. The plot of these values as a function of q_{est} is shown in Figure 10.6 (step 9). The minimum of the obtained curve is attained at $(q_{est})_{min} = 2$. Thus, $L_{min} = 2$ and, therefore, the optimal process model for $N_0 = 11$ is $\mathcal{M}_2$.

The main difficulties encountered in this simulation work were related to the local minima problems. To overcome them, a longer period of time was devoted for the function minimization calculations using all four algorithms listed above.

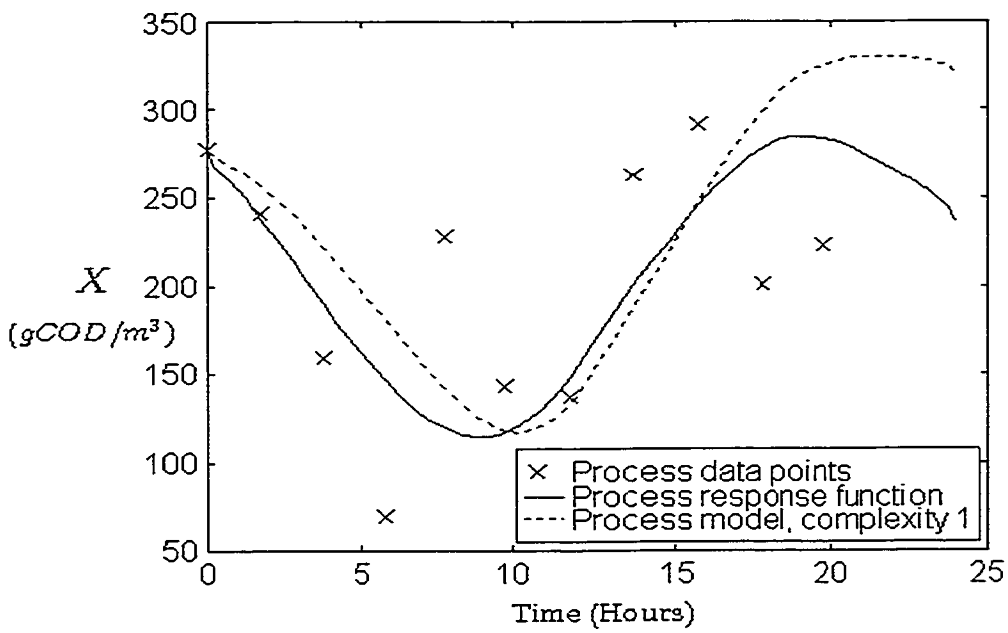

Figure 10.2: Process identification using model $\mathcal{M}_1$ (model complexity $q_{est} = 1$)

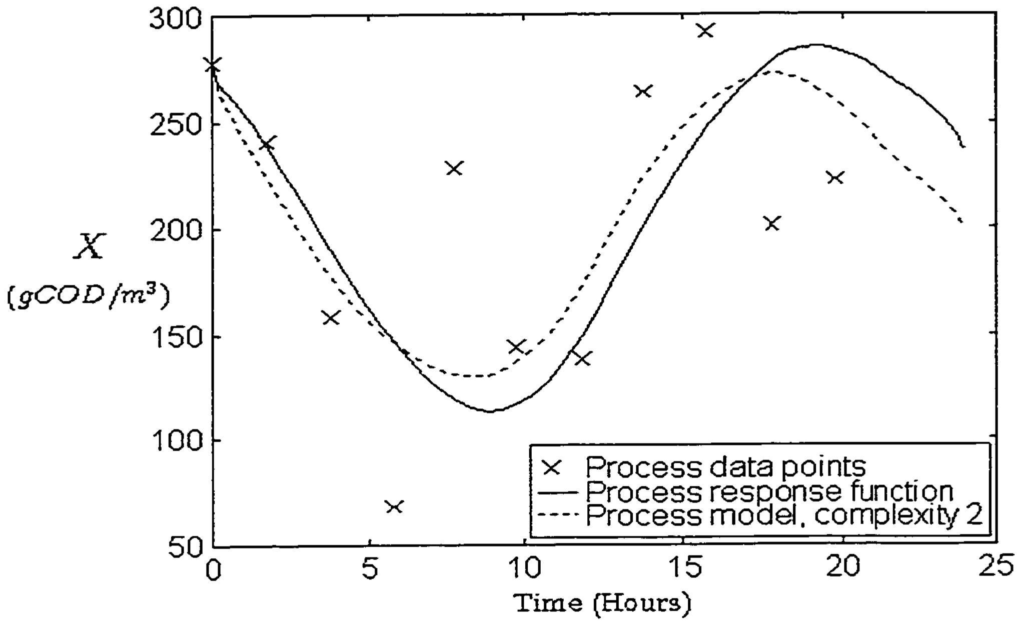

Figure 10.3: Process identification using model $\mathcal{M}_2$ (model complexity $q_{est} = 2$)

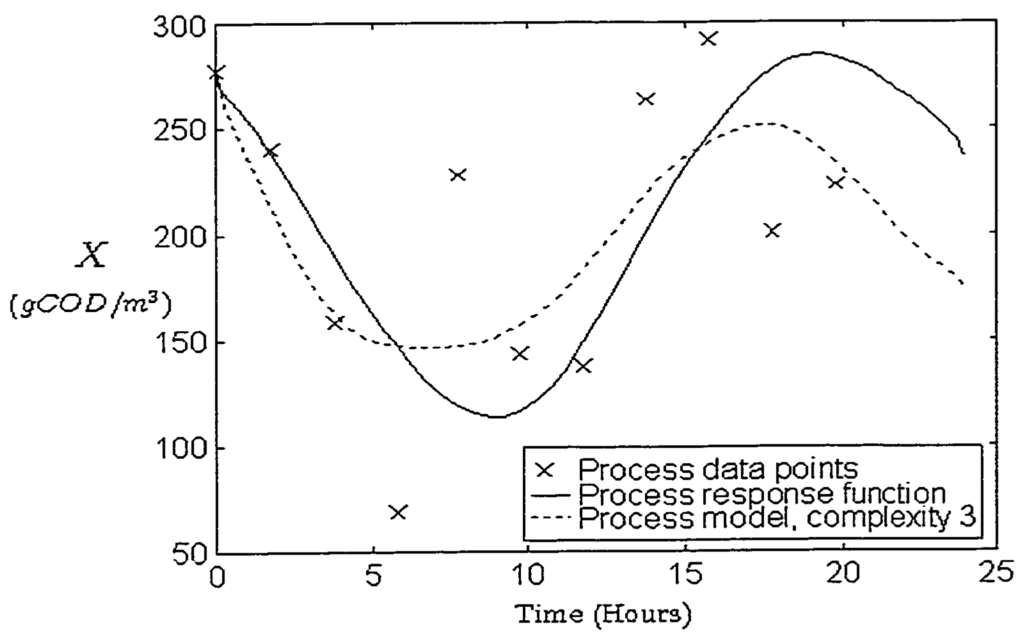

Figure 10.4: Process identification using model $\mathcal{M}_3$ (model complexity $q_{est} = 3$)

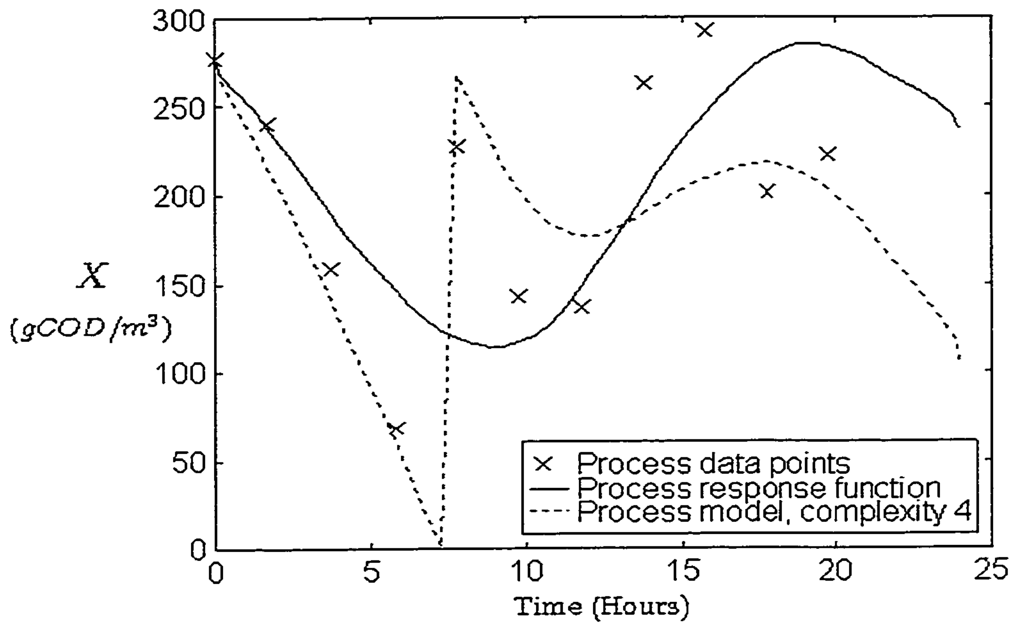

Figure 10.5: Process identification using model $\mathcal{M}_4$ (model complexity $q_{est} = 4$)

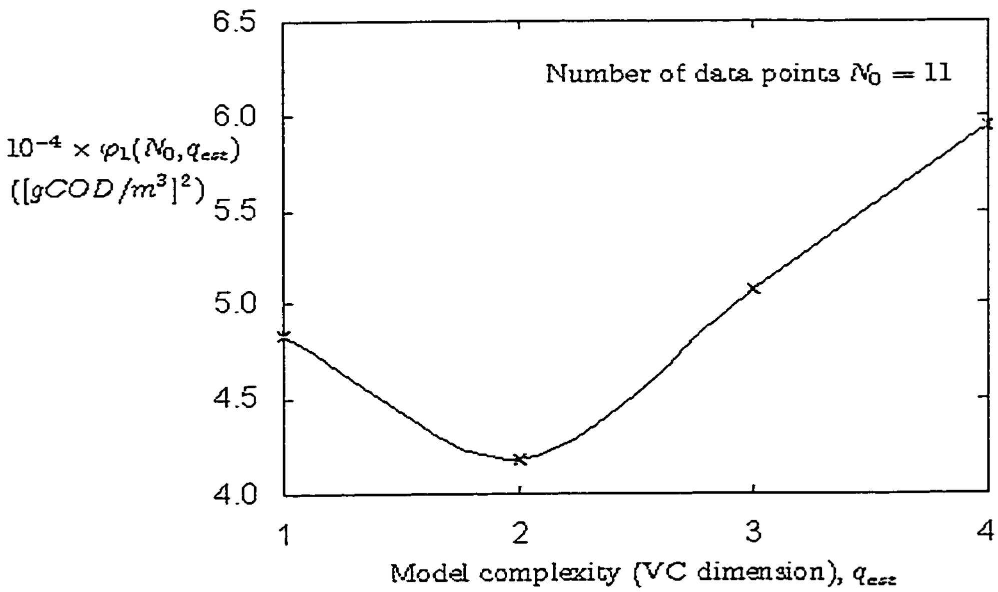

Figure 10.6: The guaranteed deviation $\varphi_1(N_0, q_{est})$ attains its minimum at the optimal process model complexity $q_{est} = 2$

Part V

CONCLUSIONS

Chapter 11

Conclusions and Recommendations for Future Research

This Chapter summarizes the thesis results and contributions (first section), and presents a list of topics for future research (second section).

11.1 Summary of Results

In this work, the foundations of a new area of research regarding biological WWT process mathematical modelling have been set. The main feature of this area is the introduction of innovative concepts and tools from emerging information modelling technologies into the traditional field of WWT process modelling.

The model identification procedure has been viewed as a learning problem or, equivalently, an information transfer from a set of real data into the process model. Model identifiability has not been considered as an essential criterion for model evaluation and selection. Instead, model evaluation has been based on one single criterion: performance. Model performance has been measured by the mathematical deviation between reality and model prediction. As for the model verifiability problem, it has been suggested to replace the condition *"measurability of all model state variables"* by a less stringent observability criterion. This criterion does not require that all state variables be measurable directly and separately, as long as the unmeasurable ones are uniquely determinable from those that are measurable.

It has been proposed in this work to shift the attention from modelling the activated sludge wastewater treatment process itself to modelling the uncertainty that underlies its dynamical behaviour. The goal is to answer questions such as: what makes uncertainty high or low? How can uncertainty be controlled and to what extent can it be reduced? The concept "process uncertainty model" has been introduced and an intuitive explanation of how to develop an uncertainty model presented.

The data overfitting phenomenon, which is encountered in the course of complex model identification has been thoroughly analyzed. Model cross validation and model parsimony have also been discussed. The fundamental object of interest in model identification has been defined as the process "general tendency" or process response function.

A mathematical framework for uncertainty modelling and management has been developed. The goal of this framework has been to construct simple inequalities to describe uncertainty. These inequalities, called uncertainty models, relate the mathematical deviation between model prediction and the real process behaviour to some simple variables that govern model performance — namely:

- the size of the data set used for model identification, N.
- the quality of these data, measured by their statistical dependency.
- the model complexity, measured by the VC dimension.
- the empirical measure of the mathematical deviation computed on the basis of the foregoing data set.

The development of these inequalities has been based on a principle called "*Inductive Principle of Empirical Risk Minimization*" ($\mathcal{IPERM}$).

The conditions of applicability of $\mathcal{IPERM}$ have been thoroughly examined in the case of the activated sludge process being described by a simple mechanistic model $\mathcal{M}$. The VC dimension of this model has been estimated and two uncertainty models have been developed for the ASP.

These two uncertainty models have been compared and the differences between them accounted for. The following result has been established: empirical data *cannot* compensate for our limited knowledge of process mechanisms, even if an infinite amount of data and computing power are made available to the model identification procedure.

Measures of process model maximal and marginal improvements have been developed. Then another result has been established: 80% of the model ($\mathcal{M}$) maximal improvement occurs at a number of data points of about $N_{80\%} \approx 15$ to 18. To achieve the other 20%, N has to be increased from the relatively small number $N_{80\%}$ to infinity.

Procedures for computing the marginal cost of process model improvement and solving the "*Practical Case n° 1*" of Chapter 1 have been developed.

A new approach to modelling the activated sludge process itself and dealing with the almost-infinite degree of complexity of the ASP behaviour has been developed. This approach consisted in constructing an *infinite series $\mathcal{NS}$ of nested ASP mechanistic models* of increasing complexity, with the following goal in mind: as the complexity of the process model is allowed to increase within this nested series $\mathcal{NS}$ of models, the process mechanisms are described better. This nested series has been developed using the *multi-substrate hypothesis*, implementing two different kinetic models: the Monod and the Tiessier models. Uncertainty models have then been developed for the obtained nested series $\mathcal{NS}$.

Another principle called "*Inductive Principle of Structural Risk Minimization*" ($\mathcal{IPSRM}$) has been introduced and implemented to solve the "*Practical Case n° 2*" of Chapter 1. Computer simulations have then been carried out to illustrate the use of $\mathcal{IPSRM}$ and confirm the theory.

11.2 Topics for Future Research

This thesis represents the first investigation of ASP uncertainty in the light of the most recent results of computational learning theory. As can be seen from the conclusions of this work, the methodology developed herein has a great deal of potential and can be implemented to address several important questions related to the task of ASP modelling and operation. However, there are several topics that still need to be investigated. Some of them are summarized below.

Local minima problems were the main obstacles encountered in the course of this research (Chapter 9). More simulation work is needed to investigate/evaluate the effect of these problems on the results of the $\mathcal{IPSRM}$.

Once the local minima issue is resolved, a study has to be carried out over a long period of time on a real plant in order to implement the tools developed in this thesis and make use of the predictions of the mathematical framework of Chapter 6.

A systematic and simple methodology (analytical or numerical) of estimating VC dimensions is still to be developed in order to facilitate the use of process uncertainty models.

In Chapter 9, two different substrate biodegradation mechanisms have been considered. An experimental and/or theoretical study is needed to elucidate the true mechanism that takes place in bioreactors.

One of the strongest points of the methodology developed in this thesis lies in the fact that the probability density function P_z need not be known. However, it is assumed that P_z is fixed. Since this function can change during a plant's life, there is a need to investigate the effect of this change on the results of the $\mathcal{IPERM}$ and $\mathcal{IPSRM}$.

The $\mathcal{IPERM}$ and $\mathcal{IPSRM}$, as they were developed by Vapnik, apply to the determination of a process response function of a scalar type only. An extension of these principles to functions of the vector type is still to be developed.

The implementation of the framework of Chapter 6 was limited to the case of heterotrophic biodegradation. An extension of this study to include the nitrification process needs to be investigated.

Appendix A

Notation and Abbreviations

$\mathcal{A}$: algorithm

ARMA: auto-regressive moving average

ASP: activated sludge process

b: decay rate coefficient

BS: bi-substrate hypothesis

c: value of the empirical risk

$\mathcal{C}$: a bound function

c_l: coefficient

CLT: Computational Learning Theory

COD: Chemical Oxygen Demand

Cor: correcting function

Ct: cost

$\mathcal{D}$: weighted root mean square deviation between functions

e: exponential function

$\mathbf{E}(\,.\,)$: expected value

$\mathcal{E}$: probabilistic environment

f: a scalar function

$\mathbf{f}$: a vector function

F: probability distribution function

$\mathcal{F}$: a set

g: process response function

gap: gap

h: decision rule

$\mathbf{h}$: a vector function

$H(. , .)$: decision rule

$\mathcal{H}$: decision rule space

$\mathcal{I}$: a measure of the information content

IAWPRC: International Association on Water Pollution Research and Control

$\mathcal{IP}$: process model improvement

$\mathcal{IPERM}$: Inductive Principle of Empirical Risk Minimization

$\mathcal{IPSRM}$: Inductive Principle of Structural Risk Minimization

$J(.)$: objective function

k: μ/Y

k_l: coefficient

K: half-saturation coefficient

l: quadratic loss function

$\mathcal{LM}$: learning machine

ln: logarithm function

M: bound on the loss function (weak prior information 1)

$\mathcal{M}$: system or process model

N : number of data points

$\aleph$: Set of all integers

$\mathcal{NS}$: nested series

$o(\,.\,)$: $f_1 = o(f_2)$ means that f_1/f_2 converges to 0

$\mathcal{OM}$: operating mode

$\mathbf{p}$: parameter vector

P: probability density function

$\mathcal{P}$: a noise pattern

PCS: plant control strategy

pdf: probability density function

Pr: probability of an event

q: value of the VC dimension

Q: flow rate

R: Risk

$\Re$: the real line (set of all real numbers)

RB: readily biodegradable

$\mathcal{RC}$: rate of convergence

s: order at which weak prior information 2 is calculated

S: substrate concentration

SB: slowly biodegradable

$\mathcal{SG}$: substrate group

sol: solution

$\mathcal{S}$: system

$\sup(a, b)$: maximum of a and b

t: time

$\mathcal{T}$: a transformer

$\mathbf{u}$: input vector

$\mathcal{UM}$: uncertainty model

$\mathbf{v}$: instance or situation vector

V: instance space

V: reactor volume

VC: Vapnik-Chervonenkis dimension

w: outcome

W: outcome space

$\mathcal{WPI}$: weak prior information

WRPE: wastewater researchers and practicing engineers

WWT: wastewater treatment

x: state variable

$\mathbf{x}$: state vector

X: biomass concentration

$\mathcal{X}$: state space

$\mathcal{X}$: random variable

X'_{ND}: total concentration of entrapped organic nitrogen

Y: yield coefficient

$\mathbf{y}$: output vector

Z: sample space $= V \times W$

Greek letters:

α: the α-mixing function

β: a positive real number

γ: real-valued function defined as $\gamma(s) = \sqrt[s]{\frac{1}{2}\left(\frac{s-1}{s-2}\right)^{s-1}}$

Γ: parameter space

δ: deviation measure on the real line

Δ: variation of a quantity

ϵ: process noise

ζ: real-valued function defined as $\zeta = 4\frac{\left[q\left(\ln\left(\frac{2N}{q}\right)+1\right)-\ln\left(\frac{\eta}{4}\right)\right]}{N}$

η: value of the probability of an event

η: energy conversion efficiency factor

ϑ: information conversion efficiency function

κ: ratio of M and empirical risk

μ: maximum specific growth rate

ξ: squared difference between model prediction and process response

ϖ: rate of process model improvement

ς: statistical dependence coefficient

τ: bound used in weak prior information 2

ϕ: real-valued function defined as $\phi(N) = 1 + \frac{\kappa\zeta(N)}{2}\left(1 + \sqrt{1 + \frac{4}{\kappa\zeta(N)}}\right)$

φ: guaranteed deviation function

Υ: a set of data

Ω: fraction of a substrate group concentration in X_{COD}

Ω': fraction of a substrate group organic nitrogen concentration in X'_{ND}

Subscripts:

$^\circ$: refers to a set that does not contain 0

∞: infinity

$+$: refers to positive numbers

A: refers to autotrophic

B: refers to biomass

asp: refers to ASP

avg: refers to average

cl: refers to close

CF: refers to continuous functions

dep: refers to dependent

emp: refers to empirical

est: refers to estimate

H: refers to heterotrophic

i: index

I: refers to inert

in: refers to influent

k: index

l: index

L: integer

loc: refers to local

min: refers to minimal

mrg: refers to marginal

n: index

ND: refers to organic nitrogen

NH: refers to ammonia nitrogen

NO: refers to nitrite/nitrate nitrogen

P: refers to particulate

pt: point

r: refers to recycle

Superscripts:

∞: infinity

$+$: refers to positive numbers

$data$: refers to process data

max: refers to maximum

$meas$: refers to measured variables

T: transposed vector

Appendix B

The Weak Prior Information: can it be circumvented?

The author was very much excited at the beginning when he started implementing the results of computational learning theory, because of the ambition of this theory to assume no prior information about the probability density function. Later on, he realized that a certain weak prior information $\mathcal{WPI}$ was still needed in order to be able to carry out any meaningful work. Then the question that arose was: would it be possible to re-construct another mathematical framework that assumes absolutely no prior information about the type of uncertainty underlying the physical system? The conclusions of the author's investigation is that not only the answer to the foregoing question is *no*. but any framework that claims to not require prior information to analyze uncertainty is simply an absurd one. James O. Berger (1985) reported this fact in very simple and concrete terms:

"$\cdots$ *The second source of non-sample information that is useful to consider is called prior information. This is information* $\cdots$ *arising from sources other than the statistical investigation.* $\cdots$

A compelling example of the possible importance of prior information was given by Savage (1961). He considered the following three statistical experiments:

1. *A lady, who adds milk to her tea, claims to be able to tell whether the tea or the milk was poured into the cup first. In all of ten trials conducted to test this, she correctly determines which was poured first.*

2. *A music expert claims to be able to distinguish a page of Haydn score from a page of Mozart score. In ten trials conducted to test this, he makes a correct determination each time.*

3. *A drunken friend says he can predict the outcome of a flip of a fair coin. In ten trials conducted to test this, he is correct each time.*

In all three situations, the unknown quantity θ is the probability of the person answering correctly. A classical significance test of the various claims would consider the null hypothesis (H_0) that $\theta = 0.5$ (i.e., the person is guessing). In all three situations this hypothesis would be rejected with a (one-tailed) significance level of 2^{-10}. Thus the above experiments give strong evidence that the various claims are valid.

In situation 2 we would have no reason to doubt this conclusion. (The outcome is quite plausible with respect to our prior beliefs). In situation 3, however, our prior opinion that this prediction is impossible (barring a belief in extrasensory perception) would tend to cause us to ignore the experimental evidence as being a lucky streak. In situation 1 it is not quite clear what to think, and different people will draw different conclusions according to their prior beliefs of the plausibility of the claim. In these three identical statistical situations, prior information clearly cannot be ignored."

Hence, when we develop uncertainty models, our main concern must be how weak is the prior information on which these models are based, and *not* whether it is possible to completely eliminate such prior information.

Bibliography

Ahmed, N. U. (1988), *Elements of finite-dimensional systems and control theory*, Longman Scientific & Technical.

Anthony, M.H.G., Biggs, N. (1992), *Computational Learning Theory: An Introduction*, Cambridge University Press, Cambridge.

Barlett, P.L., Long P.M., Williamson, R.C., (1994), "Fat-shattering and the learnability of real-valued functions", *Proceedings of the 7th Annual ACM Workshop on Computational Learning Theory*, 299-310, ACM Press.

Beck, M.B. (1986), "Identification, estimation and control of biological wastewater treatment processes", *Proc. Inst. Elec. Eng.*, **133**, 254-264.

Berger, J.O. (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, NY.

Berger, M.A. (1993), *An Introduction to Probability and Stochastic Processes*, Springer-Verlag, NY.

Benefield, L.D., Randall, C.W., (1980) *Biological Process Design for Wastewater Treatment*, Prentice Hall, NJ.

Berthouex, P.M., Hunter, W.G., Fan, R. (1985), "Characterization of treatment plant upsets", in *Instrumentation and Control of Water and Wastewater Treatment and Transport Systems*, edited by Drake, R.A.R. Pergamon Press, London, England, 287-296.

Bidstrup, S.M., Grady, C.P.L. (1987), *A User's Manual for SSSP*, Clemson University, SC.

Borrie, J.A. (1992), *Stochastic Systems for Engineers: modeling, estimation, and control*, Prentice Hall, NY.

Box, G.E.P., Jenkins, G.M. (1970), *Time series analysis, forecasting and control*, Holden-Day, San Francisco.

Bradley, R.C. (1986), "Basic properties of strong mixing conditions", in *Progress in Probability and Statistics*, edited by Eberlein, E. and Taqqu, M.S., Birkhäuser, Boston.

Capodaglio, A., Novotny, V., Fortina, L. (1992), "Modelling wastewater treatment plants through time series analysis", *Environmetrics*, **3**, 1, 99-120.

Cohen, J., Stewart, I. (1994), *The Collapse of Chaos*, Viking (Penguin Group), NY.

Côté, M., Grandjean, B.P.A., Lessard, P., Thibault, J. (1995), "Dynamic modelling of the activated sludge process: improving prediction using neural networks", *Water Research*, **29**, 4, 995-1004.

Dold, P.L., Ekama, G.A., Marais, G.v.R. (1980), "A general model for the activated sludge process", *Prog. Wat. Tech.*, **12**, 47-77.

Dold, P.L., Marais, G.v.R. (1986), "Evaluation of the general model activated sludge model proposed by the IAWPRC Task Group", *Wat. Sci. Tech.*, **18**, 6, 63-89

Doukhan, P. (1994), *Mixing : properties and examples*, Springer-Verlag, NY.

Environment Canada, 1997.

Fossard, A.J., Normand-Cyrot, D. (1993), *Systèmes non linéaires*, Masson, Paris, France.

Gall, R.A.B, Patry, G.G. (1988), *Knowledge-Based Systems in Wastewater Engineering*. McMaster File n° 5-52618. DSS Project: KE405-7-6012/01-SE. McMaster University, Department of Civil Engineering.

Gallant, S.I. (1993), *Neural Network Learning and Expert Systems*, MIT Press, Cambridge, Mass.

Grady, C.P.L., Lim, H.C. (1980), *Biological Wastewater Treatment*, Marcel Dekker, NY.

Greenwald, D. (1994), *The McGraw-Hill Encyclopedia of Economics*, McGraw-Hill, NY.

Guergachi, A. (1995), "Neuro-fuzzy control of the activated sludge process", Ph.D. research proposal, University of Ottawa, Ottawa, Canada.

Haussler, D. (1992), "Decision theoretic generalizations of the PAC model for neural net and other learning applications", *Information and Computation*, **100**, 1, 78-150.

Haykin, S. (1994), *Neural Networks, a Comprehensive Foundation*, Macmillan, NY.

Henze, M., Grady, C.P.L., Gujer, W., Marais, G.v.R. and Matsuo, T. (1987), *Activated Sludge Model n° 1*, IAWPRC, London, England.

Hiraoka, M., Tsumura, K. Fujitsa, I., Kanaya, T. (1990), "System identification and control of activated sludge process by use of autoregressive model", in *Instrumentation, Control and Automation of Water and Wastewater Treatment and Transport Systems*, edited by Biggs, R., Pergamon Press, London, Great Britain, 121-128.

Honavar, V., Uhr, L. (1994), *Artificial Intelligence and Neural Networks : steps toward principled integration*, Academic Press, Boston.

Huang, J.Y.C. (1981), "Microorganism viability", in *Application of on-line Analytical Instrumentation to Process Control*, edited by Arthur, R.M. Arthur Technology and Ann Arbor Science, 81-103.

Hydromantis, Inc. (1995), *GPS-X Technical Reference*, Hamilton, ON, Canada.

Jeppsson, U. (1993), *On the Verifiability of the Activated Sludge System Dynamics*, Licentiate's thesis, Lund Institute of Technology, Department of Industrial Electrical Engineering and Automation, Lund, Sweden.

Jeppsson, U., (1996), *Modelling Aspects of Wastewater Treatment Processes*, Lund Institute of Technology, Department of Industrial Electrical Engineering and Automation, Lund, Sweden.

Konikow, L.F., Bredehoeft, J.D. (1992), "Ground-water models cannot be validated", *Adv. Water Resour.*, **19**, 75-83.

Lens, P.N., Verstraete, W.H. (1992), "Aerobic and anaerobic treatment of municipal wastewater", in *Profiles on Biotechnology*, edited by Villa, T.G. and Abalde, J. Servicio de Publications, Universidade de Santiago, Santiago de Compostella, Spain, 333-356.

Li, M., Vitányi, P. (1993), "*An introduction to Kolmogorov complexity and its applications*", Springer-Verlag, NY.

Madigan M.T., Martinko, J.M., Parker, J. (1997), "*Biology of microorganisms*", Prentice Hall, NJ.

Marsili-Libelli, S. (1989), "Computer control of the activated sludge process", in *Encyclopedia of Environmental Control Technology - Wastewater Treatment Technology*, Volume 3, edited by Cheremisinoff, P.N., 229-270.

Metcalf and Eddy, Inc. (1991) *Wastewater Engineering*, McGraw-Hill, NY.

Monod, J. (1942), "Recherche sur la Croissance des Cultures Bacteriennes", *Herman et Cie*, Paris, France.

Mybeck, P.S. (1979), *Stochastic Models, Estimation and Control*, Academic Press, NY.

Nolasco, D.G., Patry, G.G., Olsson, G. (1989), "Knowledge Acquisition from Simulation Models", *6th Conf. on Computer Applications in Civil Eng.*, Atlanta, USA.

Novotny, V., Capodaglio, A.G. (1992), "Strategy of stochastic real-time control of wastewater treatment plants", in *Applying Instrumentation and Automation in Environmental Engineering: Water and Wastewater*, edited by Alleman, J.E., Sweeney, M.W. and Vaccari, D.A., *ISA transactions*, **31**, 1, 87-96.

Savage, L.J. (1961), *The subjective basis of statistical practice, Technical report*, Department of Statistics, University of Michigan, Ann Arbor

Slotine, J.-J. E, Li, W. (1991), *Applied nonlinear control*, Prentice Hall, NJ.

Smith. M. (1993). *Neural Networks for Statistical Modeling.* Van Nostrand Reinhold. NY.

Stark. H., Woods. J.W. (1994). *Probability. Random Processes. and Estimation Theory for Engineers.* Prentice Hall. NJ

Tiessier. C. (1936). *Ann. Physiol. Physicochem Biol..* **12.** 527.

Vapnik. N. (1982). *Estimation of Dependencies Based on Empirical Data.* Springer-Verlag. NY.

Vapnik, N. (1995). *The Nature of Statistical Learning Theory.* Springer-Verlag. NY.

Vapnik, N. (1998). *Statistical Learning Theory.* Wiley. NY.

Vapnik, N., Chervonenkis, A.Ya. (1968). "On the uniform convergence of relative frequencies of events to their probabilities". *Soviet Math. Dokl..* **9**, 915-918.

Vapnik, N., Chervonenkis. A.Ya, (1981). "Necessary and sufficient conditions for the uniform convergence of the means to their expectations". *Theory Prob. Appl..* **26**. 532-553.

Vapnik, N.. Chervonenkis. A.Ya. (1991). "The necessary and sufficient conditions for consistency of the method of empirical risk minimization". *Pattern Recogn. Image Anal..* **1** (3). 284-305.

von Altrock. C. (1995). *Fuzzy logic and NeuroFuzzy applications explained.* Prentice Hall. NJ.

Wenocur. R.S.. Dudley. R.M. (1981). "Some special Vapnik-Chervonenkis classes". *Discrete Math..* **33**. 313-318.

White. H. (1984). *Asymtotic Theory for Econometricians.* Academic Press. Orlando.

White. H. (1990). "Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings". *Neural Networks.* **3**. 535-549

White, H., Wooldridge, J. (1990), "Some results on sieve estimation with dependent observations", in *Nonparametric and semi-parametric methods in econometrics and statistics*, Cambridge University Press, NY.

Yager, R.R., Zadeh, L.A. (1994), "Fuzzy Sets, Neural Networks, and Soft Computing", Van Nostrand Reinhold, NY.

Zadeh, L.A. (1965), "Fuzzy sets", *Information and Control*, 8, 338-353.

Zheng, C., Bennett, G.D. (1995), *Applied Contaminant Transport Modeling - Theory and Practice*, Van Nostrand Reinhold, NY.