Université de Lille

Thèse de doctorat

Agrégation et méta-modélisation de procédés de méthanisation : théorie et algorithmes

Thèse de doctorat de l'Université de Lille
préparée à SUEZ

École doctorale n°631 Mathématiques-Sciences du numérique (MADIS)
Spécialité de doctorat: Mathématiques pures et appliquées.

ANTOINE PICARD

Thèse soutenue le 3 avril 2025

Composition du Jury :

Gérard Biau Professeur, Sorbonne Université	Président
Christine Keribin Professeure, Université Paris Saclay	Rapporteur
Nicolas Chopin Professeur, ENSAE	Rapporteur
Céline Hudelot Professeure, CentraleSupélec	Examinateur
Jean-Philippe Steyer Directeur de Recherche, Inrae	Examinateur
Pierre Gaillard Chargé de Recherche, Inria	Examinateur
Benjamin Guedj Directeur de Recherche, Inria et University College London	Directeur de thèse
Roman Moscoviz Ingénieur de recherche, SUEZ	Encadrant de thèse
Vincent Schmitt Ingénieur de recherche, SUEZ	Invité

Acknowledgements

Garbage in, methane out. In between, a numerical twin, PAC-Bayes generalisation bounds, surrogate risks, JIT compilation, uncertainty quantification assessment, and many things I never suspected existed not that long ago. Overall, one thesis, two advisors, three years work, too many lines of code to count, too many scraps of notes to remember, and much compute time. All of this would not have been possible without the kind help and support of many.

First, I want to thank my fellow PhD inmates. Many thanks to you, Maxime Haddouche, Antonin Schrab, Antoine Vendeville, Reuben Adams, Szilvia Ujváry, Chloé Hashimoto-Cullen, Mathieu Alain, Fredrik Hellström, Eugenio Clerico, for your never failing kindness, and for helpful discussions, whether on statistics, maths, literature, tea, opera or simply plain gossip, without which many pleasant lunches would have been dull indeed.

I warmly thank my two PhD advisers, Roman Moscoviz and Benjamin Guedj, for both professional and friendly support during these three (and a half) years. Their inputs and encouragements proved invaluable.

My colleagues and ex-colleagues from across the Channel at SUEZ, who had to bear with me and my weird Kullback–Leibler and f-divergences formulas repeated endlessly on the blackboard (once white, now a greyish blue from overlapping layers of unremovable ink), should also be much thanked. That includes of course all of SUEZ's data team, in chronological order Jean-Michel Rodrigues, Guillaume Cussonneau, Gilles Faÿ, Claire Mathieu, Marc-Antoine Giuliani, Sira Ferradans, Camille Balmond Leblanc, Stamatina Georgiou, Ilham Laatarsi, Cuichen Zhang, Jason Enguehard, Victor Bouvier, Soufiane Eddamani, Olaf Kouamo, Pierre Rebillat, Grégoire Goujon, Émilie Bouaoula; but also our kind colleagues sharing our coffee-machine, Laurent Galtier, Baptiste Borie, Zdravka Doquang; and, it goes without saying, all my colleagues from the BRS, Marion Crest, Alexandre Mallet, Francesco Novellis, Valentin Larzilliere, Silvio Riggio, Florian Paillet, Maxime Rouez, Patricia Camacho, Felipe Guilayn, Élise Crestey. I wish to thank Roman a second time for ensuring that my integration in the firm went so smoothly, his constant availability and dedication to the thesis. A special thanks also to you, Cyril Marcillhac and Mathieu Haddad, not forgetting the LBE team, for pleasant memories of the IWA AD conference last June.

I wish to express special gratitude to Vincent Schmitt, who, in addition to being a capital fellow and a tireless big-bug-hunter always willing to start the chase at literally any hour of the day and night, muscled through an early version of this manuscript and conquered many typos,

syntax problems and other fiendish mistakes.

Blessed be also those who brought me into statistics. Looking back, statistics was not at all what I expected I would do. Even when not thinking of film-making and when mathematically-minded, I did not use to take statistics for serious. This all changed thanks to three people: first Raphael Cerf, who introduced me to the beauties of probability; then Stéphane Boucheron and Claire Boyer, who made me look at statistics as the fascinating inverse problem of probability it is. Many thanks also to Matthias Löffler and Richard Nickl who put the finishing touch to this conviction.

I want to thank the members of the jury for accepting to take part in it. I sincerely hope that they will not come to regret their decision after their perusal of the manuscript is completed!

A thesis is quite a lot of work, which means less time for friends and family than one could wish. I want to thank my friends, who all rallied round during these three years: friends from secondary school, Eric Toutounji, Shu Okabe, Ulysse Silva, Julien Portal, Antoine Barre and Dmitry Chernyak; friends from later on, Arnaud Eteve, Giacomo Martinelli, Nieves Ugarte, Raphael Cousin, Alice Andres, Valentin Guergueb. Thanks also to Gwendolyn, John, Algernon and Canon Chasuble, better known off the stage as Diane-Iris Ricaud, François Trublereau, Avery Colobert and Éloi Massoulié, not forgetting Adrien Plaine, Hugo Lacoue-Labarthe, Maxime Chabert and the Brouillaud family, for joining me on this mad theatrical project three years ago. Hopefully, we will all be back on the stage together any time now.

A million thanks also to my parents, grandparents, and in particular to my wife, Manon, for constant support and love all these years! Words fail to express my gratitude.

Finally, as all recognition must not go to the living, I want to express special thanks towards P. G. Wodehouse and Jonathan Cecil for always providing me with a laugh when I've needed one.

Abstract

Anaerobic digestion technologies play a key role in the transition to cleaner energy by converting organic waste into biogas which can replace fossil natural gas. To optimize anaerobic digestion processes, a thorough understanding of the biological and physicochemical mechanisms is required. Biochemical computational models are a step in this direction but are prone to overfitting and are often specific to the process they are calibrated for, limiting their predictive power.

The PAC-Bayes framework inherently limits the risk of overfitting through its reliance on trainable generalisation guarantees. As an extended Bayesian approach, it moreover provides the opportunity to incorporate useful expert knowledge in the learning process, and inherently provides uncertainty quantification, albeit with no theoretical guarantees. However, the correlated nature of the data and the high computational cost of the simulation make the application of PAC-Bayes to anaerobic digestion challenging and require adaptation.

In its current state, expert knowledge is insufficient to construct meaningful models for new anaerobic digestion units, for which no data is available. Meta-learning strategies offer a promising way to enhance this expert knowledge by leveraging information across multiple anaerobic digestion units. By learning common features across different units, meta-learning could result in more robust models, especially for units when few, or no data is available. Previous PAC-Bayes meta-learning algorithms are however computationally intensive, and ill-fitted to the high simulation time of AD processes.

This thesis focuses on developing efficient methods for applying PAC-Bayes theory and meta-learning approaches to anaerobic digestion modelling. It contributes theoretical insights into PAC-Bayes, presents efficient implementations of both PAC-Bayes and meta-learning strategies, and provides experimental assessments of the resulting models' predictive power and uncertainty quantification.

Contents

From Meta-Learning to Methane production

Meta PAC-Bayes learning for Anaerobic Digestion

As part of the shift to cleaner energy, **A**naerobic **D**igestion (AD) technologies are a valuable waste-to-energy process. AD transforms organic waste into biogas composed mostly of methane and carbonic gas. The methane produced can be substituted to fossil natural gas. As such, AD processes play a part in France's national low-carbon strategy [Ministère de la Transition Écologique et Solidaire, 2020, Orientation D3]. A wide range of micro-organisms are involved in AD, spanning multiple functions (*i.e.* biochemical reactions they catalyse), various optimal operating conditions (*e.g.* medium or high temperatures), and sensitivity to change of operating conditions (*e.g.* pH, temperature). As such, making the most out of AD's potentialities requires solid understanding of the biological and physicochemical mechanisms that govern the process. Yet, at the industrial level, AD plants are more often than not operated using a collection of empirical rules built over time and do not take into account the biological complexity of the process [Carballa et al., 2015, De Vrieze et al., 2017]. This approach succeeded in building and operating many plants with satisfactory results, but also presents serious shortcomings. These rules:

- are conservative, leading to the design of oversized plants requiring high initial expenditures;
- can not be used to operate digesters in a flexible way, and notably do not take into account the adjustment of microbial kinetics to the type of substrate. This can lead to unnecessarily high operating costs (*e.g.* inadequate heating of the tanks) or loss of returns (*e.g.* perfectible biogas production);
- must be redefined for all new process design and new type of substrate. This necessitates an intensive research and development phase, usually performed via trial and error, which is costly and time consuming.

Biochemical models, incorporating knowledge on the biological and physicochemical mechanisms at play during AD, can attenuate the dependence on trial and error by relying on scientific

literature knowledge. The popular **A**naerobic **D**igestion **M**odel 1 (ADM1) introduced in Batstone et al. [2002a] has proved its relevance for a wide range of applications [Batstone and Keller, 2003, Flotats et al., 2006, Kalfas et al., 2006, Derbal et al., 2009, Couto et al., 2019, Donoso-Bravo et al., 2020, Li et al., 2021]. It however involves a large number of parameters, a sizeable fraction of which must be properly calibrated for the model to give an adequate representation of a given AD unit [Dittmer et al., 2021]. From a statistical point of view, the large number of parameters and flexibility of the model raise the potential issue of overfitting, which would limit the predictive power of the model. Moreover, the calibrated parameter values of models such as ADM1 are generally valid only for the specific AD process considered, limiting the possibility to extrapolate to other, or future AD units.

To fulfil modelling's goal of helping designers and operators of AD units, it is necessary to ensure that the model has predictive power and to limit the risk of overfitting. To this end, the use of stochastic models can bring valuable insights on the range of plausible predictions. Compared to deterministic models, stochastic models do not have to choose between equally well-performing models. This, in the case of AD monitoring, might imply the difference between foreseeing, or not foreseeing, a costly process failure in the near future. A principled way to construct such stochastic models is Bayesian statistics. In short, Bayesian statistics transforms prior beliefs on the most likely models into *a posteriori* beliefs through confrontation with data. This confrontation mechanism is performed through statistical modelling, which informs on how likely it is to observe the given data for each model. However, the uncertainty quantification brought by Bayesian methods is only valid under strong conditions on the model's validity, which are not satisfied for AD processes. In the realm of approximate Bayesian algorithms, PAC-Bayes strategies come with the benefit of theoretical guarantees on the model's predictive power even under model misspecification. While the typical hypotheses under which the guarantees hold, such as independent data and bounded loss, are not satisfied for AD processes, previous works (see section 1.2.3) hint that these hypotheses can be to some extent relaxed.

The starting point of Bayesian inspired methods is the prior belief, which formalizes the modeller's beliefs on the different models' plausibility. In the context of AD modelling, these beliefs can be guided by the vast literature on AD, from which ranges of model parameters value can be inferred. PAC-Bayes calibration benefits from this information, by excluding impossible models. Still, building this prior belief purely on literature reports has some limitations. First, the methodology used to measure or estimate parameter values varies between authors. Notably, modellers use diverse calibration methodologies or modify part of the computational model for specific AD set ups, leading to incomparable results. Second, the reference ranges of parameter values are typically quite wide [Rosén and Jeppsson, 2006]. As a result, the uncertainty of stochastic AD models relying solely on prior beliefs is too large to be usable, preventing its use for AD plant design. Meta-learning techniques [Pentina and Lampert, 2014, Finn et al., 2017] provide a principled way to construct prior beliefs from multiple digestion units, with the potential to overcome these shortcomings. Rather than calibrating AD units in an independent fashion, meta-learning considers a two stage learning framework where common features are learnt globally (meta-level), and specificities to each AD units are learnt individually (inner-level). In the

context of AD PAC-Bayes learning, meta-learning techniques could be applied to learn the prior beliefs from multiple digestion units, with the potential to learn correlations between different parameters and to provide more robust task-specific models. Ideally, supplementary details on the AD unit's specifications (such as the range of temperatures, the nature of substrates) could be leveraged to construct AD plant specific prior beliefs at the meta level, leading to low prediction uncertainty even without calibration.

We strive in this thesis to build a set of learning tools which can:

- convert deterministic biochemical models into stochastic models, bringing meaningful insights on the model uncertainty;
- optimise the training process for individual AD unit modelling task;
- aggregate the insights brought by the individual AD unit models into a global model through meta-learning.

Before reaching this goal, some difficulties have to be addressed. First, PAC-Bayes theory has been mostly focused on the classification-like setting, assuming independent data and bounded loss. On the other hand, AD's data is correlated, and might suffer from distribution shift (*e.g.* due to a change of substrate). Bridging the gap between PAC-Bayes assumptions and AD's data is necessary to obtain theoretically valid guarantees. Second, the uncertainty quantification inherently obtained from PAC-Bayes trained stochastic models comes with little to no theoretical guarantee[1]. Assessing whether this uncertainty quantification is meaningful in the setting of AD model calibration is therefore necessary if insights are to be drawn from it. Third, more complex AD models can be computationally intensive, calling for the design of simulation efficient PAC-Bayes calibration algorithms. Fourth, real world AD data might greatly differ from unit to unit (*e.g.* different sensors). This could hurt the meta-learning procedure, which assumes similar tasks. Finally, the computational cost of the AD models also have a direct impact on the meta-learning algorithm. These algorithms typically require multiple calibration of the same tasks, which, even for simulation efficient calibration algorithms, might prove impracticable.

Outline of the thesis

Chapter 1 provides a literature review on the three main fields of this thesis, Anaerobic Digestion, PAC-Bayes learning and Meta-Learning. Chapter 2 details theoretical contributions to PAC-Bayes, and can be read independently. Chapters 3 to 5 tackle the construction of a PAC-Bayes learning algorithm for AD and the properties of the resulting posterior distribution. Chapter 3 studies the uncertainty quantification provided by PAC-Bayes posterior on AD calibration tasks. Chapter 4 describes how the PAC-Bayes learning process can be efficiently sped up for AD modelling by improving both model and algorithm's efficiency. Chapter 5 adjusts the

[1] Apart from the case when PAC-Bayes coincides with Bayes, see Germain et al. [2016].

learning strategy to tackle real world digester data in a unified way, and to provide online monitoring. Finally, Chapter 6 presents how meta-learning can be used for AD modelling, notably in conjunction with our PAC-Bayes learning algorithm.

The output of the thesis consists in:

Papers

- On change of measures inequality for f-divergences (*arxiv preprint*), with Benjamin Guedj [Picard-Weibel and Guedj, 2022] (covered in section 2.1);
- Bayesian uncertainty quantification for Anaerobic Digestion models (Bioresource Technology, 2024), with Gabriel Capson-Tojo, Benjamin Guedj and Roman Moscoviz [Picard-Weibel et al., 2024a] (covered in chapter 3);
- Learning via Surrogate PAC-Bayes (NeurIPS, 2024), with Benjamin Guedj and Roman Moscoviz [Picard-Weibel et al., 2024b] (covered in section 4.1);
- Predicting municipal sludge volatile solids reduction in industrial anaerobic digesters based on 32 years of cumulated data (under review at the Bioresource Technology, 2025), with Danielle Trap, Damien Batstone, Roman Moscoviz and Mathieu Haddad (covered in section 6.1);
- How Good is PAC-Bayes at explaining generalisation? (under review at COLT 2025), with Eugenio Clerico, Roman Moscoviz and Benjamin Guedj [Picard-Weibel et al., 2025] (covered in section 2.2).

Implementation The following five[2] python packages were developed during the thesis:

- anaerodig (`https://pypi.org/project/anaerodig/`), containing abstract classes for Anaerobic Digestion models as well as implementations of ADM1[3] and AM2 models;
- picproba (`https://pypi.org/project/picproba/`), containing abstract classes for probability measure, families of probability measures, various methods and functions for operations on such objects (*e.g.* KL computation, integration, tensorization of measures), subclasses for Exponential Families and implementations for some specific measures and family of measures (*e.g.* Gaussian, Gamma, Exponential);
- picoptim (`https://pypi.org/project/picoptim/`), containing abstract classes for optimisation routines, and implementations of specific routines (notably a modified CMA-ES algorithm [Hansen, 2016]);

[2] A sixth package, apicutils (`https://pypi.org/project/apicutils/`), contains various helpers shared throughout the packages.

[3] The ADM1 implementation is based on Sadrimajd et al. [2021] and has been thoroughly modified for code efficiency (code cleaning, memory management, JIT compilation - see Section 4.2) and structure.

- picpacbayes (`https://pypi.org/project/picpacbayes/`), containing abstract classes for PAC-Bayes bound minimisation routines and the implementations of the routines introduced in this thesis (VarBUQ, SuPAC-CE);
- picmeta (`https://pypi.org/project/picmeta/`), containing classes for the PAC-Bayes meta-learning approach introduced in this thesis (see Section 6.2).

In addition to these packages, implementations for experiments can be found in Github repositories associated with publications.

Chapter 1

Literature review

This chapter provides a review of the literature in the three main fields of interest in this thesis: Anaerobic Digestion, PAC-Bayes, and Meta-Learning.

Contents

1.1 Anaerobic Digestion

1.1.1 Overview of Anaerobic Digestion

Creating value from organic waste

Anaerobic **D**igestion, to quote Moletta [2015], is the process which "transforms organic waste into biogas, composed for the most part of methane and carbonic gas, by a consortium of microorganisms under anaerobic conditions" (*i.e.* without oxygen). Anaerobic digestion spontaneously occurs in many environments, including swamps, animal digestive tracts, wherever there is no oxygen, organic matter, and micro organisms can develop.

The biogas produced by AD contains an important fraction of methane (CH_4), typically 50 to 75 percent of the gas. Methane is a combustible gas, and as such can be used as a source of energy. In uncontrolled environments, the methane produced by the decomposition of organic waste through AD is lost in the atmosphere, and, being a greenhouse gas, contributes to climate change. Controlled AD units, on the other hand, capture the biogas produced. This biogas can, after treatment, be injected into the natural gas grid, or generate electricity.

A short history of Anaerobic Digestion

The first record of the fact that combustible gas is produced by decaying matter appears to have been made by Van Helmont in 1630 [Gunnerson et al., 1986, Abbasi et al., 2011]. The first large scale AD applications date from the nineteenth century. The first anaerobic digester tank built in 1859 in Mumbai, India by a leper colony [Marsh, 2008, Moletta, 2015], with gas generated from human waste being used for lighting. Exeter (UK) became, in 1897, the first city to apply AD for the treatment of all the city's wastewater, with part of the methane produced reused for heating and lighting. AD processes were at first used as an efficient wastewater treatment process. Through AD, diluted and unwanted organic waste is converted into gas which separates from the wastewater. AD proved to be an efficient technology for removing organic waste from wastewater, removing as much as ten times more pollution than an aerobic system for a tank of similar size [Moletta, 2015]. [Gunnerson et al., 1986] reports that AD systems were adjusted to capture the biogas from 1920.

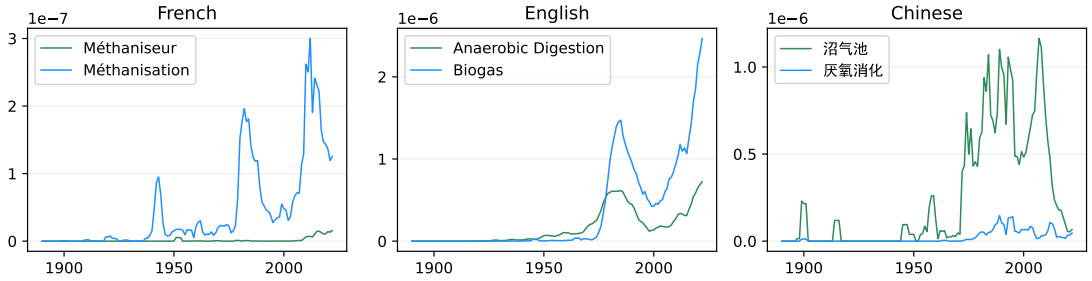


Figure 1.1: Evolution of the occurrence of Anaerobic Digestion related terms between 1890 and 2024, obtained through Google NGram Viewer, using French, English and Chinese terms (french terms: "Anaerobic Digestion plant", "Anaerobic Digestion"; chinese terms: "Anaerobic Digestion plants" (smaller scale units), "Anaerobic Digestion"). All graphs highlight the impact of the 1970s energy crisis in the raising interest in Anaerobic Digestion. A second wave of interest in French and English speaking countries starting in the early 2000s is also noticeable.

The development of AD varied considerably between countries depending on economic conditions (see Figure 1.1). In China, AD technologies were introduced in the 1920s, remained stable until 1972 (1300 digesters), then rapidly grew during the next decade due to strong government support to approximately seven million small scale digesters in order to replace firewood as fuel due to shortage [Gunnerson et al., 1986]. Similarly, biogas production through AD was driven in Thailand by the energy crisis of 1973, with eventually 300 digesters, mostly small 5 cubic meters sizes, part of which quickly fell in disuse [Gunnerson et al., 1986].

In Germany, 9000 AD plants were in operations, accounting for 80% of biogas projects in Europe Wang et al. [2022]. As of 2018, Germany was the leading biogas producer in the world[1], with 32 TWh of electricity accounting for 17% of Germany's total electricity generation from renewable source, and 17 TWh of heat generated in 2017 through biogas [Daniel-Gromke

[1] It accounted for 15.8% of the world's production in 2015, Xue et al. [2020]

et al., 2017].

In France, about 1400 AD plants were operated in 2022 [Thual et al., 2023], with half operated by farms [Gerber, 2022]. AD plants contributed 9 TWh of gas (2.5% of the annual consumption of France) and 2.5 TWh of electricity (0.6% of the national production). While the production is rapidly increasing, this represents only a fraction of the potential production of energy through renewable gases, which is estimated between 90 to 130 TWh [Thual et al., 2023].

Anaerobic Digestion process

Anaerobic Digestion involves four stages of reactions [Moletta, 2015]:

- Hydrolysis, where long chain polymers of the organic matter such as proteins, lipids, carbohydrates are decomposed into monomers;
- Acidogenesis, where the monomers produced by hydrolysis are further decomposed into volatile fatty acids, hydrogen, and carbonic gas;
- Acetogenesis, where the compounds are further decomposed into precursors of methane, *i.e.* acetate ($C_2H_3O_2$), hydrogen and carbonic gas;
- Methanogenesis, where acetate, hydrogen and carbonic gas react to produce methane, through two pathways: $C_2H_3O_2 + H_2O \longrightarrow CH_4 + HCO_3$, where acetate is decomposed in methane and bicarbonate (*acetotrophic* pathway), $4\,H_2 + CO_2 \longrightarrow 2\,H_2O + CH_4$, where hydrogen and carbonic gas combine to form methane (*hydrogenotrophic* pathway).

Each of these reactions are catalysed by specific microbial communities whose metabolism degrades the input (substrate) into the output. The speed (rate) at which the reactions occur and the communities develop depends on the bio physical conditions of the digestate, *e.g.* whether the reaction catalysed is thermodynamically feasible, the temperature, but also on factors impacting the micro organisms such as acidity, ammonia concentration or presence of heavy metals.

Optimising Anaerobic Digestion processes

The goal of an aerobic digestion unit is to convert as much organic matter as possible into biogas. This can be measured either as a fraction of expression of the **B**iochemical **M**ethane **Potential** (BMP), *i.e* the maximum amount of methane which could be produced from the feedstock used, or as the fraction of **V**olatile **S**olids (VS) removed from the influent by the AD process, the **V**olatile **S**olids **R**eduction (VSR) (the former indicator being more typical for waste AD, and the latter for sewage sludge AD).

This conversion should be as fast as possible (treat more feedstock), using as small a tank as possible (smaller construction cost). Moreover, the operating costs of the digestion process

should be controlled. These are impacted by heating of the AD tank (with mesophilic AD taking place at 30 to 38 degrees Celsius, while thermophilic AD taking place at 50 to 60 degrees Celsius) and operational difficulties, among which foaming issues Barjenbruch et al. [2000], Ganidi et al. [2009] and digester acidification Babel et al. [2004], Siegert and Banks [2005], Alavi-Borazjani et al. [2020]. Digester acidification occurs when the first three steps of the AD process occur at a significantly faster rate than the conversion of acids to methane. In such cases, acids accumulate in the digester, inhibiting microbial community contributing to methanogenesis, eventually resulting in a non reversible decrease of the microbial activity leading to process failure.

Operators and designers of AD plants can influence the process through selection of the feedstock, the feeding rate, tank temperature, pre-treatments, agitation, design and eventually additives[2]. Optimisation of AD process can be performed by better understanding of the way operational parameters impact of the process [Panigrahi and Dubey, 2019]. A way to approach such impact is through computational modelling of the process.

1.1.2 Anaerobic Digestion modelling

Main Anaerobic Digestion models

We briefly present here the main AD computational models which will be considered in this thesis, ADM1, **A**naerobic Digestion **M**odel 2 (AM2) and ProdAD. More details on ADM1 and AM2 can be found in Appendix A. ProdAD being a confidential model, the description of this model will be kept to a minimum. Some relevant metrics (*e.g.* number of parameters, simulation time) on all three models are provided in Table 1.1.

Table 1.1: Metrics of complexity for anaerobic digestion models

	AM2	ADM1	ProdAD[1]
Reactions	3	19	~ 12
State variables	6	28	~ 25
Microbial communities	2	7	9
Number of parameters[2]	5	30	~ 70
Computation time (s/simulated year)[3]	0.65	1.7	2.0

[1] Depending on the AD plant specificities, some of the actual metrics reported here for ProdAD can vary.
[2] The number of parameters reported here is the number of calibrated parameters. AM2 and ADM1 involve other parameters which are typically not calibrated (e.g. liquid-gas equilibrium parameters). The overall number of parameters would be 12 for AM2 (77 for ADM1).
[3] These computation times are for our numba Just-In-Time compiled implementations presented in 4.2, considering a single digestion tank. The original implementations were less efficient, at about 5s/year simulated for AM2, 30s/year simulated for ADM1 and 200s/year simulated for ProdAD. Computation time depends on the machine used. For AM2 and ADM1, the computation time can also vary depending on the intrant description. For other implementation, the timing can also greatly differ. For instance, Rosén and Jeppsson [2006] reports a baseline computation time of 1800 seconds/year simulated for ADM1.

[2]This last option being rarely routinely used due to the increase in operational costs.

ADM1 The standard model for anaerobic digestion, ADM1, was introduced by Batstone et al. [2002a] and subsequently steadily used in a large number of publications (*e.g.* Batstone and Keller [2003], Blumensaat and Keller [2005], Jeong et al. [2005], Flotats et al. [2006], Kalfas et al. [2006], Derbal et al. [2009], Mairet et al. [2011], Spyridonidis et al. [2018], Couto et al. [2019], Donoso-Bravo et al. [2020], Zhou et al. [2020], Baquerizo et al. [2021], Li et al. [2021], either straight or with minor adjustments [Ramirez et al., 2009, Weinrich et al., 2021, Mo et al., 2023].

ADM1 models anaerobic digestion as a biochemical reaction networks in a perfectly homogenous tank[3]. An initial disintegration step transforms the complex substrates into three secondary substrates (carbohydrates, lipids, proteins) and an inert fraction. Each secondary substrate undergoes a distinct hydrolysis step which turns them into sugars, amino acids and long chain fatty acids. These are in turn decomposed in four volatile fatty acids (valerate, butyrate, propionate, acetate), carbonic gas and hydrogen. The fatty acids are converted to acetate, and the two methanogenic pathway (through acetate and through hydrogen) contribute to the production of methane and carbonic gas. The resulting biochemical graph model involves 19 reactions, and 7 distinct microbial communities catalysing reactions. Biomass decay is also incorporated into the model.

From the stoichiometry of the reactions, the evolution of compound concentration (in unit of **C**hemical **O**xygen **D**emand (COD) per volume) through AD is modelled through the **O**rdinary **D**ifferential **E**quation (ODE)

$$\frac{\mathrm{dState}}{\mathrm{d}t} = M \times r(\mathrm{State}), \tag{1.1}$$

where State is a vector containing the concentration of the compounds, and where r is the rate of each reaction (*e.g.* how much the reaction advances during δt) and M is the (fixed) stoichiometry matrix. The values of the rates $r(\mathrm{State})$ depend both on the concentrations of the compounds (current state of the digester) and the characteristics of the microbial community which catalyse each reactions, which may vary greatly from one digester to another. These characteristics are summarized in the multi dimensional parameter γ (*i.e.*, $r(\mathrm{State}) = r(\mathrm{State}, \gamma)$). γ details both the maximum growth rates of the reactions, as well as how each reaction are inhibited depending on the state of the digester[4]. Inhibition factors considered include substrate limitation, pH, hydrogen and free ammonia. The function describing the microbial activity limitation by substrate availability is obtained through Monod's equation

$$I^{\mathrm{Monod}}(\mathrm{S}) = \frac{\mathrm{S}}{\mathrm{S} + \mathrm{K_S}}, \tag{1.2}$$

where $\mathrm{K_S}$ is the half saturation growth rate (inferred from γ) and S the substrate concentration. The rate is obtained by multiplying the reaction's maximal growth rate by all inhibition factors.

[3] Recent efforts have been dedicated to tackling non homogenous tanks, by coupling ADM1 with Computational fluid dynamics, see Tobo et al. [2020], Caillet et al. [2023], Dabiri et al. [2023]. These drastically increase the computational cost of simulations, and will not be considered in this thesis.

[4] The parameter γ can also encode some information on the stoichiometry matrix in the case of ADM1 (*e.g.* the decomposition of carbohydrates, lipids and sugars into **V**olatil **F**atty **A**cids (VFAs). Here, these parameters will be left to their default value.

Equilibriums between acid/base and liquid/gas forms of the same chemical compounds are implemented as a **D**ifferential **A**lgebraic **E**quation (DAE) system.

AM2 While ADM1 mostly superseded previous models, its large number of parameters and complexity left some room for simpler models. Amongst those, the earlier AM2 Bernard et al. [2001] proved to be a popular option for modelling AD, and is still used in recent publications (*e.g.* Hassam et al. [2015], Arzate et al. [2017], Sari and Benyahia [2021], Hajji et al. [2023], Campos-Rodríguez et al. [2022], Harker-Sanchez et al. [2024]).

AM2 considers a simplified graph to model AD, consisting in a two steps system. During acidogenesis, the organic substrate S_1 is decomposed into VFA S_2, the acidogenic bacteria X_1 and carbonic gas with fixed stoichiometry. Then the methanogenesis step turns the volatile fatty acid is decomposed into methane, carbonic gas and the methanogenic bacteria, again with fixed stoichiometry. The reaction rates follow Monod kinetics for the acidogenic bacteria, and the three parameter Haldane kinetics

$$I^{\text{Haldane}}(\mathrm{S}) = \frac{\mathrm{S}}{\mathrm{S} + \mathrm{K_S} + \frac{\mathrm{S}^2}{\mathrm{K_I}}} \tag{1.3}$$

for methanogenesis. Liquid gas transfer for carbonic gas is also incorporated in the model.

ProdAD A third AD model, ProdAD, will be considered during the thesis. ProdAD is SUEZ's proprietary model for AD. It relies on a reaction graph of similar complexity to ADM1, and involves a comparable total number of parameters[5]. For confidentiality reasons, this model will not be detailed, with no prejudice to the results presented in this thesis.

Calibration of Anaerobic Digestion models

Complex AD models involve a large number of parameters (more than 70 for ADM1), which must be determined before the model can be used. The model's parameters are specific to a given digester process (*e.g.* a given digester tank in a **W**aste**W**ater **T**reatment **P**lant (WWTP)). Indeed, the microbial community will adapt to the typical feedstock they receive and operating conditions, which differs from plant to plant.

Calibrating an AD model consists in selecting the values of these parameters for a given AD process (be it a single lab size tank reactor or a complex AD plant with multiple tanks). Typically, modellers do not calibrate all parameters of the model, but only a fraction expected to have the highest impact on the process considered. Remaining parameter values are then set to default values. For ADM1, such default values have been investigated by Rosén and Jeppsson [2006] for the AD of sewage sludge, and have become a popular choice for modellers. Some modellers do not calibrate any parameter, and directly used the reference values [Parker, 2005, de Gracia

[5]While only part of ADM1 parameters will be pre-selected for calibration in this thesis, (almost) all of ProdAD's parameters will be systematically calibrated, making the calibration of ProdAD in practice an harder task.

et al., 2006, Patón and Rodríguez, 2019, Monje et al., 2022] - although Baquerizo et al. [2021] indicates that the default model can fail to provide adequate predictions.

The parameters selected for calibration are then either manually adjusted [Li et al., 2021, Zhou et al., 2020, Spyridonidis et al., 2018, van Loosdrecht et al., 2016, Mairet et al., 2011, Derbal et al., 2009, Fezzani and Ben Cheikh, 2009, Blumensaat and Keller, 2005] until the model's predictions matches observations, or obtained by minimising the prediction error on some training data. The prevalent way to compute the prediction error is the **R**oot **M**ean **S**quare **E**rror (RMSE),

$$R(\gamma) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\mathrm{AD}(\gamma)_t - \mathrm{Obs}_t)^2}. \tag{1.4}$$

Other prediction error used include the **M**ean **A**bsolute **E**rror (MAE) [Koch et al., 2010, Weinrich et al., 2021] and mean logarithmic squared error [Wichern et al., 2009]. A wide range of optimisation techniques have been applied to minimise the prediction error, including Nelder-Mead algorithm, the secant method, particle swarm optimisation, genetic algorithm, active set optimisation (see discussion in Section 3.1 and Donoso-Bravo et al. [2011])

Limits of current modelling approach

While calibrated AD models have been validated on test data for many specific use cases (*e.g.* [Fezzani and Cheikh, 2008, Ramirez et al., 2009]), the current approach to AD modelling has some limitations. First of all, the default parameter values have only limited predictive power [Baquerizo et al., 2021], limiting the usefulness of AD models for the design of new plants. This results in empirical rules designed to tackle worst case scenarios, typically leading to less effective, but more stable, oversized, digester design.

Calibrated AD models can suffer from overfitting, *i.e.* obtain predictions adequately matching observations on the training data, but still with little predictive power for future data. This risk is heightened whenever the operational condition or feedstock characteristics differ between train and test data (*distribution shift*). Indeed, AD models can suffer from a lack of practical identifiability, that is to say that some parameter values might not be estimated with high confidence. For instance, the growth rates of AD models are the result of a maximum rate and inhibition factors. If the operating conditions remain stable, inferring whether the observed growth rate combines a high maximum rate with high inhibition or low maximum rate with no inhibition might not be practicable. Hence predictions on how a digester would react to a change of operational condition - leading, for instance, to a higher acid concentration - might not be accurate, potentially leading to poor decisions.

Quantitative information on predictions uncertainty is key to proper decision making when optimising AD processes [Dochain and Vanrolleghem, 2001, Donoso-Bravo et al., 2011]. The most common strategies to quantify uncertainty rely solely on the dataset at hand, and make no use of abundant literature information on the plausible ranges of each parameter values, or on the information brought by other datasets. Bayesian methodologies, on the other hand,

incorporate such information directly into the calibration and uncertainty quantification process, and can therefore provide a useful tool, if so far little explored tool for the modelling of AD[6].

1.2 PAC-Bayes theory

In this section, we introduce the main tool and theory which we will consider for calibrating anaerobic digestion models, **P**robably **A**pproximately **C**orrect (PAC)-Bayes. Bayesian statistics are an elegant framework which combines model calibration and uncertainty quantification. PAC-Bayes naturally extend Bayesian statistics to the learning setting. PAC-Bayes theory is built around the notion of *generalisation bounds*, guarantees on the test performance of the model. We describe how such bounds are constructed and introduce classical PAC-Bayes bounds. We then focus on efforts to overcome critical assumptions necessary to obtain valid PAC-Bayes bound, which are typically not met in the AD setting considered. We conclude by describing how PAC-Bayes bounds are used in practice as learning objectives.

1.2.1 From Bayesian Theory to PAC-Bayes learning

Research in PAC-Bayes theory has considerably grown in the past decade, and PAC-Bayes now in many respects stands as a field of its own, wholly independent from the field of Bayesian statistics. Therefore, introducing PAC-Bayes as a natural offspring of Bayesian statistics rather than as a learning oriented methodology to provide generalization guarantees might appear outdated, or even provocative. Indeed, the comprehensive introduction to PAC-Bayes by Alquier [2024] defers the connection to Bayesian statistics to the final section. We still propose to follow this first course, notably followed by Grünwald [2011], Bissiri et al. [2016], Guedj [2019], Knoblauch et al. [2022], for the following two reasons.

First, the main field of application for PAC-Bayes considered in this thesis is the calibration of AD models. Such models involve physical parameters, for which a natural prior (knowledge on the most likely parameter values) can be constructed. Indeed, the underlying assumption in the Bayesian philosophy of a joint random distribution on a model (here, the specifications of the substrates and species in the digester) and on the observations is *on the whole* valid in the setting considered; the difficulty to overcome is that we do not have access to the true statistical model, but only to an approximation.

Second, while the current focus on PAC-Bayes research appears to be deep learning applications, we argue that there has so far been little evidence indicating that PAC-Bayes generalisation bounds actually improve on the time-tested test bounds (see table 2.2). Indeed, the main assumption that prior distributions in deep learning applications should be data dependent if competitive performances are to be obtained strongly recalls the prevalent train test methodology. The impact of the prior on the PAC-Bayes generalisation guarantees will be discussed in Chapter 2.

[6] Previous work considering Bayesian flavoured routines for AD includes Martin and Ayesa [2010], Couto et al. [2019], Pastor-Poquet et al. [2019]. These works will be discussed in more details in Section 3.5.6.

For these reasons, we will first shortly introduce Bayesian inference, and present how PAC-Bayes extends Bayesian statistics to learning theory. We present how PAC-Bayes hinges on the notion of generalisation bounds, which offer both theoretical guarantees on the performance of randomised predictors and a learning objective to construct such randomised predictors. We describe the general methodology used to construct such bounds, and list some of the most popular bounds in the literature. We then focus on methods used to overcome some restrictive assumptions of classical PAC-Bayes generalisation bounds (assumptions which will typically not be valid in the AD setting considered in this thesis). Finally, we list previous strategies which have been used to train randomised predictors from PAC-Bayes bounds.

This introduction to the PAC-Bayes world leaves out rich parts of the theory, such as the oracle PAC-Bayes bounds developed notably by Catoni [2004], de-randomised bounds such as proposed by Mhammedi et al. [2019] and the online PAC-Bayes approach studied by Haddouche and Guedj [2022]. Such approaches will not be considered in our application to AD[7]. We also defer PAC-Bayes use in Meta-Learning to the Section 1.3.4 concerned with Meta-Learning.

Notations

For a measurable space $\mathcal{A}$, we denote

- $\Sigma_{\mathcal{A}}$ its σ-algebra,
- $\Pi_{\mathcal{A}}$ the set of all probability measures on $(\mathcal{A}, \Sigma_{\mathcal{A}})$,
- $\mathcal{M}_{\mathcal{A}}$ the set of all bounded measures and $\overline{\mathcal{M}}_{\mathcal{A}}$ the set of all bounded signed measures,
- $\mathcal{F}_{\mathcal{A}}$ the set of all real valued measurable functions considering Borel's σ-algebra on $\mathbb{R}$.

For two measurable spaces $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{A} \times \mathcal{B}$ denotes the product measurable space equipped with the product σ-algebra (the smallest σ-algebra containing all products of measurable subsets). $\mathcal{A}^n$ denotes the measurable space $\mathcal{A} \times \cdots \times \mathcal{A}$. Topological spaces $\mathcal{T}$ are implicitly considered as measurable spaces equipped with Borel σ algebra. For a probability measure $\mathbb{P}$ on $\mathcal{A}$, the notation $a \sim \mathbb{P}$ indicates that the random variable a follows the probability distribution $\mathbb{P}$.

For a probability measure $\nu \in \Pi_{\mathcal{A}}$ and any real or vector values ν integrable random variable X, $\nu[X] := \int X(\omega) \mathrm{d}\nu(\omega)$ denotes the expected value of X with respect to ν. With slight abuse of notations, we use $\nu[f] := \int f(X(\omega)) \mathrm{d}\nu(\omega)$ with implicit random variable X when no confusion can arise. Moreover, for a set $A \in \Sigma_{\mathcal{A}}$, $\nu(A)$ denotes $\nu[\mathbb{1}_A]$, where $\mathbb{1}_A$ is the indicator function of the set A, *i.e.* $\mathbb{1}_A(x) = 1$ if $x \in A$, else 0 (in short, $\nu(A)$ is the probability of the set A).

For a measurable function f defined on topological space, we note $\mathrm{supp}(f)$ the essential support f (*i.e.* the intersection of all closed sets such that f is null on the complement).

[7] Online PAC-Bayes could be considered for AD, considering the online nature of the observations. But the main advantage of Online PAC-Bayes is its use on a sequence of priors (inductive bias) dependent on past data. On the other hand, our approach learns a global prior from multiple data sources; which is incompatible with the online prior evolution strategy.

For two measures ν_1, $\nu_2 \in \mathcal{M}_{\mathcal{A}}$ defined on the same measurable space $\mathcal{A}$, we say that ν_1 is absolutely continuous with respect to ν_2 (denoted $\nu_1 \ll \nu_2$) if for all measurable set B, $\nu_2(B) = 0 \implies \nu_1(B) = 0$. If $\nu_1 = \nu_1^+ - \nu_1^-$ is a signed measure, we say that $\nu_1 \ll \nu_2$ if $\nu_1^+ \ll \nu_2$ and $\nu_1^- \ll \nu_2$. For (potentially signed) measure ν_1 and (positive) measure ν_2, we denote $\frac{\mathrm{d}\nu_1}{\mathrm{d}\nu_2}$ the Radon-Nikodym derivative of ν_1 with respect to ν_2 (also called density function). We remark that if ν_1 is positive, then so is $\frac{\mathrm{d}\nu_1}{\mathrm{d}\nu_2}$.

When considering measures on $\mathbb{R}^n$ equipped with Borel's σ-algebra, λ^{Leb} denotes Lebesgue's measure. The same notation is used if considering measures on a measurable subset of $\mathbb{R}^n$ of strictly positive Lebesgue mass, to denote the restriction of Lebesgue's measure on this subset.

We denote $(\mathcal{Z}, \mathcal{M}_{\mathcal{Z}})$ the measurable space of observation, and $\mathbb{P}$ probability distributions on $\mathcal{Z}$. Such $\mathbb{P}$ are called data generation mechanisms. The data is denoted $z \sim \mathbb{P}$. Note that z denotes all available observations.

The space of predictors is denoted $\mathcal{H}$. For the calibration of AD models, predictors coincide with the parameters to calibrate. When it is measurable, we use the notation π to denote a probability distribution on $\mathcal{H}$.

Bayesian statistics

The motivation of statistics is to infer, from a realisation of a random variable $z \in \mathcal{Z}$, information about the random variable's probability measure. To recover such information, a statistician makes assumptions about the probability measures which might define the random variable. Formally, one considers a subset $\mathcal{P} \subset \Pi$, assumes that $\mathbb{P}$ belongs $\mathcal{P}$, and strives to estimate $\beta(\mathbb{P})$ where $\beta : \mathcal{P} \to I$ is the required information about the probability (*e.g.* a sufficient statistics, the mean if $\mathcal{Z}$ is a vector space, the measure density). To control the performance of the inference process, one assumes that I is equipped with a (pseudo)-distance $d : I^2 \to \mathbb{R}^+$ which should satisfy $\forall a \in I,\ d(a,a) = 0$.

As an example, by observing the weather during a given month (*observations*), and assuming that the probability of its raining on a given day is independent from whether it rained on other days, and follow the same measure (*assumptions on the admissible probability measures*, here that the data is independent, identically distributed, or i.i.d. for short), one can strive to infer what is the probability of it raining during a day of that month (the information on the probability measure). Denoting $\hat{p}$ the predicted probability of rain, and p the true probability of rain, the performance of the inference can be measured for instance by $d(\hat{p}, p) = (\hat{p} - p)^2$.

Traditionally, the statistical community was split in two groups, divided on how the inference process should be carried out. The *frequentist* community relies on the use of an estimator, *e.g.* a measurable function $\hat{\beta} : \mathcal{Z} \to I$ (see for instance Lehmann and Casella [1998], Tsybakov [2009]). The quality of an estimator for a specific $\mathbb{P}$ can be assessed by its average distance to the true information, $E(\hat{\beta}, \mathbb{P}) := \mathbb{P}\left[d(\hat{\beta}(z), \beta(\mathbb{P}))\right]$. But the quality of an estimator can not be judged only on its performance for a given $\mathbb{P}$, since this measure is unknown. The minimax approach uses a worst case analysis to evaluate the performance of an indicator, *i.e.* $E(\hat{\beta}) :=$

$\sup_{\mathbb{P}\in\mathcal{P}} E(\hat{\beta}, \mathbb{P})$, and defines the minimax risk of the inference task as the infima of the risk E over all measurable functions. Hence, taking a minimax frequentist approach, the goal of a statistician is to design estimators $\hat{\beta}$ with worst case risk $E(\hat{\beta})$ achieving, or approximating the minimax risk. This estimator can be tailored-made for the class of admissible probability measures considered as well as for the actual pseudo distance considered.

The minimax approach considers all admissible probability measures on an equal footing. The *Bayesian* community, on the other hand, considers that while all members of $\mathcal{P}$ are *possible*, they might not all be equally *plausible* (see Robert [2007] for a thorough overview of the field). Let us develop the previous example. Between different years, climatic events might cause the (theoretical and unobserved) likelihood of rain in November in Paris to vary. But this (random) likelihood does not take all values with identical probability, *e.g.* it is much more probable to observe a wet November than a dry November. Formally, one considers a probability measure π_{p} on the space $\mathcal{P}$ (which must therefore be measurable). The data is assumed to be drawn according to the following two-level process: first draw the data generation distribution $\mathbb{P} \sim \pi_{\mathsf{p}}$, then draw the observation $z \sim \mathbb{P}$. This defines a probability distribution on the product space $\mathcal{P} \times \mathcal{Z}$. The inference question hence becomes: having observed z, can I infer which $\mathbb{P}$ was drawn? The Bayesian community answers this question using Bayes law, which states that for a distribution $\nu \in \Pi_{\mathcal{A}}$, and two measurable sets $A, B \in \Sigma_{\mathcal{A}}$ such that $\nu(A), \nu(B) > 0$, $\nu(B \mid A) = \frac{\nu(B)}{\nu(A)} \nu(A \mid B)$. By applying this result to $\mathcal{A} = \mathcal{P} \times \mathcal{Z}$, using the distribution ν on the random variable $(\mathbb{P}, z)$ where $\mathbb{P}$ is drawn from measure π_{p}, and, conditionally on $\mathbb{P}$, z is drawn from $\mathbb{P}$, and sets $A = \mathcal{P} \times \{z\}$, $B = \{\mathbb{P}\} \times \mathcal{Z}$, one obtains

$$\begin{aligned} \hat{\pi}(\{\mathbb{P}\}) :=& \nu(\{\mathbb{P}\} \times \mathcal{Z} \mid \mathcal{P} \times \{z\}) \\ =& \frac{\nu(\{\mathbb{P}\} \times \mathcal{Z})}{\nu(\mathcal{P} \times \{z\})} \nu(\mathcal{P} \times \{z\} \mid \{\mathbb{P}\} \times \mathcal{Z}) \\ =& \frac{\pi_{\mathsf{p}}(\{\mathbb{P}\})}{\nu(\mathcal{P} \times \{z\})} \mathbb{P}(\{z\}) \\ \propto& \mathbb{P}(\{z\}) \pi_{\mathsf{p}}(\{\mathbb{P}\}). \end{aligned} \tag{1.5}$$

Being a multiplicative constant, the renormalisation term $\nu(\mathcal{P} \times \{z\})$ can be omitted since the measure $\hat{\pi}$ is constrained to satisfy $\hat{\pi}[1] = 1$. This construction, exact if $\mathcal{P}$ and $\mathcal{Z}$ have finite cardinals, can be extended to more general settings, and is the cornerstone of Bayesian inference. The initial measure on the data generating probabilities π_{p} is called the prior distribution. The measure $\hat{\pi}$ is called the posterior distribution; it describes the updated measure on the data generating probabilities conditional on the data value. The posterior distribution defines a randomised estimator $\beta(\mathbb{P})$, $\mathbb{P} \sim \hat{\pi}$, which is completely determined by the prior distribution π_{p} and the data generating mechanisms $\mathcal{P}$. Notably, it does not take into account a distance d, as it only involves probabilistic considerations.

An intrinsic aspect of Bayesian inference is its use of randomised estimators $\hat{\pi}$. Bayesian inference maps the initial uncertainty on the possible data generation mechanisms (*e.g. November is probably going to be rainy,* e.g. *raining from half to every day*) to the up to date uncertainty

once observed data is taken into consideration (*it has been raining every day this November so far, hence the chance of rain this November is somewhere between 90% to 100%*).

By considering pushforward measures of the posterior, this notion of uncertainty can moreover be converted to any functional of the data generation mechanisms (*e.g.* the mean, any sufficient statistic). For posterior distribution $\hat{\pi}$, a *credible* region of level α is any measurable set $\mathsf{CR}_\alpha^{\mathsf{Bayes}}(\hat{\pi}) \subset \mathcal{P}$ such that $\hat{\pi}\left(\mathsf{CR}_\alpha^{\mathsf{Bayes}}(\hat{\pi})\right) = \alpha$. These are analogues of the *confidence* regions CR_α of level α, *e.g.* functions of the observation which satisfy $\mathbb{P}\left(\beta(\mathbb{P}) \in \mathsf{CR}_\alpha(z)\right) = \alpha$. While confidence regions offer pointwise guarantee for individual data generation mechanism $\mathbb{P}$, credible regions only offer an average guarantee on the prior. That is to say,

$$\pi_{\mathsf{p}}\left[\mathbb{P}\left(\mathbb{P} \in \mathsf{CR}_\alpha^{\mathsf{Bayes}}(\hat{\pi})\right)\right] = \alpha.$$

A remarkable result in Bayesian theory is that under some regularity assumptions, in the case of parametric families $\mathcal{P}$ of finite dimension describing n i.i.d. data points (*i.e.* $\mathcal{P} = \left\{p_\gamma^n \mid \gamma \in \Gamma \subset \mathbb{R}^k\right\}$), credible regions of level α are also confidence regions of the same level asymptotically for all $\mathbb{P}$ in the support of $\frac{\mathrm{d}\pi_{\mathsf{p}}}{\mathrm{d}\lambda^{\mathsf{Leb}}}$, where λ^{Leb} is the Lebesgue measure, *i.e.*

$$\forall \gamma \in \mathsf{supp}\left(\frac{\mathrm{d}\pi_{\mathsf{p}}}{\mathrm{d}\lambda^{\mathsf{Leb}}}\right), \lim_{n\to\infty} p_\gamma^n\left(\gamma \in \mathsf{CR}_\alpha^{\mathsf{Bayes}}(\hat{\pi}_n)\right) = \alpha. \tag{1.6}$$

This is a consequence of the celebrated Bernstein–von Mises theorem (see van der Vaart [2002]), which moreover implies that the posterior is asymptotically Gaussian, with variance decaying in the classic parametric rate $1/n$. While Bernstein–von Mises theorem does not hold in general in nonparametric (*i.e.* infinite dimensional) settings, some extensions have been established in that case (see Rousseau et al. [2014] for a review), providing the same asymptotic guarantees on the credible regions.

A key insight from Bernstein–von Mises theorem is that the posterior distribution can asymptotically recapture frequentist guarantees under very mild conditions on the prior distribution, namely that the prior puts at least some mass on the true data generating mechanism. This implies that Bayesian statistics is thus asymptotically motivated even when studying a single realisation of the data generation mechanisms, but also when limited prior knowledge is available. Let us give an example to illustrate this point. Consider a bag containing K biased coins, with coin C_i drawing heads with probability $0 < p_i < 1$. Draw a coin, then toss it up n times to infer the probability that drawn coin draws a head. Here, there exists a natural prior distribution on the probability of drawing head for the coin drawn, which is simply $(1/K)\sum_{k=1}^K \delta_{p_k}$. The Bayes posterior defined for that prior naturally defines confidence regions whose average coverage is α (where averaging is done by repeating the coin selection, coin tossing operating). Unfortunately, one might not have the luxury of repeating the procedure, and hence the average guarantees might not be of any use. Furthermore, one might not be allowed to look ahead inside the bag and examine the coins, that is to say one might not known the probabilities p_i and hence the natural prior might not be accessible. The Bernstein–von Mises theorem partially overcomes these difficulties by stating that for *any* prior distribution on $]0,1[$ absolutely

continuous with respect to Lebesgue measure with density of full support will yield a posterior distribution with asymptotic (in the number of draws n) guarantees, regardless of the content of the bag, regardless of the coin which was drawn.

While the knowledge (or even the existence) of the natural prior is not requisite for Bayesian statistics, the knowledge of the laws of the data generation mechanism is imperative for the construction of the posterior distribution. Consider a parametrization $\gamma \in \Gamma$ of $\mathcal{P} = \{\mathbb{P}_\gamma\}$. We further assume that there exists a measure $\overline{\mathbb{P}}$ such that $\forall \gamma \in \Gamma$, $\mathbb{P}_\gamma \ll \overline{\mathbb{P}}$. Denote $L(\gamma, z) := \frac{\mathrm{d}\mathbb{P}_\gamma}{\mathrm{d}\overline{\mathbb{P}}}(z)$ the likelihood function. Then the informal eq. (1.5) becomes

$$\begin{aligned} \frac{\mathrm{d}\hat{\pi}}{\mathrm{d}\pi_{\mathsf{p}}}(\gamma) &= \frac{L(\gamma, z)}{\int_{\gamma \in \Gamma} L(\gamma, z) \mathrm{d}\pi_{\mathsf{p}}(\gamma)} \\ &\propto L(\gamma, z). \end{aligned} \tag{1.7}$$

From (Bayesian) statistics to learning theory

The classic statistical inference framework considers a setting where the admissible laws, here encapsulated in the likelihood function, are known in advance. In a typical parametric setting for instance, the data generation mechanism could be of form $f(\gamma, X)$ for some known function f and random variable X (*e.g.* $z \sim \mathcal{N}(\gamma, \mathsf{Id})$), and the goal would be estimate the value of γ. The learning framework, on the other hand, considers an arbitrary (unknown) data generation mechanism, a set of predictors and a learning objective. The goal of a typical learning algorithm would be to select the predictor achieving the best average performance on the data generation mechanism, from a single realisation of this mechanism. Learning theory strives to obtain guarantees for the selection process (nearly) regardless of the actual data generation mechanism, from the properties of the learning algorithm (*e.g.* train/validation split, cross validation) and of the predictors (*e.g.* Vapnik-Chervonenkis dimension).

Let us discuss this distinction on an AD calibration problem. We are given an initial state State_0, some input data Feed, some observation data Obs, which constitutes our data $z = (\mathsf{Feed}, \mathsf{State}_0, \mathsf{Obs})$. We further have an anaerobic digestion model AD which involves a (multidimensional) parameter $\gamma \in \Gamma$ and outputs predictions through $\mathsf{Pred} = \mathsf{AD}(\mathsf{Feed}, \mathsf{State}_0, \gamma)$. The classical statistics point of view of this problem would be to assume that the observations z are generated through $(\mathsf{Feed}, \mathsf{State}_0) \sim \mathbb{P}_1$, $\mathsf{Obs} \sim \mathsf{AD}(\mathsf{Feed} + \varepsilon_{\mathsf{Feed}}, \mathsf{State}_0 + \varepsilon_{\mathsf{State}_0}, \gamma^*) + \varepsilon_{\mathsf{Obs}}$ with $(\varepsilon_{\mathsf{Feed}}, \varepsilon_{\mathsf{State}_0}, \varepsilon_{\mathsf{Obs}})$ being some (partially) known noise signals (*e.g.* Gaussian noise). The strong assumption is that the model AD is correct, *i.e.* it exactly describes the inner mechanism of anaerobic digestion. The learning point of the view of the problem does not assume that the model AD is correct. The typical goal would be to construct the best approximation of the data Obs in the form of $\mathsf{AD}(\mathsf{Feed}, \mathsf{State}_0, \gamma)$. That is to say, for a prediction error function $\mathsf{Error}(\mathsf{Pred}, \mathsf{Obs})$, find γ such that the random variable $\mathsf{Error}\left(\mathsf{AD}\left(\mathsf{Feed}, \mathsf{State}_0, \gamma\right), \mathsf{Obs}\right)$ is the lowest possible (either on average, or considering quantiles, or for any other relevant assessment).

Statistical inference can be compared to a detective novel. From a given list of suspects (the

set of possible data generating mechanisms, $\mathcal{P}$) and some evidence (the data z), a detective either selects the culprit (point estimation) or draws as tight a net as possible around the culprit (confidence region). Success for point estimation consists in, if not putting the blame squarely on the guilty party, at least getting one of his neighbour into jail. Success for building confidence regions consists in putting the culprit in prison with sufficiently high probability, while minimising the number of innocent casualties. Pushing this analogy to the extreme, the Bayesian prior would encode the prejudices of the detective against each suspects.

The learning framework, on the other hand, can be compared to a hiring process. From a list of applicants (predictors), a recruiter strives to hire the best employee from their performance (prediction loss) on a sample of representative tasks of their future work (training data). The goal of the recruiter is to select the employee which will have the best average performance on future tasks, and not only those of the hiring process (avoid overfitting).

As the focus of learning theory is the selection of a predictor, predictors h in the learning framework play a role analogue, although distinct, to the probability distributions $\mathbb{P} \in \mathcal{P}$ in statistical inference. Similarly, the prediction loss $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$, which measures compatibility between a predictor and an observation, is analogue to the likelihood function. Hence an extension of Bayesian statistics would transform a prior distribution on the predictors to a posterior distribution, by shifting the measure weight from predictors with low compatibility to predictors with high compatibility. A straightforward way to define generalised posterior is to modify eq. (1.7) and for decreasing function $C : \mathbb{R} \to \mathbb{R}$, consider

$$\mathrm{d}\hat{\pi}(h) \propto C(\ell(h,z))\mathrm{d}\pi_{\mathsf{p}}(h). \tag{1.8}$$

Classical choices of C include $C_\lambda(x) = \exp(-\lambda^{-1}x)$, which define the so-called Gibbs posteriors:

$$\frac{\mathrm{d}\hat{\pi}_\lambda}{\mathrm{d}\pi_{\mathsf{p}}}(h) \propto \exp(-\lambda^{-1}\ell(h,z)). \tag{1.9}$$

Gibbs posterior take their name in analogy of Gibbs distribution (*a.k.a.* Boltzmann distribution) in statistical physics, which gives probability $\exp(-E(S)/RT)$ to observing a state S of energy $E(S)$ at temperature T. Following this analogy, the factor λ is often called the temperature of the distribution (although it is also frequent that "temperature" denotes λ^{-1} in PAC-Bayes literature). In the limit where the temperature freezes to 0, the Gibbs posterior becomes a Dirac on the empirical risk minimiser, *i.e.* the minimiser of $\ell(\cdot, z)$ (in the case where the minima is attained on a set m, the Gibbs posterior becomes the measure π_{p} conditioned on the event that the draw belongs to m). On the other hand, $\hat{\pi}_\infty = \pi$. Hence the temperature plays the role of a learning rate, with low temperature indicating high confidence in the data.

Gibbs posterior are strongly related to the notion of η-generalized posteriors in Bayesian statistics (appearing under various names in Ibrahim and Chen [2000], Grünwald [2012], Grünwald and van Ommen [2017], Grünwald and Mehta [2020], Vovk [1990], McAllester [2003] - we use the name given by Grünwald in his series of articles). Using the notations from eq. (1.7), η-generalized posteriors are defined as $\mathrm{d}\hat{\pi}_\eta \propto \mathsf{L}(\cdot, z)^\eta \mathrm{d}\pi_{\mathsf{p}}$. Considering $\ell = -\log(\mathsf{L})$ bridges

the learning oriented Gibbs posterior and statistics oriented η-generalized posteriors, who coincide. For such a loss, Grünwald [2012] showed that η-generalized posteriors concentrate on an optimal approximation of the model under model misspecification *if* $\eta < \eta_C \ll 1$ (*i.e.* if $1 \ll \lambda$). This essentially indicates that the hypothesis that the data generation is known can be traded against convergence speed.

Let us develop our AD example. Let us consider our statistical model with the further assumption that there is no noise on the feed and initial state data ($\varepsilon_{\mathsf{Feed}} = 0$, $\varepsilon_{\mathsf{S}_0} = 0$). We further consider that the noise on the observation data is Gaussian, i.i.d., with variance σ^2. For such a model, the posterior distribution on Γ can be defined as $\mathrm{d}\hat{\pi} \propto \exp\left(-\frac{\|\mathsf{AD}(\gamma)-\mathsf{Obs}\|^2}{2\sigma^2}\right)\mathrm{d}\pi_{\mathsf{p}}$. This matches the Gibbs posterior for the mean square error loss $\ell(\gamma,\ (\mathsf{Feed},\ \mathsf{State}_0,\ \mathsf{Obs})) = \frac{1}{N}\|\mathsf{AD}(\gamma) - \mathsf{Obs}\|^2$ for temperature $\lambda = 2\sigma^2/N$. The results on tempered posteriors hints that selecting a higher temperature ($\lambda \gg 2\sigma^2/N$) will be sufficient to overcome the loss of the statistical model.

Formally, shifting from Bayesian statistics to Bayesian for learning theory thus brings two questions:

- What kind of guarantees can we obtain for learning posteriors?
- Can we define a principled way to design prior to posterior transforms in the learning setting?

PAC-Bayes extension

From now on, $\mathbb{P}$ will stand for the unknown data distribution mechanism. The observation z is drawn from $\mathbb{P}$. The empirical risk function R is defined as $R(h) = \ell(h,z)$. This depends implicitly on the observed data z. The risk function $\tilde{R}$ is defined as the average (with respect to the observed data z) of the empirical risk, $\tilde{R}(h) = \mathbb{P}\left[\ell(h,z)\right]$. In other words, the empirical risk is a random field indexed by the predictor space, and the risk the pointwise average.

PAC-Bayes (Probably Approximately Correct Bayes) theory answers the two questions above through the notion of PAC generalisation bounds. Let $f : \mathcal{Z} \to \mathbb{R}$. A PAC guarantee for the random variable $f(z)$ is a bound of form $\mathbb{P}\left(\mathbb{P}\left[f(z)\right] > B(f(z),\delta)\right) < \delta$ which should hold for $0 < \delta \leq 1$. That is to say, with (high) probability $1-\delta$, the unobserved and unknown average $\mathbb{P}\left[f(z)\right]$ is smaller than a function of its empirical counterpart $f(z)$ and the required confidence level δ. In a similar fashion, a PAC-Bayes bound is a guarantee of form

$$\mathbb{P}\left(\forall \pi \in \Pi\ s.t.\ \pi \ll \pi_{\mathsf{p}}, \pi\left[\tilde{R}\right] \leq \mathsf{PB}(\pi, R, \pi_{\mathsf{p}}, \delta)\right) \geq 1-\delta. \tag{1.10}$$

A remarkable aspect of PAC-Bayes bounds is that they offer guarantees for all possible probability measures π absolutely continuous[8] to a common measure π_{p} (called the PAC-Bayes

[8]This condition of absolute continuity can be removed in some PAC-Bayes bounds (*e.g.* Wasserstein PAC-Bayes, Total Variation penalized PAC-Bayes), when the behaviour of the bound on the singular part of π with respect to π_{p} (in Lebesgue decomposition sense, see Theorem 6.10 in Rudin [1986]) can be controlled by a worst case analysis.

prior or simply prior) *simultaneously*, that is to say on the same high probability event. A consequence of that is that the PAC bound PB can be optimised on the measure π. Hence the PAC-Bayes bound offers a learning objective in the form of

$$\text{Find } \hat{\pi} \in \arg \min_{\pi \in \Pi,\ \pi \ll \pi_{\mathsf{p}}} \mathrm{PB}\left(\pi, R, \pi_{\mathsf{p}}, \delta\right), \tag{1.11}$$

for which eq. (1.10) guarantees that with probability $1-\delta$, the risk posterior average $\hat{\pi}\left[\tilde{R}\right]$ is bounded by the observable $\mathrm{PB}(\hat{\pi}, R, \pi_{\mathsf{p}}, \delta)$. In other words, optimising a PAC-Bayes bound will be robust to overfitting with a probability higher than $1-\delta$, no matter how complex the risk landscape and adverse the data generation mechanism. The PAC-Bayes convention is to call any prospective probability measure π on which the PAC-Bayes bound is used a posterior distribution.

While a PAC-Bayes bound naturally defines a learning objective, it might not always be possible to compute the minimiser of the bound, or even to sample from it. We will develop in chapters 3 and 4 how the minimiser can be efficiently approximated, first for a specific bound, and then for a generic bound, when restricting the search to a parametric subset of measures. We remark that as the PAC-Bayes bound holds for any measure π which may depend on the data z, the PAC generalisation guarantees also hold for any approximation of the minimiser, or even for any algorithm outputting a probability measure.

Remark 1.1

Generalised posteriors as defined through eq. (1.8) and posteriors defined as minimiser of eq. (1.11) are related. Indeed, for PAC-Bayes bound which can be rewritten as

$$\mathrm{PB}\left(\pi, R, \pi_{\mathsf{p}}, \delta\right) = \widetilde{\mathrm{PB}}(\pi, \pi\left[R\right], \pi_{\mathsf{p}}, \delta), \tag{1.12}$$

satisfying, $\forall f \in \mathcal{F}_{\Gamma}$, $\forall r \in \mathbb{R}$

$$\widetilde{\mathrm{PB}}(\pi_f^*, r, \pi_{\mathsf{p}}, \delta) \leq \widetilde{\mathrm{PB}}(\pi, r, \pi_{\mathsf{p}}, \delta), \tag{1.13}$$

where

$$\frac{\mathrm{d}\pi_f^*}{\mathrm{d}\pi_{\mathsf{p}}} = \pi_{\mathsf{p}}\left[\frac{\mathrm{d}\pi}{\mathrm{d}\pi_{\mathsf{p}}} \mid f\right], \tag{1.14}$$

it follows that the minimiser of the PAC-Bayes bound is of form eq. (1.8). Note that eq. (1.13) can be thought of as a data processing inequality, and is satisfied for most PAC-Bayes bounds.

Proof. For any posterior π, Equation (1.13) implies that the posterior π_R^* achieves a lower PAC-Bayes bound since $\pi_R^*\left[R\right] = \pi\left[R\right]$. Hence any minimiser $\hat{\pi}$ of the PAC-Bayes bound must be a fixed point of $\pi \mapsto \pi_R^*$. Hence $\frac{\mathrm{d}\hat{\pi}}{\mathrm{d}\pi}$ must be R measurable, which implies the decomposition eq. (1.8). □

We remark however that there is an important limit in this relation between two forms: the function C is allowed to depend on the risk, but also on the prior. We will study in Chapter 2 the form of the minimiser of PAC Bayes bounds for a subset of PAC-Bayes bound, and show that it involves a renormalisation factor which is, in the general case, more involved than in the Bayesian case (see the second part of Theorem 2.1). Hence the sampling strategies developed for Bayes posteriors might not be applicable. A remarkable exception are Gibbs posteriors, which minimise Catoni's PAC-Bayes bound (Catoni [2004], Alquier [2024]). As the proof of this bound is remarkably straightforward and instructive in the canonical way to construct a PAC-Bayes bound, we give here a short proof.

1.2.2 Constructing PAC-Bayes generalisation bounds

A Probably Approximately Correct Bayesian generalisation bound is constructed by combining two ingredients [Bégin et al., 2016, Guedj, 2019, Alquier, 2024]:

- A concentration inequality, which measures the deviations between the empirical risk and risk on a fixed, data independent randomised predictor,
- A change of measure inequality, which informs on the cost of moving from the integral of a function with respect to a first measure to the integral of the same function to another measure.

Concentration inequalities

A concentration inequality is a bound on the tails of function of many independent random variables (see Boucheron et al. [2013] for an overview of the field). Well studied examples of concentration concern the average of independent random variables, where Kolmogorov's strong law of large number (see *e.g.* Feller [1957]) implies that $\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mathbb{P}\left[X_i\right]\right) \to_{n\to\infty} 0\right) = 1$ for $\mathbb{P}$ drawing independent X_i satisfying $\sum_{i=1}^{\infty} i^{-2}\mathbb{V}\left[X_i\right] < \infty$. The goal of concentration inequalities is to quantify the deviations of the functional (here the empirical average) with respect to a reference for a non asymptotic number of observations n. In the sum of random variables example, this translates into obtaining a bound on

$$\mathbb{P}\left(\sum_{i=1}^{n}\left(X_i - \mathbb{P}\left[X_i\right]\right) \geq t\right).$$

The most famous bound on the probability of observing deviations of a random variable to its mean is certainly Markov's inequality, which states that for all $t > 0$ and any real valued random variable X,

$$\mathbb{P}\left(|X| \geq t\right) \leq \frac{1}{t}\mathbb{P}\left[|X|\right]. \tag{1.15}$$

Markov's inequality's proof is remarkably simple, being a straightforward consequence of Lebesgue's integral's monotonicity and the inequality $|X| \geq t\mathbb{1}\left(|X| \geq t\right)$. While the proof is simple,

the implications of Markov's inequality are far reaching. A simple way to construct new concentration inequalities consists in choosing a non decreasing, positive function $f : \mathbb{R} \to \mathbb{R}_+$, and use Markov's equality on $f(X)$. As f is non decreasing, $\mathbb{P}(X \geq t) \leq \mathbb{P}(h(X) \geq f(t))$ (with equality holding if f is increasing). Using Markov's inequality, the right hand side can be upper bounded by $\frac{1}{f(t)}\mathbb{P}[f(X)]$. A useful refinement of this technique consists in considering a family of increasing functions $\mathcal{F}$; since the upper bound holds for all function in $\mathcal{F}$, it holds for the infimum and hence

$$\mathbb{P}(X \geq t) \leq \inf_{f \in \mathcal{F}} \left\{ \frac{1}{f(t)} \mathbb{P}[f(X)] \right\}. \tag{1.16}$$

This program is used to establish two famous concentration inequalities. For the random variable X, consider the random variable $|X - \mathbb{P}[X]|$. Applying Equation (1.16) to this random variable using for $\mathcal{F}$ the singleton containing the square function (which is increasing on $\mathbb{R}^+$), it follows that $\mathbb{P}(|X - \mathbb{P}[X]| \geq t) \leq \frac{1}{t^2}\mathbb{V}[X]$, taking for convention that $\mathbb{V}[X] = \infty$ if X does not admit a second moment. This inequality is known as Chebychev's inequality. Using the set of functions $\mathcal{F} = \{x \mapsto \exp(\lambda x) \mid \lambda \geq 0\}$, we obtain $\mathbb{P}(X \geq t) \leq \inf_{\lambda \geq 0} \exp(-\lambda t)\mathbb{P}\left[e^{\lambda X}\right]$, known as Chernov's bound.

These two inequalities are particularly well behaved when the random variable is a sum of independent random variables, *i.e.* $X = \sum_{i=1}^n X_i$. In this case, Chebychev's inequality yields a bound on the probability of observing small deviations of form $\frac{t}{\sqrt{n}}$ between the empirical mean and mean of form

$$\mathbb{P}\left(|X - \mathbb{P}[X]| \geq \sqrt{n}t\right) \leq \frac{1}{nt^2} \sum_{i=1}^n \mathbb{V}[X_i].$$

On the other hand, Chernov's bound implies an exponential decay in large deviations (constant) between the empirical mean and mean, of form

$$\mathbb{P}(X - \mathbb{P}[X] \geq nt) \leq \exp\left(-n \inf_{\lambda \geq 0} \left\{ \lambda t - \frac{1}{n} \sum_{i=1}^n \log\left(\mathbb{P}\left[\exp\left(\lambda\left(X_i - \mathbb{P}[X_i]\right)\right)\right]\right) \right\} \right).$$

In particular, for i.i.d. distributed X_i, noting $\xi(\lambda) = \log\left(\mathbb{P}\left[\exp\left(\lambda\left(X_1 - \mathbb{P}[X_1]\right)\right)\right]\right)$, it follows that

$$\mathbb{P}(X - \mathbb{P}[X] \geq nt) \leq \exp\left(-n\xi^*(t)\right),$$

where ξ^* is the Legendre transform of ξ.

Legendre transform As Legendre transforms play an important part in Chapter 2, we recall here some properties on Legendre transforms (see section 3.3 of Boyd and Vandenberghe [2004] for an introduction on finite dimensional spaces, and section 2.1 of Barbu and Precupanu [2012] for the general case). For $\mathcal{E}$ a topological vector space on $\mathbb{R}$, noting $\mathcal{E}^*$ its dual space (*i.e.* the space of continuous linear forms on $\mathcal{E}$) and $\langle \cdot, \cdot \rangle : E \times E^* \to \mathbb{R}, z, y \mapsto y(z)$, the Legendre

transform of any convex function $f : E \to \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ is defined as

$$f^* : \begin{array}{l} E^* \to \overline{\mathbb{R}}, \\ y \mapsto \sup_{z \in E} \langle z, y \rangle - f(z). \end{array} \tag{1.17}$$

This transform satisfies $(f^*)^* = f$ if f is a proper convex function ($\exists z \in \mathcal{E}, f(z) \neq \infty$) and lower semi continuous (inferior limit of any $f(x_i)$ such that $x_i \to x$ is $f(x)$). If f is not lower semi continuous, the bi-conjugate $(f^*)^*$ is the "lower semi continuous version" of f, *i.e.* the largest convex, lower semi continuous function smaller than f. Moreover, $\forall a \in \partial f(b)$, then $f^*(a) = \langle b, a \rangle - f(b)$ (the minima is achieved for $z = b$). Moreover, the sub differentiates of f and f^* are inverse of one another, *i.e.* $\forall z \in E$, $x \in \partial f(z) \implies z \in \partial f^*(x)$. As a consequence, if f and f^* can be differentiated, this implies that $f'^{-1} = f^{*\prime}$.

Change of measure

A change of measure inequality controls the average of a real valued random variable X with respect to a first measure π_1 by the average of $f(X)$ with respect to a second measure π_2, for some function f, and a pseudo distance between the two measures. Formally, it is a bound of form $\pi_1[X] \leq \mathsf{CM}(\pi_2[f(X)], \pi_1, \pi_2)$ which should hold for all π_1 integrable random variable.

The most famous change of measure inequality is due to Csizár, Donsker and Varadhan [Csiszár, 1975, Donsker and Varadhan, 1975], and implies that $\forall \pi_1$, $\forall \pi_2$, $\forall X$ bounded,

$$\pi_1[X] \leq \log(\pi_2[\exp(X)]) + \mathsf{KL}(\pi_1, \pi_2) \tag{1.18}$$

where KL denotes the **K**ullback–**L**eibler divergence (KL) divergence

$$\mathsf{KL}(\pi_1, \pi_2) = \begin{cases} \pi_2\left[\frac{\mathrm{d}\pi_1}{\mathrm{d}\pi_2} \log\left(\frac{\mathrm{d}\pi_1}{\mathrm{d}\pi_2}\right)\right] & \pi_1 \ll \pi_2 \\ +\infty & \text{else} \end{cases}, \tag{1.19}$$

Moreover, Csizár-Donsker-Varadhan's identity implies that the bound is tight, in the sense that $\forall \pi_2$, $\forall X$ bounded, $\exists \pi_1$ such that the inequality is an equality. This can be interpreted in terms of Legendre transform: the convex function $\pi_1 \mapsto \mathsf{KL}(\pi_1, \pi_2)$ has Legendre transform $X \mapsto \log(\pi_2[\exp(X)])$.

A PAC-Bayes recipe

Let us combine Ciszár-Donsker-Varadhan's change of measure with Hoeffding concentration. For any distribution $\pi \ll \pi_{\mathsf{p}}$, using $X = \lambda^{-1}(\tilde{R} - R)$ for $\lambda > 0$, Equation (1.18) implies that $\forall \pi_{\mathsf{p}}, \pi \ll \pi_{\mathsf{p}}$,

$$\pi\left[\tilde{R} - R\right] \leq \lambda \log\left(\pi_{\mathsf{p}}\left[\exp\left(\lambda^{-1}\left(\tilde{R} - R\right)\right)\right]\right) + \lambda \mathsf{KL}(\pi, \pi_{\mathsf{p}})$$

We now require a high probability bound (on the data generation mechanism) on the prior average exponential moment $\pi_{\mathsf{p}}\left[\exp\left(\lambda^{-1}\left(\tilde{R}-R\right)\right)\right]$. Using Markov inequality, one knows that

$$\begin{aligned}\mathbb{P}\left(\pi_{\mathsf{p}}\left[\exp\left(\lambda^{-1}\left(\tilde{R}-R\right)\right)\right]>\epsilon\right) &\leq \frac{1}{\epsilon}\mathbb{P}\left[\pi_{\mathsf{p}}\left[\exp\left(\lambda^{-1}\left(\tilde{R}-R\right)\right)\right]\right] \\ &\leq \frac{1}{\epsilon}\pi_{\mathsf{p}}\left[\mathbb{P}\left[\exp\left(\lambda^{-1}\left(\tilde{R}-R\right)\right)\right]\right]\end{aligned}$$

where we used Fubini's theorem for positive random variables to swap the integrals. Choosing ϵ such that the the right hand side is δ, *i.e.* $\epsilon = \delta^{-1}\pi_{\mathsf{p}}\left[\mathbb{P}\left[\exp\left(\lambda^{-1}(\tilde{R}-R)\right)\right]\right]$, this implies that $\forall\pi_{\mathsf{p}}$, with probability $1-\delta$, $\forall\pi \ll \pi_{\mathsf{p}}$,

$$\pi\left[\tilde{R}-R\right] \leq \lambda\log\left(\pi_{\mathsf{p}}\left[\mathbb{P}\left[\exp\left(\lambda^{-1}\left(\tilde{R}-R\right)\right)\right]\right]\right) - \lambda\log(\delta) + \lambda\mathsf{KL}(\pi,\pi_{\mathsf{p}})$$

In the setting where the risk is the average of n independent, identically distributed random point losses which are bounded between 0 and 1, Hoeffding's inequality states that for all γ, $\mathbb{P}\left[\exp\left(\lambda^{-1}\left(\tilde{R}(\gamma)-R(\gamma)\right)\right)\right] \leq \exp\left(\frac{1}{8\lambda^2 n}\right)$. This implies that

$$\log(\pi_{\mathsf{p}}\left[\mathbb{P}\left[\exp\left(\lambda^{-1}\left(\tilde{R}-R\right)\right)\right]\right]) \leq \frac{1}{8\lambda^2 n}.$$

Hence we obtain, $\forall\lambda>0$, $\forall\pi_{\mathsf{p}}$,

$$\mathbb{P}\left(\forall\pi\ll\pi_{\mathsf{p}},\ \pi\left[\tilde{R}\right] \leq \pi\left[R\right] + \lambda\mathsf{KL}(\pi,\pi_{\mathsf{p}}) - \lambda\log(\delta) + \frac{1}{8n\lambda}\right) \geq 1-\delta$$

This defines a PAC-Bayes bound of form

$$\mathsf{PB}_\lambda(\pi,R,\pi_{\mathsf{p}},\delta) = \pi\left[R\right] + \lambda\mathsf{KL}(\pi,\pi_{\mathsf{p}}) - \lambda\log(\delta) + \frac{1}{8n\lambda}. \tag{1.20}$$

This bound, first established in Catoni [2004], will from now on be called Catoni's bound[9]. This bound has two remarkable properties. First of all, it defines a learning objective which does not depend on the wanted confidence level δ. Indeed, the learning objective can be understood as a penalized empirical risk minimisation objective, with the empirical risk of randomised predictor π simply being its average $\pi\left[R\right]$, and the penalisation $\mathsf{KL}(\pi,\pi_{\mathsf{p}})$. Second, the minima of PB_λ is the Gibbs posterior with temperature λ (see Alquier [2024]).

Let us take a step back. For a fixed predictor γ, Chernoff concentration inequality implies that $\mathbb{P}\left(\tilde{R} \leq R + \sqrt{-\frac{\log(\delta)}{2n}}\right) \geq 1-\delta$. For a fixed, randomised predictor π_{p}, the concentration inequality becomes $\mathbb{P}\left(\pi_{\mathsf{p}}\left[R\right] \leq \pi_{\mathsf{p}}\left[\tilde{R}\right] + \sqrt{-\frac{\log(\delta)}{2n}}\right) \geq 1-\delta$. Catoni's PAC-Bayes bound implies that this point wise guarantee can be extended to all measures π *simultaneously* by adding an extra penalisation term $\mathsf{KL}(\pi,\pi_{\mathsf{p}})$. Essentially, the PAC-Bayes strategy consists in

[9] We remark that this is far from the only, nor the tighter PAC-Bayes bound contributed by O. Catoni. It is however probably the simplest, and the objective it defines also arises naturally in other PAC-Bayes bounds.

using concentration to obtain a generalisation bound on the prior, which is uniformly transformed to a bound on all measures through change of measure, which penalizes measures far away from the initial guess.

1.2.3 PAC-Bayes bounds: a bestiary

Classical PAC-Bayes bounds

We give here a small sample of the variety of PAC-Bayes bounds. Early PAC-Bayes theorems were concerned with the setting where the risk R is an average of n-i.i.d. bounded loss, *i.e.* $R(\gamma) = \frac{1}{n}\sum_{i=1}^{n} \ell(\gamma, z_i)$ with z_i i.i.d. observations, and the loss satisfying $0 \leq \ell \leq 1$.

The first PAC-Bayes theorems, due to Shawe-Taylor and Williamson [1997], McAllester [1999a] focus respectively on the generalisation abilities of a single ideal predictor in a classification setting and on posteriors of form $\pi \propto \mathbb{1}_U \pi_{\mathsf{p}}$. McAllester [1999b] introduced the notion of PAC-Bayes posterior distribution and established the first proper PAC-Bayes bound:

$$\mathbb{P}\left(\forall \pi \in \mathcal{F}(\pi_{\mathsf{p}}),\ \pi\left[\tilde{R}\right] \leq \pi\left[R\right] + \sqrt{\frac{\mathsf{KL}(\pi, \pi_{\mathsf{p}}) - \log\left(\delta\right) + \frac{5}{2}\log(n) + 8}{2n-1}}\right) \geq 1 - \delta \tag{1.21}$$

where $\mathcal{F}(\pi_{\mathsf{p}})$ is the set of *pruned* distributions, *i.e.* distributions which must satisfy $\frac{\mathrm{d}\pi}{\mathrm{d}\pi_{\mathsf{p}}}(\gamma) < 1 \implies \frac{\mathrm{d}\pi}{\mathrm{d}\pi_{\mathsf{p}}}(\gamma) = 0.0$. In short, if the posterior distribution trusts predictor γ less than the prior, it can not trust it at all. It can be shown that the minimiser of the bound on all measures is a Gibbs posterior for a data dependent temperature [Alquier, 2024], which is *not* a pruned distribution.

The following decade improved these results by removing this assumption, as well as obtaining tighter bounds. Langford and Seeger [2001], Seeger [2002], Maurer [2004] considered a variant of the generalisation gap $\tilde{R} - R$. For $0 \leq p_0, p_1 \leq 1$, let us define $\mathsf{kl}(p_0, p_1) := \log\left(\frac{p_0}{p_1}\right) p_0 + \log\left(\frac{1-p_0}{1-p_1}\right)(1-p_0)$ (this the KL divergence between two Bernoulli distributions of parameters p_0 and p_1). Then, by studying the concentration properties of $\mathsf{kl}(\tilde{R}, R)$, Maurer [2004] established that

$$\mathbb{P}\left(\forall \pi \ll \pi_{\mathsf{p}}, \mathsf{kl}(\pi\left[R\right], \pi\left[\tilde{R}\right]) \leq \frac{\mathsf{KL}(\pi, \pi_{\mathsf{p}}) + \log\left(2\sqrt{n}\right) - \log\left(\delta\right)}{n}\right) \geq 1 - \delta. \tag{1.22}$$

As the goal of the PAC-Bayes bound is to upper bound the average true risk, one can define $\mathsf{kl}^{-1}(r, p) = \sup(q \leq 1, \mathsf{kl}(q, p) \leq r)$. This definition implies that $\mathsf{kl}(q, p) \leq r \implies q \leq \mathsf{kl}^{-1}(r, p)$, which completes the definition of **M**aureer-**L**angford-**S**eeger's bound (MLS)[10] PAC-Bayes bound:

$$\mathsf{PB}_{\mathsf{MLS}}(\pi, R, \pi_{\mathsf{p}}, \delta) = \mathsf{kl}^{-1}\left(\frac{\mathsf{KL}(\pi, \pi_{\mathsf{p}}) + \log\left(2\sqrt{n}\right) - \log\left(\delta\right)}{n}, \pi\left[R\right]\right). \tag{1.23}$$

[10] The initial result by Langford and Seeger [2001], Seeger [2002] was discussed only for binary classification. Maurer [2004] stated the general and slightly improved version of the bound.

Pinsker's inequality implies that the bound (1.23) is always tighter than eq. (1.21) (see Pérez-Ortiz et al. [2021b]). Building on this extended notion of generalisation gap, Germain et al. [2009] proposed the use of a generic convex function $\Delta : [0,1] \times [0,1] \to \mathbb{R}$ and showed that

$$\mathbb{P}\left(\begin{array}{l}\forall \pi \ll \pi_{\mathsf{p}}, \\ \Delta\left(\pi\left[R\right], \pi\left[\tilde{R}\right]\right) \leq \mathsf{KL}(\pi, \pi_{\mathsf{p}}) + \log\left(\frac{1}{\delta}\pi_{\mathsf{p}}\left[\mathbb{P}\left[\exp(\Delta(R, \tilde{R}))\right]\right]\right)\end{array}\right) \geq 1 - \delta. \tag{1.24}$$

This naturally opens the question of which choice of function Δ yields the tightest bound. Foong et al. [2021] shows that essentially, no choice of Δ can improve on $\mathrm{PB}_{\mathrm{MLS}}$ beyond removing the $\log(\sqrt{n})$ term, and conjecture the existence of a Δ obtaining this optimistic rate.

While powerful, the generalized MLS bound involves the incomputable $\mathbb{P}\left[\exp\left(\Delta\left(R, \tilde{R}\right)\right)\right]$. Bégin et al. [2016] further show that the expectation with respect to the unknown data generating mechanism can be factored out by upper bounding

$$\pi_{\mathsf{p}}\left[\mathbb{P}\left[\exp\left(\Delta\left(R, \tilde{R}\right)\right)\right]\right] \leq \sup_{r \in [0,1]} \sum_{k=0}^{n} \binom{n}{k} \exp\left(n\Delta\left(k/n, r\right)\right).$$

Bégin et al. [2016] investigated penalisation terms between prior and posterior distribution beyond the classic KL divergence, and obtained PAC-Bayes bounds with Rényi divergence penalisation

$$\mathcal{D}_{\alpha}(\pi_1, \pi_2) := \begin{cases} \frac{1}{\alpha - 1} \log\left(\pi_2\left[\left(\frac{\mathrm{d}\pi_1}{\mathrm{d}\pi_2}\right)^{\alpha}\right]\right) & \pi_1 \ll \pi_2 \\ \infty & \text{else} \end{cases}. \tag{1.25}$$

The resulting bounds are of form

$$\mathbb{P}\left(\begin{array}{l}\forall \pi \ll \pi_{\mathsf{p}}, \\ \log\left(\Delta\left(\pi\left[R\right], \pi\left[\tilde{R}\right]\right)\right) \leq \mathcal{D}_{\alpha}(\pi, \pi_{\mathsf{p}}) + \log\left(\frac{1}{\delta}\pi_{\mathsf{p}}\left[\mathbb{P}\left[\Delta\left(R, \tilde{R}\right)^{\frac{\alpha}{\alpha-1}}\right]\right]\right)\end{array}\right) \geq 1 - \delta. \tag{1.26}$$

Viallard et al. [2023] proposed a PAC-Bayes bound based on Wasserstein penalisation, valid under the assumption that the risk is Lipschitz. In the Lipschitz risk setting, any high probability generalisation guarantee of form $\tilde{R}(\gamma_0) \leq R(\gamma_0) + c(\delta)$ can be simultaneously extended to all predictors γ through $\tilde{R}(\gamma) \leq R(\gamma) + c(\delta) + Ld(\gamma, \gamma_0)$. Viallard et al. [2023] shows that this property can be extended to randomised predictors, replacing the distance $d(\gamma, \gamma_0)$ by the Wasserstein distance (using Rubinstein–Kantorovich duality, see *e.g.* Edwards [2011]).

Beyond the bounded risk assumption

For the past decade, some attention has been paid to obtaining PAC-Bayes bound beyond the bounded loss assumption. We give here a quick overview of known results.

Haddouche et al. [2021] used the notion of self-bounding function to extend Germain et al. [2009] to the case where the loss might not be uniformly bounded on all predictors, but can be

bounded individually for all predictors. The resulting PAC-Bayes bound (Theorem 4) essentially coincide with Catoni's bound with a more involved generalisation term on the posterior[11]. This resulting PAC-Bayes bound still involves the classical KL penalisation.

Other publications extending PAC-Bayes to unbounded risk consider penalisation beyond the KL. Rényi divergence penalised bounds from Bégin et al. [2016] are valid under the assumption that the generalisation gap $\Delta(R, \tilde{R})$ has bounded moments of finite order. For comparison, the corresponding results from Germain et al. [2009] require a stronger exponential moment assumption. In a similar approach, Alquier and Guedj [2018] obtained PAC-Bayes bound for specific f-divergences, which similarly require only bounds on moments of finite order on the generalisation gap $\tilde{R} - R$. We remark that while this bounded moment assumption is much less restrictive than the point wise bound for the loss of Haddouche et al. [2021], only this second assumption can be reasonably checked without making virtually unverifiable assumptions on the data generating mechanism.

Beyond independence

The theory of concentration inequality is based on the notion of independence between observations. In many cases however, this assumption is inadequate; for the AD setting of interest, the state of the digester at a given time is highly dependant on the state of the digester the day before, if only because what is inside the digester has only partially been renewed. Ralaivola et al. [2010] obtains a PAC-Bayesian bound for dependent observations by considering fractional covers of a dependency graph (*i.e.* weighted set of subsets of a graph where each subset contains only independent random variables). They obtain a MLS like PAC-Bayes bound where the number of observations is in essence replaced by the number of observations divided by the fractional cover, *i.e.*

$$\mathsf{PB}_{\text{non i.i.d}} = \mathsf{kl}^{-1}\left(\frac{\chi^*}{n}\left(\mathsf{KL}(\pi, \pi_{\mathsf{p}}) + \log\left(\frac{n}{\chi^*} + 1\right) - \log(\delta)\right), \pi\left[R\right]\right). \tag{1.27}$$

By construction, the fractional number $n / \tilde{X}^*$ is smaller than the largest number of independent random variables in the graph. Hence the rate in the bound obtained by Ralaivola et al. [2010] is worse than the rate that would be obtained by disregarding part of the observations (the generalisation guarantee might still be better, since the empirical risk on all data might be lower than the empirical risk on a subset of the data). Moreover, the dependency graph condition might still appear a too restrictive condition for many applications, *e.g.* for time series where observations, no matter how distant in time, are still dependent. Ralaivola et al. [2010] shows that by using an independent block decomposition technique, (1.27) can be used to obtain PAC-Bayes bound for β-mixing stationary process (in essence, processes such that the data generation and the data generation knowing what happened n days in the past have vanishing total variation distance $\beta(n)$), although only for low enough confidence (whether this can be

[11] Note that theorem 3.4 in Haddouche et al. [2021] replaces the classic PAC-Bayes temperature into a rate α acting on the number of observations n, *i.e.* $\lambda = n^{-\alpha}$

arbitrarily low depends on whether $\beta(n)$ goes to 0 faster than $1/n$).

Alquier and Guedj [2018] consider α-mixing to bound the square moment of the generalisation gap. This results in a χ^2 penalised PAC-Bayes bound which is valid for dependent data and unbounded loss, but requires a bound on a moment strictly higher than 2 of the loss, which is typically unknown. For bounded loss, they obtain a bound matching their corresponding $i.i.d.$ bound up after correcting n for the sum of all α-mixing coefficient.

1.2.4 PAC-Bayes in practice

Data-Dependent priors

A key restriction of the bounds presented so far is that the prior distribution is independent from the data. Alquier [2024] remarks that strategies to adapt the PAC-Bayes framework to data dependent priors are nearly as old as PAC-Bayes [Seeger, 2002, Catoni, 2004, Parrado-Hernández et al., 2012]. Still, the use of data dependent prior proved to be a major breakthrough for applied PAC-Bayes. Using an union bound argument for the selection of a data-dependent prior, Dziugaite and Roy [2017] obtained non vacuous generalization bounds (strictly less than 1 generalisation guarantees for classification) for deep neural networks. Dziugaite et al. [2021], Pérez-Ortiz et al. [2021b] obtained tighter generalisation guarantees by building a data-dependent prior through a simple train prior, train posterior data split. This strategy results in valid generalisation bounds, since the PAC-Bayes bound is applied to a subset of the data independent from the data used to train the prior. The performance of this strategy was investigated by Pérez-Ortiz et al. [2021b,a] and showed that data dependent prior greatly improve the generalisation guarantees.

Using a similar data splitting approach, Mhammedi et al. [2019] showed that rather than using part of the data to train the prior, and the second part to train the posterior, one could train two priors, each on part of the data, and train the posterior on the whole data with the average penalisation on the two priors (a similar strategy can be found in Parrado-Hernández et al. [2012]). This idea can be extended to K splits of the data defining K partially data dependent prior - somewhat in the spirit of cross validation splitting - and finally obtain a PAC-Bayes bound involving the aggregated penalisation and the average posterior performance on *all* data points. Such a strategy was also pursued by Haddouche et al. [2021], Viallard et al. [2023]. The key ingredient of this strategy being an union bound on the events of each bound failing, the dependency of the bound with the failure probability δ is changed to δ/K, and the number of data n is reduced.

Dziugaite and Roy [2018] considered differential privacy to allow using data-dependent prior. They show that for any data-dependent prior construction technique which is ε-differentially private[12], the following MLS-like PAC-Bayes bound is obtained for bounded losses:

$$\mathrm{PB}_{\text{Diff. Privat}} = \mathrm{kl}^{-1}\left(\frac{\mathrm{KL}(\pi,\pi_{\mathrm{p}}) + \log\left(4\sqrt{n}\right) - \log\left(\delta\right)}{n} + \frac{\varepsilon^2}{2} + \varepsilon\sqrt{\frac{\log(4/\delta)}{2n}}\right). \tag{1.28}$$

[12] *i.e.* the change of a single observation impacts the prior mass by at most a factor $\exp(\varepsilon)$

Dziugaite and Roy [2018] then use the fact that Gibbs posteriors are $2\lambda^{-1}/n$-differentially private to obtain PAC-Bayes bound when the prior is a Gibbs posterior. The authors report non-vacuous, if rather loose, generalisation guarantees for classification using neural networks.

The construction of data-dependent priors and its implications for the learning abilities of PAC-Bayes algorithms is discussed in Section 2.2.

Sampling from the posterior

We now consider the problem of evaluating the posterior distribution. As noted above, PAC-Bayes generalisation bounds offer a learning objective, *i.e.* a posterior can be inferred from a PAC-Bayes bound and a prior by minimising the generalisation bound. It remains to be seen how this bound can be minimised. In the general case, the objective is defined on the non parametric set of all measures absolutely continuous with respect to the prior, and there is no guarantee that the objective is convex with respect to the posterior.

A classic approach consists in 'constructing' the posterior distribution by drawing i.i.d. samples of arbitrary size from the posterior (or a close approximation). Starting from the form eq. (1.8), one can use techniques developed to draw samples from Bayes posteriors by simply replacing the likelihood by its generalized counterpart. Such techniques include rejection sampling (see *e.g.* Casella et al. [2004]) or Metropolis-Hastings Monte Carlo Markov chain [Metropolis et al., 1953, Hastings, 1970]. This last technique is practical when sampling from the Gibbs posteriors, as they construct samples distributed as the posterior (asymptotically) even if the renormalisation constant $\pi_{\mathsf{p}}\left[\exp(-\lambda R)\right]$ can not be computed or correctly approximated[13]. A refinement of Metropolis-Hastings is the Monte-Carlo adjusted Langevin dynamic (MALA) algorithm, which combines a discretized version of the stochastic differential equation $\mathrm{d}\gamma = \nabla \log\left(\frac{\mathrm{d}\hat{\pi}}{\mathrm{d}\lambda^{\mathsf{Leb}}}\right)(\gamma)\mathrm{d}t + \sqrt{2}\mathrm{d}W$ (where W is a standard Wiener process and λ^{Leb} the Lebesgue measure) and the acceptance rule of Metropolis-Hastings. As the stochastic differential equation involves the gradient of the log density, evaluation of multiplicative factors such as the renormalisation constant are not necessary. The Langevin diffusion is exponentially ergodic with invariant distribution $\hat{\pi}$ [Roberts and Tweedie, 1996], which intuitively suggests that the resulting Monte-Carlo procedure will have higher acceptance ratio and faster convergence. Guarantees for the discretized version are studied in Roberts [2002]. This approach was notably used by Dalalyan and Tsybakov [2012b] to sample from a Gibbs posterior. Other posterior sampling strategies can be found in Guedj and Alquier [2013], Guedj and Robbiano [2018].

Variational PAC-Bayes

A second PAC-Bayes learning approach consists in building posterior distributions belonging to a known, parametric set of measures $\mathcal{P} = \left\{\pi_\theta, \theta \in \Theta \subset \mathbb{R}^K\right\}$ (*e.g.* Gaussian, Gaussian

[13] Note that rejection sampling could theoretically be used, but that for temperature λ, an average of $\exp(\lambda^{-1})$ draws (and evaluations of R) are required to construct a single sample. Temperatures yielding the best guarantees tend to be low. Even for the moderate $\lambda = 0.1$ temperature, drawing 50 i.i.d. samples from the posterior would require on average more than one million draws from the prior.

mixtures, Gaussian with fixed variance, uniform distributions on sets). This approach brings practical advantages: first of all, this yields a full description of the measure rather than a partial view brought by the sample, and as such, informative functional of the measure such as mean, variance, modes can be directly inferred rather than approximated. Such description does not require a choice of sample size - this sample size must moreover be large to obtain satisfactory insights on the measure when the dimension of the predictor space is large. On a practical side, such description can also be easier to store than numerous samples. Finally, this full description of the distribution can be re used if the optimisation procedure is to be restarted (*e.g.* if the risk changes as new data accumulates or if the prior changes).

Constructing approximation of the posterior distribution belonging to a parametric set is a well studied strategy in Bayesian statistics called Variational Bayes [Beal, 2003]. Noting $\hat{\pi}$ the posterior distribution, the approximate posterior belonging to the set $\mathcal{P}$ is defined as the best approximation as measured by KL divergence of $\hat{\pi}$ by a measure belonging to $\mathcal{P}$, *i.e.*

$$\hat{\pi}_{\mathcal{P}} = \arg\inf_{\pi\in\mathcal{P}} \mathrm{KL}(\pi, \hat{\pi}). \tag{1.29}$$

While some interest has been raised by generalising the above formula by considering divergences beyond KL [Opper and Winther, 2000, Minka, 2001, Li and Turner, 2016, Jaiswal et al., 2020], PAC-Bayes offers an alternative generalisation which consists of simply limiting the optimisation space of the generalisation bound to the set $\mathcal{P}$. Knoblauch et al. [2022] investigated the difference between the two approaches[14] and concluded that the PAC-Bayes approach is the more robust approach of the two. Indeed, from a theoretical guarantee viewpoint, it is clear that the constrained minimisation approach attains the best trade-off between restraining the posterior to a set while maintaining as tight guarantees on test performance as possible. We remark that the two approaches coincide when considering the standard KL variational inference and Catoni's PAC-Bayes bound eq. (1.20). As a result, techniques constructing a posterior distribution by minimising a PAC-Bayes bound on a parametric space will be called Variational PAC-Bayes approaches.

To construct a Variational PAC-Bayes posterior, **G**radient **D**escent (GD) approaches have been investigated. Virtually all publications considered Gaussian measures both for the prior and posterior. For KL penalised PAC-Bayes objective, Gaussian distributions benefit from closed form expressions. The remaining difficulty is the computation of the gradient of the posterior risk average $w, \sigma \mapsto \pi_{w,\sigma}[R]$. For linear classifiers, Germain et al. [2009] obtained explicit gradient formula using Gaussian distribution of unit variance. Dziugaite and Roy [2017] learnt a homoschedastic Gaussian posterior distribution on neural networks, using a single posterior sample to estimate the gradient of the risk average ($\partial_{w,\sigma}\pi_{w,\sigma}[R] \simeq \partial_{w,\sigma}R(w+\zeta\sigma)$ with $w+\zeta\sigma \sim \pi_{w,\sigma}$). This estimate is unbiased *assuming* that one can swap expectation and differentiation[15]. This technique was extended in Pérez-Ortiz et al. [2021b] and dubbed "PAC-Bayes

[14] To be exact, they considered Discrepancy Variational Inference objectives defined by generalising eq. (1.29) to other divergences (considering Gibbs posterior as target) and Generalised Variational Inference consisting in optimising PAC-Bayes like objective $\pi[R] + D(\pi, \pi_{\mathsf{p}})$

[15] Note that this is a *strong* assumption in the classification setting. Notably, this implies that one can not train

with Backprop", as it relies on the classic back propagation algorithm used to estimate derivatives of neural networks [Rumelhart et al., 1986]. A variant for a specific, non differentiable activation function is proposed in Letarte et al. [2019]. Considering Gaussian posteriors for a linear regression setting Haddouche et al. [2021] uses an original strategy consisting of learning the posterior average as the classic empirical risk minimiser and choosing the posterior variance through grid search.

1.2.5 Synthesis

PAC-Bayes learning extends the Bayesian statistics framework to the learning setting. The distinctive feature of both Bayes and PAC-Bayes is their use of randomised estimators, leading to a natural notion of uncertainty. These randomised estimators are built by confronting a priori belief on the most likely predictors to data, penalizing predictors which fail to adequately model it.

Having access to abundant literature informing on the likely values of its parameters, AD modelling can fully benefit from Bayesian methods. As exact statistical modelling is intractable and of limited validity, PAC-Bayes offers a principled way to apply such methods. We develop in Chapters 3 and 4 techniques to apply PAC-Bayes calibration to AD models, building on the Variational PAC-Bayes strategy described in Section 1.2.4. In Section 2.1, we extend on section 1.2.3 and show how the bounded risk assumption can be relaxed by extending the change of measure inequality of section 1.2.2. In section 2.2, we analyse the impact of the prior on the generalisation guarantees offered by PAC-Bayes, giving an interpretation to the use of data-dependent priors described in section 1.2.4.

1.3 Meta-Learning

Meta-Learning, also known as lifelong learning or learning-to-learn, is an extension of learning theory concerned with learning the hyperparameter of classic learning strategy (*e.g.* the architecture of a neural network, the initial parameter) rather than the parameters (*e.g.* the neural network weights). The expression meta-learning has been used by various learning communities Hospedales et al. [2022], and as a result encompasses a variety of use cases, some considering on the fly modifications of hyperparameters during a training algorithm's procedure (*single-task* meta-learning), others concerned with leveraging information between different dataset (*multi-task* meta-learning). As in Vanschoren [2019], we will use the term meta-learning exclusively for the multi-task setting[16].

The key motivation of meta-learning is that many learning use cases share a lot of common structure. Let us consider for instance image classification task; while such problems as classifying digits from black and white images (as in the celebrated MNIST benchmark) and

directly on the classification error, but only using a smooth loss (*e.g.* cross entropy), *even though* the posterior average $\theta \mapsto \pi_\theta [R]$ is differentiable with respect to θ for the classification error.

[16] The *single-task* meta-learning will be referred to as hyperparameter optimisation.

classifying different species of birds involve different classes and have distinct end goals, humans perform both these tasks through a similar process of detaching the main image from the background and analysing shapes and curvatures. Such prior knowledge that the tasks share some structure is already leveraged by the type of network architecture which would be typically used to perform them (*e.g.* convolutional neural networks). One could go a step further, and used a trained multi-purpose network, to extract meaningful features from the images, which would then be used as input data for a task specific learning procedure. When data is scarce in a given task (*few shot learning*), using a pre-learnt representation of the inputs can mitigate the risk of overfitting by limiting the number of model parameters which are fitted[17]. This strategy can result in learning algorithms which are simpler to train for a new given task (*i.e.* part of the training was pre-computed at the meta stage), and lead to higher performance (*i.e.* similarly to human behaviour, the model can correctly classify after observing very few (as low as one) input data, relying on the inductive bias brought by the meta-training). Meta-Learning is the field studying, formalizing and improving such strategies.

1.3.1 Learning from multiple datasets

Notations

Following Hospedales et al. [2022], we formalize Meta-Learning by considering a meta parameter[18] space Φ and defining a *learning task* $\mathcal{T}$ as a triplet of data $\mathcal{D} \in \mathbb{D}$, a predictor space Γ and an assessment function $L : \Phi \times \mathbb{D} \to \mathbb{R}$. This assessment function assesses how a meta parameter ϕ behaves on the task data. For instance, splitting the data $\mathcal{D}$ in a train and test dataset[19], resulting in respectively a train and test risk $(R^{\text{train}}, R^{\text{test}})$, considering a learning algorithm $\mathcal{A}$ with hyperparameter ϕ, a natural assessment function would be $L(\phi, \mathcal{D}) = R^{\text{test}}(\mathcal{A}(\phi, \mathcal{D}^{\text{train}}))$. The algorithm $\mathcal{A}$ is called the inner learning algorithm. The meta parameter ϕ can in practice play two roles in the inner learning task, either speeding up the learning procedure or incorporating inductive bias. In both cases, the performance of the meta parameter on the task is formalized in the assessment function.

In meta-learning, one assumes to have access to a collection of learning tasks $(\mathcal{T}_i)_{i\in[1,N_T]}$. We further assume that these learning tasks are drawn independently from an unknown task generation distribution ρ[20]. To illustrate this point, one can consider an animal species classification task. The task generation distribution picks at random an animal, then constructs a dataset containing images of different subspecies of this animal. This constitutes the task data. For a

[17] This constitutes *transfer learning*, which can be considered either as a precursory to modern meta-learning strategies or a simple meta-learning strategy in its own right. The meta-learning formalisation presented here encompasses such strategies.

[18] We focus here on *parametric* meta-learning strategies, *i.e.* one should think of the meta parameter space Φ as a regular subset of $\mathbb{R}^K$ for some K. *However*, we do not require this, and hence also cover non parametric meta-learning algorithms such as metric-based meta-learning.

[19] We remark that contrary to most meta-learning formalisations, we do not require such a split in the data. This is in order to encompass the PAC-Bayes meta-learning strategy developed in this thesis.

[20] The assumption that the tasks are independent would not be valid for some *learning to optimise* strategies, where the next task depends on the previous task and the assessed meta parameter (*e.g.* when a single dataset is considered).

given task, a neural network with K layers is considered as a learning algorithm. The weights of the first k layers are parametrized by ϕ, while the architecture of the remaining layers may be task dependent (notably adjust for the number of classes), and its weights are learnt using GD on the task training data. The performance of the resulting network is assessed on the task test data. This train/test procedure using a given training algorithm defines the assessment function. Our meta-learning data consists of multiple such datasets.

Leveraging knowledge

Similar to standard learning, the end goal of meta-learning is the construction of a metaparameter ϕ^* minimising a functional of the distribution of the assessment function push-forward, *i.e.* the distribution of $L(\phi, \mathcal{D})$ with $(\mathcal{D}, \Gamma, L) \sim \rho$. Typically, the average performance is considered, leading to the meta-learning objective

$$\text{minimise } \tilde{R}^{\text{meta}}(\phi) := \rho\left[L(\phi, \mathcal{D})\right]. \tag{1.30}$$

Finding ϕ^* achieving this minima is the outer learning task, or meta-learning task. The formulation eq. (1.30) can be interpreted as a standard learning task where the predictors are ϕ, Φ the predictor space, data points the tasks, and the loss is the test loss of the inner trained model with meta parameter ϕ. Hence meta-learning can be thought of as a learning process for predictors which are already learners, *e.g.* as a two level learning procedure.

As in standard learning, solving Equation (1.30) is hampered by having only access to a finite number of tasks. The objective can be replaced by an empirical counterpart $R(\phi) := \frac{1}{N}\sum_{i=1}^{N} L_i(\phi, \mathcal{D}_i)$. To assess overfitting, the tasks are often split between training, validation and test tasks.

While our formalisation casts meta-learning as a standard learning procedure, meta-learning aims to take advantage of the specific form of the predictors (themselves learning procedure) in order to construct meta-learning algorithms. We give examples of such strategies in the following section.

1.3.2 Meta-Learning strategies

Metric based approach

Metric based meta-learning approaches learn a metric which is then used to build a non parametric predictor. This strategy has proved to be efficient for few shot image classification tasks. Metric based meta-learning is similar to nearest neighbours algorithms using a meta trained metric[21]. The Nearest Neighbours algorithm (k-NN) is a popular non parametric learning algorithm which, in its simplest form, involves as meta parameters the number of neighbours and metric [Fix and Hodges, 1989]. Contrary to most learning algorithm, it does not involve a train-

[21] Hence Φ is the set of all similarity function on the input space, which is shared between all learning tasks.

ing phase (no optimisation). In a classification setting, it predicts the class of an input point x by considering the majority class of the k closest neighbours of x in the training dataset.

The motivation behind the k-NN algorithm is that two inputs close to one another should belong to the same class. Hence if sufficient observation data is available, the closest points to any new data point should belong to the true class; averaging can help mitigate noise and outliers. Even when this assumption is correct however, the k-NN algorithm can still fail to provide adequate results when the dimension of the input space is large. Indeed, the assumption behind the k-NN motivation is that any point in the input space has many close neighbours in the observation data. Put another way, the input observations should provide a good paving of the input space. But as the dimension grows, the number of spheres of a given radius $\epsilon < 1$ required to cover a sphere of radius 1 grows exponentially. Hence for large dimension (*e.g.* images), the number of observations is typically much too small for k-NN to function using the standard euclidean distance.

Metric based meta-learning approaches solve this difficulty by meta-learning a similarity function $m : \mathcal{X} \times \mathcal{X} \to [0,1]$. This function is then turned into a learning algorithm for a task T containing data (x_i, y_i) by predicting

$$\hat{y}(x) = \frac{1}{N} \sum_{i=1}^{N} m(x, x_i) y_i.$$

As discussed above, similarity functions based on standard euclidean distance would fail to provide adequate classification for images. Meta-learnt similarity, on the other hand, can provide more adequate results if for instance the input data all remain close to a low dimensional manifold. In this case, the manifold can be adequately paved by the input observations, and similarity functions based on the manifold metric could provide adequate classification.

Koch [2015] defined its similarity function as $m(x_1, x_2) = \sigma(\alpha \cdot |f_w(x_1) - f_w(x_2)|)$ where f_w is a neural network with weight w, α is some weight and σ is the sigmoid function. This constitutes a Siamese network (since both input x_1 and x_2 pass through networks f_w whose weights are coupled). The weights w and α are trained to minimise a cross entropy error with ridge penalisation on the weights w on mini batch samples (which behave as a learning task). The resulting model was able to perform one shot twenty ways classification on unseen handwritten alphabets (Omniglot dataset, Lake et al. [2015]) with a high accuracy of 0.92 (versus 0.22 for 1-NN). The trained similarity matrix was transferred to the task of digit recognition using MNIST dataset, and resulted in one shot learning performance of 0.7 (*without* any training on MNIST), versus 0.27 for 1-NN.

Vinyals et al. [2016] follows a similar strategy, but considered a different form of similarity function which encodes the values of the task input and considers a different network for the input data points and query datapoints (*matching networks*). The authors report 0.94 accuracy for one shot twenty ways classification on Omniglot (0.72 on MNIST). The authors also showed the applicability of this strategy for language modelling.

Optimisation based approaches

A core aspect of the metric based approaches to meta-learning is that the inner learning algorithm cannot be adjusted, but is, by design, an average of the closest neighbours. The resulting nonparametric predictor might neither be the most practical, nor the most efficient predictor. For instance, physics or biology based models can incorporate valuable expert knowledge (*e.g.* flow conservations), which improve the model's robustness; the calibrated model parameter yields readily interpretable information on the system, in a way which would not be possible for the meta-learnt similarity function.

In contrast, optimisation based meta-learning is a meta-learning framework motivated to minimise empirical counterparts of eq. (1.30) on a parametric meta parameter space $\Phi \subset \mathbb{R}^K$.

Finn et al. [2017] proposed a meta-learning framework which makes no assumption on the type of model learnt, **M**odel **A**daptative **M**eta **L**earning (MAML). MAML considers as learning algorithm $\mathcal{A}$ the optimisation of a training objective R^{train} on a parametric predictor space Γ using gradient descent. The meta parameter which is learnt is the initialisation of the gradient descent (*i.e.* the meta parameter space coincides with the predictor space). The rationale behind MAML is that a *good initialisation* in a meta-learning setting is an initialisation which can be quickly (*e.g.* in a few steps) be adapted to most learning tasks. Assuming access, for each training task, to a training and test set leading to train loss and test loss R_i^{train}, R_i^{test}, MAML considers meta gradient steps of form

$$\begin{aligned} \gamma^*(\phi) &= \phi - \eta_{\text{inner}} \nabla R_i^{\text{train}}(\phi) \\ \Delta\phi &= -\frac{\eta_{\text{meta}}}{N} \sum_{i=1}^{N} \partial_\phi R_i^{\text{test}}(\gamma_i^*(\phi)). \end{aligned} \tag{1.31}$$

Variants of MAML include cases where a fixed number of gradient steps are performed by the inner algorithm $\mathcal{A}$. Indeed, the core idea of MAML can be adapted to a generic learning algorithm $\gamma^*(i) = \mathcal{A}(\phi, R_i^{\text{train}})$, where ϕ might no longer belong to the predictor space. The key difficulty consists in the computation of the derivative of the solution γ_i^* with respect to the meta parameter. For the original MAML (and multiple gradient steps variant), this can exactly be done at the cost of computing the second derivative of the training risk with respect to the predictor. Finn et al. [2017] reports this computation proves to be the main computational bottleneck when the predictors are neural networks, but that omitting second order computations did not result in loss of performance. As such, the meta training step simplifies to $\Delta\phi \propto -\sum_{i=1}^{n} (\nabla R_i^{\text{test}})(\gamma_i^*(\phi))$. Nichol et al. [2018] proposed to go a step further; an inner solution γ^* is evidence that the initialisation should be shifted from γ to γ^*, resulting in the *Reptile* meta update step of $\Delta\phi \propto \sum_{i=1}^{N} (\gamma^*(\phi) - \phi)$. The authors report performance for meta-learning tasks similar to those reported for MAML. Finn and Levine [2018] followed up on Finn et al. [2017] and notably showed that MAML was robust to using a different learning procedure on test tasks[22].

[22] More precisely, considering more gradient steps on the test task than during the meta training procedure did not lead to overfitting.

1.3.3 Conditional Meta-Learning

Meta-Learning strategies such as MAML learn an inductive bias which is shared between all tasks. Such an inductive bias might lead to less than optimal performance when considering *heterogenous* sets of tasks. For instance, if the learning tasks consist of five tasks of animal species classification, three tasks of car brand classification and eighteen tasks of handwriting recognition, finding an inductive bias for all tasks simultaneously might be less effective than finding an inductive bias for each cluster independently. On the other hand, manually or automatically clustering the learning tasks might neither be practicable, nor desirable, as some information could still be shared between the learning tasks.

Conditional meta-learning (Denevi et al. [2020, 2022], Wang et al. [2020], with similar strategies explored in Jerfel et al. [2019], Rusu et al. [2019], Vuorio et al. [2019]) proposes to meta-learn no longer single inductive bias, but a function which, from the task at hand, returns a specific inductive bias. Wang et al. [2020] considered a task-adaptive structured meta-learning framework. The inductive bias for a new task $\mathcal{T}$ is learnt by minimising a weighed average of the performance of the meta parameter ϕ on the meta training datasets $\mathcal{T}_i$. The weights inform on the similarity between the task at hand $\mathcal{T}$ and the previous training task $\mathcal{T}_i$; similarly to the metric based approach, this similarity function is learnt during meta training[23]. Building on Wang et al. [2020], Denevi et al. [2020] considers the task of learning inductive bias in penalized linear regression where side information is available. Noting s_i the side information of task $\mathcal{T}_i$, this result in the conditional meta-learning objective

$$\rho \left[\tilde{R} \left(\mathcal{A} \left(\tau \left(s \right), R \right) \right) \right], \tag{1.32}$$

where the task generating mechanisms returns a triplet $\tilde{R}, R, s$ and τ is the (learnt) mapping between side information and inductive bias[24]. For Lipschitz and convex loss and input data drawn for all task on a bounded set, Denevi et al. [2020] can control the excess risk of the strategy τ by a 'variance' term $\rho \left[\left\| \left(\arg\inf \tilde{R} \right) - \tau(s) \right\|^2 \right]$. Such analysis motivates the conditional approach, since for the unconditional meta-learning approach result in a constant value τ leading to a much larger upper bound on the excess risk than the oracle conditional τ (the conditional expectation of the true risk minimiser knowing the side information s). Noting that the optimal constant value for $\tau(s)$ is the expectation of the true risk minimiser, the difference between the upper bound between the oracle unconditional and conditional approach is the difference between the variance and average conditional variance. A similar approach is studied in Denevi et al. [2022] to meta-learn low dimensional linear models in a conditional fashion (*i.e.* the low dimensional space will be a function of side information).

[23] The difference with the metric based approach is that the similarity is not constructed on the input data space, but between whole training datasets using a reproducing Kernel such as Maximum Mean Discrepancy.

[24] The learning algorithm is the minimisation of $R(\gamma) + \lambda \left\| \gamma - \tau(s) \right\|^2$ (bias regularized empirical risk minimisation)

1.3.4 PAC-Bayes meta-learning

Bayesian statistics and meta-learning share a common underlying hypothesis: tasks are generated according to a probability distribution. Bayesian statistic focuses on inference from observations from a single task, relying on prior knowledge on how likely observing each task is. The Bayesian inference technique is fully determined by this prior distribution, using Bayes law to infer the posterior distribution. In practice, perfect prior knowledge is often unattainable, and Bayesian theory strives to establish conditions implying asymptotic guarantees for generic priors, valid for a given task. However, the performance of Bayesian inference for non asymptotic behaviour can be greatly influenced by the choice of the prior distribution, which gives strong inductive bias to the inference procedure.

Learning the prior

The idea that the inductive bias in Bayesian statistics can bridge the gap between the generalisation abilities of humans and learning algorithms in the low data regime dates back to at least Tenenbaum [1998, 1999], Fe-Fei et al. [2003]. As a motivation of meta-learning is the construction of learning algorithm able to perform well with few data, Bayesian approaches have been applied to meta-learning from the inception of the latter. Lake et al. [2013, 2015] build on Hierarchical Bayes model to define the "Hierarchical Bayes Program Learning." After training the meta parameters of the hierarchical Bayes model on a train dataset (*i.e.* the meta-learning learns how human written characters are generated), the meta testing performs one shot classification by assessing the likelihood that the query character and the test task characters are identical (and returning the most likely option). The resulting approach obtained state of the art one shot classification result for the Omniglot dataset, with 0.952 correct prediction in 1 one shot 20 way trials[25]. While effective, this approach requires careful engineering of the Hierarchical Bayes model and heavily relies on expert insights on the way the data is generated. Moreover, the prior is learnt by pooling all training data, rather than considering different test tasks. Such an approach might not be reproducible for other settings. Yoon et al. [2018] has adapted MAML algorithm to learn a prior - as a set of initial points and prior hyperparameter - for a inner algorithm using Stein Variational Gradient Descent[26] [Liu and Wang, 2016]. Grant et al. [2018] interprets MAML as approximately a Hierarchical Bayes procedure. When the loss function is a log likelihood, Grant et al. [2018] interprets MAML's learning objective as an approximation of the likelihood of observing the tasks for a hierarchical model $\phi \sim \pi_{\text{meta}}$, $\gamma \sim \pi_{\phi}$, $\mathcal{D} \sim \pi_{\gamma}$ - MAML corresponding to the special case where the prior π on ϕ corresponds to a Dirac distribution. Building on this strategy, Finn et al. [2018] proposes a MAML strategy to meta-learn a Gaussian prior on the meta parameter ϕ called PLATIPUS. This incorporates uncertainty quantification for the meta-learnt test task, which the authors moreover show to be useful in active

[25] While this is higher than the performances reported for matching networks and Siamese networks, Lake et al. [2013] has access to extra information such as the number of strokes, the order in which they were written, etc...

[26] As the hyperparameters of the variance of the prior distribution are included in the particle points, the strategy learns a *distribution* on the prior rather than a single point prior.

learning settings (where additional data points can be queried). Bauer et al. [2017], Zhang et al. [2021b] also investigate the task of learning a prior for the final layer of classification problems. Other works linking Bayesian methods and meta-learning include Gordon et al. [2019], Patacchiola et al. [2020], Zhang et al. [2021a]. More recently, Zhang et al. [2024] considered Bayesian meta-learning to a real world application in personalized room temperature regulation.

Two fold PAC-Bayes: hyperpriors, hyperposteriors

PAC-Bayes approaches to meta-learning have also been explored Pentina and Lampert [2014], Amit and Meir [2018], Liu et al. [2021], Rezazadeh [2022], Rothfuss et al. [2023], Zakerinia et al. [2024], Ding et al. [2024]. These approaches share a common strategy which differs from the Bayesian meta-learning community. Instead of learning a prior distribution for Bayesian inspired learning procedure, the PAC-Bayes meta-learning community advocates a two-level PAC-Bayes approach to learning, resulting in hyperprior and hyperposterior distributions (probability measures on probability measures) at the meta level. Indeed, starting from the interpretation of meta-learning as a two stage learning procedure whose objective is the minimisation of eq. (1.30) on the meta predictor ϕ, the standard PAC-Bayes strategy can be used to obtain generalisation guarantees at the meta level, using a randomised meta parameter. For independent tasks considering bounded loss, the assessment function of the meta parameter L is itself a bounded function, and hence the empirical meta risk is itself a sum of bounded losses, which can readily be treated through PAC-Bayes, resulting, for a hyperprior $\pi_{\mathsf{p}}^{\text{meta}}$, in

$$\pi^{\text{meta}}\left[\tilde{R}^{\text{meta}}\right] \leq \pi^{\text{meta}}\left[R^{\text{meta}}\right] + \lambda^{\text{meta}}\mathsf{KL}(\pi^{\text{meta}}, \pi_{\mathsf{p}}^{\text{meta}}) + \frac{1}{8\lambda^{\text{meta}}N_T} + \lambda^{\text{meta}}\log(\delta) \tag{1.33}$$

holding with probability δ for all hyperposterior π^{meta} (we here use Catoni's PAC-Bayes bound (1.20) for simplicity's sake). Here we account for the finite number N_T of task. To avoid setting aside test data during each task, a PAC-Bayes *inner learning* algorithm can be used as well. In this case, the empirical meta risk R^{meta} can be upper bounded by the average PAC-Bayes generalisation guarantees between the different task. As the outer algorithm is PAC-Bayes, this results in an average over the performance of the PAC-Bayes bound trained using a *randomised* prior. The outer PAC-Bayes learning algorithm implies a KL penalisation between the hyperprior (prior distribution on priors) and hyperposterior (meta-learnt posterior distribution on priors). Refining the above proof sketch, Pentina and Lampert [2014] proposes a PAC-Bayes meta-learning upper bound of the meta generalisation gap of form

$$\left(\frac{1}{\sqrt{N_T}} + \frac{1}{N_T\sqrt{n}}\right)\mathsf{KL}(\pi^{\text{meta}}, \pi_{\mathsf{p}}^{\text{meta}}) + \frac{\sum_{i=1}^{N_T}\pi^{\text{meta}}\left[\mathsf{KL}(\pi_i, \pi_{\mathsf{p}})\right]}{N_T\sqrt{n}} + O\left(\frac{1-\log(\delta)}{\sqrt{n}} + \frac{1-\frac{1}{n}\log(\delta)}{\sqrt{N_T}}\right)$$

holding simultaneously for all hyperposterior π^{meta} and all learning strategy for the computation of the task posteriors π_i. Amit and Meir [2018] obtains a McAllester PAC-Bayes meta-learning bound improving the dependency on the KL penalisation from $\mathrm{KL}(\pi^{\text{meta}}, \pi_{\mathsf{p}}^{\text{meta}})$ to $\sqrt{\mathrm{KL}(\pi^{\text{meta}}, \pi_{\mathsf{p}}^{\text{meta}})}$. The authors implement their learning approach using variational PAC-Bayes with Gaussian distributions with diagonal variance at both level[27]. Liu et al. [2021] explores other PAC-Bayes bounds, and implements their strategy for similar hyperdistributions. Rothfuss et al. [2020, 2022, 2023] study a similar double PAC-Bayes strategy considering a Catoni like objective similar to Pentina and Lampert [2014]. Considering as inner learning algorithm the Gibbs posterior, the authors remark that the meta-objective can be optimised in closed form and result in a Gibbs like hyperposterior, where each individual loss is $\pi \mapsto \log\left(\pi\left[\exp\left(-\lambda^{-1}R_i\right)\right]\right)$. They implement their strategy using Stein Variational Gradient Descent at the meta level (resulting in a sample of priors at the test task time), considering hyperdistributions outputting diagonal Gaussian priors and a centred, spherical Gaussian hyperprior over the prior hyperparameters. Guan and Lu [2022] applies PAC-Bayes bound with faster convergence rates (MLS bound) to the meta-learning setting. Ding et al. [2024] adapts the double PAC-Bayes strategy to the setting where the number of observations is smaller in test task than in the train task, by removing samples from the training task. Rezazadeh [2022] proposes a generic PAC-Bayes meta-learning strategy combining the generalized generalisation gap approach of Germain et al. [2009] with the split between meta generalisation and in task generalisation of the previous PAC-Bayes meta-learning approaches.

Farid and Majumdar [2021] remarks that "none of the approaches mentioned above report numerical values for generalization bounds, even for relatively simple problems. Here, we empirically demonstrate that prior approaches tend to provide either near-vacuous or loose bounds even in relatively small-scale settings." While this statement was made a few years ago, it holds for the ulterior publications cited here, which reports the meta test performance rather than the meta generalisation bounds. In response, Farid and Majumdar [2021] combined a uniform stability argument to analyse the generalisation performance of the inner algorithm and PAC-Bayes analysis at the meta level, which obtains non vacuous meta generalisation guarantees for a synthetic problem and a few shot learning benchmark[28]. The meta-PAC-Bayes learning strategy of Amit and Meir [2018], MLAP, resulted in comparable test task loss but had vacuous meta-learning generalisation bound.

1.4 Bridging the fields

AD is a well established technology for the valorisation of waste into biogas. Insights on AD processes can be obtained through biochemical models, such ADM1. Such models must how-

[27] The hyperdistributions draw a prior by drawing the mean and log standard deviations from Gaussian with diagonal variance.

[28] Mini-Wiki [Balcan et al., 2019], with 1, 3 and 5 shots, 4 ways few shot learning task. The authors report a classification error generalisation guarantee for 1 shot tasks of 0.5 compared to an average test task error of 0.4 (0.39 for the smallest reported test task error achieved by MAML). Surprisingly, the reported generalisation guarantees *increased* with the number of shots.

ever be calibrated to obtain adequate predictive power. Due to the flexibility of the models, the calibration process may suffer from overfitting; moreover, limited practical identifiability may lead to lack of robustness of the calibrated model.

PAC-Bayes learning techniques are a promising way to alleviate the shortcomings of standard calibration approaches. Through the use of generalisation bounds as learning objectives, PAC-Bayes inherently controls the risk of overfitting. By returning a stochastic predictor, Bayesian techniques perform calibration and uncertainty quantification jointly; this uncertainty takes into account the lack of identifiability, increasing robustness. Unfortunately, these two appealing features might not apply to AD. Indeed, most PAC-Bayes learning objectives are mostly designed for the bounded loss, independent data setting, and therefore do not hold for AD data. Contrary to exact Bayesian inference however, PAC-Bayes does not benefit from uncertainty quantification guarantees on the model parameter. We will address part of these limitations by constructing more generic PAC-Bayes bounds (Chapter 2 and experimentally assessing the uncertainty quantification provided by PAC-Bayes posteriors (Chapters 3 and 4). We will show in Chapter 5 that PAC-Bayes can be successfully applied to monitor real-world AD plants.

PAC-Bayes learning relies on beliefs on the plausibility of the parameter values encoded in the prior distribution. In the context of AD, the abundant literature can guide the design of this distribution. Still, the plausible ranges of parameter values prove to be too wide to provide an informative model in the absence of data - moreover, information on the potential correlations between parameter values is usually missing, resulting in a possibly pessimistic representation of uncertainty. As such, the prior distribution can not be used to help design new plants. In order to both improve the PAC-Bayes trained posteriors, and to provide usable AD models for the design of new plants, we'll consider the task of meta-learning the prior distribution in Chapter 6, using an "optimisation based" meta-learning approach.

Chapter 2

Contributions to PAC-Bayes theory

In this chapter, we cover the theoretical contributions of the thesis to PAC-Bayes, namely

- an extension of PAC-Bayes bound beyond Kullback–Leibler penalisation and an investigation on how penalisation impacts the prior moment of the generalisation gap involved in the generalisation bound;
- a framework to study requirements on the prior risk behaviour necessary to reach a certain generalisation guarantee, which can be applied to most PAC-Bayes bounds.

Contents

PAC-Bayes generalisation bounds offer appealing features for the calibration of physical models such as Anaerobic Digestion. Indeed, they offer a principled methodology to control the test error of the learnt models by providing generalisation guarantees, hence effectively informing on the risk of overfitting. Moreover, they offer a convenient and generic methodology to introduce expert knowledge in the learning procedure, in the form of the prior distribution. PAC-Bayes theory is the foundation on which we build learning algorithms for AD models in this thesis.

This chapter is dedicated to theoretical results related to PAC-Bayes theory. The first orientation considered was bridging the gap between the usual hypotheses required in PAC-Bayes theory (risk boundedness, independent observations) and the nature of the data involved in AD models (where the typical risk is the unbounded root mean square error, and the available data are time series), and follows the path of such works as Ralaivola et al. [2010], Seldin et al. [2012], Bégin et al. [2016], Alquier and Guedj [2018]. The results obtained extend those of Alquier and Guedj [2018] and enable the construction of PAC-Bayes bound on generic moment requirements on the generalisation gap [Picard-Weibel and Guedj, 2022]. This work was conducted during the end of 2021 and early 2022, at the beginning of the thesis; I became aware that similar bounds had been independently obtained by Ohnishi and Honorio [2021], using a similar strategy, after drafting the results. The refined bound introduced in Section 2.1.4 is original.

The second orientation is the study on minimal requirements on the prior necessary to reach a target generalisation bound. We provide a protocol to analyse generic PAC-Bayes bound and obtain quantile conditions on the prior empirical risk which must be satisfied to reach a given generalisation level. This type of analysis is, as far as we are aware, new. We connect these requirements to the study of PAC-Bayes in the popular deep learning settings, and essentially show that tight generalisation can only occur for informed priors. These results were obtained in collaboration with Eugenio Clerico [Picard-Weibel et al., 2025].

2.1 From risk assumptions to penalisation: a f-divergence outlook

Classic PAC-Bayes bounds rely on two hypotheses:

- the risk is a mean of n independent losses ℓ_i,
- each independent loss is positive and bounded.

These hypotheses are perfectly natural when the learning task at hand is classification; in this case, the natural loss is the classification loss, valuing 0 if the point is adequately labelled, else 1, which satisfies the second hypothesis, and, under the typical assumption that the label at point x is drawn from a probability measure l_x and that the point x_i are independent, the first assumption is also met. Similarly, the assumption is met for the regression of a bounded function $y = f(x)$ from independent x_i, using bounded predictors and a mean square error or mean absolute error loss.

These assumptions are however not met when the observations are unbounded time series and the error is measured in term of RMSE. This is generally the case for AD model calibration use cases. The observation data which we will consider consists in multiple auto correlated time series, and the prevalent loss in the field is RMSE.

Adjusting PAC-Bayes to time series and unbounded loss has received some attention in the past (see Section 1.2.3). Under mixing assumptions on the time series, both Ralaivola et al. [2010] and Alquier and Guedj [2018] recover PAC-Bayes bounds by applying a corrective factor to the number of observations. This corrective factor quantifies how much the time series mixing, and the resulting apparent number of observations can be thought of as the number of nearly independent points in the dataset.

The unbounded loss setting was notably studied by Bégin et al. [2016] and Alquier and Guedj [2018]. Both these work extend PAC-Bayes beyond the usual KL penalisation considered in classic bounds. This indeed is not wholly surprising considering the proof structure of the bounds; the change of measure inequality of Czisar, Donsker, Varadhan involves the expectation of the exponential of the generalisation gap (or of a functional of the gap); hence this change of measure leads to non trivial result only under an *exponential* moment requirement for the generalisation gap. Considering f-divergences involving the q moment penalisation of the ratio of density, Alquier and Guedj [2018] obtains a change of measure inequality involving the q^* moment of the generalisation gap, where $1/q^* = q/(q-1)$. This change of measure inequality is essentially a consequence of Holder's inequality, remarking that

$$\begin{aligned}\pi\left[D\right] - \pi_{\mathsf{p}}\left[D\right] &= \pi_{\mathsf{p}}\left[\left(\frac{d\pi}{d\pi_{\mathsf{p}}} - 1\right) D\right] \\ &\leq \pi_{\mathsf{p}}\left[\left(\frac{d\pi}{d\pi_{\mathsf{p}}} - 1\right)^{q}\right]^{1/q} \pi_{\mathsf{p}}\left[D^{q^*}\right]^{1/q^*}.\end{aligned}$$

We show that this connexion between the choice of f-divergence and bound requirement

is valid for generic f-divergence, bridging the gap between the exponential moment and finite order moment. By replacing Holder's inequality with Jensen–Shannon's inequality, we construct PAC-Bayes objectives penalized by any f-divergence, involving a f^* moment term on the generalisation gap. f^* here denotes the Legendre transform of the convex function f, and recaptures the exponential requirement for KL divergence and the conjugate moment for the f-divergences considered in Alquier and Guedj [2018]. We further show that this form of objective can be optimised on a single positional parameter to recover the exact Legendre transform of the f-divergence, implying that the bound is tight, and that the moment hypothesis must therefore be satisfied. We further introduce a scale parameter to further tighten the bound, which can be understood as a temperature parameter. We show that the optimisation on both these degrees of freedom can still be understood as a Legendre transform on the bound. For some tractable f-divergences, we show how this optimisation process can be done. Finally, we show how PAC-Bayes bounds can be constructed under specific moment assumptions on the loss.

2.1.1 From Legendre Transforms to PAC-Bayes objectives

We recall Csizár-Donsker-Varadhan [Csiszár, 1975, Donsker and Varadhan, 1975] change of measure, valid for all bounded[1] measurable function D:

$$\log\left(\pi_{\mathsf{p}}\left[\exp D\right]\right) = \sup_{\pi\in\Pi}\left\{\pi\left[D\right] - \mathsf{KL}(\pi,\pi_{\mathsf{p}})\right\} \tag{2.1}$$

with the convention that $\infty - \infty = -\infty$.

Let us first remark that the space $\Pi = \Pi_{\mathcal{H}}$ is an affine space of the vector space consisting of all bounded signed measures $\overline{\mathcal{M}}_{\mathcal{H}}$. The continuous dual of this space is the set of all bounded functions [Fichtenholz and Kantorovitch, 1934]. Note that the function $\overline{\mathsf{KL}} : \pi \mapsto \mathsf{KL}(\pi,\pi_{\mathsf{p}})$ is a convex function on Π. By extending it to $+\infty$ for all signed measures in the complementary of Π to $\overline{\mathcal{M}}$, it remains a convex function. Since $D, \pi \mapsto \pi[D]$ is the extension of the scalar product (see Hildebrandt [1934], section IV.5, Theorem 1), it follows that

$$D \mapsto \sup_{\pi\in\Pi}\left\{\pi\left[D\right] - \mathsf{KL}(\pi,\pi_{\mathsf{p}})\right\}$$

is the Legendre Transform of $\overline{\mathsf{KL}}$, denoted KL^*. Hence Csizár-Donsker-Varadhan can be written more compactly as

$$\overline{\mathsf{KL}}^* : D \mapsto \log\left(\pi_{\mathsf{p}}\left[\exp D\right]\right). \tag{2.2}$$

By construction, the Legendre transforms satisfy Fenchel–Young's inequality:

$$\pi\left[D\right] \leq \overline{\mathsf{KL}}^*(D) + \overline{\mathsf{KL}}(\pi) \tag{2.3}$$

This splits the control between the bilinear form $\pi\left[D\right]$ in a term depending only on D and term

[1] Remarkably, Csizár-Donsker-Varadhan change of measure remains correct even for unbounded continuous measurable function.

depending only on π. This is of practical interest in constructing PAC-Bayes bound. Here, D should be understood as a generalized form of the generalisation gap $\tilde{R} - R$; if one can control $\overline{\mathrm{KL}}^*(D)$ using a concentration inequality, one controls the average generalisation gap *for all* posterior simultaneously, since $\overline{\mathrm{KL}}(\pi)$ no longer involves the generalisation gap. This strategy leads to such PAC-Bayes bounds as those given by Catoni [2004], Maurer [2004], Langford and Seeger [2001], Germain et al. [2009].

A key insight is that this strategy does not rely on the form of penalisation considered. That is to say, for any real valued function $P(\pi)$, one can define its convex conjugate as

$$P^* : D \mapsto \sup_{\pi \in \Pi} \pi[D] - P(\pi) \tag{2.4}$$

and obtain the resulting Fenchel–Young inequality

$$\pi\left[D\right] \leq P^*(D) + P(\pi) \tag{2.5}$$

for all probability measure π. This can even be done with no restriction on the function D apart from measurability, by using as convention that $\pi\left[D\right] = -\infty$ whenever D is not in $\mathrm{L}^1(\pi)$. Moreover, the bound also holds if one replaces $P^*(D)$ by an upper bound $\overline{P}^*(D)$. This gives strong incentive to build PAC-Bayes bounds for penalisations P such that tight upper bounds of $\overline{P}^*$ can be constructed.

2.1.2 Upper bound of f-divergence's Legendre transform

The f-divergences prove to be such penalisations. For a convex function $f : \mathbb{R}_+ \to \mathbb{R}$ such that $f(1) = 0$, the f-divergence $\mathcal{D}_f$ on probability measures on $\mathcal{H}$ is defined as

$$\mathcal{D}_f(\pi_1, \pi_2) = \begin{cases} \pi_2\left[f\left(\frac{\mathrm{d}\pi_1}{\mathrm{d}\pi_2}\right)\right] & \pi_1 \ll \pi_2 \text{ and } \pi_2\left[\left|f\left(\frac{\mathrm{d}\pi_1}{\mathrm{d}\pi_2}\right)\right|\right] < \infty \\ +\infty & \text{else.} \end{cases} \tag{2.6}$$

Since f is convex, one can use Jensen's inequality to show that $\mathcal{D}_f(\pi_1, \pi_2) \geq f(1) = 0$. Moreover, $\mathcal{D}_f(\pi, \pi) = \pi\left[f(1)\right] = 0$. Hence $\mathcal{D}_f$ defines a notion of proximity between measures; since it is not symmetric and might not satisfy the triangular identity, it is not properly a distance, but only a divergence.

The most popular f-divergence is without doubt KL's divergence, which is obtained for the function $f(x) = x\log(x)$. We remark that the f-divergence $\mathcal{D}_f$ does not identify the convex function f, that is to say that there exists multiple convex functions f_1, f_2 such that $\mathcal{D}_{f_1} = \mathcal{D}_{f_2}$. These are actually defined by a simple relationship: if $\mathcal{D}_{f_1} = \mathcal{D}_{f_2}$, there exists $c \in \mathbb{R}$ such that $\forall x, f_1(x) = f_2(x) + c(x-1)$.

Proof. We prove this result in the case $|\mathcal{H}| > 1$. Note that if $|\mathcal{H}| = 1$, only one probability measure exists, and hence the result does not hold (all f lead to the same operator).

The functional $f \mapsto \mathcal{D}_f$ defined on measurable (possibly non convex, non positive) f is linear. Hence solving $\mathcal{D}_{f_1} = \mathcal{D}_{f_2}$ amounts to solving $\mathcal{D}_f = 0$ (here on condition that $f(1) = 0$, but this condition can be removed).

Since $|\mathcal{H}| > 1$, $\mathcal{H}$ contains at least 2 elements h_1 and h_2. Then, using measures $\pi_a = a\delta_{h_1} + (1-a)\delta_{h_2}$ and $\pi_b = b\delta_{h_1} + (1-b)\delta_{h_2}$, one obtains $\forall 0 < a < 1$, $\forall 0 < b < 1$,

$$bf\left(\frac{a}{b}\right) + (1-b)f\left(\frac{1-a}{1-b}\right) = 0.$$

Consider the case where $a \neq b$, and denote $x = \frac{a}{b}$, $y = \frac{1-a}{1-b}$. By inverting this system, one obtains $b = \frac{y-1}{y-x}$, $a = x\frac{y-1}{y-x}$. Note that for all $0 < x < 1 < y$, this results in $0 < b < 1$ and $0 < a < 1$. Hence for all $0 < x < 1 < y$, f satisfies

$$(y-1)f(x) + (1-x)f(y) = 0.$$

By fixing $y = 2$, this implies that $\forall 0 < x < 1$, $f(x) = (x-1)f(2)$. Notably, $f(1/2) = -\frac{1}{2}f(2)$, and hence, now setting $x = 1/2$, for all $y > 1$, $f(y) = -2(y-1)f(1/2) = (y-1)f(2)$. Hence for all $x \neq 1$, $f(x) = (x-1)f(2)$. Then considering $\mathcal{D}_f(\pi, \pi)$, it follows that $f(1) = 0$, completing the proof[2]. □

We will show later that this lack of unicity has a profound impact on the bound.

Let us consider the f-divergence originating from function f. Since f is convex, it has a convex conjugate f^*, defined for all $y \in \mathbb{R}$ by

$$f^*(y) = \sup_{x \geq 0} xy - f(x). \tag{2.7}$$

Then f and f^* satisfy Fenchel–Young's inequality. Hence for all x, y,

$$xy \leq f(x) + f^*(y). \tag{2.8}$$

This implies that $\forall \pi \ll \pi_{\mathsf{p}}$, for all measurable real valued function D,

$$\begin{aligned}
\pi[D] &= \pi_{\mathsf{p}}\left[D\frac{\mathrm{d}\pi}{\mathrm{d}\pi_{\mathsf{p}}}\right] \\
&\leq \pi_{\mathsf{p}}\left[f^* \circ D + f\left(\frac{\mathrm{d}\pi}{\mathrm{d}\pi_{\mathsf{p}}}\right)\right] \\
&\leq \pi_{\mathsf{p}}[f^* \circ D] + \mathcal{D}_f(\pi, \pi_{\mathsf{p}}).
\end{aligned}$$

This is a Fenchel–Young's inequality of form (2.5). A consequence is that the Legendre transform of $\overline{\mathcal{D}_f}$ is bounded by $D \mapsto \pi_{\mathsf{p}}[f^* \circ D]$.

[2]We have implicitly assumed that singletons $\{h\}$ for $h \in \mathcal{H}$ are measurable. This assumption can be relaxed to $|\Sigma_{\mathcal{H}}| > 2$ (*i.e.* the sigma algebra is not limited to the empty set and the whole set), and by considering a set $\overline{h} \notin \{\varnothing, \mathcal{H}\}$. Then let $h_1 = \overline{h}$, $h_2 = \overline{h}^c$ and define δ_{h_1} as any probability measure on h_1, δ_{h_2} as any probability measure on h_2. The proof is then valid with no further adjustments.

Remark 2.1: Some properties of Legendre transforms

We list here some useful properties of f^*, valid for any convex function f such that $f(1) = 0$. We note $\partial f(x)$ the sub differential of f at x.

- $\forall x$, $f^*(x) \geq x$,
- Define $f'(0) := \inf \cup_{t>0} \partial f(t)$. If $f'(0) > -\infty$, f^* is constant on $]-\infty, f'(0)]$ and values $f^*(f'(0))$. As a consequence, $\partial f^*(x) = \{0\}$ on the interval $]-\infty, f'(0)[$.
- $\inf_{x\in\mathbb{R}} \cup \partial f^* = 0$. As a consequence, f^* is non decreasing.
- If f is such that f^* is differentiable, then $f^*(t) = tf^{*\prime}(t) - f \circ f^{*\prime}(t)$.
- Define $f'(\infty) = \sup \cup_{t>0} \partial f(t)$. If $f'(\infty) < \infty$, then $\forall x > f'(\infty)$, $f^*(x) = \infty$.

A consequence of the first and third property is that for lower bounded D, $\pi_{\mathsf{p}}[D] < \infty$ must be satisfied for $\pi_{\mathsf{p}}[f^*(D)] < \infty$ to be verified.
A consequence of the second property is that if $f'(0) > -\infty$, the moment on D involves a threshold on the values on D.
A consequence of the last property is that Fenchel–Young's inequality is trivial whenever D is not bounded by $f'(\infty)$. In other words, penalisation with f divergences such that $\liminf f(x)/x < \infty$ leads to the requirement that D is π_{p} almost surely bounded (if not, the change of measure is trivial).

2.1.3 Refinement of the upper bound

Let us evaluate our upper bound on the KL divergence. As stated above, the KL divergence is a f-divergence for $f(x) = x\log(x)$. The Legendre transform of this function is $f^*(x) = \exp(x-1)$. As a result, we obtain the upper bound

$$\overline{KL}^*(D) \leq \exp(-1)\pi_{\mathsf{p}}[\exp(D)].$$

This proves to be a very loose upper bound. The upper bound we have just constructed is *exponentially* larger than the true value. The result so far is too loose to be usable in practice, as it will lead to poor rates in the confidence level.

As noted above, various convex functions f define the same f-divergence. But since the transform f to f^* is bijective (indeed, the bi-conjugate recovers the initial function for proper convex function), these various f do not define the same moment f^*. This offers a degree on freedom on which our bound can be minimised. Noting $f_c : x \mapsto f(x) + c(x-1)$, the convex conjugate of f_c^* can be inferred from the convex conjugate of f and values $f_c^*(x) = f^*(x-c) + c$ (to prove this, notice that $xy - f_c(x) = c + (x(y-c) - f(x))$, and notice that the minima of the second term is by definition $f^*(x-c)$). Hence we can deduce that

$$\overline{\mathcal{D}_f}^*(D) \leq \inf_{c\in\mathbb{R}} \pi_\mathsf{p}\left[f^*(D-c)\right] + c. \tag{2.9}$$

Does this bridge the gap for the Kullback–Leibler's Legendre transform and its upper bound? In this case, it is easy to optimise on the positional parameter c, since the bound becomes $\exp(-1-c)\pi_\mathsf{p}\left[\exp(D)\right] + c$. The minima is obtained for $c = \log(\pi_\mathsf{p}\left[\exp(D)\right]) - 1$ and exactly recovers the formula from Csizár-Donsker-Varadhan. Hence minimisation on a single parameter enabled us to move from a quite loose bound to the tightest additive bound achievable (since for all bounded D, there exists a probability measure π for which the inequality becomes an equality).

We remark that the same degree of freedom could have been obtained by replacing D by the function $D - c$, and noting that an upper bound on $\pi[D-c]$ translates into an upper bound on $\pi[D]$ by adding c to both sides.

Is this supplementary degree of freedom sufficient in the general case? We show in the following theorem that it is indeed so for generic f-divergence under mild assumptions.

Theorem 2.1: Legendre transform of f-divergences

Consider a differentiable convex function $f : \mathbb{R}_+ \to \mathbb{R}$ such that $f(1) = 0$ and a prior measure π_p on $\mathcal{H}$.
Then for any positive measurable function D,

$$\overline{\mathcal{D}_f}^*(D) := \sup_{\pi\ll\pi_\mathsf{p}} \left\{\pi[D] - \mathcal{D}_f(\pi,\pi_\mathsf{p})\right\} \leq \inf_{c\in\mathbb{R}} \pi_\mathsf{p}\left[f^*\left(D-c\right)\right] + c. \tag{2.10}$$

If f is such that f^* is differentiable, with continuous derivative and D such that $\exists c_1 \in \mathbb{R}$ such that $1 \leq \pi_\mathsf{p}\left[f^{*\prime}(D-c_1)\right] < \infty$,

$$\overline{\mathcal{D}_f}^*(D) = \inf_{c\in\mathbb{R}} \pi_\mathsf{p}\left[f^*\left(D-c\right)\right] + c. \tag{2.11}$$

Moreover, any c^* such that $\pi_\mathsf{p}\left[f^{*\prime}(D-c^*)\right] = 1$ is a minimiser of the right hand side of (2.11), while the probability measure defined by $\frac{\mathrm{d}\pi^*}{\mathrm{d}\pi_\mathsf{p}} = f^{*\prime}(D-c^*)$ maximises Equation (2.4).

Theorem 2.1. Starting from (2.9), we know that

$$\sup_{\pi\in\Pi} \left\{\pi\left[D\right] - \mathcal{D}_f(\pi,\pi_\mathsf{p})\right\} = \overline{\mathcal{D}_f}^*(D) \leq \inf_{c\in\mathbb{R}} \left\{\pi_\mathsf{p}\left[f^*(D-c)\right] + c\right\}. \tag{2.12}$$

Let us assume for the moment that there exists $c^* \in \mathbb{R}$ such that $\pi_\mathsf{p}\left[f^{*\prime}(D-c^*)\right] = 1$. Since $f^{*\prime}$ takes non negative values, we can define the probability measure π^* such that $\frac{\mathrm{d}\pi^*}{\mathrm{d}\pi_\mathsf{p}} =$

$f^{*\prime}(D-c^*)$. Therefore

$$\begin{aligned}\overline{\mathcal{D}_f}^*(D) &= \sup_{\pi\in\Pi}\left(\pi[D]-\mathcal{D}_f(\pi,\pi_{\mathsf{p}})\right)\\ &\geq \pi_{\mathsf{p}}\left[Df^{*\prime}(D-c^*)-f\circ f^{*\prime}(D-c^*)\right]\\ &\geq \pi_{\mathsf{p}}\left[f^*(D-c^*)\right]+c^*\\ &\geq \inf_c\left\{\pi_{\mathsf{p}}\left[f^*(D-c)\right]+c\right\},\end{aligned}$$

which implies equality from Equation (2.12). There only remains to prove the existence of c^*. Define $M: c\mapsto \pi\left[f^{*\prime}(D-c)\right]$. Our assumptions guarantee that $\exists c_1$ such that $M(c_1)\geq 1$. If $M(c_1)=1$, this concludes the result. If $M(c_1)>1$, notice that M is non increasing, and for all $c>c_1$, is bounded by $M(c_1)<\infty$. Since $f^{*\prime}(x)\to_{x\to-\infty}0$, it follows by Lebesgue's dominated convergence theorem that $M(c)\to_{c\to\infty}0$. Moreover, since $f^{*\prime}$ is continuous, it follows that $M(c)$ is continuous on $]c_1,\infty[$ [Schilling, 2005]. Hence, by the intermediate value theorem, it follows that there exists $c^*>c_1$ such that $M(c^*)=1$. □

Remark 2.2: Some intuition on the proof of Theorem 2.1

The proof of Theorem 2.1 is based on the knowledge of the form of the minimiser in the Legendre transform definition. It is possible to motivate such a form by considering a Lagrange multiplier. Starting from the definition of the Legendre transform, one can reframe the minimisation problem on the probability measure π as a minimisation problem on a positive function g, of the criteria

$$\sup_{c\in\mathbb{R}} L(c,g)=\pi_{\mathsf{p}}[Dg-f\circ g]-c(\pi_{\mathsf{p}}[g]-1)$$

where c is a Lagrange multiplier. Considering a perturbation function δg, one obtains for $\epsilon\to 0$

$$L(c,g+\epsilon\delta g)-L(c,g)=\epsilon(\pi_{\mathsf{p}}\left[(D-f'\circ g-c)\delta g\right])+o(\epsilon).$$

This implies that for the optima g^*, $D-c=f'\circ g^*$ where $g^*>0$, and $D-c-f'\circ g^*\geq 0$ wherever $g^*=0$. This implies that $g^*=f^{*\prime}\circ(D-c)$ for all points. Moreover, the Lagrange multiplier must be such that $\pi_{\mathsf{p}}[g^*]=1$, hence that $\pi_{\mathsf{p}}\left[f^{*\prime}(D-c)\right]=1$. For such a c^*, the value of the objective is

$$\begin{aligned}\pi_{\mathsf{p}}[Dg^*-f\circ g^*] =&\pi_{\mathsf{p}}\left[Df^{*\prime}\circ(D-c^*)-f\circ f^{*\prime}\circ(D-c^*)\right]\\ =&\pi_{\mathsf{p}}\left[(D-c^*)f^{*\prime}\circ(D-c^*)-f\circ f^{*\prime}\circ(D-c^*)\right]+c^*\\ =&\pi_{\mathsf{p}}\left[f^*(D-c^*)\right]+c^*.\end{aligned}$$

Hence the motivation for the value of the c achieving the lower bound.

Remark 2.3

A consequence of Theorem 2.1 is that the Legendre transform must satisfy for all π_{p}, π, and f satisfying the assumptions of Theorem 2.1

$$\forall c \in \mathbb{R}, \overline{\mathcal{D}_f}^*(D+c) = \overline{\mathcal{D}_f}^*(D) + c. \tag{2.13}$$

Proof. First of all, if $\exists c_1$ such that $1 < \pi_{\mathsf{p}}\left[f^{*\prime}(D-c_1)\right] < \infty$, the condition is also valid for $D+c$, using $\tilde{c}_1 = c_1 + c$. Then using Equation (2.11),

$$\begin{aligned}
\overline{\mathcal{D}_f}^*(D+c) &= \inf_{\tilde{c}\in\mathbb{R}} \left(\pi_{\mathsf{p}}\left[f^*(D+c-\tilde{c})\right] + \tilde{c}\right) \\
&= \inf_{\bar{c}\in\mathbb{R}} \left(\pi_{\mathsf{p}}\left[f^*(D-\bar{c})\right] + \bar{c}\right) + c \\
&= \overline{\mathcal{D}_f}^*(D) + c.
\end{aligned}$$

□

2.1.4 Further improvement for regular f

The optimisation problem involved in Equation (2.9) is not always practicable, and it might be necessary to use the bound given by some approximation of c^*. This motivates the search for tighter bounds of $\overline{\mathcal{D}}_f^*$ at a given c, that is to say functions $\tilde{\mathcal{D}}_{f,c}^*$ such that

$$\overline{\mathcal{D}}_f^*(D) \leq \tilde{\mathcal{D}}_{f,c}^*(D) \leq \pi_{\mathsf{p}}\left[f^*(D-c)\right] + c.$$

We show that when f is twice differentiable and such that $1/f''$ is concave, such tighter bounds can be constructed.

Theorem 2.2: A tighter upper-bound for regular f

Consider a twice differentiable convex function $f : \mathbb{R}^+ \to \mathbb{R}$, such that $f(1) = 0$ and $1/f''$ is concave. Then, noting

$$\tilde{\Delta}_f(D) := \begin{cases} f^* \circ f'\left(\pi_{\mathsf{p}}\left[f^{*\prime} \circ D\right]\right) - f'\left(\pi_{\mathsf{p}}\left[f^{*\prime} \circ D\right]\right) & \text{if } \pi_{\mathsf{p}}\left[\left|f^{*\prime} \circ D\right|\right] < \infty \\ 0 & \text{else,} \end{cases}$$

one can upper bound the Legendre transform of the f-divergence for D lower bounded by

$$\overline{\mathcal{D}}_f^*(D) \leq \inf_{c\in\mathbb{R}} \pi_{\mathsf{p}}\left[f^* \circ (D-c)\right] - \tilde{\Delta}_f(D-c) + c. \tag{2.14}$$

Proof. Let us remark that upper bounding $\mathcal{D}_f^*(D)$ by a function $G(D)$ is equivalent to establish a Young–Fenchel inequality replacing $\mathcal{D}_f^*(D)$ by $G(D)$, since

$$G(D) + \mathcal{D}_f(\pi, \pi_{\mathsf{p}}) \geq \pi[D] \Rightarrow G(D) \geq \sup_{\pi} \pi[D] - \mathcal{D}_f(\pi, \pi_{\mathsf{p}}) = \mathcal{D}_f^*(D).$$

We will thus prove a Young–Fenchel version of Equation (2.14) for $c = 0$. Replacing D by $D - c$ in the resulting Young–Fenchel inequality the implies the inequality for all c.

The proof starts with Lemma 14.2 in Boucheron et al. [2013], which states that for any f convex, twice differentiable on $\mathbb{R}_+^*$ such that $\frac{1}{f''}$ is concave, for any $Z > 0$ such that $f(Z)$ is π_{p}-integrable, then

$$\pi_{\mathsf{p}}\left[f(Z)\right] - f\left(\pi_{\mathsf{p}}\left[Z\right]\right) = \sup_{T \neq 0} \left\{ \pi_{\mathsf{p}}\left[\left(f'(T) - f'(\pi_{\mathsf{p}}\left[T\right])\right)(Z - T) + f(T)\right] - f\left(\pi_{\mathsf{p}}\left[T\right]\right)\right\}$$

where the supremum is taken on all non negative π_{p}-integrable random variables T. The maximum is achieved for $T = Z$.

For $\pi \ll \pi_{\mathsf{p}}$ such that $\mathcal{D}_f(\pi, \pi_{\mathsf{p}}) < \infty$, $Z = \frac{\mathrm{d}\pi}{\mathrm{d}\pi_{\mathsf{p}}}(\omega)$ for $\omega \sim \pi_{\mathsf{p}}$ is π_{p} integrable. Hence Lemma 14.2 implies

$$\begin{aligned}
\mathcal{D}_f\left(\pi, \pi_{\mathsf{p}}\right) &= \sup_{T \neq 0} \left\{ \pi\left[f'(T)\right] + \pi_{\mathsf{p}}\left[-Tf'(T) + (T-1)f'(\pi_{\mathsf{p}}\left[T\right]) + f(T) - f(\pi_{\mathsf{p}}\left[T\right])\right]\right\} \\
&= \sup_{T \neq 0} \left\{ \pi\left[f'(T)\right] - \left(\pi_{\mathsf{p}}\left[f^* \circ f'(T)\right] - f^* \circ f'\left(\pi_{\mathsf{p}}\left[T\right]\right) + f'(\pi_{\mathsf{p}}\left[T\right])\right)\right\}
\end{aligned}$$

where we used $xf'(x) - f(x) = f^* \circ f'(x)$ twice to obtain the second equality. Consider the change of variable $D = f' \circ T$ which maps π_{p}-integrable random variables to the set of functions

$$\mathcal{T} := \left\{ D \mid \forall h \in \mathsf{Supp}\left(\pi\right),\ f'(0) \leq D(h) \leq f'(\infty),\ \pi_{\mathsf{p}}\left[\left|f^{*\prime} \circ D\right|\right] < \infty \right\}.$$

Note that this change of variable is well defined, since f is twice differentiable implies that f' has inverse $f^{*\prime}$. The bound becomes

$$\mathcal{D}_f\left(\pi, \pi_{\mathsf{p}}\right) = \sup_{D \in \mathcal{T}} \pi\left[D\right] - \left(\pi_{\mathsf{p}}\left[f^* \circ D\right] - \tilde{\Delta}_f(D)\right).$$

We extend outside of $\mathcal{T}$ by checking the behaviour of the bound when some conditions are broken. We can first relax the condition that $D \geq f'(0)$ since $f^{*\prime} \circ D = f^{*\prime}(\max(f'(0), D))$ and $f^* \circ D = f^*(\max(f'(0), D))$. Then, we can relax the hypothesis that $f^{*\prime} \circ D$ is π_{p} integrable. If it is not the case, then $\tilde{\Delta}_f(D) = 0$ and the Young–Fenchel inequality from Equation (2.10) implies that $\mathcal{D}_f(\pi, \pi_{\mathsf{p}}) \geq \pi[D]$. Hence no element higher than $\mathcal{D}_f(\pi, \pi_{\mathsf{p}})$ is added in the sup,

and hence this does not increase the sup. Thus

$$\begin{aligned}\mathcal{D}_f(\pi, \pi_{\mathsf{p}}) &= \sup_{D \in \mathcal{T}} \left\{ \pi[D] - \left(\pi_{\mathsf{p}}[f^* \circ D] - \tilde{\Delta}_f(D) \right) \right\} \\ &= \sup_{\substack{D < f'(\infty), \\ \pi_{\mathsf{p}}[D] < \infty}} \left\{ \pi[D] - \left(\pi_{\mathsf{p}}[f^* \circ D] - \tilde{\Delta}_f(D) \right) \right\} \\ &= \sup_{D} \left\{ \pi[D] - \left(\pi_{\mathsf{p}}[f^* \circ D] - \tilde{\Delta}_f(D) \right) \right\}.\end{aligned}$$

This implies Young–Fenchel's inequality, which implies the result. □

Remark 2.4

Equation (2.14) gives a better approximation that equation Equation (2.10). Indeed, by definition,

$$f^*(t) = \sup_{x>0} xt - f(x) \geq t$$

since $f(1) = 0$. Hence $\forall t$, $t - f^*(t) \leq 0$. This implies that $\forall D$, $\tilde{\Delta}_f \geq 0$.

Remark 2.5

Consider a convex function f and D satisfying both the assumptions of Theorem 2.1 and Theorem 2.2. Then the inequality in (2.14) is an equality. Notably, it is met for any c^* satisfying the condition given in Theorem 2.1.

Proof. This is a consequence of the fact for a given c, the right hand side of Equation (2.14) is upper bounded by the right hand side of Equation (2.10), and lower bounded by $\overline{D}_f^*$. Since the minima of the right hand side of Equation (2.10) matches $\overline{D}_f^*$, so must the minima of the right hand side of Equation (2.14). □

If one considers Kullback–Leibler, the evaluation of the right hand side of Equation (2.14) gives $-\log(\pi_{\mathsf{p}}[\exp(-D)])$ for any c, exactly matching the exact Legendre transform. This shows that the refined bound can significantly improve on the standard bound.

Remark 2.6

The condition $1/f''$ concave implies that

$$\liminf f^*(t)/t^2 > 0.$$

This can be interpreted as a requirement that D has at least second order moment for Theorem 2.2 to yield non vacuous bounds.

Proof. Since f is convex and twice differentiable, it follows that $f'' >= 0$. Therefore, the concave function $1/f''$ is concave and positive on $\mathbb{R}_+^*$.

Let us show that this implies that $1/f''$ is non decreasing. Suppose that there exists $x_1 > x_2 > 0$, $1/f''(x_1) < 1/f''(x_2)$, the concavity of $1/f''$ implies that for all $x > x_1$,

$$\frac{1}{f''(x)} \leq -\left(\frac{1}{f''(x_2)} - \frac{1}{f''(x_1)}\right)\frac{x - x_2}{x_1 - x_2} + \frac{1}{f''(x_2)}.$$

As the right hand side goes to $-\infty$ as $x \to \infty$, this is impossible, and hence $1/f''$ must be non decreasing.

There thus exists $t_0 > 0$, $\alpha = 1/f''(t_0) > 0$ such that for all $t > t_0$, $1/f''(t) \geq \alpha$. Hence $f''(t) \leq \alpha^{-1}$, which implies $f'(t) \leq \alpha^{-1}(t - t_0) + f'(t_0)$, and hence, using the fact that $f^{*\prime}$ is increasing and the inverse of f', that $t \leq f^{*\prime}\left(\alpha^{-1}(t - t_0) + f'(t_0)\right)$. Thus for all $t > f'(t_0)$, we have $\alpha t + t_0 - \alpha f'(t_0) \leq f^{*\prime}(t)$. By integration, it follows that $f^*(t)/t^2 \geq \alpha/2 + O(1/t)$ for all $t > f'(t_0)$. Taking $t \to \infty$ concludes the proof. □

2.1.5 A temperature degree of freedom

Another way to improve the resulting bound is the introduction of a scale degree of freedom. Due to its close relationship to the Gibbs temperature in the case where the f-divergence is the KL divergence, we call this degree of freedom the temperature and note it λ. This degree of freedom can be introduced in two equivalent ways; either by replacing the convex function f by λf, or by replacing the generalised generalisation gap D by $\lambda^{-1}D$. In both cases, this result in an extended form of the bound as

$$\pi_{\mathsf{p}}[D] \leq \lambda \pi_{\mathsf{p}}\left[f^* \circ (\lambda^{-1}D)\right] + \lambda \mathcal{D}_f(\pi, \pi_{\mathsf{p}}).$$

We now give the most general form of Young–Fenchel's inequality with f-divergence penalisation in the following theorem.

Theorem 2.3

For π_{p} a probability measure on $\mathcal{H}$, for f a convex function such that $f(1) = 0$, then for any lower bounded, measurable function D, for all $\lambda > 0$, for all $c \in \mathbb{R}$,

$$\pi[D] \leq \lambda \pi_{\mathsf{p}}\left[f^*(\lambda^{-1}(D - c))\right] + c + \lambda \mathcal{D}_f(\pi, \pi_{\mathsf{p}}). \tag{2.15}$$

Moreover, if f is twice differentiable such that $1/f''$ is concave, then

$$\pi[D] \leq \lambda \pi_{\mathsf{p}}\left[f^*(\lambda^{-1}(D - c))\right] - \tilde{\Delta}(\lambda^{-1}(D - c)) + c + \lambda \mathcal{D}_f(\pi, \pi_{\mathsf{p}}). \tag{2.16}$$

Remark 2.7

To use Theorem 2.3 to bound $\pi[D]$ simultaneously for all π, then there must exist c such that the f^* moment of $\lambda^{-1}(D-c)$ is upper bounded. This implies that in the case of KL, the exponential moment assumption can not be weakened, since $f^*(t) = \exp(t-1)$.

Theorem 2.3 states that we can control the average of the generalised generalisation gap D over all probability measures π from two terms : a measure of the distance between π and π_{p}, and what is morally a moment of the random variable with respect to π_{p}. These two terms offer a trade-off between the type of penalisation considered - controlled by how fast f grows - and the strength of the moment assumption - controlled by how fast f^* grows (see Figure 2.1). The more the f-divergence discriminates between π and π_{p}, the weaker is the moment needed. On the other hand, if strong moment assumptions can be made on the random variable, one can control its mean over π for a larger set of probability measures.

The bounds can be optimised on two degrees of freedom, the positional parameter c and a scale parameter λ. Theorem 2.1 implies that the first optimisation factor can recover the optimal additive bound for regular f and generalisation gap with bounded $f^{*\prime}$ moments[3]. Moreover, it states that optimisation on c is related to the normalisation problem for the probability measure reaching the upper bound (the maximiser in the definition of the Legendre transform).

Although the bounds can be optimised on two degrees of freedom, it might not be possible to apply this double optimisation procedure. We do not have clear arguments to favour optimising with respect to c over optimising with respect to λ or vice-versa. We note however that most of the bounds we examined proved easier to optimise on the scale parameter rather than on the positional degree of freedom.

Remark 2.8

For all $f \in \mathcal{F}$, the reverse f-divergence $\nu, \mu \mapsto \mathcal{D}_f(\mu, \nu)$ is a f-divergence for $\tilde{f}(t) = t \times f(1/t)$. Therefore, Theorem 2.3 provides change of measure inequalities for the reverse f-divergence.

Remark 2.9

Embedded in the bounds of Theorem 2.3 is the condition that $\lambda(D-c) < f'(\infty)$ π_{p}-almost surely (see Remark 2.1) As a consequence, whenever $f'(+\infty) \neq +\infty$, the generalisation gap D needs to be upper bounded for the bounds to be useable. Whenever this is the case, we find it good practice to choose f such that $f'(\infty) = 0$ and as such, we can apply the bound to $\lambda(D - D_{\max} - c)$ for $\lambda > 0, c > 0$.

[3]We conjecture that this assumption can be relaxed, in the sense that if $\forall c$, $\pi_{\mathsf{p}}\left[f^{*\prime}(D-c)\right] = \infty$, then the Legendre transform is infinite. Whether there are settings where $\exists c_1$ such that $\pi_{\mathsf{p}}\left[f^{*\prime}(D-c_1)\right] < 1$ but not c_2 such that $\pi_{\mathsf{p}}\left[f^{*\prime}(D-c_2)\right] > 1$ and how the bound would behave in such cases is also an open question.

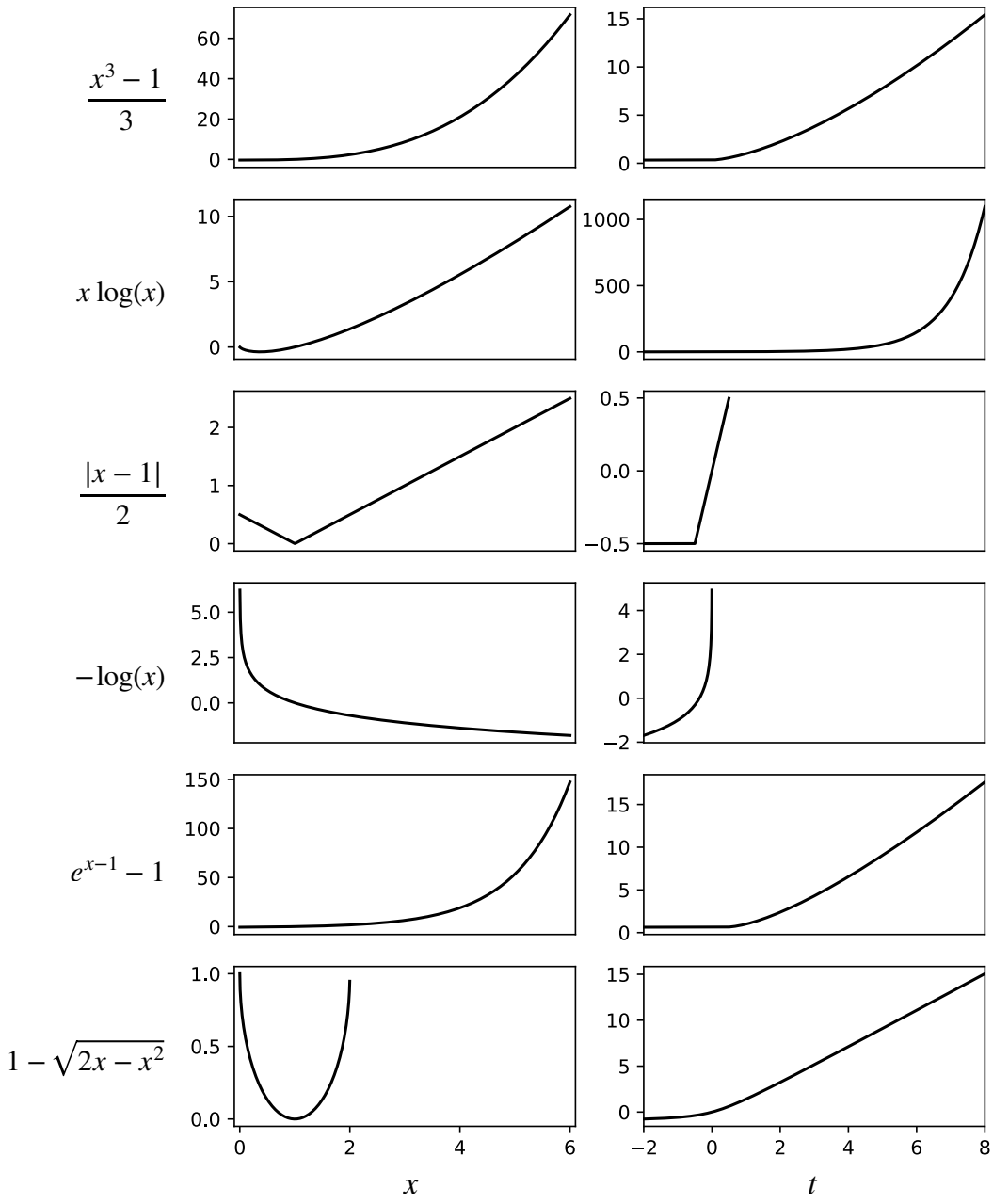


Figure 2.1: Various convex functions f satisfying $f(1) = 0$ (on the left column) and their corresponding Legendre transform f^* (on the right column). The first row corresponds to a power 3 divergence, resulting in f^* behaving asymptotically as $t^{3/2}$. The second row corresponds to Kullback–Leibler divergence, and result in an exponential Legendre transform. The third and fourth rows describe weak penalisation ($f'(\infty) < \infty$), leading to f^* with upper bounded support (note that the third row corresponds to the Total variation distance). The fifth and sixth rows describe strong penalisation, resulting in slowly increasing f^*.

Remark 2.10

One can reinterpret the minimisation on λ and c for every bound of the form (2.15) in term of Legendre transforms. Indeed, for $\lambda > 0$ and $c \in \mathbb{R}$, consider

$$L_c : \lambda \mapsto \lambda \pi_{\mathsf{p}} \left[f^* \circ (\lambda^{-1}(D - c)) \right],$$
$$L_\lambda : c \mapsto \lambda \pi_{\mathsf{p}} \left[f^* \circ (\lambda^{-1}(D - c)) \right].$$

Then both L_c and L_λ are convex functions, and the minimisation of the bound on λ yields

$$\pi[D] \leq -L_c^*(-\mathcal{D}_f(\pi, \pi_{\mathsf{p}})) + c,$$

while the minimisation of the bound on c yields

$$\pi[D] \leq -L_\lambda^*(-1) + \lambda \mathcal{D}_f(\pi, \pi_{\mathsf{p}}).$$

Moreover, if

$$L : \lambda, c \mapsto \lambda \pi_{\mathsf{p}} \left[f^* \circ (\lambda^{-1}(D - c)) \right]$$

is convex, then the bound can be interpreted as

$$\pi[D] \leq -L^* \left(\begin{pmatrix} -\mathcal{D}_f(\pi, \pi_{\mathsf{p}}) \\ -1 \end{pmatrix} \right).$$

A similar argument can be used for the bound of form (2.16), although in this case, the functions L_c and L_λ might not be convex any longer.

Remark 2.11

To define the Legendre transform of f in Equation (2.7), we consider a suprema on $x \in \mathbb{R}_+$. This is equivalent to extending f to $\mathbb{R}$ by setting $f(x) = +\infty$ for all $x < 0$ (any negative x is thus ruled out since it leads to $-\infty$ in the bound). As noted in Remark 2.1, this introduces a threshold at $f'(0)$ in the values of f^*; that is to say, $f^*(x) = -\liminf_{x \to 0} f(x) := f(0)$ for all $x \leq f'(0)$. Hence the functional D can be replaced by $\max(D, f'(0))$.

This threshold can be problematic when trying to optimise the bounds on the two degrees of freedom. A way to obtain more tractable bounds is to consider other convex extensions of f to $\mathbb{R}$ in the definition of f^*. If $f'(0) > -\infty$, f can be extended for $x < 0$ by $\tilde{f}(x) = f(0) + xf'(0)$. If $f''(0) < \infty$ moreover, $\tilde{f}(x) = f(0) + xf'(0) + \frac{x^2}{2} f''(0)$ for all $x < 0$ also provides a convex extension of f. Specific f might also have natural extensions (*i.e.* power functions). Since Young–Fenchel's inequality remains valid for these $\tilde{f}^*$, the upper bounds of form (2.15) also remain valid. While these will result in looser bounds,

the added tractability might result in better bound after optimisation.

2.1.6 Application to Learning Theory

Some more PAC-Bayesian bounds

We now explore how Theorem 2.3 can be leveraged in learning theory. So far, the change of measure was performed for any function D. To obtain PAC-Bayes bounds, one can replace D by the generalisation gap $D = \tilde{R} - R$ (or $D = \max(\tilde{R} - R, 0)$ if the risk is not bounded) in Equation (2.15) to obtain, $\forall c \in \mathbb{R}$, $\forall \lambda \in \mathbb{R}_+$, $\forall \pi_{\mathsf{p}}$, with probability higher than $1-\delta$, $\forall \pi \ll \pi_{\mathsf{p}}$

$$\pi\left[\tilde{R}\right] \leq \pi\left[R\right] + \frac{\lambda \mathbb{P}\left[\pi_{\mathsf{p}}\left[f^* \circ (\lambda^{-1}(D-c))\right]\right]}{\delta} + \lambda \mathcal{D}_f(\pi, \pi_{\mathsf{p}}) + c.$$

This is simply a consequence of Markov's inequality on the moment term. Note that in this expression, the degree of freedom λ and c must be set before using Markov's inequality on the term of the right hand side independent of π. While the optimisation on λ depends on the value of π and hence couples the two terms of the Young–Fenchel inequality, the optimal choice on c is independent on π and can therefore be put into the expected value. This yields the improved PAC-Bayes bound, stating that $\forall \lambda \in \mathbb{R}_+$, $\forall \pi_{\mathsf{p}}$, with probability higher than $1-\delta$, $\forall \pi \ll \pi_{\mathsf{p}}$

$$\pi\left[\tilde{R}\right] \leq \pi\left[R\right] + \frac{\mathbb{P}\left[\inf_c \lambda \pi_{\mathsf{p}}\left[f^* \circ (\lambda^{-1}(D-c))\right] + c\right]}{\delta} + \lambda \mathcal{D}_f(\pi, \pi_{\mathsf{p}}).$$

Note that concentration inequalities other than Markov can be used to improve the resulting PAC-Bayes bound. For instance, for $\mathcal{D}_f = \mathrm{KL}$, Catoni's bound (1.20) can be recovered using a Chernoff bound. When dependent data is considered, adapted concentration inequalities should be used to bound the quantiles of

$$\inf_c \pi_{\mathsf{p}}\left[\lambda f^* \circ (\lambda^{-1}(D-c))\right] + c.$$

The general form considered in Theorem 2.3 can also be leveraged to obtain tighter PAC-Bayes bound using the generalised generalisation gap approach developed by Bégin et al. [2016]. Considering generalisation gaps of form $D(\omega) = \Delta(\tilde{R}(\omega), R(\omega))$ with Δ a convex function, one can "inverse" the Δ function through $\Delta^{-1}(t, y) = \sup\{x \mid \Delta(x, y) \leq t\}$. It then follows from Jensen's inequality and Theorem 2.3 that $\forall \lambda > 0$, $\forall \pi_{\mathsf{p}}$, $\forall \delta$, with probability at least $1-\delta$, $\forall \pi \ll \pi_{\mathsf{p}}$,

$$\pi[\tilde{R}] \leq \Delta^{-1}\left(\frac{\mathbb{P}\left[\inf_c \pi_{\mathsf{p}}\left[f^*\left(\lambda^{-1}\Delta\left(\tilde{R}, R\right) - c\right)\right] + c\right]}{\delta} + \lambda \mathcal{D}_f(\pi, \pi_{\mathsf{p}}), \pi\left[R\right]\right) \tag{2.17}$$

where once again, Markov's inequality can be replaced with a more strategic concentration inequality, and a fixed c might be used for convenience.

Remark 2.12

The change of measures inequality of Theorem 2.3 are valid *for all* generalised generalisation gaps D. As a consequence, it follows that the bound can *theoretically* be optimised on the convex function Δ, and even on all f-divergences, resulting in

$$\forall \pi \ll \pi_{\mathsf{p}}, \pi[\tilde{R}] \leq \inf_{f \in \mathcal{F}} \inf_{\Delta} \Delta^{-1} \left(\pi_{\mathsf{p}} \left[f^* \left(\Delta \left(\tilde{R}, R \right) \right) \right] + \mathcal{D}_f(\pi_{\mathsf{p}}, \pi), \pi[R] \right) \tag{2.18}$$

where $\mathcal{F}$ is the set of all convex functions of $\mathbb{R}_+$ to $\mathbb{R}$ such that $f(1) = 0$, and the minima on Δ is taken on all lower bounded convex functions. Note that this formulation recovers both degrees of freedom, and should be quite tight.
However, to obtain a PAC-Bayes bound, it is necessary to upper bound the quantiles of the right hand side *simultaneously* for all posterior distributions π. This, in the general case, prevents optimisation on a degree of freedom whenever the optimal value depends on the posterior distribution (this couples the value of the bound to π) - and the analysis is blocked when there is no closed form expression for the minima (the coupling is unknown). Hence such optimised forms as Equation (2.18) are in the general case of little use in obtaining PAC-Bayes bounds. For some tractable bounds, it is possible to upper bound the quantiles for all π *after* optimisation on the temperature degree of freedom λ (see Section 2.1.7).

From moment assumption to penalisation

We now consider a setting where the generalised generalisation gap D is fixed. We study in this section the task of reverse engineering assumptions on the M-moment of the generalisation gap into PAC-Bayes bound. In plain words, the question we are trying to answer is whether we can transform an assumption of form $\forall h \in \mathcal{H}$, $\mathbb{P}\left[M \circ D(h)\right] \leq \alpha$ into a PAC-Bayes bound. We show that such a strategy is indeed possible if M goes faster to infinity than linearly, and study how the form of M impacts the bound.

Let us assume that $D > 0$, and that the function M satisfies $\lim_{t \to +\infty} M(t)/t = +\infty$ and $M(t) > t$. Denote M_- the lower convex envelope of M. Note that M_- also satisfies $\lim_{t \to +\infty} M(t)/t = +\infty$ (since the conditions imply that for all $a > 0$, $\exists b$ and t_a such that $M(t) > at + b \; \forall t > t_a$). Since for all values of D, $0 \leq M_- \leq M$, it follows that $\mathbb{P}\left[M_- \circ D\right] \leq \alpha$. Using $f = M_-^* - M_-^*(1)$ in Theorem 2.3 implies that

$$\pi[D] \leq \pi_{\mathsf{p}}\left[M_- \circ D + M_-^*(1) \times D\right] + \mathcal{D}_{M_-^* - M_-^*(1)}.$$

Using the fact that $M_-(t) > t$ (inherited from the condition on M), this implies that $\forall \pi$

$$\pi[D] \leq (1 + M_-^*(1)) \pi_{\mathsf{p}}\left[M_- \circ D\right] + \mathcal{D}_{M_-^* - M_-^*(1)}(\pi, \pi_{\mathsf{p}}).$$

Finally, using Fubini in conjunction with Markov's inequality implies that with probability at least

$1-\delta$, for all π,

$$\pi\left[D\right] \leq \frac{(1+M_{-}^{*}(1))\alpha}{\delta} + \mathcal{D}_{M_{-}^{*}-M_{-}^{*}(1)}(\pi, \pi_{\mathsf{p}}). \tag{2.19}$$

If D is of form $\Delta(\tilde{R}, R)$, the same Jensen argument as in Equation (2.17) can be used, leading to

$$\pi\left[\tilde{R}\right] \leq \Delta^{-1}\left(\frac{(1+M_{-}^{*}(1))\alpha}{\delta} + \mathcal{D}_{M_{-}^{*}-M_{-}^{*}(1)}(\pi, \pi_{\mathsf{p}}), \pi\left[R\right]\right) \tag{2.20}$$

holding simultaneously for all π with probability at least $1-\delta$.

2.1.7 Some change of measure inequalities

Standard f-divergence

We apply Theorem 2.1 to the most popular f-divergences found in the literature. Table 2.1 presents a summary of all the resulting change of measure inequalities. Note that the optimal values of λ and the Legendre transform f^* are gathered, when available, in Table B.1 which is deferred to Appendix B.

Equation (2.14) recovers the exact Legendre transform of the KL divergence. The bound is also quite tight for Pearson's χ^2- divergence. Indeed, Theorem 2.1 implies that for D such that $D > 0$ and $\pi_{\mathsf{p}}\left[D\right] \leq 1$, $\overline{\mathcal{D}}_f^*(D)$ has closed form expression $\frac{1}{2}\mathbb{V}_{\pi_{\mathsf{p}}}\left[D\right] + \pi_{\mathsf{p}}\left[D\right]$ (see proof in Appendix B.1).

The bounds presented in Table 2.1 are coherent with those obtained independently by Ohnishi and Honorio [2021]. The last three are, to the best of our knowledge, the first change of measure inequalities obtained for these f-divergences.

For KL, one recovers the change of measure inequality established by Csiszár [1975] and Donsker and Varadhan [1975]. That bound is the starting point of the proof of the general PAC bound established by Bégin et al. [2016], which recovers bounds obtained in Langford and Seeger [2001], McAllester [2003], Catoni [2007] and Alquier et al. [2016].

For power p divergences with $p > 1$, only moments of order $\frac{p}{p-1} = q$ for D are needed rather than exponential moments, considerably lessening the assumptions needed on the loss l and the underlying data distribution. When $1 < p \leq 2$, the bounds we propose improve on those obtained in Alquier and Guedj [2018]. Indeed, these last bounds exactly match those we obtain through Equation (2.15) for $c = 0$ after minimisation on λ, which is looser than Equation (2.16) which we consider. The bounds for $1 < p \leq 2$ can be slightly simplified, noticing that

$$\pi_{\mathsf{p}}\left[D_{+}^{q}\right] - \pi_{\mathsf{p}}\left[D_{+}^{\frac{q}{p}}\right]^{p} \leq \pi\left[D^{q}\right] - \pi_{\mathsf{p}}\left[D^{\frac{q}{p}}\right]^{p}.$$

For all the remaining f-divergences, $f'(\infty) < \infty$. Therefore the Legendre transforms of these f-divergences only take real values on bounded functions D. The bounds are of the form $D_{\max}$ minus a term involving the moment of $D_{\max} - D$.

Table 2.1: Bounds for typical f-divergence. We denote $D_+ := \max(D, 0)$. For power-divergences, q is such that $\frac{1}{q} + \frac{1}{p} = 1$. For Lin's measure, f_ϑ is given by $f_\vartheta(t) = (\vartheta t \log(t\vartheta) - (\vartheta t + 1 - \vartheta)\log(\vartheta t + 1 - \vartheta) - \vartheta \log(\vartheta)$.

f-div	$f(t) =$	$\pi[D] \leq \dots$	c, λ
KL	$t\log(t)$	$\lambda \log \pi_{\mathsf{p}}\left[\exp\left(\lambda^{-1}D\right)\right] + \lambda \mathsf{KL}(\pi, \pi_{\mathsf{p}})$	$\lambda > 0$
Power-p, $1 < p \leq 2$	$t^p - 1$	$\pi_{\mathsf{p}}\left[D_+^{q-1}\right]^{p-1} + \left(\pi_{\mathsf{p}}\left[D_+^q\right] - \pi_{\mathsf{p}}\left[D_+^{\frac{q}{p}}\right]^p\right)^{\frac{1}{q}} \mathcal{D}_{f_p}(\pi, \pi_{\mathsf{p}})^{\frac{1}{p}}$	
Power-p, $1 < p$	$t^p - 1$	$c + \pi_{\mathsf{p}}\left[(D-c)_+^q\right]^{\frac{1}{q}} \left(1 + \mathcal{D}_{f_p}(\pi, \pi_{\mathsf{p}})\right)^{\frac{1}{p}}$	$c \in \mathbb{R}$
Pearson χ^2	$t^2 - 1$	$\pi_{\mathsf{p}}[D_+] + \mathbb{V}_{\pi_{\mathsf{p}}}[D_+]^{\frac{1}{2}} \chi^2(\pi, \pi_{\mathsf{p}})^{\frac{1}{2}}$	
Power-p, $0 < p < 1$	$1 - t^p$	$D_{\max} + c - \pi_{\mathsf{p}}\left[(D_{\max} - D + c)^q\right]^{\frac{1}{q}} \left(1 - \mathcal{D}_{f_p}(\pi, \pi_{\mathsf{p}})\right)^{\frac{1}{p}}$	$c > 0$
Power-p, $p < 0$	$t^p - 1$	$D_{\max} + c - \pi_{\mathsf{p}}\left[(D_{\max} - D + c)^q\right]^{\frac{1}{q}} \left(1 + \mathcal{D}_{f_p}(\pi, \pi_{\mathsf{p}})\right)^{\frac{1}{p}}$	$c \geq 0$
TV	$\lvert t-1\rvert / 2$	$D_{\max} + \pi_{\mathsf{p}}\left[\max\left(D - D_{\max}, -\lambda\right)\right] + \lambda \mathsf{TV}(\pi, \pi_{\mathsf{p}})$	$\lambda > 0$
Squared Hellinger	$1 - \sqrt{t}$	$D_{\max} + c - \left(1 - H^2(\pi, \pi_{\mathsf{p}})\right)^2 \pi_{\mathsf{p}}\left[\frac{1}{D_{\max} - D + c}\right]^{-1}$	$c > 0$
Reverse Pearson	$t^{-1} - 1$	$D_{\max} + c - \frac{\pi_{\mathsf{p}}\left[\sqrt{c + D_{\max} - D}\right]^2}{1 + \chi^2(\pi_{\mathsf{p}}, \pi)}$	$c > 0$
Reverse KL	$-\log(t)$	$D_{\max} + c - \exp\left(\pi_{\mathsf{p}}\left[\log\left(D_{\max} - D + c\right)\right]\right) - \mathsf{KL}\left(\pi_{\mathsf{p}}, \pi\right))$	$c > 0$
Lin's measure ($\vartheta \in]0, 1[$)	$f_\vartheta(t)$	$D_{\max} + c$ $- \lambda(1-\vartheta)\pi_{\mathsf{p}}\left[\log\left(1 - \exp\left(\lambda^{-1}\vartheta^{-1}(D - D_{\max} - c)\right)\right)\right]$ $+ \lambda\left(L_\vartheta(\pi, \pi_{\mathsf{p}}) + (1-\vartheta)\log(1-\vartheta) - \vartheta\log(\vartheta)\right)$	$\lambda > 0$, $c > 0$
Jensen-Shannon	$f_{\vartheta=\frac{1}{2}}(t)$	$D_{\max} + c - \lambda\pi_{\mathsf{p}}\left[\frac{1}{2}\log\left(1 - \exp\left(2\lambda^{-1}(D - D_{\max} - c)\right)\right)\right]$ $+\lambda \mathsf{JS}(\pi, \pi_{\mathsf{p}})$	$\lambda > 0$, $c > 0$
Vincze-Le Cam	$\frac{2-2t}{t+1}$	$2\left(D_{\max} + c\right) + \pi_{\mathsf{p}}[-D] - \frac{4\pi_{\mathsf{p}}\left[\sqrt{c + D_{\max} - D}\right]^2}{2 + \mathsf{VC}(\pi, \pi_{\mathsf{p}})}$	$c > 0$
	$e^{t-1} - 1$	$\pi_{\mathsf{p}}\left[(D-c)_+ + \lambda\right)\log(((D-c)_+ + \lambda))]$ $-(1 + \log(\lambda))\pi_{\mathsf{p}}\left[(D-c)_+ + \lambda\right)]$ $+ c + (e-1)\lambda + e\lambda\mathcal{D}_f(\pi, \pi_{\mathsf{p}})$	$c \in \mathbb{R}$ $\lambda > 0$

For the power divergences with $0 < p < 1$, let us remark that when $\mathcal{D}_{f_p}(\pi, \pi_{\mathsf{p}}) \to 0$, the bound is optimised for $c \to \infty$, while when $\mathcal{D}_{f_p}(\pi, \pi_{\mathsf{p}}) \to 1$, the bound is optimised for $c \to 0$. A similar behaviour is observed for power divergences with $p < 0$. It seems important to pick adequately $c\left(\mathcal{D}_{f_p}(\pi, \pi_{\mathsf{p}})\right)$ if one wishes to obtain tight bounds for all π.

For total variation, let us first remark that since the generator $f(x) = \frac{|x-1|}{2}$ is not differentiable at $x = 1$, it can not be approximated by a sequence of convex functions such that $1/f_n''$ is concave[4]. It is possible to minimise the bound on c, but we could not compute the optimal scale parameter.

Vincze-Le Cam's bound somewhat stands out as it involves $2D_{\max}$ rather than $D_{\max}$. This is explained by the fact that the bound is not derived directly from Theorem 2.3, but results from Remark 2.11, extending f to $t \in (-1, 0)$ by $f(t) = \frac{2-2t}{t+1}$.

Change of measures with strong penalisation

The strength of the penalisation considered in Equation (2.19) depends on the strength of the moment assumption considered. Stronger penalisation will result in looser moment assumption, while on the other hand, strong moment assumption leads to weaker penalisation. The usual KL divergence being obtained for $f(x) = x\log(x)$ which grows slowly to infinity, it involves strong exponential moments on the generalisation gap. The trade-off between moment assumption and penalisation is also apparent for the power f-divergence, where p power f-divergence results in the conjugate q moment. Choices of f such that $f(\infty) < +\infty$ (which implies that no choice of f is super linear) leads to strict upper bounded generalisation gap requirement. On the other hand, choosing the fast growing $f(x) = e^{x-1} - 1$ leads to the mild requirement of $x\log(x)$ bounds. Note that all moment requirements must be stronger than the first order moment, since $f^*(x) \geq x$ whenever $f(1) = 0$.

Bégin et al. [2016], Alquier and Guedj [2018] broke from the traditional bounded or bounded exponential moment requirement by obtaining bounds involving finite p-moments for all $p > 1$. We go a step further by introducing two bounds involving strong penalisation and resulting in a $x\log(x)$-moment requirement or a first order moment requirement.

Our first bound considers an exponential f-divergence applied to positive generalisation gaps. For $D \geq 0$, $\forall \pi \ll \pi_{\mathsf{p}}$,

$$\pi\left[D\right] \leq \pi_{\mathsf{p}}\left[D\log(D)\right] - \pi_{\mathsf{p}}\left[D\right]\log\left(\pi_{\mathsf{p}}\left[D\right]\right) + \log \pi_{\mathsf{p}}\left[\exp\left(\frac{\mathrm{d}\pi}{\mathrm{d}\pi_{\mathsf{p}}}\right)\right]\pi_{\mathsf{p}}\left[D\right]. \tag{2.21}$$

Proof. Consider $f(x) = \exp(x-1) - 1$ on $\mathbb{R}_+$, and $f(x) = \infty$ on $\mathbb{R}_-^*$. This results in $f^*(t) = t\log(t) + 1$ for all $t > 1/e$, and $1 - \frac{1}{e}$ for $t < \frac{1}{e}$. This function can be upper bounded by $t\log(t) + 1$ for all $t > 0$. Considering $D \geq 0$, $c = 0$ and $\lambda > 0$, and using Equation (2.10) in

[4]Whenever f' is not continuous at $x_0 > 0$, then f'' is a Dirac mass at x_0 and therefore $1/f''(x_0) = 0$. It follows that $1/f''$ has a local minima at x_0 since $f'' \geq 0$, and therefore it can not be concave for any reasonable approximation.

conjugation with this upper bound, this results in

$$\begin{aligned}\pi[D] &\leq \lambda\pi_{\mathsf{p}}\left[\lambda^{-1}D\log(\lambda^{-1}D)+1\right]+\lambda\pi_{\mathsf{p}}\left[\exp\left(\frac{\mathrm{d}\pi}{\mathrm{d}\pi_{\mathsf{p}}}-1\right)-1\right]\\ &\leq \pi_{\mathsf{p}}[D\log(D)]-\log(\lambda)\pi_{\mathsf{p}}[D]+\lambda\pi_{\mathsf{p}}\left[\exp\left(\frac{\mathrm{d}\pi}{\mathrm{d}\pi_{\mathsf{p}}}-1\right)\right].\end{aligned}$$

This bound holds for all $\lambda > 0$ and is minimised for $\lambda^* = \frac{\pi_{\mathsf{p}}[D]}{\pi_{\mathsf{p}}\left[\exp\left(\frac{\mathrm{d}\pi}{\mathrm{d}\pi_{\mathsf{p}}}-1\right)\right]}$, yielding

$$\pi[D] \leq \pi_{\mathsf{p}}[D\log(D)]-\pi_{\mathsf{p}}[D]\log(\pi_{\mathsf{p}}[D])\pi_{\mathsf{p}}[D]\log\left(\pi_{\mathsf{p}}\left[\exp\left(\frac{\mathrm{d}\pi}{\mathrm{d}\pi_{\mathsf{p}}}\right)\right]\right).$$

□

Our second bound involves a custom made penalisation which forces the ratio of density $\frac{\mathrm{d}\pi}{\mathrm{d}\pi_{\mathsf{p}}}$ to be upper bounded; that is to say, $\exists r_{\max}$, $\forall r > r_{\max}$, $f(r) = \infty$. To obtain tractable expressions, we construct a convex f such that $f'(0) = -\infty$ and $f'(r_{\max}) = \infty$. An instance of such U shaped convex function is the lower half circle, resulting in

$$f_{\mathsf{U}}(x) = \begin{cases} 1-\sqrt{1-(1-x)^2} & x \in [0,2] \\ \infty & \text{else.}\end{cases}$$

By rescaling this function, we obtain for $r_{\max} > 1$ the convex function $f_{r_{\max}}(x) = f_{\mathsf{U}}\left(\frac{2x}{r_{\max}}\right) - f_{\mathsf{U}}\left(\frac{2}{r_{\max}}\right)$. The Legendre transform of f_{U} has closed formed expression

$$f^*_{\mathsf{U}}(t) = t + |t|\sqrt{\frac{t^2}{1+t^2}} - 1 + \sqrt{\frac{1}{1+t^2}},$$

which results in

$$\begin{aligned}f^*_{r_{\max}}(t) =& \frac{r_{\max}}{2}\left(t+|t|\sqrt{\frac{r_{\max}^2t^2}{4+r_{\max}^2t^2}}-1-\sqrt{\frac{r_{\max}^2}{4+r_{\max}^2}}\right)\\ &+2\sqrt{\frac{1}{4+r_{\max}^2t^2}}-2\sqrt{\frac{1}{4+r_{\max}^2}}.\end{aligned}$$

While somewhat involved, $f^*_{r_{\max}}(t)$ behaves as $r_{\max}t - C_{r_{\max}}$ for $t \to \infty$, and as $-C_{r_{\max}}$ for $t \to -\infty$. The asymptotic for large values recaptures the non-penalized change of measure for $\frac{\mathrm{d}\pi}{\mathrm{d}\pi_{\mathsf{p}}} \leq r_{\max}$ and $D \geq 0$, $\pi[D] \leq r_{\max}\pi_{\mathsf{p}}[D]$. The penalized bound improves on this behaviour by adding some flexibility. Moreover, since ${f^*_{r_{\max}}}'(\infty) = r_{\max}$, and $0 \geq f^{*\prime}$, it follows that $\forall c, D$, $D + c$ is $f^{*\prime}$ integrable, and moreover, for all D such that $\pi_{\mathsf{p}}[D > \infty] > 0$, $\exists c, \pi_{\mathsf{p}}\left[{f^*_{r_{\max}}}'(D+c)\right] = 1$. Hence we can apply the second statement of Theorem 2.1, and guarantee that minimising our bound in c recovers the true f divergence.

2.1.8 Perspectives

Change of measures with very weak penalisation

Csizár-Donsker-Varadhan's change of measure (2.1) implies that the generalisation gap must have exponential moments to provide non trivial bounds. This condition is looser than the classic PAC-Bayes assumption that the risk is bounded. This raises the question of whether more efficient PAC-Bayes bounds could be built for looser penalisation than KL, leading to a strict bounded risk requirement. Obtaining such competitive bound would necessitate carefully designing "slow" convex function with tractable Legendre transform.

Legendre transform of the entropy

For some choices of f, it might be convenient to trade-off some tightness on the bound for more tractable expressions. A possible way to gain tractability could be to study the Legendre transform of the f-entropy, which is defined as

$$\mathcal{E}_{f,\pi_\mathsf{p}} : P \mapsto \pi_\mathsf{p}\left[f \circ P\right] - f\left(\pi_\mathsf{p}\left[D\right]\right).$$

The f-entropy collapses to the f-divergence between π and π_p when evaluated for $P = \frac{\mathrm{d}\pi}{\mathrm{d}\pi_\mathsf{p}}$, since $f(1) = 0$. While the f-entropy might not be convex, an upper bound of $\mathcal{E}^*_{f,\pi_\mathsf{p}}$ still results in an upper bound of $\mathcal{D}^*_{f,\pi_\mathsf{p}}$. More generally, any extension of the f-divergence to a larger space can be used to upper bound $\mathcal{D}^*_{f,\pi}$.

Change of measure inequalities for Variational PAC-Bayes

Variational PAC-Bayes strategies (see Section 1.2.4) construct posterior distribution belonging to a parametric family of probability measures. The definition of the Legendre's transform of the f-divergence, on the other hand, involves a worst case analysis performed on *all* probability measures. Modifying the f-divergence to return ∞ outside of the variational family results in a decrease of the Legendre transform of this operator, leading to tighter bounds. Whether this decrease is significant or not would conceivably depend on the form of the variational family considered. Whether tractable expressions of the Legendre transform can be obtained remains uncertain. We expect the analysis to be more involved, and to depend on the form of the variational family. Moreover, the modified f-divergence might no longer be a convex operator if the exponential family is not a convex set (*e.g.* exponential families are usually not convex sets).

In a similar spirit, tighter change of measure inequalities can be constructed by considering other forms of constraints. For instance, one may limit the search to the pruned posterior considered in McAllester [1999b] (see Section 1.2.3). Another option could be to only consider posterior distributions belonging to a f-divergence sphere (*i.e.* disregarding all posterior distribution at distance more than $\tilde{r}$). For instance, considering Catoni's PAC-Bayes bound (1.20) for a fixed temperature λ in a bounded risk, all posteriors such that $\mathrm{KL}(\pi, \pi_\mathsf{p}) \geq \lambda^{-1}$ result in a

vacuous bound, and can therefore be disregarded. The same questions and limitations on the potential improvement and tractability occur.

Finding approximately optimal λ and c

As discussed in Section 2.1.6, an appropriate choice of c and λ is necessary to obtain tight inequalities. In most cases, we could not explicitly compute which values are optimal, especially for c. Getting some theoretical or practical insight on how to pick these parameters in such a way as to obtain nearly optimal bounds is an exciting future avenue.

To approximate the optimal c, a strategy can consist in considering an idealized case where the generalisation gap takes a single, known value. For instance, for $D = \tilde{R} - R$, the generalisation gap at a given predictor is a random variable of mean 0, and if the number of observations n is high, should have variations of order $O(n^{-\frac{1}{2}})$. Replacing D by 0 transforms the intractable renormalisation equation $\pi_{\mathsf{p}}\left[f^{*\prime}(D-c)\right] = 1$ by $f^{*\prime}(-\tilde{c}) = 1$ which has solution $\tilde{c} = -f'(1)$ (or more generally $c = \overline{D} - f'(1)$ when D fluctuations close to $\overline{D}$).

2.1.9 Conclusion

PAC-Bayes generalisation relies on change of measure inequalities to transfer a concentration inequality on a fixed probability measure to all probability measures simultaneously. Additive change of measure inequalities can naturally be interpreted as Legendre transform of a penalisation term. In this section, we have studied how these Legendre transform can be upper bounded for a generic class of penalisation, f-divergence, which extends on the classic KL penalisation. The analysis shows a trade-off between the penalisation considered, and the assumptions which will be required on the risk. Weaker penalisation, which allows constructing posterior distribution further away from the prior, is paid for by stronger moment on the generalisation gap, and hence stronger assumptions on the risk.

Computing the exact Legendre transform of the penalisation involves optimisation on a single degree of freedom. This optimisation has no closed form expression in the general case, involving an intractable renormalisation condition. This makes the construction of tight PAC-Bayes bound with a generic f-divergence change of measure difficult. In this respect, the classic KL penalisation represents a sweet spot. First, the exact Legendre transform has a closed form expression as the renormalisation condition is tractable. Second, it involves exponential moment of the generalisation gap, matching the form used in Chernov's concentration inequalities. Finally, closed form expressions are available for the computation of the KL divergence for popular family of distributions such as Gaussian, facilitating the computation of the bound and its derivative for Variational PAC-Bayes settings. For these reasons, classic KL penalized PAC-Bayes bounds will be considered for AD modelling applications.

2.2 Impact of the PAC-Bayes prior on generalisation bounds

Our second contribution to PAC-Bayes theory focus on the evaluation of generic PAC-Bayes bounds. We study conditions for a PAC-Bayes bound to provide an acceptable generalisation guarantee. Our analysis shows that the optimal generalisation guarantee only depends on distribution of the risk from the prior distribution. As such, a target generalisation level is only achievable if the prior puts sufficient weight on high performing predictors. We connect such requirements to the prevalent technique of using data-dependent priors in deep-learning PAC-Bayes applications, and discuss the implications for the claim that PAC-Bayes explains generalisation. Finally, we investigate the relationship between the achievable generalisation and the quantiles of the prior risk for Catoni's PAC-Bayes bound (1.20).

2.2.1 Generalisation guarantees

A celebrated strategy to evaluate the performance of a trained predictor on future, unknown data consists in:

- splitting the available data in a train set and a test set,
- constructing a predictor using only data in the train set,
- evaluating the performance of the predictor on the test set (empirical test performance),
- using concentration inequalities to infer, from this finite sample performance on the test set, an upper bound on the average performance on the data generation mechanism (test performance).

In the case of i.i.d. data, this approach yields perfectly valid high probability bounds on the performance of the learner on future data. This comes at the cost of sacrificing a fraction of the data when training. To obtain tight bounds on the test performance, this fraction can not be too small - for n_{test} samples reserved for the test sets, tight concentration inequalities will have deviations of order n_{test}^{-1} when the test error is almost 0 and $n_{\text{test}}^{-1/2}$ in the general case. In practice, 10 to 35 percent of the data is reserved for testing.

Strategies have been devised in order to circumvent the need to set aside data at training time. These concentrate on properties of the potential predictors. Intuitively, the higher the dimension of these potential predictors, the higher is the risk of overfitting, and hence the difficulty of obtaining bounds on the test performance without training data. However, one can easily construct low dimensional models with high overfitting potentials. For instance, $h_{\omega,a,\phi}(x) = a\cos(wx+\phi)$ is a simple three dimensional space of predictors which can be used to perfectly fit any observations $(x_i, y_i)_{i\leq N}$ for an arbitrary number of measurements; while the space of predictors $h_{(\alpha_i)_{i\in[0,d]}}(x) = \sum_{i=0}^{d}\alpha_i x^i$ can only fit all observations if the number of sample N is less than $d+1$. The dimension of the predictor space is hence a misleading proxy of the flexibility of the predictor space. Vapnik and Chervonenkis [1971] introduced the notion of **V**apnik–**C**hervonenkis (VC) dimension, which captures this level of flexibility: the VC dimension

is the maximum number of points which can be exactly fitted by a predictor. This notion proved in an important breakthrough in the analysis of the generalisation ability based, not on the nature of the predictor selection, nor on specifications of the data generation mechanism, but solely on the space of predictors alone. For a class of predictors of VC dimension d_{VC}, a standard result for losses bounded by 1 is that $\tilde{R}(\hat{h}) \leq R(\hat{h}) + \epsilon((d_{\text{VC}} - \log(\delta))/n, R(\hat{h}))$ holds for all estimators $\hat{h}$ with probability $1 - \delta$. The generalisation gap ϵ moreover recaptures the usual rates of $\epsilon(\xi_n, 0) = O(\xi_n)$, $\epsilon(\xi_n, r) = O(\sqrt{\xi_n})$. A proof of this result can be found in Vapnik [2000], chapter 3.7. As a consequence, any algorithm selecting a predictor in a space of finite VC dimension for classification (more generally with bounded loss) can be trained using *all* the data and still obtain an upper bound on the test performance with tight rate. As these rate involve the ratio between the VC dimension and the number of samples, the classic curse of dimensionality is transformed into a curse of VC-dimensionality: the upper bounds will be vacuous if the number of observations n is lower than the VC dimension. Unfortunately, the VC dimension of popular predictor spaces such as neural network, while finite, is typically prohibitive (*e.g.* of order $N \log(N) L + NL^2$ for a neural network with N parameters and L layers involving polynomial activation, see Bartlett et al. [1998]). On the other hand, overwhelming evidence has accumulated, showing that carefully training a neural network can result in a small test error, even in the deep learning setting where the number of parameters (and VC dimension) far exceeds the number of data [Zhang et al., 2017]. Indeed, rather than considering the training of smaller network as an imperative for adequate test guarantees, practitioners now perceive the ability to fit any signal (*i.e.* VC dimension higher than the number of data) as an appealing feature of a network architecture [Bubeck and Sellke, 2021]. It appears there is a steep gap between the theoretical and observed generalisation gaps.

PAC-Bayes theory has generated much excitement due to its ability to obtain guarantees on test performance while training on all available data, with no restriction on the space of predictors considered. While VC dimension considers the overall overfitting ability of a class of predictor and derives from it a PAC bound holding simultaneously for all predictors (and hence for any algorithm), the essence of the PAC-Bayes strategy is to share the generalisation abilities of a fixed predictor to all other predictors through penalisation. An analogue of this strategy can be found in the case where the loss is Lipschitz with respect to the predictor parameter; in this setting, the generalisation ability of data independent parameter γ_0 can be communicated to any parameter γ by remarking that $\tilde{R}(\gamma) \leq R(\gamma_0) - \tilde{R}(\gamma_0) + R(\gamma) + 2L \|\gamma - \gamma_0\|$. From the generalisation gap on a single fixed parameter γ_0, one can bound the generalisation gap on all parameters simultaneously, by paying a penality for parameters far away from γ_0. If the risk decreases faster than the increase of this penality, one can construct an estimator with improved test performance guarantees. PAC-Bayes removes the requirement for Lipschitz loss by considering an initial guess spread out on all predictors, *i.e.* a randomised predictor.

In the Lipschitz analogy considered, small bounds on the test performance can only be obtained if the initial predictor guess γ_0 is not too far away from a predictor guess with small empirical risk. To obtain a bound on the test performance lower than ε, there must exists a predictor whose empirical risk is lower than ε in the ball centred on γ_0 of radius $\varepsilon/(2L)$. This

implies that the empirical risk of the initial guess must be small enough, that is to say less than $\frac{3}{2}\epsilon$[5].

2.2.2 PAC-Bayes minima as a prior risk pushforward functional

Does similar restrictions apply in the PAC-Bayes setting? The surge of interest of PAC-Bayes was motivated in its ability to obtain non vacuous generalisation guarantees for deep neural networks as first described in Dziugaite and Roy [2017]. To obtain non vacuous bounds on the test performance of their trained network, Dziugaite and Roy [2017] relied on a data-dependent prior selected using a union bound argument. This rekindled interest in the use of data dependent priors to improve PAC-Bayes bounds, which had been previously studied in *e.g.* Ambroladze et al. [2006], Parrado-Hernández et al. [2012]. Dziugaite et al. [2021] considered splitting the data into a *prior training* set and a *posterior training* set (an idea already present in Parrado-Hernández et al. [2012]), and reported improved PAC-Bayes bound. Pérez-Ortiz et al. [2021b] followed a similar strategy and reported tight generalisation guarantees on MNIST dataset. Such a strategy strongly hints that a "good" prior is prerequisite if a "good" posterior is to be obtained.

Let us first remark that contrary to the Lipschitz case, the guarantees on the posterior test performance can massively improve on the test performance guarantees of the prior. Indeed, let us consider a setting with only two classifiers γ_0, γ_1. Consider $N = 500$ data, and assume that one classifier is perfect, and the other always wrong half the time ($\ell(\gamma_0) = 0$, $\ell(\gamma_1) = .5$). Then for a prior giving equal weight to both classifier, a posterior trained through Catoni's bound with temperature 0.01 obtains a generalisation guarantee of 0.062 (for a confidence level of 0.95). This improves on the average performance of the prior (0.25), and is thus always better than *any* test bound on the prior. Moreover, a prior considering a single classifier achieving a risk of 0.1 will result in generalisation guarantees of 0.155; while the prior risk average is lower in this second case, it results in a higher test bound. The information brought by the prior risk average proves too crude to be informative on the generalisation ability. A more fine grained description of the prior must be introduced to study its impact on the posterior test guarantees.

For a wide class of PAC-Bayes bound, the minimisation of the bound depends only on the *pushforward* measure of the prior on the empirical risk. For PAC-Bayes bounds of form Equation (1.12) satisfying the data processing inequality (1.13), we have shown that any optimal posterior distribution must have density with respect to the prior which is R measurable. This implies that the minimisation of the bound can be restricted to such posteriors. Let us moreover assume that the impact of the prior can be limited to a divergence term, *i.e.* there exists a

[5] In fact, this strategy can *never* yield a tighter bound than the one that could be obtained for the prior, since $R(\gamma_0) \leq R(\gamma) + L\,\|\gamma - \gamma_0\|$. The same negative result also holds for the Wasserstein PAC-Bayes bounds of Viallard et al. [2023] (Theorem 1) - training the bound on the posterior π will result in a bound that is provably worse than the one obtained for the mixture of partially data dependent prior. To train the posterior in practice, the authors lessen the impact of the Wasserstein distance by adding a multiplicative factor (which is tantamount to penalized regression in the Lipschitz case). Although the bounds are of limited practical interest, the analysis can therefore still serve as a way to motivate useful objectives.

function f (not necessarily convex) such that

$$\mathsf{PB}(\pi, R, \pi_{\mathsf{p}}, \delta) = \widetilde{\mathsf{PB}}\left(\pi\left[R\right], \pi_{\mathsf{p}}\left[f\left(\frac{\mathrm{d}\pi}{\mathrm{d}\pi_{\mathsf{p}}}\right)\right], \delta\right) \tag{2.22}$$

with convention that the bound values ∞ whenever $\pi \not\ll \pi_{\mathsf{p}}$. Then it follows (see proof of Theorem 2.4) that the minimisation of this bound amounts to the minimisation on the density function $g : \mathbb{R} \to \mathbb{R}^+$ defined on the risk, of the criteria

$$\mathsf{PB}_{\mathbb{R}}(g, \pi_{\mathsf{p}}^{\#R}, \delta) := \widetilde{\mathsf{PB}}\left(\pi_{\mathsf{p}}^{\#R}\left[\mathsf{Id} \times g\right], \pi_{\mathsf{p}}^{\#R}\left[f \circ g\right], \delta\right) \tag{2.23}$$

under the constraint that $\pi_{\mathsf{p}}^{\#R}\left[g\right] = 1$. Here, $\pi^{\#R}$ denotes the pushforward of the prior by the risk, *i.e.* the probability measure of $R(\gamma)$ for $\gamma \sim \pi$, and Id denotes the identity function $\mathsf{Id}(x) = x$ on $\mathbb{R}$. Hence, rather than constructing a posterior distribution on the *predictors*, the minimisation of such PAC-Bayes bounds amounts to searching for a posterior distribution on *risk values*. We summarize this in the following theorem:

Theorem 2.4

Any PAC-Bayes bound following the form of Equation (2.22), satisfying the data processing inequality of Equation (1.13) attains a minimal value fully determined by the pushforward measure $\pi^{\#R}$, *i.e.* there exists a function $\mathsf{PB}^{\mathsf{min}}$ such that

$$\inf_{\pi \ll \pi_{\mathsf{p}}} \mathsf{PB}(\pi, R, \pi_{\mathsf{p}}, \delta) = \mathsf{PB}^{\mathsf{min}}(\pi_{\mathsf{p}}^{\#R}, \delta). \tag{2.24}$$

Proof. Reusing notations and the proof of Remark 1.1, we remark that for any posterior distribution $\pi \ll \pi_{\mathsf{p}}$, the probability measure π_R^* achieves a lower PAC-Bayes bound. Moreover, the random variable $R(\gamma)$ where $\gamma \sim \pi_R^*$ is absolutely continuous with respect to the random variable $R(\gamma)$, $\gamma \sim \pi_{\mathsf{p}}$. Noting $\pi^{*,\#R}$ and $\pi_{\mathsf{p}}^{\#R}$ their respective measure, one can note $g_\pi = \frac{\mathrm{d}\pi^{*,\#R}}{\mathrm{d}\pi_{\mathsf{p}}^{\#R}}$ the Radon-Nikodym derivative. Since the Radon-Nikodym derivative of the conditional posterior and the prior $\frac{\mathrm{d}\pi_R^*}{\mathrm{d}\pi_{\mathsf{p}}}$ is moreover R measurable by definition of π_R^*, we can identify $\frac{\mathrm{d}\pi_R^*}{\mathrm{d}\pi_{\mathsf{p}}} = g_\pi \circ R$. Hence

$$\begin{aligned}
\mathsf{PB}(\pi, R, \pi_{\mathsf{p}}, \delta) &\geq \mathsf{PB}(\pi_R^*, R, \pi_{\mathsf{p}}, \delta) \\
&= \widetilde{\mathsf{PB}}\left(\pi_R^*\left[R\right], \pi_{\mathsf{p}}\left[f\left(\frac{\mathrm{d}\pi_R^*}{\mathrm{d}\pi_{\mathsf{p}}}\right)\right], \delta\right) \\
&= \widetilde{\mathsf{PB}}\left(\pi_{\mathsf{p}}\left[R\frac{\mathrm{d}\pi_R^*}{\mathrm{d}\pi_{\mathsf{p}}}\right], \pi_{\mathsf{p}}\left[f\left(\frac{\mathrm{d}\pi_R^*}{\mathrm{d}\pi_{\mathsf{p}}}\right)\right], \delta)\right) \\
&= \widetilde{\mathsf{PB}}\left(\pi_{\mathsf{p}}\left[R \times g_\pi \circ R\right], \pi_{\mathsf{p}}\left[f \circ g_\pi \circ R\right], \delta)\right) \\
&= \widetilde{\mathsf{PB}}\left(\pi_{\mathsf{p}}^{\#R}\left[\mathsf{Id} \times g_\pi\right], \pi_{\mathsf{p}}^{\#R}\left[f \circ g_\pi\right], \delta\right) \\
&=: \mathsf{PB}_{\mathbb{R}}(g_\pi, \pi_{\mathsf{p}}^{\#R}, \delta).
\end{aligned}$$

Thus the minima of the left hand side over probability measures $\pi \ll \pi_{\mathsf{p}}$ is greater than the minima of the right hand side over all densities for $\pi_{\mathsf{p}}^{\#R}$, *i.e.* real valued functions $g : \mathbb{R} \to \mathbb{R}_+$ such that $\pi_{\mathsf{p}}^{\#R}[g] = 1$. To prove the reverse inequality, one should only notice that the series of inequality above is an equality whenever $\pi_R^* = \pi$, and that for any density function g, the probability measure π_g such that $\frac{\mathrm{d}\pi_g}{\mathrm{d}\pi_{\mathsf{p}}} = g \circ R$ satisfies that condition.

Noting $\mathrm{PB}^{\min} := \arg\inf_{g \text{ dens. } \pi_{\mathsf{p}}^{\#R}} \mathrm{PB}_{\mathbb{R}}(g, \pi_{\mathsf{p}}^{\#R}, \delta)$ thus concludes the proof □

For clarity's sake, we now focus on the case where the risk takes value in $[0,1]$. In this case, the density g becomes a function of $[0,1]$ to $\mathbb{R}^+$. To stress this assumption, we use the notation $\mathrm{PB}_{[0,1]}$ rather than $\mathrm{PB}_{\mathbb{R}}$ from now on. We expect that most of the findings could be extended to the case where the risk takes positive values.

Remark 2.13

Equation (2.23) states that the minimisation of a PAC-Bayes bound involves searching not for a probability measure on the potentially high dimensional $\mathcal{H}$, but on a probability measure on real values. For KL penalized PAC-Bayes bound such as Catoni and MLS, this implies that the divergence term can be understood as a divergence between two probability measures on the same space of dimension 1, irrespectively of the dimension of the predictor space.

This partially explains why the Kullback–Leibler term does not empirically increase when the number of parameter increases, a non intuitive property which startled experimenters (see section 5.5 in Pérez-Ortiz et al. [2021a] and section 7.7 in Pérez-Ortiz et al. [2021b])[a]. The resulting Kullback–Leibler term depends on the amount of mass put on high risk predictors which need to be moved to low risk predictors before the average risk becomes small enough (*i.e.* before the cost of moving more mass exceeds the average risk improvement). Our interpretation of the smaller KL for larger networks paradox is the following. The priors in Pérez-Ortiz et al. [2021b,a] are learnt on a prior learning set. Since the training algorithm for learning the prior has good generalisation ability for all architectures and larger networks result in lower test risk (this is empirically observed and *not explained*), the prior pushforward of the risk on the PAC-Bayes learning data puts more weight on smaller risks for larger networks. Since the risk is on the whole lower, shifting the same amount of weight result in a smaller decrease of the average risk. Hence the trade-off between shifting weight and diminishing the average risk is met earlier, that is to say for a smaller KL value.

For priors which are not data dependent, the same analysis implies that an increase in the dimension of the parameter space should result in a larger KL value *if* the increase of dimensionality can be understood as an increase of confounding factors, *e.g.* if the typical prior risk is increased, though the smallest prior risk might decrease (more flexibility).

[a] Note that as the two articles cited below consider variational approximations of the posterior, the performance of the optimal posterior is no longer a functional of the risk pushforward. Still, if the optimal variational

posterior obtains nearly optimal result, our analysis should hold.

The form (2.22) and assumption (1.13) are valid for PAC-Bayes bounds such as Catoni (eq. (1.20)), Maurer–Langford–Seeger (eq. (1.22), the Rényi divergence penalised bounds of Bégin et al. [2016] (eq. (1.26)) and all the f-divergence penalised bounds described in the previous section. It is however not valid for bounds involving the variance landscape of empirical risk such as Tolstikhin and Seldin [2013], although the above reasoning can be adjusted (the density g must then be defined on both the risk and variance).

PAC-Bayes minima as non decreasing functionals

What consequences can one draw from this? Theorem 2.4 states that the generalisation guarantee is wholly determined by the performance of a random predictor drawn from the prior, that is to say by the prior empirical risk, and *not* any other characteristics of the predictor space *or* the data generating mechanism. Let us consider two spaces of predictors $\mathcal{H}_1$, $\mathcal{H}_2$, each equipped with a prior π_1, π_2. If, at a fixed sample S, the empirical risks R_1 (defined on $\mathcal{H}_1$) and R_2 (defined on $\mathcal{H}_2$) are such that ${\pi_1}^{\#R_1} = {\pi_2}^{\#R_2}$, then the test guarantees will be identical. Similarly, consider a data generating mechanism outputting two datasets z_1, z_2. Denote $R_1 = \ell(\cdot, z_1)$ and $R_2 = \ell(\cdot, z_2)$. If the measures $\pi^{\#R_1}$ and $\pi^{\#R_2}$ are equal, then the test guarantees are equal. This notably holds in the case where z_2 does not contain any signal (*e.g.* random label case): for the test guarantees of the posterior to be non vacuous, the prior empirical risk must be better behaved (with high probability) in the case where there is some signal.

What minimal properties are required on the prior empirical risk in order to obtain a "good" test guarantees on the posterior? Is there a way to measure and compare prior empirical risk in terms of known quantities? A first remark consists in noting that the stochastic order on measures on $[0,1]$ should be preserved. That is to say, if $\forall x \in [0,1]$, $\pi_1^{\#R}([0,x]) \geq \pi_2^{\#R}([0,x])$, we expect the bound on $\pi_1^{\#R}$ to improve on the bound on $\pi_2^{\#R}$. This is indeed the case for PAC-Bayes bound of form (2.22), as long as the bound is a decreasing function of the posterior average.

Theorem 2.5

Consider any PAC-Bayes satisfying the hypotheses of Theorem 2.4, and risk functions taking values in $[0,1]$.
If the PAC-Bayes bound is a decreasing function of the posterior average of the empirical risk, the function $\mathrm{PB}^{\min}$ is an increasing function of its first argument for the stochastic order, *i.e*

$$\pi_1^{\#R} \leq_{\text{stoch}} \pi_2^{\#R} \implies \mathrm{PB}^{\min}(\pi_1^{\#R}, \delta) \leq \mathrm{PB}^{\min}(\pi_2^{\#R}, \delta). \tag{2.25}$$

Proof. Consider two probability measures $\pi_1^{\#R} \leq_{\text{stoch}} \pi_2^{\#R}$. Then there exists a (potentially randomised) function G such that $G(X_1) \sim X_2$ for $X_1 \sim \pi_1^{\#R}$ and $X_2 \sim \pi_2^{\#R}$ which satisfies $G(x) \geq x$ almost surely (see below). Consider any density g for $\pi_2^{\#R}$. Then $\pi_1^{\#R}[f \circ g \circ G] =$

$\pi_2^{\#R}[f \circ g]$ while $\pi_1^{\#R}[\mathsf{Id} \times g \circ G] \leq \pi_1^{\#R}[G \times g \circ G] = \pi_2^{\#R}[\mathsf{Id} \times g]$. Hence for all density g,

$$\widetilde{\mathsf{PB}}\left(\pi_1^{\#R}[\mathsf{Id} \times g \circ G], \pi_1^{\#R}[f \circ g \circ G], \delta\right) \leq \widetilde{\mathsf{PB}}\left(\pi_2^{\#R}[\mathsf{Id} \times g], \pi_2^{\#R}[f \circ g], \delta\right). \tag{2.26}$$

Moreover, $\pi_1^{\#R}[g \circ G] = \pi_2^{\#R}[g] = 1$, so that $g \circ G$ is a valid density for $\pi_2^{\#R}$. This implies for any density g for $\pi_2^{\#R}$, one can construct a randomised density $g \circ G$ for $\pi_1^{\#R}$ achieving a lower PAC-Bayes bound.

The possible randomness of G is a technicality designed to overcome cases where $\pi_1^{\#R}$ takes "fewer" values than $\pi_2^{\#R}$ (*e.g.* $\pi_1^{\#R}$ is a Dirac on 0 and $\pi_2^{\#R}$ follows a uniform measure between $[0,1]$). Using randomised density in Equation (2.23) is equivalent to allowing the ratio of density g to depend on an observation ω from a second probability space Ω. By extending the space of predictor from $\mathcal{H}$ to $\mathcal{H} \times \Omega$ and using the PAC-Bayes bound on the new space of predictor with prior π_1, one can extend the search of density in Equation (2.23) to randomised densities (these are just going to lead to looser bounds since optimal posterior are R measurable, while these are R, ω measurable). This implies that

$$\begin{aligned}
\inf_{\pi \ll \pi_1} \mathsf{PB}(\pi, R_1, \pi_1, \delta) &= \inf_{g \text{ rdm. dens } \pi_1^{\#R}} \widetilde{\mathsf{PB}}\left(\pi_1^{\#R}[\mathsf{Id} \times g], \pi_1^{\#R}[f \circ g], \delta\right) \\
&\leq \inf_{g. \text{ dens } \pi_2^{\#R}} \widetilde{\mathsf{PB}}\left(\pi_1^{\#R}[\mathsf{Id} \times g \circ G], \pi_1^{\#R}[f \circ g \circ G], \delta\right) \\
&\leq \inf_{g. \text{ dens } \pi_2^{\#R}} \widetilde{\mathsf{PB}}\left(\pi_1^{\#R}[\mathsf{Id} \times g], \pi_1^{\#R}[f \circ g], \delta\right) \\
&= \inf_{\pi \ll \pi_2} \mathsf{PB}(\pi, R_2, \pi_2, \delta).
\end{aligned}$$

We now show how G can be constructed. In the case where both measures $\pi_1^{\#R}$ and $\pi_2^{\#R}$ are dominated by Lebesgue measure, one can construct a deterministic G. Note cdf_i the cumulative distribution function of $\pi_i^{\#R}$ (*i.e.* $\mathrm{cdf}_1(x) = \pi_1^{\#R}([0,x])$). Then consider $G = \mathrm{cdf}_2^{-1} \circ \mathrm{cdf}_1$, where cdf_2^{-1} is defined through $\mathrm{cdf}_2^{-1}(q) = \inf(x, \mathrm{cdf}_2(x) \geq q)$ (the quantile function). If $\pi_i^{\#R}$ has density with respect to Lebesgue, it follows that $\mathrm{cdf}_1(X)$, $X \sim \pi_1^{\#R}$ has distribution $U[0,1]$, which implies that $G(X)$, $X \sim \pi_1^{\#R}$ has distribution $\pi_2^{\#R}$. Moreover, the stochastic ordering implies that $\mathrm{cdf}_1 \geq \mathrm{cdf}_2$, which in turn implies that $G(x) \geq x$ if the $\pi_2^{\#R}$ has density with respect to Lebesgue.

When Lebesgue is no longer a dominating measure, one can consider a randomised function G by considering δ the random variable defined as $\delta = \mathrm{cdf}_2^{-1}(U) - \mathrm{cdf}_1^{-1}(U)$. The stochastic order guarantees that $\delta \geq 0$ surely. Noting that $\mathrm{cdf}_1^{-1}(U)$ has distribution $\pi_1^{\#R}$, and $\mathrm{cdf}_2^{-1}(U)$ has distribution $\pi_2^{\#R}$, it follows that the function $G(x) = x + \delta$ is such that $G(X_1) = X_2$. □

The requirement that the PAC-Bayes bound decrease with the posterior average is met by all bounds in the literature, and is thus very mild. Stochastic dominance is a partial order (*i.e.* for two measures $\pi_1^{\#R}$ and $\pi_2^{\#R}$, we might have $\pi_1^{\#R} \not\preceq_{\text{stoch}} \pi_2^{\#R}$ and $\pi_1^{\#R} \not\succeq_{\text{stoch}} \pi_2^{\#R}$), and as such might not be usable to analyse the behaviour of the bound on two distinct settings. However, Theorem 2.5 still implies that, if one considers a subset of priors $\tilde{P}_{[0,1]}$, the smallest

bound achievable is attained (if it is attained) by a prior $\tilde{\pi_{\mathsf{p}}}^{\#R}$ which is a minima, *i.e.* there is no $\pi_{\mathsf{p}}^{\#R} \in \tilde{\mathcal{P}}_{[0,1]}$ such that $\tilde{\pi_{\mathsf{p}}}^{\#R} \leq_{\text{stoch}} \pi_{\mathsf{p}}^{\#R} \in \tilde{\mathcal{P}}_{[0,1]}$. Notably, if $\tilde{\mathcal{P}}_{[0,1]}$ has a minima which can be compared to every element in $\tilde{\mathcal{P}}_{[0,1]}$, then it automatically follows that it minimises the PAC-Bayes bound (*i.e.* it is the prior yielding the smallest PAC-Bayes bound).

Quantile requirements on the prior

What sort of conditions can one consider to build subset of priors? A naive approach would be to define $\tilde{\mathcal{P}}^{\mu}_{[0,1]}$ as the set of all priors with same average empirical risk $\mu \in [0,1]$. Unfortunately, one can easily verify that $\pi_1^{\#R} \leq_{\text{stoch}} \pi_2^{\#R}$ and $\pi_1^{\#R}[\mathsf{Id}] = \pi_2^{\#R}[\mathsf{Id}]$ implies that $\pi_1^{\#R}$ and $\pi_2^{\#R}$ have identical distributions, and hence every element of $\tilde{\mathcal{P}}^{\mu}_{[0,1]}$ is minimal - Theorem 2.5 does not help the search for the smallest value of a generic PAC-Bayes bound[6]. A more rewarding form of conditions are quantile requirement $\pi_{\mathsf{p}}^{\#R}([r,1]) \geq q$ for some $r \in [0,1]$, $q \in [0,1]$. We will denote $\tilde{\mathcal{P}}^{r,q}_{[0,1]}$ the set of all probability measures satisfying this condition. Then $\tilde{\mathcal{P}}_{[0,1]}$ has a single minima, which can be compared to all element of the set: this is the scaled Bernoulli distribution $r\mathcal{B}(q)$ (see Figure 2.2). We summarize the implications of Theorem 2.5 on such set of priors in the following corollary:

Corollary 2.1

For any PAC-Bayes bound PB satisfying the assumptions of Theorem 2.5, for any prior π_{p}, then

$$\mathrm{PB}^{\min}(\pi_{\mathsf{p}}^{\#R}, \delta) \geq \max_{r \in [0,1]} \mathrm{PB}^{\min}(r\mathcal{B}(\pi_{\mathsf{p}}^{\#R}([r,1])), \delta),$$

where $r\mathcal{B}(q)$ is the rescaled Bernoulli with success probability q.

Proof. For any $r \in [0,1]$, the distribution $\pi_{\mathsf{p}}^{\#R}$ stochastically dominates $r\mathcal{B}(\pi_{\mathsf{p}}^{\#R}([r,1]))$. Applying Theorem 2.5 concludes the proof. □

Let us denote $\mathrm{PB}^{\min}(r,q,\delta) := \mathrm{PB}^{\min}(r\mathcal{B}(q),\delta)$. It follows from Corollary 2.1 that to obtain a generalisation guarantee at most $\mathrm{PB}^{\min}(r,q,\delta)$ at level δ, the prior *must* put more than q weight on values smaller than r. To obtain that generalisation bound with probability at least $1-\delta$, this property must hold with probability at least $1-\delta$ on the data generating mechanism. This implies the following protocol to evaluate the conditions required to obtain a certain generalisation level:

[6] Further analysis on Catoni's bound shows that the minima is reached on Bernoulli distribution in that setting - this implies that this must also be the case for MLS's bound, since MLS's bound minimiser minimises Catoni's bound for some temperature. Such analysis is of slight interest: PAC-Bayes bound are sensitive to the amount of mass put on the lowest achievable risk, and not on the mass repartition on high risk values.

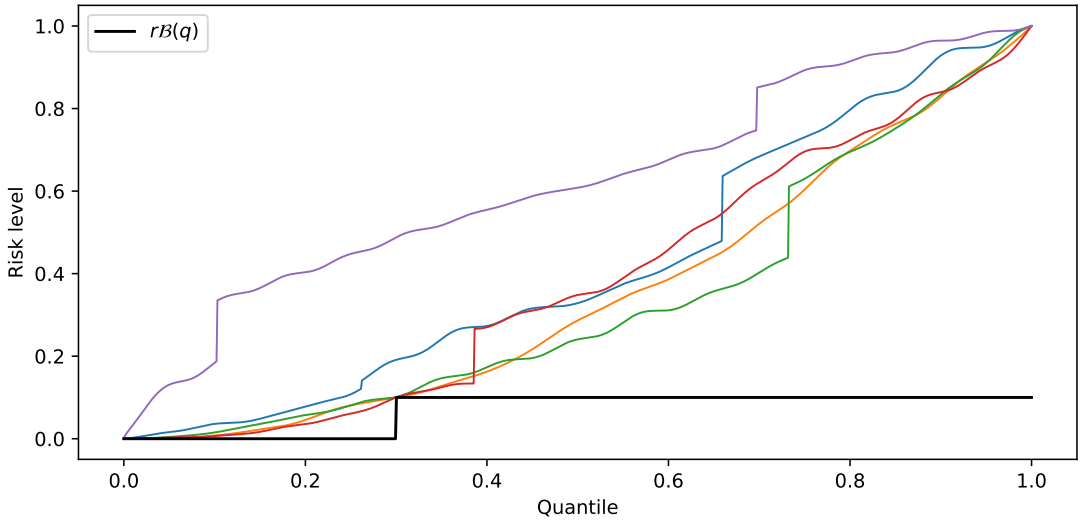


Figure 2.2: Quantile functions of various random variables on $[0,1]$ putting more than $q = 0.7$ mass on risk values higher than $r = 0.1$. This condition is equivalent with requiring that the quantile function remains above the quantile function of the scaled Bernoulli $0.1\mathcal{B}(0.7)$, which is equivalent with being stochastically smaller than this distribution.

Corollary 2.2

For a PAC-Bayes bound PB satisfying the assumptions of Theorem 2.5, a confidence level δ and a target generalisation level $G \in \mathbb{R}^+$, define

$$Q(r, G, \delta) = \inf \left\{ q, \mathrm{PB}^{\min}(r, q, \delta) > G \right\} \tag{2.27}$$

with convention that the infimum of an empty set is 1. Let $\overline{Q}(r, G, \delta)$ denote $1 - Q(r, G, \delta)$. Then any prior π_{p} satisfying $\mathrm{PB}^{\min}(\pi_{\mathrm{p}}^{\#R}, \delta) < G$ *must* satisfy

$$\pi_{\mathrm{p}}^{\#R}\left([0, r[\right) \geq \overline{Q}(r, G, \delta). \tag{2.28}$$

Proof. Suppose that $\pi_{\mathrm{p}}^{\#R}$ satisfies $\mathrm{PB}^{\min}(\pi_{\mathrm{p}}^{\#R}, \delta) < G$ and that there exists r such that the inequality does not hold. Since, if $Q(r, G, \delta) = 1$, the inequality automatically holds, this implies that $Q(r, G, \delta) < 1$, and hence that the set in the definition of Q is not empty. Then $\pi_{\mathrm{p}}^{\#R}\left([r, 1]\right) > Q(r, G, \delta)$ and thus $\pi_{\mathrm{p}}^{\#R} \geq_{\mathrm{stoch}} r\mathcal{B}(q)$ for some $1 > q > Q(r, G, \delta)$. This implies that $\mathrm{PB}^{\min}(\pi_{\mathrm{p}}^{\#R}, \delta) \geq \mathrm{PB}^{\min}(r, q, \delta)$. Since, for all $q_1 \geq q_2$, $r\mathcal{B}(q_1) \geq_{\mathrm{stoch}} r\mathcal{B}(q_2)$, for all $q > Q(r, G, \delta)$, $\mathrm{PB}^{\min}(r, q, \delta) \geq G$. Hence we have

$$\mathrm{PB}^{\min}(\pi_{\mathrm{p}}^{\#R}, \delta) \geq G,$$

which contradicts the assumption that $\mathrm{PB}^{\min}(\pi_{\mathrm{p}}^{\#R}, \delta) < G$. This concludes the proof. □

Corollary 2.2 states that conditions on the prior empirical risk quantiles necessary to achieve a target generalisation bound are summarized in function Q, fully determined by the function $P : r, q \mapsto \mathrm{PB}^{\min}(r, q, \delta)$. This implies our PAC-Bayes bound prior requirements investigation protocol: first evaluate P, then invert the formula to obtain the quantile requirements $\overline{Q}$.

We now apply this protocol to the tractable Catoni's bound.

2.2.3 Catoni's bound prior requirements

Let us consider Catoni's bound (1.20) at a given temperature λ. Using the fact that Gibbs posteriors are minimiser of Catoni's bound, one infers that the minimal value of the bound is

$$\mathrm{PB}^{\min}_{\mathrm{Cat}}(r, q, \delta) = -\lambda \log\left((1-q) + q\exp\left(-r\lambda^{-1}\right)\right) + \frac{1}{8\lambda n} - \lambda\log(\delta). \tag{2.29}$$

This implies that the quantile requirement function has closed form expression derived from

$$Q_{\mathrm{cat},\lambda}(r, G, \delta)) = \min\left(1, \max\left(0, \frac{1 - \exp\left(-\lambda^{-1}G + \frac{\lambda^{-2}}{8n} - \log(\delta)\right)}{1 - \exp\left(-\lambda^{-1}r\right)}\right)\right). \tag{2.30}$$

Notably, the quantile requirement $\overline{Q}_{\mathrm{cat},\lambda}$ is 1 whenever G is unreachable, *i.e.* whenever $G \leq \frac{1}{8\lambda n} - \lambda\log(\delta)$.

A few useful observations can be drawn from Equation (2.30). First of all, the quantile requirement function $\overline{Q}_{\mathrm{cat},\lambda}$ is very sensitive to the choice of λ. Moreover, $\overline{Q}_{\mathrm{cat},\lambda}$ quickly saturates when $\lambda \ll r$ (see Figure 2.3).

Finally, $\overline{Q}_{\mathrm{cat},\lambda}$ is increasing in r, and always smaller than the maximal quantile requirement at λ,

$$\begin{aligned}\overline{Q}_{\mathrm{max}}(\lambda, G, \delta, n) :&= \min\left(1, \exp\left(-\lambda^{-1}G + \frac{\lambda^{-2}}{8n} - \log(\delta)\right)\right)\\ &\geq \overline{Q}_{\mathrm{cat},\lambda}(r, G, \delta, n),\end{aligned}$$

for all $r \in [0, 1]$. This value is reached when $\lambda r \to \infty$. To factor out the temperature parameter, we minimise on it and define $\overline{Q}_{\mathrm{max}}(G, \delta, n) := \inf_{\lambda>0} Q_{\mathrm{max}}(\lambda, G, \delta, n)$ the maximal quantile requirement. This is reached for $\lambda_{\mathrm{opt}} = 1/(4G \times n)$, and equals

$$\overline{Q}_{\mathrm{max}}(G, \delta, n) = \min(1, \exp(-2G^2 n - \log(\delta))).$$

Essentially, if we consider small temperatures $\lambda \leq \lambda_{\mathrm{max}} \ll 1$, $\overline{Q}_{\mathrm{max}}(G, \delta, n)$ will be smallest quantile condition for all r such that $\lambda_{\mathrm{max}} \ll r$.

On the other hand, the quantile requirement $\overline{Q}_{\mathrm{cat},\lambda}$ saturates to 1 for all values of r if the

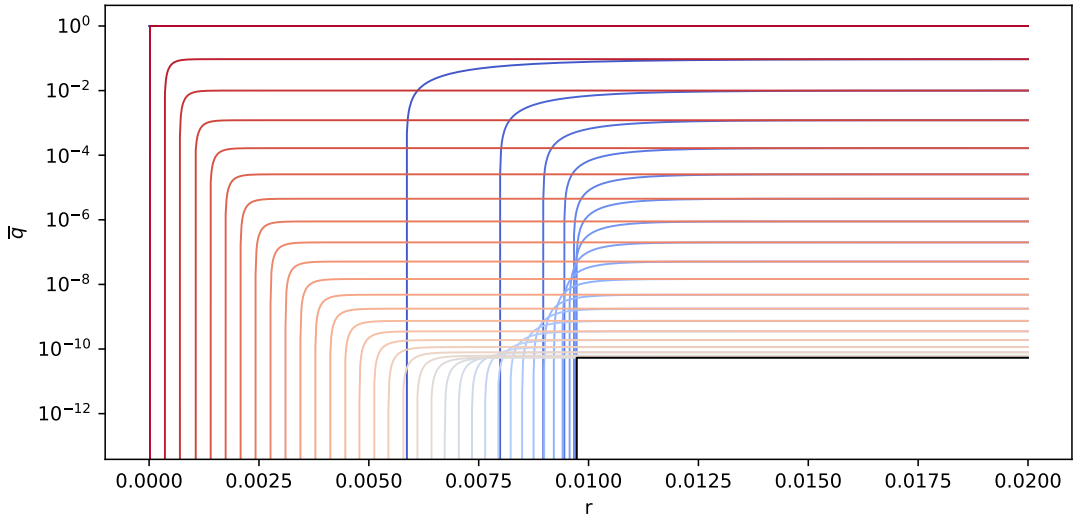


Figure 2.3: Evaluation of $\overline{Q}_{\text{cat},\lambda}$ as a function of r for different temperatures. The target generalisation gap is fixed to $G = 0.015$, the number of observations to 60 000 and the confidence level to $1 - \delta = 1 - 0.035$. 40 temperatures between λ_{min} and λ_{max} are assessed (blue denotes lower temperature, red larger temperature). In black, the minima of the risk requirement over all temperatures is plotted - this exhibits a phase transition at $r = G - 2\sqrt{\frac{-\log(\delta)}{8n}} \simeq 0.009714$. This graph implies that no prior putting less than 5e-11 mass on predictors with risk smaller than 0.01 can hope to obtain a generalisation guarantee higher than 0.015 valid with confidence level of 0.965 by training Catoni's bound on datasets with 60 000 samples (such as MNIST).

temperature λ is not in the range

$$[\lambda_{\text{min}}, \lambda_{\text{max}}] = \left[\frac{1}{4n\left(G + \sqrt{G^2 + \frac{\log(\delta)}{2n}}\right)}, \frac{1}{4n\left(G - \sqrt{G^2 + \frac{\log(\delta)}{2n}}\right)} \right].$$

If $G \leq \sqrt{\frac{-\log(\delta)}{2n}}$, this range is no longer defined. In this case, no temperature can yield a quantile requirement less than 1 for any r, that is to say, the generalisation guarantee G is unreachable using Catoni's PAC-Bayes bound. Note that this rate of generalisation guarantee is *not* optimal, since for perfect priors, a rate of $\log(\delta)/n$ is expected. Hence tighter PAC-Bayes bound should lead to smaller quantile requirement.

The fast saturation of quantile requirement at fixed λ suggests the analysis of the quantile requirement for the λ_{opt} minimising the asymptotic requirement as an approximation of the optimal quantile requirement for each r (*i.e.* where the optimal temperature is allowed to depend on r), which is less tractable. The comparison between the two approximations is given in Figure 2.4. For the example considered, the quantile requirement for λ_{opt} recaptures the minimal quantile requirement level for r large enough, but exhibits a smoother and more conservative

behaviour for small r values.

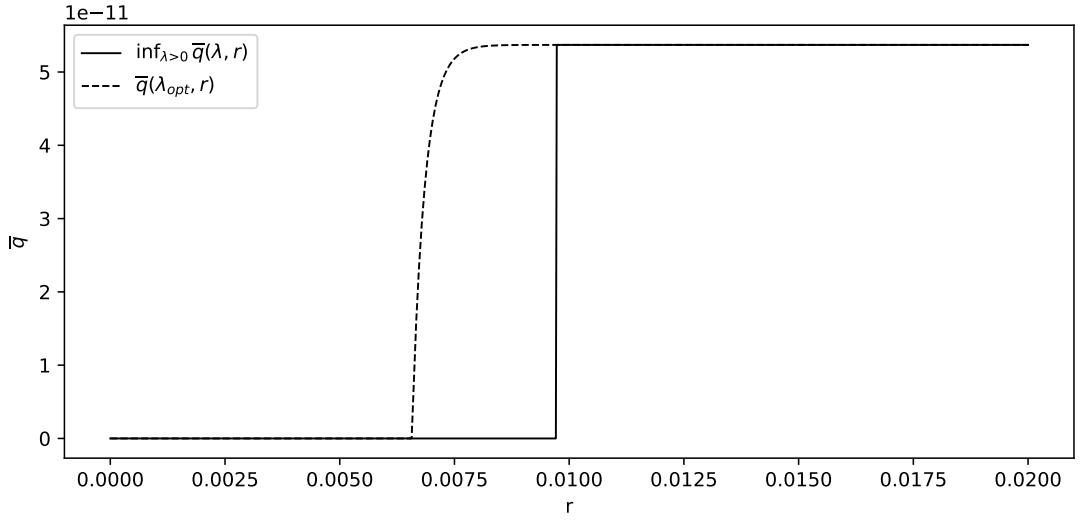


Figure 2.4: Comparison of $\overline{Q}_{\text{cat},\lambda_{\text{opt}}}$ and $\inf_\lambda \overline{Q}_{\text{cat},\lambda_{\text{opt}}}$. The target generalisation gap is fixed to $G = 0.015$, the number of observations to 60000 and the confidence level to $1-\delta = 1-0.035$. While $\overline{Q}_{\text{cat},\lambda_{\text{opt}}}$ provides a good approximation of the point wise minima for large value of r, it leads to more conservative quantile requirements for r close to r_{thresh}.

We can remark that for any r such that there exists λ in the range $[\lambda_{\text{min}}, \lambda_{\text{max}}]$ satisfying $r > G - \frac{\lambda^{-1}}{8n} + \lambda \log(\delta)$, the minima of the bound on λ is 0. The maxima of this lower bound on r is for $\lambda^{-1}_{\text{thresh}} = \sqrt{-8n\log(\delta)}$ (which always belongs to the range if G is reachable) and values $r_{\text{thresh}} = G - 2\sqrt{\frac{-\log(\delta)}{8n}}$. Hence, r values yielding non trivial quantile requirements for all temperatures should be higher than r_{thresh}.

Then, for $r > r_{\text{thresh}}$,

$$\begin{aligned}
\inf_{\lambda>0} \overline{Q}_{\text{cat},\lambda}(r, G, \delta, n) &= \inf_{\lambda_{\text{min}} \leq \lambda \leq \lambda_{\text{max}}} \max\left(0, \min\left(1, \frac{\overline{Q}_{\text{max}}(\lambda, G, \delta, n) - \exp(-\lambda^{-1} r)}{1 - \exp(-\lambda^{-1} r)}\right)\right) \\
&\geq \min\left(1, \inf_{\lambda_{\text{min}} \leq \lambda \leq \lambda_{\text{max}}} \frac{\overline{Q}_{\text{max}}(\lambda, G, \delta, n) - \exp(-\lambda^{-1} r)}{1 - \exp(-\lambda^{-1} r)}\right) \\
&\geq \min\left(1, \inf_{\lambda_{\text{min}} \leq \lambda \leq \lambda_{\text{max}}} \frac{\overline{Q}_{\text{max}}(G, \delta, n) - \exp(-\lambda^{-1} r)}{1 - \exp(-\lambda^{-1} r)}\right) \\
&\geq \min\left(1, \frac{\overline{Q}_{\text{max}}(G, \delta, n) - \exp(-\lambda^{-1}_{\text{max}} r)}{1 - \exp(-\lambda^{-1}_{\text{max}} r)}\right).
\end{aligned}$$

where we use the fact that the since $\overline{Q}_{\text{max}} \leq 1$, the function $\lambda \mapsto \frac{\overline{Q}_{\text{max}}(G,\delta,n) - \exp(-\lambda^{-1}_{\text{max}} r)}{1 - \exp(-\lambda^{-1}_{\text{max}} r)}$ is decreasing.

Let us consider $r_{\text{opt}}(\epsilon)$ such that $-\lambda^{-1}_{\text{max}} r_{\text{opt}}(\alpha) \leq \log(\overline{Q}_{\text{max}}(G, \delta, n)) + \log(\epsilon)$ and $r_{\text{opt}}(\epsilon) >$

r_{thresh}. Then it follows that

$$\begin{aligned}\inf_{\lambda>0} \overline{Q}_{\text{cat},\lambda}(r_{\text{opt}}(\alpha), G, \delta, n) &\geq \min\left(1, \frac{(1-\epsilon)\overline{Q}_{\text{max}}(G,\delta,n)}{1-\epsilon\overline{Q}_{\text{max}}(G,\delta,n)}\right) \\ &\geq \min\left(1, (1-\epsilon)\overline{Q}_{\text{max}}(G,\delta,n)\right).\end{aligned}$$

We summarize this choice of quantile to evaluate in the following theorem:

Theorem 2.6

To achieve a generalisation certificate of G at level $1-\delta$ using Catoni's bound of any temperature $\lambda > 0$ with n observations, then the prior must be at least q_ϵ mass on predictors with empirical risk smaller or equal than r_ϵ with

$$r_\epsilon = \max\left(G - 2\sqrt{\frac{-\log(\delta)}{8n}}, \frac{\left(2G^2n + \log(\delta/\epsilon)\right)\left(G + \sqrt{G^2 + \frac{\log(\delta)}{2n}}\right)}{-2\log(\delta)}\right)$$

$$q_\epsilon = (1-\epsilon)\exp(-2G^2n - \log(\delta)).$$

for all $0 < \epsilon$, if $G \geq \sqrt{\frac{-\log(\delta)}{2n}}$, and $r = 0$, $q = 1$ else.

Remark 2.14

Theorem 2.6 might not always provide the tightest guidelines possible. For large G values, the analysis might prove to be quite loose. For instance, taking the values of Figure 2.3, one obtains for $\epsilon = 0.05$ the value $r_\epsilon = 0.115$, a far cry from $r_{\text{thresh}} = 0.0097$.

2.2.4 Implications for PAC-Bayes in deep learning

The analysis conducted in the previous sections formalize and quantify the intuition that to obtain good generalisation guarantees, the prior must put sufficient mass on high performing predictors. This questions the use of PAC-Bayes learning for the deep learning setting where it is commonly observed that there is no natural prior. A PAC-Bayes bound is valid for any data generating mechanism, including the case where there is no signal. Expecting the PAC-Bayes bound to behave better when there is some signal implies belief in the fact that the prior risk is better behaved in that case, that is to say put more weights on high performing predictors. This is at odds with the notion that the prior is simply a numerical intermediary. Let us for a moment consider the classification through Deep neural network case. In the limit of infinitely wide, fully connected, Neural Network, Gaussian prior distributions on the weights result in a last layer distributed according to a Gaussian Process. Moreover, as the number of layers increase, the

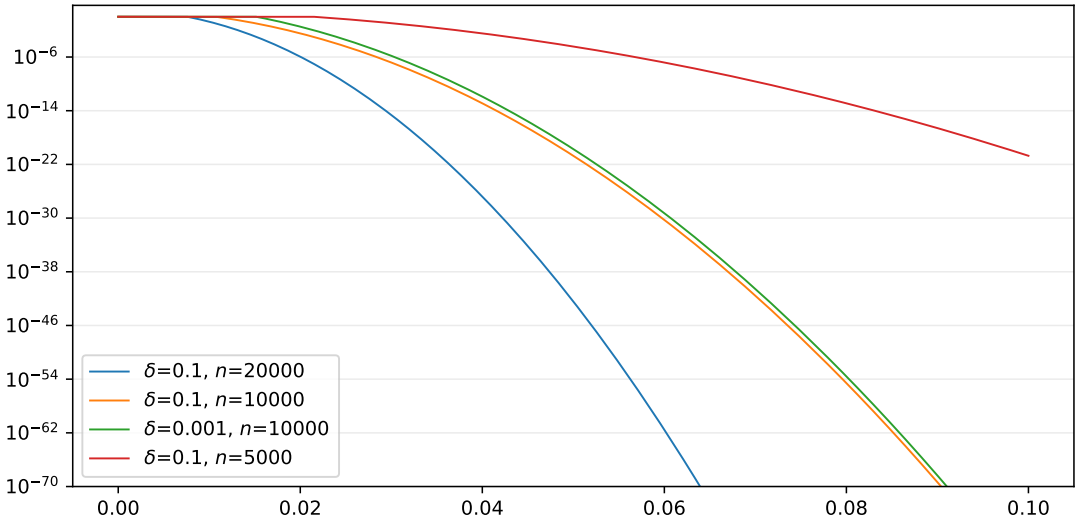


Figure 2.5: Evaluation of $\overline{Q}_{min}$ as a function of G for different δ and n values. The minimal quantile requirement decreases more than exponentially fast as the target generalisation bound grows. The speed of decrease is mostly driven by the number of observations.

covariance structure of the Gaussian process becomes increasingly flat [Lee et al., 2018]. In the limit where the number of layers is also infinite, this results in the predictor predicting the same class for all inputs with probability 1. Due to symmetry between the classes, the prior risk measure becomes the sum of Dirac distribution $\frac{1}{C}\sum_{i=1}^{C}\delta_{1-n_i/N}$ where C is the number of class and n_i is the number of data points of class i. The minimal prior risk is thus the majority class vote, leading to very loose generalisation guarantees. On the other hand, a prior leading to a fully random prediction (*i.e* the prediction for each x_i is independent and draws a class uniformly at random) would obtain identical the same prior risk whether there is signal or not, and hence, by virtue of theorem 2.4, also yield vacuous bounds.

A typical strategy to overcome the lack of natural prior is the construction of a data dependent prior, typically a Gaussian centred on the empirical risk minimiser for the first part of the data. While it leads to valid, non vacuous generalisation bounds, such a strategy raises a few questions. First of all, to lead to improved generalisation guarantees, the prior constructed using the first part of the data must generalize on the second part of the data. Corollary 2.2 shows that the data-dependent prior must put sufficient mass on high performing predictors for the second part of the data. Hence such a strategy relies heavily on the assumption that the generalisation performance of the empirical risk minimiser is satisfactory, *without* explaining why it is so. In a such cases, PAC-Bayes may be said to witness the generalisation abilities of the predictors considered (typically neural networks trained through **S**tochastic **G**radient **D**escent (SGD)) without providing any explanation. Moreover, the construction of such priors involves an arbitrary choice of variance term which, if chosen low enough, essentially forces the posterior to remain approximately equal to the prior. Hence the amount of learning which takes place in the PAC-Bayes

learning stage is essentially determined by a user-chosen criteria. Last but not least, almost all experimental results show that the generalisation bound obtained by PAC-Bayes is higher than the generalisation bound obtained by the data dependent prior mean using test bounds. Using the test performance of the prior mean reported in Pérez-Ortiz et al. [2021b,a], we compared test bounds obtained by from an extended form of Hoeffding's lemma (see Appendix B.2) with the PAC-Bayes generalisation certificate reported. For all but one case in the fourteen reported, the generalisation guarantee obtained by the test bound improved on the PAC-Bayes bound (see Table 2.2).

Table 2.2: Comparison of the generalisation guarantees obtained using PAC-Bayes with a data dependent prior and test bounds on the prior mean. Lines Spambase, Bioresponse, Har, Electricity, Mammography and MNIST 1 come from Pérez-Ortiz et al. [2021a] (table 2, without validation). Lines MNIST FCN and MNIST MCN corresponds to FCN and CNN architecture respectively, trained using criteria f_{quad} from table 1 in Pérez-Ortiz et al. [2021b]. Lines CIFAR-nL-p come from table 5 in the same source, with nL indicating the number of layers and p the fraction of data used to train the prior. The objective function considered is the one resulting in the tightest PAC-Bayes bound. In every case, the loss considered is the classification error. The PAC-Bayes bound and test bound are computed for a confidence level of 0.965.

Database	PAC-Bayes bound	Test bound	Test score	n_{valid}
Spambase	0.140	**0.0941**	0.077	1840
Bioresponse	0.318	**0.291**	0.261	1500
Har	0.035	**0.0307**	0.024	4119
Electricity	**0.223**	0.2290	0.221	18124
Mammography	0.022	**0.0202**	0.015	4473
MNIST 1	0.034	**0.0274**	0.025	30000
MNIST FCN	0.0279	**0.0224**	0.0202	30000
MNIST CNN	0.0155	**0.0120**	0.0104	30000
CIFAR-9L-50	0.2901	**0.2583**	0.2518	30000
CIFAR-9L-70	0.2377	**0.2249**	0.2169	18000
CIFAR-13L-50	0.2127	**0.1973**	0.1914	30000
CIFAR-13L-70	0.1758	**0.1649**	0.1578	18000
CIFAR-15L-50	0.1954	**0.1744**	0.1688	30000
CIFAR-15L-70	0.1667	**0.1560**	0.1490	18000

A similar analysis of the caveats of explaining generalisation through PAC-Bayes bound using data dependent prior is also shared by Lotfi et al. [2022]. The authors were able to obtain non vacuous generalisation guarantees for deep neural networks by applying various techniques such as quantization of the networks, dimension reduction and the construction of a universal prior favouring networks which can be compressed. Such techniques were also investigated by Zhou et al. [2019]. The authors motivate their choice of universal prior by remarking that it gives little weight to predictors which are selected during overfitting (see appendix A.1 of Lotfi et al. [2022]), leaving more probability mass for useful predictors. The key limitation of the PAC-Bayes objectives considered in this section is that they do not distinguish between predictors achieving the same risk. The choice of universal prior overcomes this limitation by incorporating an heuristic on the generalisation potential of each predictor. We believe that combining the weight requirement of Corollary 2.2 with such informed universal priors could be a promising way to

analyse the generalisation abilities of specific families of predictors. The discrepancy between state of the art results obtained by test bounds methods and the generalisation guarantees reported by Lotfi et al. [2022] shows that there are still exciting challenges to a comprehensive understanding of generalisation in deep learning, using PAC-Bayes or not.

2.2.5 Perspectives

Section 2.2.3 analysed the consequences of Corollary 2.2 on the tractable Catoni's bound (1.20). PAC-Bayes deep learning applications rely on other PAC-Bayes bound such as MLS [Pérez-Ortiz et al., 2021b,a]. Using a tighter PAC-Bayes bound should result in looser quantile requirements. MLS's bound can be interpreted as a modified Catoni's objective with temperature optimisation (see Section 4.1.5). As such, the PAC-Bayes bound for prior risk of rescaled Bernoulli form can be computed in the form of a minima on the temperature of a modified version of Equation (2.29). Preliminary analysis showed that Bernoulli distribution led to an optimal choice of temperature of 0 for a wide range of scale and probability, resulting in trivial quantile requirements for a wide range of generalisation target. Our interpretation of this is that Bernoulli distribution, assuming a strictly positive mass on perfect predictors, proved too optimistic an approximation of the prior risk distribution to lead to a meaningful analysis. A perspective for the analysis of MLS bound would be to consider other comparison distributions, *e.g.* truncated beta of form $\min(\beta, r)$, investigating the behaviour of the prior risk cumulative distribution function near 0.

2.2.6 Conclusion

The minimal generalisation level achieved by (almost all) PAC-Bayes bounds is an increasing function of the prior risk distribution. As such, a PAC-Bayes generalisation guarantee can only achieve a given target if the prior risk puts enough mass on low-risk predictors. This mass requirement involve small mass (*e.g.* of order 1e-10 in the MNIST inspired example). While such conditions appear tame for such problems as AD modelling involving a natural prior and a parameter space of moderate dimension, they might be more formidable in deep learning settings. The prevailing practice of using data-dependent priors in PAC-Bayes deep learning illustrates that uninformed priors fail to put sufficient mass on low-risk predictors. On the other hand, the fact that data-trained priors do put sufficient mass on these predictors illustrate, rather than explain the generalisation guarantees of deep learning algorithms - moreover, this practice resulted in looser generalisation bounds than the counterparts test bounds on all but one reported applications. Our PAC-Bayes bound analysis framework provides a way to quantify the conditions on the prior performance, and could offer an interpretation of recent breakthrough of informed, but data-independent prior, in terms of prior mass distributed to useful (*e.g.* not modelling noise) predictors.

2.3 General conclusion

Our two theoretical PAC-Bayes contributions addressed two distinct goals. The construction of new change of measure inequalities presented in section 2.1 serves as a first step towards the design of new PAC-Bayes generalisation bounds suited to more complex assumptions on the risk than the classical bounded assumption. It exhibited a trade-off between the type of risk assumption and the form of the penalisation involved in the PAC-Bayes bound which is of theoretical interest. These change of measure inequalities can be optimised on a single degree of freedom. This optimisation process is however intractable in the general case, hampering the design of adequate concentration inequalities for the tightest possible change of measure. As the KL divergence is both a classic choice and has simple a closed form expression between Gaussian distributions, we will consider KL penalised PAC-Bayes objective for the calibration of AD process, enforcing the classic bounded risk assumption using a soft minima on the residuals as detailed in Section 3.2.4.

Our second contribution is a framework designed to derive conditions on the prior in order to achieve a target generalisation level. This framework can be applied to most PAC-Bayes bounds in the literature, with the notable exception of the bound by Tolstikhin and Seldin [2013]. For these bounds, our key observation is that the optimal generalisation guarantee is fully determined by the prior risk distribution. No other properties of the predictors impact the final generalisation guarantee. Hence PAC-Bayes objectives can only succeed in building tight generalisation guarantees if a reasonable prior - one that puts sufficient mass on low risk predictors - is available. While this is questionable for the deep learning setting, the results of Lotfi et al. [2022] show that such priors can be constructed for some deep neural networks at least. We stress that the existence of a reasonable prior is usually not an issue for physical models, where the dimension is usually moderate and a natural prior exists. For the AD calibration tasks motivating this thesis, natural priors will be inferred from the literature; we proceed to analyse the behaviour of the resulting posterior distributions in the next chapter.

Chapter 3

Uncertainty quantification for Anaerobic Digestion process: the PAC-Bayes way

In this chapter, we focus on the application of PAC-Bayes to Anaerobic Digestion. Our goal is to assess the uncertainty quantification provided by PAC-Bayes posteriors. The main results in this chapter were obtained in collaboration with Roman Moscoviz, Benjamin Guedj and Gabriel Capson-Tojo and published in Bioresource Technology [Picard-Weibel et al., 2024a]. Code for this section can be found in the project's repo, `https://github.com/APicardWeibel/ADUncertaintyQuantification/`. Part of the 'aduq' source code it contains was merged and updated in the packages 'anaerodig', 'picoptim' and 'picpacbayes'.

Contents

A key motivation for the practical use of Bayesian statistics is that estimation and uncertainty quantification is performed in a unified way. Bayesian methods return a posterior distribution which informs on the most plausible parameter values after confronting prior beliefs to observations. When the data is generated from a statistical model belonging to a known parametric set, Bayesian statistics state that the posterior distribution will concentrate close to the true parameter and that moreover, Bayesian credible regions will asymptotically be frequentist confidence intervals, that is to say provide adequate uncertainty quantification (see Section 1.2.1).

This result is no longer valid whenever one no longer assumes that the possible ways the data may be generated are known. That is typically the case for biology applications, where the modelling involves approximation and leaves out some mechanisms deemed to have little impact, when the noise pattern is unknown or when the exact likelihood would be intractable. AD modelling suffers from all these difficulties; computational models[1] provide only a macroscopic description of the phenomena involved in AD and typically rely on unrealistic hypotheses

[1] The term AD model will be used for *computational* models rather than *statistical* models. We detail in eq. (3.7) how these computational models provide statistical models which can be used as a basis for uncertainty quantification.

(*e.g.* infinitely stirred reactor), the form of the measurement noise on the observations and feed description might be intricate (*e.g.* measuring low level signal might be more difficult, different methodologies may be used to measure the same quantity, the data might have undergone some complex cleaning), and moreover, even assuming that the noise pattern is known and that the model is accurate, evaluating the likelihood requires evaluating how the noise on the feed data is processed by the model, a task which might not be tractable.

PAC-Bayes theory overcomes some of the limitations of Bayesian theory by providing generalisation guarantees on the average test performance of randomised predictors. The minimisation of PAC-Bayes bounds leads naturally to the construction of PAC-Bayes posteriors, which in there turn can be leveraged to define PAC-Bayes credible regions. This provides a form of uncertainty quantification on the predictors. However, the classic Bayesian result providing asymptotic coverage no longer holds for the PAC-Bayes framework. This is no surprise; as PAC-Bayes extends Bayes to the learning framework, there no longer is any true parameter to recover.

Still, in the Anaerobic Digestion use case considered, parameter values are not simply computational intermediates; they are of independent interest as they inform on the biological and physical characteristics of the system considered. Obtaining adequate uncertainty quantification on these parameter values proves moreover to be crucial to correctly understand a digester, *e.g.* when assessing a response to new operational conditions. Indeed, popular AD models from the literature are known to have identifiability issues. That is to say, for a given operational condition, two sets of parameters with widely differing values γ_1 and γ_2 might lead to similar predictions. Consider for instance, the Monod equation, parametrized by $(\mathrm{K_S}, \mu_{\max})$, computing the growth rate as $\mu = \mu_{\max}\mathrm{S}/(\mathrm{S} + \mathrm{K_S})$. In the limit where the substrate concentration S is constant, the equation leads to the same growth rate μ for all couples $(\mathrm{K_S}, \mu(1 + \mathrm{K_S}/\mathrm{S}))$. To infer how the digester would react to a sudden increase of substrate concentration, it is necessary to keep track of the joint uncertainty on the maximum growth rate and inhibition parameter.

The following chapter is dedicated to the empirical study of the performance of the uncertainty quantification provided by the PAC-Bayes credible regions for AD models. The PAC-Bayes uncertainty quantification is benchmarked against the prevailing uncertainty quantification procedures found in AD literature.

3.1 Uncertainty quantification for AD models

Historically, modellers interested in AD have dedicated much effort over optimisation techniques, in order to efficiently calibrate AD models such as ADM1 [Bernard et al., 2001] and AM2 [Batstone et al., 2002a]. Typically, only a small number (less than 10) of parameters are calibrated, while the remaining parameters are set to some default value - either guessed by the modeller, or found in the literature (*e.g.* Rosén and Jeppsson [2006]). This trend is particularly noticeable since ADM1 and other models involving a large number of parameters were introduced.

These were found to be, in most settings, only sensitive to a fraction of parameters. The selected parameters are then either modified by hand until the output predictions visually match observations [Blumensaat and Keller, 2005, Ramirez et al., 2009, Schoen et al., 2009, Derbal et al., 2009, Mairet et al., 2011, Zhou et al., 2020, Li et al., 2021], or are set by minimising some objective function, *e.g.*, some weighted mean of squared residuals. This calibration strategy coincides with **E**mpirical **R**isk **M**inimisation (ERM). Various optimisation techniques have been explored in order to efficiently calibrate AD models: some require approximating the gradients and potentially higher order derivatives, such as sequential quadratic programming [Sales-Cruz and Gani, 2004, Aceves-Lara et al., 2005], Levenberg-Marquardt [Garcıa-Ochoa et al., 1999, Deveci and Çiftçi, 2001, Lokshina et al., 2001, Aceves-Lara et al., 2005], trust-region reflective Feldman et al. [2017], Odriozola et al. [2019], Regueira et al. [2021], while others can be said to be derivative free, such as Nelder-Mead [Mösche and Jördening, 1999, Ruel et al., 2002, Guisasola et al., 2009, Biernacki et al., 2013, Donoso-Bravo et al., 2013, Weinrich et al., 2021], the secant method [Batstone et al., 2003, Kalfas et al., 2006, Batstone et al., 2008, Poggio et al., 2016, Zhao et al., 2019], particle swarm optimisation Donoso-Bravo et al. [2011], simulated annealing [Haag et al., 2003, Kovalovszki et al., 2017] or genetic algorithms [Jeong et al., 2005, Wichern et al., 2009, Fatolahi et al., 2020]. Other, rarer techniques, were also tried (we refer to Donoso-Bravo et al. [2011] for a more exhaustive list of techniques).

Most of these methods rely on local search of the parameter, and should converge to a local minima. Multiple start procedures can then be used to increase the probability of finding a good approximation of the global minima. Therefore, while the choice of an optimiser is of practical interest, it can be argued that it impacts the computation cost of calibration rather than its accuracy. Even though the performance of the calibrated model should not depend on the specific optimisation technique used, the objective function remaining the same, more emphasis is put on fitting techniques than on other features that do actually have an influence on outputs. The form of the objective function is critical in favouring certain output behaviour, with mean square residuals being more likely than mean absolute residuals to force outputs to fit what could essentially be outliers Donoso-Bravo et al. [2011].

From a statistical viewpoint, the calibration of a number of parameters comparable to the number of observations is a high dimensional estimation problem, where special attention must be paid to the risk of overfitting. The prevalent methodology, by only fitting part of the parameters, mitigates that risk by lowering the effective dimension. Still, due to the non linear nature of the modelling, it is possible that sets of parameters, far apart, yield similar, nearly optimal outputs; empirical risk minimisation will favour the one which was able to model the measurement noise better, thus still overfitting.

Insight on the performance of the calibrated model is therefore essential if it is to be used with confidence. Rieger et al. [2012] indicate good practices for conducting validation, notably using different datasets, working under different conditions and operational parameters. In practice, observations are often scarce, preventing experimenters to fully validate their model [Dochain and Vanrolleghem, 2001]. This is notably the case when operating conditions are different for training and testing datasets. This assumption is called dataset or distribution shift in learning

theory [Quinonero-Candela et al., 2008]. In this setting, how much the predictions will diverge from the truth as time goes by cannot be inferred through validation, since validation data is assumed to be missing. **U**ncertainty **Q**uantification (UQ) methods are essential tools to assess the quality of the calibration. UQ is the quantitative analysis of the impact that sources of randomness have on the calibration process. This randomness originates from the measurements' noise as well as the stochastic nature of the future influent. UQ methods estimate how far the calibrated parameter values as well as the predictions of the calibrated model diverge from the truth, using statistical theory. Ideally, UQ on the predictions should be robust to distribution shift, warning the user whenever the calibration is no longer valid through an increase in predictions uncertainty.

Unfortunately, careful assessment of uncertainty is far from systematic in AD modelling. Still, techniques have been used to quantify uncertainty for AD model calibration. The three prevalent methods in the field are

- based on **F**isher's **I**nformation **M**atrix (FIM) (a.k.a. Cramér-Rao's lower bound or information inequality, see Chapter 2 in Lehmann and Casella [1998]), notably used by Lokshina et al. [2001], Haag et al. [2003], Guisasola et al. [2009], Donoso-Bravo et al. [2013],
- a statistically motivated threshold which we call Beale's method from the name of the author of its first description [Beale, 1960], notably used by Batstone et al. [2003, 2008], Spyridonidis et al. [2018][2],
- bootstrapping [López and Borzacconi, 2010, Odriozola et al., 2019, Regueira et al., 2021].

These methods infer uncertainty on the parameters from residuals of the calibrated model. They are generally used to provide UQ only for parameters which are fitted, *e.g.*, those selected by a sensitivity analysis routine. The remaining parameters are fixed at some default value, as they are not deemed to have sufficient influence on predictions. But those parameters actually have large uncertainty, since the data is not able to discriminate between two widely different values. As the sensitivity of the model's response to each parameter depends on operating conditions, the quality of extrapolations of the model on new conditions can only be known if uncertainty is quantified on all parameters, or if these operating conditions have been previously validated [Rieger et al., 2012].

No fitting method can perfectly infer all parameters whenever data is insufficient. But keeping track of uncertainty for all parameters can prevent overconfidence in predictions under all scenarios, without having to validate it first. Much attention has been raised on the question of the potential lack of practical identifiability of AD models Ruel et al. [2002], Flotats et al. [2006], van Loosdrecht et al. [2016], Poggio et al. [2016], Pastor-Poquet et al. [2019], Zhao et al. [2019], Weinrich et al. [2021]. For instance, Poggio et al. [2016] notes that "Mathematical models for biotechnological processes, such as ADM1, include many parameters with uncertain values,

[2] A variant where the threshold is arbitrary is also used by Weinrich et al. [2021].

and relatively few measured outputs, which in turn makes them hard to calibrate due to structural/practical identifiability issues (Dochain and Vanrolleghem, 2001). Attempts to fit all the parameters simultaneously usually result in very low confidence in the estimated parameters. Therefore, only the parameters describing the hydrolysis rate and the substrate kinetic fractionation were selected for calibration, as they are the most sensitive when hydrolysis rate limiting conditions occur." Similarly, Weinrich et al. [2021] states that "unique identification (structural or practical identifiability) of both (maximum growth rate and Monod constant) parameters is not possible, as previously described by Holmberg (1982) or Dochain et al. (1995). Thus, application of literature or standard values for practical and clear parameter estimation was examined." While such technique forcing identification of parameters which can not be identified (low sensitivity on the dataset) or can not be jointly identified (parameters compensating one another) can result in models giving accurate predictions on similar datasets, they will lead to models with low robustness to change of input. We advocate a change of paradigm; rather than forcing identification of the model by not fixing some parameter values to some default values, we propose the use of a comprehensive description of all equally plausible parameter values, that is to say a unified description of the joint uncertainty on all parameters.

Bayesian methods inherently tackle the uncertainty on all parameters throughout the training process, by performing calibration and UQ jointly. The uncertainty on parameters with little sensitivity impact on the model is controlled through the prior distribution, which encodes expert knowledge. The prior is twisted into the posterior distribution through confrontation with the observations, concentrating on sets of parameters likely to have generated them. The calibrated model is no longer deterministic, but stochastic. Identifiability issues are resolved by uncertainty management rather than using a single default value. No further parameter selection is necessary, as the Bayesian posterior should leave parameters having no impact on the predictions unchanged[3].

While uncommon, Bayesian flavoured techniques have already been used in the context of AD modelling [Martin et al., 2011, Couto et al., 2019, Pastor-Poquet et al., 2019]. All these Bayesian inspired algorithms output the uncertainty in the form of a sample. Statistical theory shows that satisfactory description of a generic distribution requires a number of samples increasing at a more than exponential rate with the number of parameters fitted. Indeed, the non-parametric estimation of the density function of a probability distribution from an i.i.d. sample is a well studied problem [Goldenshluger and Lepski, 2013]. The mean squared error minimax rate of estimation depends on the smoothness class of the density function. For Nikol'skii classes of function of smoothness s, which generalize Holder spaces for density functions, McDonald [2017] shows that the minimax rate of estimation for the L_2 norm has rate $\left(\frac{n^s}{d_\Gamma{}^{d_\Gamma}}\right)^{\frac{1}{2s+d_\Gamma}}$. This implies that, in order to maintain a similar (asymptotic) quality of estimation for each dimension, the number of samples should increase more than exponentially with the number of dimensions.

[3]This property is valid for exact Bayesian posteriors as well as PAC-Bayes posteriors defined as minimiser for the PAC-Bayes bound *if* the prior distribution draws each parameter dimension independently. This property remains correct for Variational Approximation for variational family which can maintain this independence. In practice, the training process can create some spurious learning on parameters having no impact on the model. This will be corrected in chapter 5 through a thresholding technique.

This heavily restricts the applicability of such algorithms for highly parametrized models.

Variational Bayesian methods (see [Hinton and van Camp, 1993, Beal, 2003] and section 1.2.4) learn the best approximation of the posterior amongst a class of distributions (*e.g.*, Gaussians, Gaussians mixture). This structured posterior can be fully assessed using fewer model evaluations. The flexibility of the posterior (covariance structure, multimodality) is governed by the distribution class considered. For AD modelling, Gaussian posteriors with block diagonal covariance present an interesting trade off. It explores non trivial correlation structure while limiting the number of hyperparameters and incorporating expert insight on parameter interactions.

We introduce a new methodology for the joint calibration and UQ of AD model called **Var**iational PAC-Bayes **U**ncertainty **Q**uantification routine (VarBUQ), designed to apply Variational PAC-Bayes approach to computationally intensive AD models. Its performance is compared to the prevailing ad hoc UQ routines, considering two AD models of varying complexity - AM2 [Bernard et al., 2001] and ADM1, [Batstone et al., 2002a]-, using six synthetic datasets describing three different operating conditions and two sets of parameters monitoring the prior distribution's inductive bias. Such synthetic data allowed to assess the ability of the UQ methods to recover the true parameter values, as well as their performance on test datasets. Special attention was paid to the robustness of each UQ methods to distribution shift.

3.2 Construction of the benchmark datasets

3.2.1 Synthetic data versus real world data

In order to assess the ability of the uncertainty quantification routine to find the correct set of parameter values (noted γ^*), such a correct set of parameter values should exist, and be known. As AD models only provide an approximation of the mechanisms involved during AD, it could be argued that for real-world datasets such a correct set has limited existence. For instance, while the resistance of a microbial community to low pH values can be measured and provides useful insight on the data, the inhibition curve used in models are only approximate, and hence no value can be said to be strictly correct. As a result, even if accepting the existence of a correct parameter, measuring its values would involve various methodological choices. This in turn would introduce potential intractable biases in the analysis of the uncertainty quantification routines, making its conclusions questionable (*if a UQ routine fails, is it because it is inappropriate for the context of AD models or because the parameter it is supposed to recapture was inappropriately measured?*).

To avoid such shortcomings, we considered synthetic datasets, generated directly from the AD model considered in the benchmark (*i.e.* ADM1, AM2) with a *known* parameter. Such a strategy factors out potential shortcomings of the model when modelling AD processes, and focuses on its inner potential for overfitting and interaction with UQ routines. This provides a conservative analysis: if the UQ routine is unable to encompass the true parameter values when the data is generated from the model, there is little likelihood that it would fare better when

considering real world data.

3.2.2 Dataset generation procedure

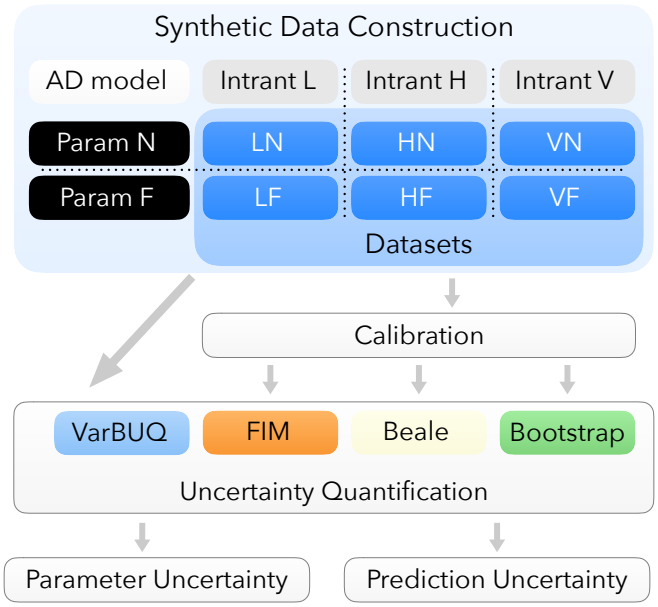


Figure 3.1: Uncertainty Quantification analysis methodology. For each Anaerobic Digestion model (AM2, ADM1), six datasets were evaluated, spanning a choice of two parameters and three intrant descriptions. After calibration, three ad-hoc uncertainty quantification methods (Fisher's information, Beale criteria, bootstrap) were assessed and compared to VarBUQ's joint calibration and uncertainty quantification, both on their ability to encompass the true set of parameter values and to quantify uncertainty on the predictions.

General dataset characteristics

For both AD models, the system modelled consisted in a single tank digestion system with a constant liquid phase measuring 3 400 cubic meters and a gas phase measuring 300 cubic meters. The temperature inside the digester was assumed constant at 308.15 Kelvin. The digester was assumed to be a perfectly homogenous system (infinitely stirred).

The frequency of the data was set to represent realistic monitoring data - respectively 24 hours and 6 hours between each data point for observations (*e.g.*, biogas quality) and influent data (*e.g.*, mass flow). 280 days were simulated, with the first 70% (196 days) used for calibration. The remaining part was used to assess the validity of the UQ methods on the predictions.

Model implementation

The implementation of ADM1 model was adapted from the PyADM1 source code of Sadrimajd et al. [2021].

The implementation of AM2 model was written from the formulation given in Bernard et al. [2001], with slight modifications to benefit from pH measurements and include microbial mortality. The original description of AM2 defines the concentration of CO_2 as $CO_2 = C + S_2 - Z$ [equation 19 from Bernard et al., 2001]. This is implied from equations $Z \simeq B + S_2$ and $C = CO_2 + B$ [respectively equations 10 and 3 from Bernard et al., 2001].

Since accurate pH measurements were assumed to be available, an alternative formulation, $CO_2 = C/(1 + 10^{pH-pKb})$, was implemented. It is a straightforward consequence of equations $K_b = \frac{[H^+]B}{CO_2}$ and $C = CO_2 + B$ [respectively equations 5 and 3 from Bernard et al., 2001].

The growth rate formulas [equations 33 and 34 in Bernard et al., 2001] were also modified to account for mortality, as described in Hassam et al. [2015]. Equations 6 and 7 from the latter source were implemented.

Both models were implemented in Python and are available in the python package anaerodig[4].

Generation of the input data

For each AD model, three type of input data, denoted L, H, and V (for *Low*, *High* and *Varying* dynamism of the intrant) were constructed. These differ in terms of minimum **H**ydraulic **R**etention **T**ime (HRT). HRT, defined as the ratio of the liquid phase volume and the feed rate (in volume per unit of time), measures the time required for the whole tank content to have been replaced and thus directly impacts the mixing time of observation data such as concentrations.

ADM1 For ADM1, the substrate was supposed to contain only carbohydrates, proteins and lipids. The concentrations of the remaining quantities were set to 0. Each of three concentration time series in the intrant was constructed independently using sums of 200 random sinusoids. The amplitudes and phase were drawn using uniform distributions between 0 and 1, and 0 and 2π respectively, while the frequencies were drawn using uniform distributions between $f_{\min}$ and $f_{\max}$. The sum was then renormalised to have a given mean and an amplitude (measured as the maximum difference between the maximum or minimum and the mean) of Amp times the mean value.

The feed rate was constructed in a similar manner, changing the renormalisation procedure to provide maximum and minimum values $Q_{\max}, Q_{\min}$. The feed rate of dataset V was obtained

[4]https://pypi.org/project/anaerodig/. The implementations can be accessed through the methods "simulate" of the classes "anaerodig.pyadm1.ADM1Dig" and "anaerodig.pyam2.AM2Dig". The code used to conduct the experimental results presented in this chapter corresponds to an earlier implementation of the AD models, which can be accessed in the project's repo. The new implementation differs in benefiting from just-in-time compilation (see chapter 4), a shared abstraction for AD models, and some minor code improvements which should not impact this chapter's main results.

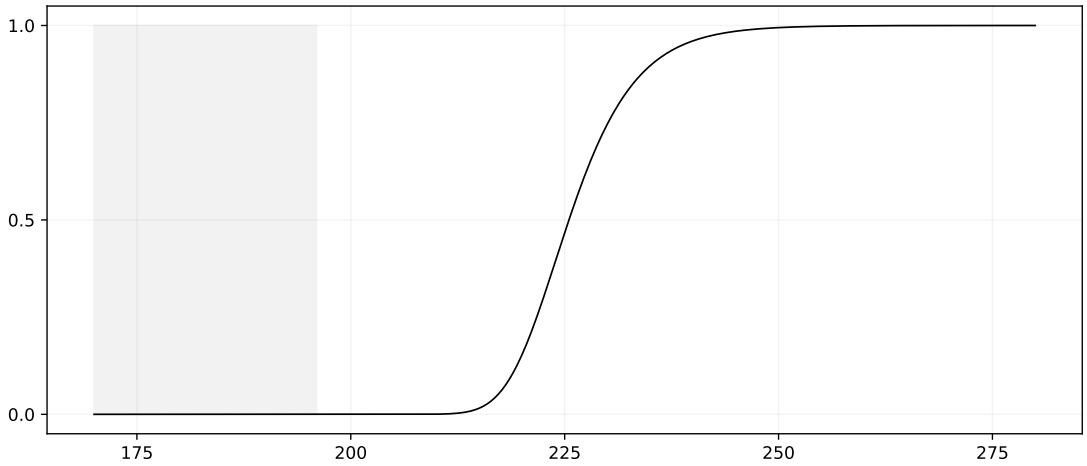


Figure 3.2: Ramp applied to the feed rate for V (*Varying intrant dynamism*) datasets. The grey zone corresponds to the training data.

from the feed rate of the L dataset by multiplying it by the ramp

$$f(t) = \left(\frac{\tanh\left(\frac{t-215}{10}\right) + 1}{2} \right)^6,$$

shifted and scaled such that $f(-\infty) = 1$ and $f(\infty) = \frac{\mathrm{Q_{max,H}}}{\mathrm{Q_{max,L}}}$ (the raw ramp is represented in Figure 3.2). This ensures that the maximum feeding rate for the test set remains below $\mathrm{Q_{max,H}}$, and that the feed data of the L and H are almost identical on the training set (relative difference smaller than $1.1e-10$). The maximum and minimum feed rates are inferred from minimum and maximum HRT, using the change of variable $Q = V/\mathrm{HRT}$ where V is the liquid phase volume.

Table 3.1: Influent dataset construction hyperparameters

	X_{ch}	X_{pr}	X_{li}	f_{max}	f_{min}	Amp	$\mathrm{HRT_{min}}$	$\mathrm{HRT_{min}/HRT_{max}}$
L	10	20	3	0.01	0.002	0.2	30	0.9
H	10	20	3	0.02	0.002	0.4	10	0.7

The concentrations of carbohydrates, proteins and lipids (X_{ch}, X_{pr}, X_{li}) are in kgCOD m^{-3}. The frequencies f_{max} and f_{min} are in Day^{-1} and the HRT in Day.

The values of the hyperparameters used to generate L and H datasets are tabulated in table 3.1.

AM2 The intrant description for AM2 was obtained by converting intrant description for ADM1 using the conversion methodology proposed by Hassam et al. [2015]. The pH of the influent, necessary for the conversion, was set to 6.0. Note that a fresh realisation of the random ADM1

feed generation process was used to generate the intrant description for AM2 (but that the L and V feed data are coupled for each AD model).

3.2.3 Generation of the true sets of parameters

Table 3.2: Default parameter values (γ_0), uncertainty ($\sigma_{\log\gamma}$) and generated parameters (γ_N, γ_F) for ADM1 and AM2 models

	γ_0	$\sigma_{\log\gamma}$	γ_N	γ_F
k_{dis} (Day^{-1})	0.5	0.24	0.35	0.23
$k_{hyd,\,CH}$ (Day^{-1})	10	0.12	10.0	5.2
$k_{hyd,\,PR}$ (Day^{-1})	10	0.12	8.6	11.9
$k_{hyd,\,LI}$ (Day^{-1})	10	0.24	23.4	2.4
$k_{m,\,su}$ (Day^{-1})	30	0.12	37.7	47.7
$k_{m,\,aa}$ (Day^{-1})	50	0.12	44.6	48.7
$k_{m,\,fa}$ (Day^{-1})	6	0.24	5.2	4.47
$k_{m,\,c4+}$ (Day^{-1})	20	0.12	21.2	17.6
$k_{m,\,pro}$ (Day^{-1})	13	0.12	12.7	9.9
$k_{m,\,ac}$ (Day^{-1})	8	0.12	10.7	10.2
$k_{m,\,h2}$ (Day^{-1})	35	0.12	27.9	28.8
k_{dec} (Day^{-1})	0.02	0.12	0.0157	0.0145
$K_{S,\,IN}$ (kMole m^{-3})	1e-4	0.046	9.3e-5	9.46e-5
$K_{S,\,su}$ (kgCOD m^{-3})	0.5	0.12	0.427	0.29
$K_{S,\,aa}$ (kgCOD m^{-3})	0.3	0.046	0.31	0.29
$K_{S,\,fa}$ (kgCOD m^{-3})	0.4	0.24	0.389	0.58
$K_{S,\,c4}$ (kgCOD m^{-3})	0.2	0.24	0.300	0.399
$K_{S,\,pro}$ (kgCOD m^{-3})	0.1	0.12	0.115	0.065
$K_{S,\,ac}$ (kgCOD m^{-3})	0.15	0.12	0.147	0.24
$K_{S,\,H2}$ (kgCOD m^{-3})	7e-6	0.12	6.6e-6	6.1e-6
$K_{I,\,H2,\,fa}$ (kgCOD m^{-3})	5e-6	0.046	5.01e-6	5.60e-6
$K_{I,\,H2,\,c4+}$ (kgCOD m^{-3})	1e-5	0.046	1.06e-5	1.1e-5
$K_{I,\,H2,\,pro}$ (kgCOD m^{-3})	3.5e-6	0.046	3.15e-6	3.4e-6
$K_{I,\,NH3}$ (kMole m^{-3})	1.8e-3	0.046	1.88e-3	2.1e-3
$pH_{UL:LL\,aa}$	1.5	0.12	1.16	1.36
$pH_{LL\,aa}$	4.0	0.12	5.11	4.4
$pH_{UL:LL\,ac}$	1.0	0.046	1.02	1.0
$pH_{LL\,ac}$	5.0	0.046	4.53	4.8
$pH_{UL:LL\,h2}$	1.0	0.12	0.52	1.6
$pH_{LL\,h2}$	5.0	0.046	5.5	6.1

(a) ADM1

	γ_0	$\sigma_{\log\gamma}$	γ_N	γ_F
μ_{1max} (Day^{-1})	1.2	0.2	1.23	0.99
μ_{2max} (Day^{-1})	0.74	0.48	1.05	1.32
K_{S_1} (gCOD L^{-1})	7.1	0.32	7.51	16.0
K_{S_2} (mmol L^{-1})	9.28	0.48	11.0	4.97
K_{I2} (mmol L^{-1})	256	0.4	196	345

(b) AM2

For each AD model, two sets of true parameters were constructed, denoted γ_N^* and γ_F^*, describing two levels of prior credibility. Using the prior distribution (see definition in section 3.3.2), a parameter γ was drawn at random. The parameter γ_N^* was then constructed by rescaling γ in such a way that γ was on the boundary of the prior credible region for level 0.05 (respectively 0.95 for γ_F^*). Using the fact that the prior is, in log-space[5], a Gaussian $\mathcal{N}(\mu_p, \Sigma_p)$, credible

[5]The log transform is applied component wise.

regions were defined as

$$\mathsf{CR}_{\mathsf{prior}}(\alpha) = \left\{ \gamma \mid \left(\log(\gamma) - \mu_p\right)^T \Sigma_p^{-1} \left(\log(\gamma) - \mu_p\right) \leq Q_{\chi^2(d_\Gamma)}(\alpha) \right\} \tag{3.1}$$

with $Q_{\chi^2(d_\Gamma)}$ the quantile function of the $\chi^2(d_\Gamma)$ distribution and d_Γ the dimension of the parameter. This led to

$$\gamma_N^* = \exp\left(\mu_p + \sqrt{\frac{Q_{\chi^2(d_\Gamma)}(0.05)}{\left(\log(\gamma) - \mu_p\right)^T \Sigma_p^{-1} \left(\log(\gamma) - \mu_p\right)}} \left(\log(\gamma) - \mu_p\right) \right).$$

Redrawing a fresh γ from the prior and replacing 0.05 by 0.95 defined the parameter γ_F^*[6].

The values of γ_N^* and γ_F^* for each AD model are provided in table 3.2.

Generation of the digester states

The true states of the digester system were obtained by running the AD model considered using the feed data Feed and the true parameter γ^*. The initial state of the system was obtained through steady state simulations. Defining $\widetilde{\mathsf{Feed}} = (\mathsf{Feed}_0, \ldots, \mathsf{Feed}_0)$ (with initial feed description repeated for 150 days hence higher than three hydraulic retention time), and starting from an arbitrary initial state $\widetilde{\mathsf{State}}_0$, the initial state was obtained as the prediction $\mathsf{State}_0 := \mathsf{AD}(\gamma^*; \widetilde{\mathsf{Feed}}, \widetilde{\mathsf{State}}_0)_{t_{\mathsf{max}}}$ at the end of the simulation. The true states of the digester, obtained by $\mathsf{AD}(\gamma^*, \mathsf{Feed}, \mathsf{State}_0)$, are denoted AD^*.

Application of noise to the data

The feed, initial state and observation data were noised in log space using uniform noise, *i.e.*

$$\begin{aligned}
\mathsf{Obs} &= \mathsf{AD}^* \times \exp(\mathcal{U}(-\sigma_O, \sigma_O)) \\
\widehat{\mathsf{Feed}} &= \mathsf{Feed} \times \exp(\mathcal{U}(-\sigma_F, \sigma_F)) \\
\widehat{\mathsf{State}}_0 &= \mathsf{State}_0 \times \exp(\mathcal{U}(-\sigma_O, \sigma_O)).^7
\end{aligned}$$

The observation noise level was set to $\sigma_O = 0.15$ while the feed noise level was set to $\sigma_F = 0.08$. These values are hidden from the calibration and uncertainty quantification routines.

A total of 12 datasets were thus constructed, spanning 2 AD models (ADM1, AM2), 3 types of intrant dynamism (L, H, V), 2 levels of inductive bias brought by the prior (N, F for *near* and *far*). The dataset constructed using the intrant L and parameter N is denoted LN.

[6] When defining a full dataset (*i.e.* a combination of feed data and a parameter), the procedure defining γ^* was repeated until the AD model computation does not fail (both for the raw and noisy feed data). Model failure arose when the model predicted acidification of the digester, which led to numerical instabilities. Such cases would not be representative of the behaviour of a typical plant.

[7] As Obs, Feed are matrices and State_0 is a vector, independent draws of the noise are applied to each component (*e.g.* $\mathsf{Obs}_{i,k}$) of the matrices and vector.

3.2.4 Construction of the risk function

The risk function was obtained by comparing the predictions of the AD model for a given parameter to the observation data. The risk considered was essentially a RMSE for the log predictions, similar to the objective in Wichern et al. [2009]. Using log predictions removed the need to renormalize the different types of observations (volatile fatty acid concentrations, gas production), was coherent with the methodology used to noise the data and offered an interpretation of the risk as the average relative prediction error. However, it may result in a single observation point having large impact on the overall risk if the signal is close to 0. To mitigate this, an offset $\eta = 10^{-8}$ was added to both predictions and observations before computing the log. Finally, to prevent a single point from having too much impact, the log residuals passed through a modified softplus function, smoothly approximating $x \mapsto \min(|x|, R_{\max})$ with $R_{\max} = 3$. We call this function a softclip function (see eq. (3.3)). For real world use case, such a mechanism would also make the fitting procedure more robust to outliers, by capping the maximum impact of a single observation. We note that replacing the softclip function by a function f which behaves as $f(t) \sim_0 t$ and $f(t) \sim_\infty \sqrt{t}$ would result in an objective similar to the classic RMSE for large values, but offering the robustness of mean absolute error to outliers. Our choice should provide even higher robustness. The saturation level of 3 translates into a factor of 20 between prediction and observation, indicating grossly inadequate predictions. When the ratio between observation and prediction is between $[1/3, 3]$ (*i.e.* log-residuals in $[-1, 1]$), the softclip function has negligible impact.

For some set of parameters, the ordinary differential equation solver involved in the AD model might fail to find a solution. In such cases, the set of parameters was given a risk of $R_{\max}$.

The type of observations (noted obs) considered available and used for the risk evaluation for AM2 were the concentrations of soluble compounds (denoted $\mathrm{S_1}$, $\mathrm{S_2}$ in the original paper) and the gas flows ($\mathrm{q_M}$ and $\mathrm{q_C}$). For ADM1, these were VFA concentrations ($\mathrm{S_{va}}$, $\mathrm{S_{bu}}$, $\mathrm{S_{pro}}$, $\mathrm{S_{ac}}$), concentration of inorganic nitrogen ($\mathrm{S_{IN}}$), gas flows ($\mathrm{q_{gas}}$, $\mathrm{q_{CH4}}$) and partial pressures ($\mathrm{p_{gas,CH4}}$, $\mathrm{p_{gas,CO2}}$). This leads to a total of $N_{\mathsf{AM2}} = 784$ train observations for AM2 (resp. $N_{\mathsf{ADM1}} = 1764$ train observations).

All in all, the risk function is defined as

$$R(\gamma) = \sqrt{\frac{1}{N} \sum_{t,\mathsf{obs}} \mathsf{softclip}\left(\log\left(\frac{\mathsf{AD}_{t,\mathsf{obs}}(\gamma, \widehat{\mathsf{Feed}}, \widehat{\mathsf{State}_0}) + \eta}{\mathsf{Obs}_{t,\mathsf{obs}} + \eta}\right)\right)^2}, \tag{3.2}$$

with the softclip function defined as

$$\mathsf{softclip}(r) = \frac{1 + \exp(-2R_{\max})}{2} \times \log\left(\frac{1 + \exp(2R_{\max})}{1 + \exp(2(R_{\max} - |r|))}\right). \tag{3.3}$$

The impact of the softclip function can be inferred from its plot (see Figure 3.3). Neglecting the effect of the noise on the influent and the noise on the initial state on the risk, the true parameter

γ^* has a train risk distributed as the root of the average of N squared uniform, which should concentrate around $\sigma/\sqrt{3} = 0.087$ (95% upper bound of 0.089 for AM2, 0.88 for ADM1). Higher values indicate that the noise on the influent and initial state cannot be neglected.

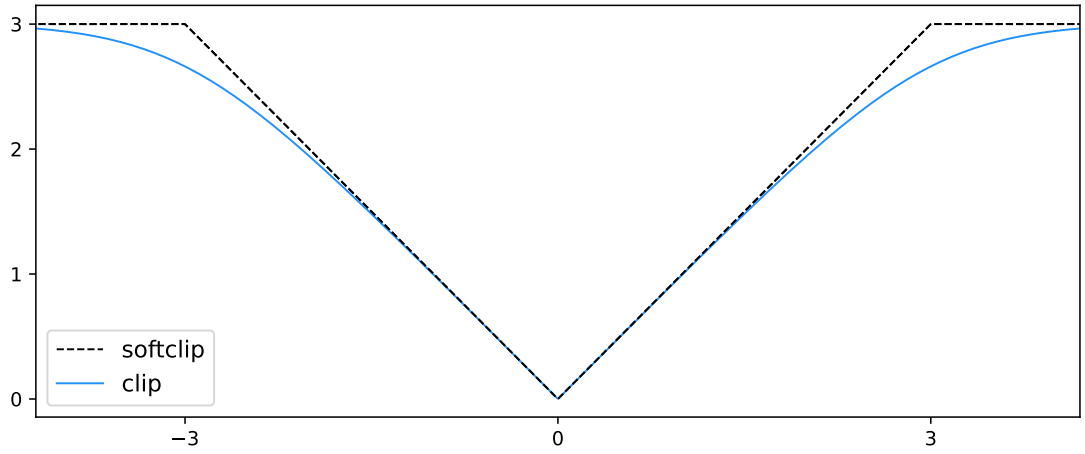


Figure 3.3: Soft-ceiling correction applied to residuals. The function is designed to cap the residual values to 3 while having little impact on residuals smaller than 1.

3.3 Variational Bayes with sample memory

We detail here our implementation of the Variational PAC-Bayes strategy for the calibration of AD models, VarBUQ.

The code for VarBUQ routine used for experiments can be found in the project's repo[8]. Pseudo code for the calibration strategy is given in Algorithm 1. An updated implementation is also available in the picpacbayes package[9].

3.3.1 Variational family

As discussed in section 1.2.4, the Variational PAC-Bayes approach considers a PAC-Bayes objective PB and defines the posterior distribution as the minimiser of the bound on a parametric family of distribution $\mathcal{P}$, *i.e.* $\hat{\pi} = \arg\inf_{\pi\in\mathcal{P}} \mathrm{PB}(\pi, R, \pi_\mathrm{p}, \delta)$.

This results in a posterior distribution whose entire form is known, simplifying interpretation, storage and reuse of the posterior distribution, and computation of functional of the posterior (*e.g.* mean, mode, variance). The choice of the variational family $\mathcal{P}$ controls the flexibility of the possible posterior distribution considered during optimisation; this will trade off the final generalisation guarantee obtained achieved and the difficulty of optimising the bound.

[8] 'optim_VI_wc' function from 'aduq.bayes' module for the general algorithm, while ADM1 (resp. AM2) specific code can be found in 'aduq.pyadm1.var_buq' module for ADM1 specific code (resp 'aduq.pyam2.var_buq' module).

[9] 'picpacbayes.GradientBasedPBayesSolver' class.

Gaussian distributions are a popular choice of variational family [Dziugaite and Roy, 2017, Pérez-Ortiz et al., 2021b] for PAC-Bayes objectives. The resulting posterior distribution is defined through the mean of the Gaussian, which can be used as a deterministic model, and its covariance structure, which offers insight on the uncertainty on the model. Different assumptions can be made on the structure of the covariance matrix to further restrict the variational family, from fixed variance hypothesis ($\{\mathcal{N}(\mu, \Sigma_0) \mid \mu \in \mathbb{R}^{d_\Gamma}\}$), to homoscedastic variance ($\{\mathcal{N}(\mu, \sigma^2 \mathrm{Id}) \mid \mu \in \mathbb{R}^{d_\Gamma}, \sigma \in \mathbb{R}_+\}$), to diagonal variance ($\{\mathcal{N}(\mu, \mathrm{diag}(\sigma^2)) \mid \mu \in \mathbb{R}^{d_\Gamma}, \sigma \in \mathbb{R}_+^{d_\Gamma}\}$, to any variance ($\{\mathcal{N}(\mu, \Sigma) \mid \mu \in \mathbb{R}^{d_\Gamma}, \Sigma \geq 0\}$).

For the AD models considered, parameters can be grouped by the reaction on which they have direct impact. This motivates the use of Gaussian distributions with a block diagonal covariance matrix, where each block consists of a group of parameters. This effectively limits the interactions between the parameter dimensions in the posterior distribution, and results in a much smaller optimisation space. To simplify the analysis, the covariance is required to be strictly positive definite (*i.e.* no 0 eigenvalues).

While Gaussian distributions draw values in $\mathbb{R}^{d_\Gamma}$, the values of the parameters considered in AD are always positive. To enforce positivity of the parameters, the Gaussian distributions define the log parameters. For ADM1, the parameters describing the pH inhibitions, $\mathrm{pH}_{\mathrm{LL}}$ (lower limit) and $\mathrm{pH}_{\mathrm{UL}}$ (upper limit), must also be ordered. To enforce that, log values for $\mathrm{pH}_{\mathrm{UL}}$ and $\mathrm{pH}_{\mathrm{UL:LL}} := \mathrm{pH}_{\mathrm{UL}} - \mathrm{pH}_{\mathrm{LL}}$ are drawn.

Thirteen groups, covering 30 parameters[10], are considered for ADM1:

- $\mathrm{k}_{\mathrm{hyd,\,CH}}$, $\mathrm{k}_{\mathrm{hyd,\,PR}}$, $\mathrm{k}_{\mathrm{hyd,\,LI}}$, $\mathrm{k}_{\mathrm{dis}}$
- $\mathrm{k}_{\mathrm{m,\,aa}}$, $\mathrm{K}_{\mathrm{S,\,aa}}$
- $\mathrm{k}_{\mathrm{m,\,fa}}$, $\mathrm{K}_{\mathrm{S,\,fa}}$, $\mathrm{K}_{\mathrm{I,\,H2,\,fa}}$
- $\mathrm{k}_{\mathrm{m,\,c4+}}$, $\mathrm{K}_{\mathrm{S,\,c4+}}$, $\mathrm{K}_{\mathrm{I,\,H2,\,c4+}}$
- $\mathrm{k}_{\mathrm{m,\,pro}}$, $\mathrm{K}_{\mathrm{S,\,pro}}$, $\mathrm{K}_{\mathrm{I,\,H2,\,pro}}$
- $\mathrm{k}_{\mathrm{m,\,ac}}$, $\mathrm{K}_{\mathrm{S,\,ac}}$
- $\mathrm{k}_{\mathrm{m,\,h2}}$, $\mathrm{K}_{\mathrm{S,\,h2}}$
- $\mathrm{k}_{\mathrm{m,\,su}}$, $\mathrm{K}_{\mathrm{S,\,su}}$
- $\mathrm{pH}_{\mathrm{LL\,ac}}$, $\mathrm{pH}_{\mathrm{UL:LL\,ac}}$, $\mathrm{K}_{\mathrm{I,\,nh3}}$
- $\mathrm{pH}_{\mathrm{LL\,aa}}$, $\mathrm{pH}_{\mathrm{UL:LL\,aa}}$
- $\mathrm{pH}_{\mathrm{LL\,h2}}$, $\mathrm{pH}_{\mathrm{UL:LL\,h2}}$
- $\mathrm{K}_{\mathrm{S,\,IN}}$

[10]Note that the implementation in our package 'anaerodig' contains up to 78 parameters. For the analysis conducted in this study, the 30 parameters most often calibrated are considered.

- k_{dec}

Two groups are considered for AM2:

- μ_{1max}, K_{S_1}
- μ_{2max}, K_{S_2}, K_{I2}

A code source for the abstract description of probability measures and parametric families of probability measures was developed during the thesis and packaged in picproba. The variational family corresponding to Gaussian distribution with block diagonal covariance matrix (with fixed blocks) is encoded in the class 'picproba.BlockDiagGaussianMap'. To simplify optimisation on this class, an over-parametrized implementation of the class was defined:

- for m blocks of dimension $d_1, \dots, d_m$, the parameter θ consists of m tuples $(\theta_{\mu_i} \in \mathbb{R}^{d_i}, \theta_{M_i} \in \mathbb{R}^{d_i, d_i})$.
- the distribution π_θ is defined as

$$\pi_\theta = \mathcal{N}\left(\begin{pmatrix}\theta_{\mu_1}\\ \vdots \\ \theta_{\mu_m}\end{pmatrix}, \begin{pmatrix}\theta_{M_1}\theta_{M_1}^t & & \\ & \ddots & \\ & & \theta_{M_m}\theta_{M_m}^t\end{pmatrix}\right).$$

This implies that for Lebesgue almost all θ of dimension $d_\Theta := \sum_{i \leq m} d_i(d_i+1)$, the distribution π_θ is well defined (since the covariance is always positive semi definite by design, positive definite whenever all θ_M are full rank, and that non full rank matrices have 0 Lebesgue mass). The actual dimension of the variational family is $\sum_{i \leq m} d_i(d_i+3)/2$, hence potentially almost twice smaller.

Similarly to the definition of the credible regions for the prior detailed in Equation (3.1), the credible regions of level α for the posterior $\pi_\theta := \mathcal{N}(\mu_\theta, \Sigma_\theta)$ is defined as

$$CR_{\text{PAC-Bayes}}(\alpha) := \left\{\gamma \mid (\log(\gamma) - \mu_\theta)^T \Sigma_\theta^{-1} (\log(\gamma) - \mu_\theta) \leq Q_{\chi^2(d_\Gamma)}(\alpha)\right\}.$$

3.3.2 Construction of the prior

VarBUQ requires a description of the a priori belief on the values of parameters in the form of the prior distribution. Following the Bayesian paradigm, this inductive bias should encode which parameter combinations are most likely to be observed for the task at hand. This can be inferred from values reported in the literature. Moreover, to simplify computations, we consider a prior distribution belonging to our variational family. This is not a requirement, but will prove to lead to much more tractable analysis. Here we neglect all potential correlations between the different parameters, resulting in a prior distribution which is a Gaussian with diagonal covariance structure, on the log parameter.

For both AD models, the prior distribution consists in multivariate Gaussian distributions with a diagonal covariance structure on the log parameters.

As such, the global prior draws each parameter independently from one another. This follows the recommendations of Tsigkinopoulou et al. [2017]. The individual priors are constructed using previous description of default values and uncertainty [Rosén and Jeppsson, 2006, Batstone et al., 2002a, Bernard et al., 2001]. Rosén and Jeppsson [2006] reported the potential ranges of the parameter values in term of three relative levels (30%, 100%, 300%). We defined the log standard deviation of the parameter by applying a factor of 0.4 to this relative level (*e.g.* $\sigma = 0.12$ for a parameter of relative level 30%). The factor is chosen so that 99% of the prior draws will be in the range provided. The resulting standard deviations are specified in Table 3.2.

3.3.3 PAC-Bayes objective

We use Catoni's bound (1.20) as a learning objective, *i.e.*

$$\mathsf{PB}_{\mathsf{Cat},\lambda}(\theta) = \mathsf{PB}_{\mathsf{Cat},\lambda}(\pi_\theta, R, \pi_\mathsf{p}, \delta) = \pi_\theta\left[R\right] + \lambda \mathsf{KL}(\pi_\theta, \pi_\mathsf{p}) + \frac{R_{\max}}{8n\lambda} + \lambda \log(\delta).$$

This is motivated by different considerations. First of all, Catoni's bound offers the most tractable objective, as the KL divergence between Gaussians has closed form expression

$$\begin{aligned} &\mathsf{KL}(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)) \\ =&\frac{1}{2}\left[(\mu_2 - \mu_1)^T \Sigma_2^{-1}(\mu_2 - \mu_1) + \langle \Sigma_2^{-1}, \Sigma_1 \rangle - \log\left(\left|\Sigma_1 \Sigma_2^{-1}\right|\right) - d_\Gamma\right], \end{aligned} \tag{3.4}$$

where $|M|$ denotes the determinant of matrix M. Second, due to its additive objective, Catoni's bound can be interpreted as a penalized regression task similar to the Lasso [Tibshirani, 1996], and as such is easy to interpret. Third, the variational PAC-Bayes setting for Catoni's bound coincides with the classic variational Bayes setting, replacing the posterior with Gibbs posterior. While this is not an advantage *per se*, this connects our PAC-Bayes algorithm to a popular learning strategy. Fourth, the optimisation of Catoni's objective can be seen as a gateway to the optimisation of other KL-penalized PAC-Bayes objectives such as MLS's bound (this point will be actively discussed in Section 4.1.5). But contrary to MLS's objective, the optimisation of Catoni's bound does not directly depend on the chosen PAC-Bayes confidence level δ and number of observations. This allows learning the simpler objective

$$\widetilde{\mathsf{PB}}(\theta) = \pi_\theta\left[R\right] + \lambda \mathsf{KL}\left(\pi_\theta, \pi_\mathsf{p}\right). \tag{3.5}$$

Moreover, the tighter rates of MLS PAC-Bayes bound come into effect whenever the prediction error of the posterior is close to 0 - in our setting, the prediction error will always remain higher than 0.08, for which both bounds should exhibit the similar $\frac{1}{\sqrt{n}}$ rates for carefully chosen temperature. Finally, the f-divergence penalized alternatives to Catoni's objective studied in Chapter 2 are designed to tackle settings where non standard moment assumptions on the risk

are available. For real world AD use case, little would be known on the actual way the data is generated, making it difficult to select one f-divergence in practice. Moreover, for complex f-divergence, closed form expressions of the divergence might not be available; the expression, and its gradient, would then have to be approximated, resulting in higher computation time and noisier gradient estimates.

3.3.4 Choice of PAC-Bayes temperature

The choice of the PAC-Bayes temperature was done a priori. The criteria used to define this PAC-Bayes temperature is based on Catoni's generalization bound (1.20), which controls the posterior true risk by Catoni's objective (3.5) plus $\frac{R_{\max}}{\lambda 8n} + \lambda \log\left(\delta^{-1}\right)$. This implies vacuous generalisation guarantees if the PAC-Bayes temperature chosen is too low.

The PAC-Bayes temperature was chosen in such a way that $\frac{1}{\lambda 8N} < 0.1$, which implied for ADM1 that $\lambda \geq 0.0007$ and for AM2 that $\lambda \geq 0.0016$. A safety margin was added, and PAC-Bayes temperatures of $\lambda_{\mathsf{ADM1}} = 0.001$ and $\lambda_{\mathsf{AM2}} = 0.002$ were used. For ADM1, PAC-Bayes temperatures two and eight times larger were also investigated. It should be stressed that this a priori choice of PAC-Bayes temperature is debatable, since Catoni's bound assumptions are not met: data is not independent, nor identically distributed, and the empirical risk is not a sum of contributions, but the square root of a sum.

Gradient estimation

The computation of the posterior was performed by minimising Catoni's objective (3.5) using accelerated gradient descent. Noting $\ell(\theta, \gamma) := \log\left(\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\pi_{\mathsf{p}}}(\gamma)\right)$ the log-likelihood of the predictor γ and $\widetilde{\mathsf{KL}}(\theta) = \mathsf{KL}\left(\pi_\theta, \pi_0\right)$, one can rewrite the objective (3.5) as

$$\widetilde{\mathsf{PB}}(\theta) = \int R(\gamma) \exp(\ell(\theta, \gamma)) \mathrm{d}\pi_{\mathsf{p}} + \lambda \widetilde{\mathsf{KL}}(\theta).$$

Since the risk is positive and upper bounded by 3, it follows that the derivative of the first term with respect to θ can be obtained by differentiation under the integral sign if the function $\theta \mapsto \int \exp\left(\ell\left(\theta, \gamma\right)\right) \mathrm{d}\pi_{\mathsf{p}}$ can be differentiated under the integral sign. This is implied if there is a neighbourhood V_θ of θ such that the $\sup_{\theta \in V_\theta} \partial_\theta \ell(\theta, x) \in \mathsf{L}^1(\pi_\theta)$ (see Theorem 6.28 in Klenke [2020]). The family of measures considered here are Gaussian with block diagonal covariance, and the parametrisation consists of multiple parametrisation for each block with no interactions. As a result, differentiation under the integral is possible if and only if it is possible for all block parametrisation. Let us consider a one block scenario with parametrisation θ_μ, θ_M. Then the log likelihood is of form $C - \log\left(|\theta_M|\right) - 1/2\left(\gamma - \theta_\mu\right)^T \left(\theta_M \theta_M^t\right)^{-1} \left(\gamma - \theta_\mu\right)$. This results in a differentiate

$$\begin{aligned}
\partial_{\theta_\mu} \ell(\theta, \gamma) &= (\theta_M \theta_M^t)^{-1} (\gamma - \theta_\mu) \\
\partial_{\theta_M} \ell(\theta, \gamma) &= -(\theta_M \theta_M^t)^{-1} \theta_M + (\theta_M - \gamma)(\theta_M \theta_M^t)^{-2} (\theta_M - \gamma)^t \theta_M.
\end{aligned}$$

Since the parametrisation considers θ_M such that $\theta_M\theta_M^t$ is invertible (the distribution is positive definite), this implies that there exists a neighbourhood of θ_M such that the absolute value of the eigenvalues of the resulting covariance is lower bounded by ϵ. This in turn implies that the supremum of $\partial_\theta \ell(\theta, \gamma)$ on a neighbourhood of θ is integrable with respect to π_θ, and hence that

$$\begin{aligned}\nabla\widetilde{\mathsf{PB}}(\theta) &= \int R(\gamma)\exp(\ell(\theta,\gamma))\partial_\theta\ell(\theta,\gamma)\mathrm{d}\pi_\theta + \lambda\nabla\widetilde{\mathsf{KL}}(\theta)\\ &= \pi_\theta\left[R\partial_\theta\ell(\theta,\cdot)\right] + \lambda\nabla\widetilde{\mathsf{KL}}(\theta).\end{aligned}$$

For Gaussian distributions, the KL divergence expression (3.4) can be directly differentiated. The derivative of the mean risk term is expressed as an expectation, which can be approximated using an i.i.d. sample of $(\gamma_i)_{i\in[1,K]}$ from π_θ. This gives an unbiased estimator, with variance decaying as $\frac{1}{\sqrt{K}}$. This slow decay of the variance implies that the sample size should be somewhat large to obtain good approximations. Moreover, since the sample $(\gamma_i)_{i\in[1,K]}$ has to be sampled from a specific distribution, the naive unbiased estimator needs to reevaluate a high number of risks at each step. If the gradient steps are small, the samples from the previous distribution $\pi_{\theta_{t-\tau}}$ should be similar to samples from the current distribution π_{θ_t} - implying that the information from the previous samples is meaningful. To reuse samples from previous evaluations, notice that

$$\begin{aligned}\pi_\theta\left[R\partial_\theta\ell\left(\theta,\cdot\right)\right] =& \pi_{\tilde\theta}\left[R\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\pi_{\tilde\theta}}\partial_\theta\ell(\theta,\cdot)\right]\\ =& \pi_{\tilde\theta}\left[R\exp(\ell(\theta,\cdot)-\ell(\tilde\theta,\cdot))\partial_\theta\ell(\theta,\cdot)\right].\end{aligned}$$

This implies that using the sample $(\gamma_{t-\tau,k})_{k\in[1,K]}$ drawn i.i.d. from $\pi(\theta_{t-\tau})$,

$$\dot\theta_{t,\tau} := K^{-1}\sum_{k=1}^{K} R(\gamma_{t-\tau,k})\exp\left(\ell(\theta_t,\gamma_{t-\tau,k})-\ell(\theta_{t-\tau},\gamma_{t-\tau,k})\right)\nabla\ell(\theta_t,\gamma_{t-\tau,k})$$

is an unbiased estimator of $\nabla\pi_\theta\left[R\right]$. The variance of this estimator is impacted by the fluctuations of the random variable $\exp\left(\ell\left(\theta_t,\gamma\right)-\ell\left(\theta_{t-\tau},\gamma\right)\right),\ \gamma\sim\pi_{\theta_{t-\tau}}$. The average of this random variable is one, indicating full mass transfer between the distributions when exploring all the predictor space. For distributions far apart (*e.g.* in KL sense) however, most of the draws from $\pi_{\theta_{t-\tau}}$ will fall in low mass regions for π_{θ_t}, while the rare draws hitting large mass regions for π_{θ_t} will be given large mass transfer to compensate; hence the random draws of the ratio of density will have large fluctuations, impacting the performance of the gradient estimator. This phenomena increases with the dimension. To mitigate this phenomena somewhat, we consider an exponentially decreasing weight schedule when aggregating our different unbiased gradient estimator, and limit the number of estimators. The final estimate of the gradient is constructed as

$$\dot\theta_t = \frac{1-\beta}{1-\beta^{\tilde\tau_{\max}+1}}\sum_{\tau=0}^{\tilde k_{\max}}\beta^k\dot\theta_{t,\tau},$$

where $0 \leq \beta \leq 1$ and $\tau_{\max} = \min(\tau_{\max}, t)$.

The gradient descent is momentum accelerated, resulting in step $\mathrm{d}\theta_t = \mu \mathrm{d}\theta_{t-1} + (1-\mu)\eta\dot{\theta}_t$.

Algorithm 1 Variational PAC-Bayes (VarBUQ)

Require: $\theta_0 \in \Theta$, $\pi_{\mathsf{p}} \in \mathcal{P}$, $R \in \mathcal{M}(\mathcal{H})$, $K \in \mathbb{N}^*$, $\mu \in [0,1]$, $\beta \in [0,1]$, $\lambda \in \mathbb{R}_+$, $\eta \in \mathbb{R}_+$
  $t \leftarrow 0$, pb $\leftarrow \infty$, $\dot{\theta} \leftarrow 0$
  while not converged **do**
    $kl \leftarrow \mathsf{KL}(\pi_{\theta_t}, \pi_{\mathsf{p}})$
    $\delta_{\mathsf{KL}} \leftarrow \partial_\theta \mathsf{KL}(\pi_\theta, \pi_{\mathsf{p}})(\theta_t)$
    for $k \in [1, K]$ **do**
      $\gamma_{t,k} \sim \pi_{\theta_t}$
      $r_{t,k} \leftarrow R(\gamma_{t,k})$, $l_{t,k} = \log\left(\frac{\mathrm{d}\pi_{\theta_t}}{\mathrm{d}\pi_{\mathsf{p}}}(\gamma_{t,k})\right)$ ▷ Eval. new risk
    end for
    $\overline{r} \leftarrow K^{-1}\sum_{k=1}^K r_{t,k}$ ▷ Estim. post. mean risk
    $\hat{\sigma}_R \leftarrow \sqrt{(K-1)^{-1}\sum_{k=1}^K (r_{t,k} - \overline{r})^2}$ ▷ Estim. post. risk deviation
    if pb $< \overline{R} + \lambda kl + \mathsf{Tol} \times \hat{\sigma}_R$ **then** ▷ Step Removal procedure
      $t \leftarrow t-1$
      pb $\leftarrow \infty$, $\dot{\theta} \leftarrow 0$
    else
      pb $\leftarrow \overline{r} + \lambda kl$ ▷ Updt. Objective
      $\delta_R \leftarrow 0$, $\omega_{\mathsf{tot}} \leftarrow 0$
      for $\tau \in [\max(0, t - \tau_{\max}), t)]$ **do**
        $\delta_{R,\tau} \leftarrow \sum_{k=1}^K r_{\tau,k} \frac{\mathrm{d}\pi_{\theta_t}}{\mathrm{d}\pi_{\mathsf{p}}}(\gamma_{\tau,k}) \exp(-l_{\tau,k})\, \partial_{\theta_t}\left(\log(\frac{\mathrm{d}\pi_{\theta_t}}{\mathrm{d}\pi_{\mathsf{p}}}(\gamma_{\tau,k})\right)$ ▷ Gradient estimate
        $\delta_R \leftarrow \delta_R + \beta^{t-\tau}\delta_{R,\tau}$ ▷ Exp. Schedule
        $\omega_{\mathsf{tot}} \leftarrow \omega_{\mathsf{tot}} + \beta^{t-\tau}$
      end for
      $\delta_R \leftarrow \delta_R / (\omega_{\mathsf{tot}} \times K)$ ▷ Renormalise
      $\delta\theta \leftarrow \mu\delta\theta + (1-\mu)\eta\,(\delta_R + \lambda\delta_{\mathsf{KL}})$ ▷ Momentum
      $t \leftarrow t+1$ ▷ Update
      $\theta_t \leftarrow \theta_{t-1} + \delta\theta$
    end if
  end while

Step removal procedure

The sample $(\gamma_{i,t})_{i\in[1,K]}$ can be used to estimate the average empirical risk for distribution θ_t. In the small step ($\eta \ll 1$) and small sample size regime, the standard deviation of the average empirical risk can be significantly higher than the step difference of the objective $|\mathsf{PB}(\theta_t) - \mathsf{PB}(\theta_{t-1})|$, so that refusing all steps resulting in $\widehat{\mathsf{PB}}(\theta_t) > \widehat{\mathsf{PB}}(\theta t - 1)$ can harm the efficiency of the algorithm (*i.e.* about half of the steps might be wrongly discarded).

Still, it is possible to remove steps for which it is highly probable that the objective is increased, by checking whether $\widehat{\mathsf{PB}}(\theta_t) > \widehat{\mathsf{PB}}(\theta_{t-1}) + \mathsf{Tol} \times \sigma_t$ where σ_t is the empirical standard deviation of $(R(\gamma_{k,t}))_{k\in[1,K]}$ and Tol is the 90% quantile of the normal distribution.

If a step t is removed, the current distribution is set back to θ_{t-1}. We also break the momentum of the gradient descent, since it may reasonably be responsible for the problematic step size, and hence the speed is reset to $\dot{\theta}_t = 0$. The sample of predictors $(\gamma_{k,t})_{k\in[1,K]}$ is erased.

3.3.5 Warm start approach

As the prior belongs to the variational family considered for posterior optimisation, it can be used as a starting point for VarBUQ. This initialisation was used for the calibration of AM2. For ADM1 however, the optimisation process started from the prior struggled to concentrate around low risk predictors. The initialisation was therefore constructed using a warm start approach, which builds a distribution by greedily concentrating its mass on low risk predictors. This approach is based on the calibration strategy of Leurent and Moscoviz [2022].

The warm start algorithm constructs the initial posterior distribution through iteratively assessing predictors, and inferring a distribution from the best performing predictors. At each step, a fixed number K of samples are drawn from the current distribution, and the empirical risks evaluated. The M predictors achieving the lowest risks are selected amongst all predictors evaluated so far. A new distribution is constructed from these predictors. This distribution is the Gaussian with mean the selected predictors average, and a diagonal variance where variance $\sigma^2_{i,i}$ corresponds to the sample variance on each dimension. The mean and variance are computed for the log parameter values. The procedure stops when the objective (1) starts increasing[11]. This procedure is initialized using the prior.

3.4 Comparison of Uncertainty Quantification routines

Three UQ routines are considered to benchmark VarBUQ: FIM, Beale and Bootstrap. For Bootstrap, the implementation was based on Regueira et al. [2021]. A comparison of the three routines is provided in the supplementary material. Bootstrap was not evaluated for ADM1, due to excessive computation time.

Contrary to PAC-Bayes joint UQ and calibration, these UQ routines are carried out after model calibration. This calibration was performed by minimising the risk function. As these methods do not follow the Bayesian paradigm, they do not require constructing a prior distribution and therefore can be easier to implement.

[11] This approach should converge to a Dirac distribution, as this is the case in the setting where there is no signal (in that case, the sequence $\log(\sigma_n)$ is a random walk with negative bias)

Algorithm 2 Warm start approach for VarBUQ

Require: $(\sigma_0^i)_{i\leq d}$, $(\mu_0^i)_{i\leq d_\Gamma}$, R, λ, M, K
  $t \leftarrow 0$, $\bar{s} \leftarrow \infty$
  while not done **do**
    $\pi \leftarrow \mathcal{N}\left(\begin{pmatrix}\mu_t^1\\ \vdots \\ \mu_t^{d_\Gamma}\end{pmatrix}, \begin{pmatrix}\sigma_t^1 & \dots & 0\\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_t^{d_\Gamma}\end{pmatrix}\right)$
    $\gamma_{t,1}, \dots, \gamma_{t,M} \sim \pi$
    $\bar{r} \leftarrow \frac{1}{p}\sum_{i=1}^{M} R(\gamma_{t,i})$
    $\tilde{s} \leftarrow \bar{r} + \mathsf{KL}(\pi, \pi_0)$ ▷ Approximate PAC-Bayes objective
    if $\tilde{s} > \bar{s}$ **then**
      break
    else
      $t \leftarrow t+1$
      $\bar{s} \leftarrow \tilde{s}$
      $\tilde{\gamma}_1, \dots, \tilde{\gamma}_K \leftarrow \arg\inf_K (R(\gamma_{i,j}))_{i\in[1,M], j\in[1,t-1]}$ ▷ K best parameters γ
      $\mu_t \leftarrow \frac{1}{K}\sum_{k=1}^{K}\tilde{\gamma}_k$
      $\sigma_t^i \leftarrow \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(\tilde{\gamma}_k^i - \mu_k^i\right)^2}$
    end if
  end while

3.4.1 Calibration strategy

Sensitivity analysis

Selecting parameters to calibrate through sensitivity analysis can mitigate the risk of overfitting. The more parameters are selected, the smaller the empirical risk of the calibrated model will be. For AM2, the number of parameters being low (5), all parameters are calibrated. ADM1, involving 30 parameters in the implementation studied here, is often calibrated only on a subpart of its parameters.

A global sensitivity analysis was conducted in the case of ADM1 datasets to select parameters which were thereafter calibrated. The routine used was based on Morris method [Morris, 1991], and adjusted to account for multidimensional responses. Consistent with the calibration, the analysis was conducted on the logarithm of the parameters. The plausible ranges were inferred from the prior, using the intervals $[\log(\gamma^i) - 2\sigma^i, \log(\gamma^i) + 2\sigma^i]$ (the notation γ^i indicating the i-th component of the vector γ). These intervals were further discretized into 8 levels. From this grid of parameter values, chains of parameters (γ_d) are constructed as described in Morris method, such that (γ_d) and (γ_{d+1}) differ only in their d-th component, and their d-th components are neighbours in the grid. The initial parameter value is drawn uniformly at random on the grid. The impact of parameter component i is then assessed as

$$\text{sensitivity}_i = \sqrt{\frac{1}{N}\sum_{k,t}\text{softclip}\left(\log\left(\frac{\text{AD}_{k,t}(\gamma_{i+1})+\eta}{\text{AD}_{k,t}(\gamma_i)+\eta}\right)\right)^2}, \tag{3.6}$$

matching the formula used to compute the risk.

A total of 96 chains are constructed, and the sensitivity values for each component averaged over these draws. Parameters whose sensitivity value were above 0.025 were selected for calibration.

Calibration methodology

Calibration is achieved by minimising the risk (ERM). The same learning algorithm was used for both AD models. Since computing gradients for ADM1 is expensive, a gradient-free minimisation technique based on **C**ovariance **M**atrix **A**daptation **E**volution **S**trategy (CMA-ES) algorithm [Hansen, 2016] is used to perform the minimisation procedure[12]. The initial parameter values are the default ones for mesophilic digester specified by Rosén and Jeppsson [2006] for ADM1 and Bernard et al. [2001] for AM2 (see table 3.2), while the initial covariance matrix matches the covariance of the prior distributions used for VarBUQ. The optimisation procedure is performed in log space. The maximal values are limited to ten times the default value (*i.e.* γ^i is replaced by $\min(\gamma^i, \gamma_0^i)$).

For the main calibration procedure, the solver was deemed to have converged when the average decrease of the risk over the thirty last optimisation steps was below 10^{-8}, or when the covariance diagonal elements all satisfy $\sqrt{\Sigma_{i,i}} < 10^{-8}$. The maximum number of optimisation step was set to 250.

The calibrated parameter is denoted $\hat{\gamma}$ (with $\hat{\gamma}_{LN}$ denoting the calibrated parameter for dataset LN).

Code for the calibration algorithm can be found in 'picoptim' ('CMAOptimizer' class)[13].

3.4.2 Fisher's information matrix

FIM encodes the insight yielded by the observations on the parameters' values. A key motivation behind its use is Fréchet–Cramér–Rao's bound, also known as the information bound or Cramér–Rao's bound, which states that unbiased estimators necessarily have a variance greater than the inverse FIM, computed at the true set of parameters [Lehmann and Casella, 1998]. In linear regression with Gaussian noise setting, this lower bound is achieved when the number of observations exceeds the number of parameters.

Similar to the standard Bayesian framework, Fisher's information requires statistical modelling. From the computational model, the statistical model is constructed by assuming that the observation data is generated from the AD model, after undergoing log-Gaussian multiplicative

[12]The implementation corresponds to the 'CMAOptimizer' class in our picoptim package (`https://pypi.org/project/picoptim/`. An older, but similar, version was used to conduct the analysis, for which code can be found in the project's repo.

[13]The exact implementations used here are slightly older. They can be found in the article's Github project, in 'aduq' library ('optim_cma' function from 'aduq.optim' for general implementation, 'optim_cma_adm1' and respectively 'optim_cma_am2' from respectively 'aduq.pyadm1.optim' and 'aduq.pyam2.optim' for ADM1 (resp. AM2) specific implementations)

noise. As such, the statistical model is

$$\mathsf{Obs}_{t,\mathsf{obs}} = \mathsf{AD}_{t,\mathsf{obs}}(\gamma^*, \widetilde{\mathsf{Feed}}, \widetilde{\mathsf{State}}_0) \times \exp(\varepsilon_{t,\mathsf{obs}}), \;\; \varepsilon_{t,\mathsf{obs}} \sim \mathcal{N}(0, \sigma^2). \tag{3.7}$$

Note that this model differs in two aspects from the true statistical model from which the data was generated. First, Equation (3.7) assumes that the observation data is generated using the true feed signal and initial state Feed and State_0 rather than with their noisy counterparts. Second, the true structure of the noise is log uniform rather than log Gaussian.

For the statistical model (3.7), Fisher's information matrix is defined as

$$\mathrm{FIM}(\gamma^*) = \sigma^{-2} J_{\mathsf{AD}}(\gamma^*) J_{\mathsf{AD}}(\gamma^*)^T, \tag{3.8}$$

where $J_{\mathsf{AD}}(\gamma)_{i,j} = \frac{\partial \log(\mathsf{AD})_j}{\partial \gamma_i}$ is the Jacobian matrix of $\log \circ \mathsf{AD}$. Assuming the estimator $\hat{\gamma}$ is unbiased, Fréchet–Cramér–Rao's bound states that its covariance is lower bounded by

$$\mathbb{V}\left[\hat{\gamma}\right] \geq \mathrm{FIM}^{-1}(\gamma^*) = \sigma^2 \left(J_{\mathsf{AD}}(\gamma^*) J_{\mathsf{AD}}(\gamma^*)^T \right)^{-1}.$$

Since neither the exact γ^* used to generate the data, nor the noise level σ are assumed to be known, both are replaced with estimates. All in all, the uncertainty is encoded in the following covariance on the parameters

$$\hat{\Sigma}_{\mathsf{FIM}} := \hat{\sigma}^2 \left(J_{\mathsf{AD}}\left(\hat{\gamma}\right) J_{\mathsf{AD}}\left(\hat{\gamma}\right)^T \right)^{-1}. \tag{3.9}$$

The covariance was used to construct confidence regions shaped as ellipsoids,

$$\mathsf{CR}_{\mathrm{FIM}}\left(\alpha\right) := \left\{ \gamma \mid \left(\gamma - \hat{\gamma}\right) \hat{\Sigma}_{\mathsf{FIM}}^{-1} \left(\gamma - \hat{\gamma}\right) \leq Q_{\chi^2(d_\Gamma)}\left(\alpha\right) \right\},$$

where $Q_{\chi^2(d_\Gamma)}$ is the quantile function of a $\chi^2\left(d_\Gamma\right)$ distribution. If the estimator followed a Gaussian behaviour, as would be the case in linear regression, these confidence regions would hav coverage α.

For AM2, the derivative of the AD model with respect to the distribution's parameters was obtained through the resolution of an ODE, which can be efficiently performed while computing the predictions. Indeed, AM2 is fully described by an ODE, where the successive states $\mathsf{State}(t)$ of the digester satisfy

$$\begin{cases} \frac{\mathrm{d}\mathsf{State}}{\mathrm{d}t}(t) = \mathsf{AM2}(\gamma, \mathsf{Feed}_t, \mathsf{State}(t)) \\ \mathsf{State}(0) = \mathsf{State}_0. \end{cases}$$

As such, the derivative of the state with respect to the parameters satisfy the following differential equation

$$\begin{cases} \frac{\mathrm{d}^2\mathsf{State}}{\mathrm{d}t\mathrm{d}\gamma}(t,\gamma) = \partial_1 \mathsf{AM2}(\gamma, \mathsf{Feed}_t, \mathsf{State}(t)) + \partial_3 \mathsf{AM2} \; \frac{\mathrm{d}\mathsf{State}}{\mathrm{d}\gamma}(t,\gamma) \\ \frac{\mathrm{d}\mathsf{State}}{\mathrm{d}\gamma}(0,\gamma) = 0. \end{cases}$$

The second ODE can then be solved efficiently while solving the first ODE. Such a method can not be used directly for ADM1 since corrections are applied to the ODE results. As such, for ADM1, the derivative was computed using a two-points symmetric finite difference scheme. This second approach occasionally failed due to instabilities in the behaviour of the ODE solver, and the maximum time step of the ODE was manually adjusted until a suitable estimation of the derivative was found.

Equation (3.9) implies that $\hat{\Sigma}_{\text{FIM}}$ is non negative. However, since Fisher's information involves parameters whose typical values differ by several order of magnitude, the ratio between the highest and lowest eigenvalues (conditioning number) is usually large ($> 10^{16}$), and therefore, numerical approximations can result in estimated small, negative eigenvalues. As a precaution, all eigenvalues of FIM were raised to 10^{-8} at least. This implies that the covariance has all its eigenvalues upper bounded by 10^{8}.

While Cramér–Rao's bound is valid only for unbiased estimator, the ERM is generally a biased estimator. Moreover, Cramér–Rao's bound is a lower bound on the variance, which could be significantly larger. Finally, Cramér–Rao's bound is valid only if the statistical model is correct, which is not the case here. As such, the uncertainty quantification provided by FIM comes with no theoretical guarantee.

The implementation of FIM's procedure can be found in the project's repo (function 'fim' in 'aduq.uncertainty' module).

3.4.3 Beale's criteria

Beale's criteria (see Beale [1960] and Dochain and Vanrolleghem [2001], Section 6.7.3) describes uncertainty on the parameter values in the form of confidence regions which can have any shape. These are defined by considering all the sets of parameters whose risk are below a statistically motivated threshold. This threshold is constructed by considering the linear regression with Gaussian noise case. Considering as the risk function the sum of squared residuals R^2, and noting $\hat{\gamma}$ the minimiser of the risk, the statistic $\frac{N-d_\Gamma}{d_\Gamma}\frac{R^2(\gamma^*)-R^2(\hat{\gamma})}{R^2(\hat{\gamma})}$ follows an $F(d_\Gamma, N-d_\Gamma)$ distribution. From this threshold, we construct the confidence region

$$\text{CR}_{\text{Beale}}(\alpha) := \left\{\gamma \mid R^2(\gamma) \leq T_\alpha := R^2(\hat{\gamma})\left(1 + \frac{d_\Gamma}{N-d_\Gamma} Q_{F(d_\Gamma, N-d_\Gamma)}(\alpha)\right)\right\} \tag{3.10}$$

where $Q_{F(d_\Gamma, N-d_\Gamma)}$ is the quantile function of the $F(d_\Gamma, N-d_\Gamma)$ distribution, is a confidence region of level α. Beale's UQ consists in using these confidence regions, even in non linear setting. The overfitting ability of non linear model might be much more important than in the linear case. As such, the resulting $R^2(\hat{\gamma})/R^2(\gamma^*)$ might be much smaller than in the linear case, resulting in overconfidence. Hence Beale's method does not come with theoretical guarantees.

In order to estimate the confidence region, the line search procedure advocated in Dochain and Vanrolleghem [2001] was implemented and performed on the log parameters (see fig. 3.4). The covariance obtained through the inverse FIM was used to obtain an initial guess of the uncertainty region. To avoid numerical aberrations, parameters whose uncertainty in log was

larger than 6 were set to the optimised value, and uncertainty was not ascertained on them. An initial sample of points $(\gamma_i)_i$ on the boundary of FIM's confidence region was drawn at random (2048 points for ADM1, 5120 points for AM2). A line search algorithm was used to solve $R^2(x(\gamma_i - \hat{\gamma}) + \hat{\gamma}) = T_\alpha$ for each γ_i. If the line search algorithm failed to find a parameter achieving the threshold with a precision of less than $0.01 \times (T_\alpha - R^2(\hat{\gamma}))$ in less than 20 steps, the point was removed from the set. The final set $(x_i(\gamma_i - \hat{\gamma}) + \hat{\gamma})$ was used as the approximative boundary of Beale's confidence region.

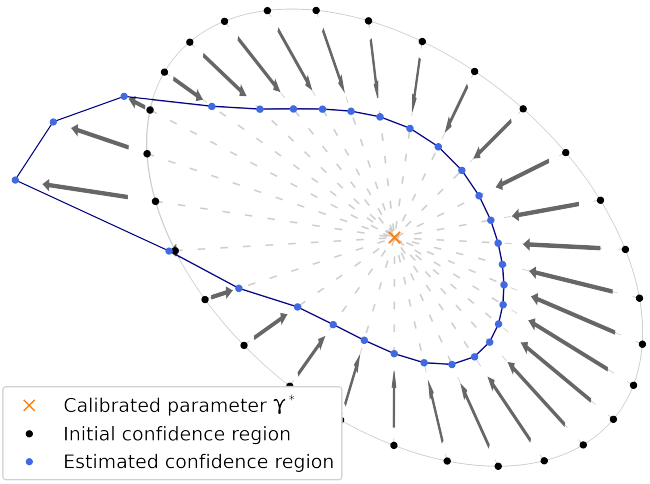


Figure 3.4: Estimation of Beale's uncertainty quantification in a two dimensional setting. The ellipse corresponding to Fisher's information matrix confidence region boundary is used as an initial guess of Beale's confidence region boundary. For each point, a line search procedure is performed to find Beale's criteria.

The line search procedure might fail to adequately represent the theoretical confidence region for Beale's procedure. First of all, the line search procedure assumes that the confidence region is connected, which might not be the case (for instance, if they are multiple local minima of similar minimum level). More generally, the line search procedure will fail to give adequate results if all functions $x \mapsto R(x\delta\gamma + \hat{\gamma})$ are not increasing on $\mathbb{R}_+$ for all $\delta\gamma$. In such cases, the result of the line search will depend on the initial value. We represent in fig. 3.5 some cases in which the approximation of Beale's confidence region fails.

The implementation of Beale's routine is available in the project's repo (function 'beale_boundary' in 'aduq.uncertainty' module).

3.4.4 Residual bootstrapping

Residual bootstrapping is designed to perform UQ when the noise structure is unknown. After calibration, the residuals are assumed to be representative of the noise, and samples of the noise are used to generate new observations. These are in turn transformed into samples

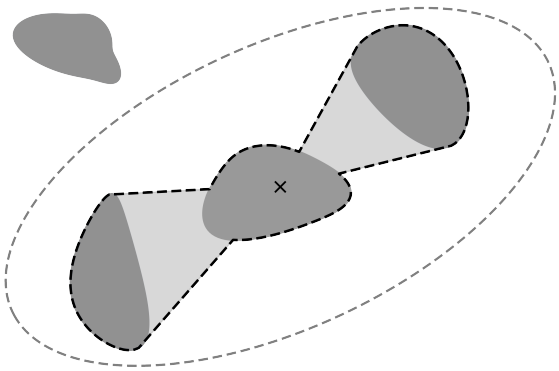


Figure 3.5: Issues of the line search procedure estimating Beale's confidence regions. The dark grey zones describe the theoretical confidence region, while the dotted line represent the approximated confidence region. The ellipse represents the initial confidence region found through Fisher's information method. Four disconnected regions define the theoretical regions, only three of which are encompassed in the approximation. Light gray area, which do not belong to the theoretical region, are added in the approximation.

of sets of parameters through recalibration. As such, the residual bootstrapping procedure describes uncertainty in the form of a sample of sets of parameters.

The bootstrap procedure used is similar to the one used in Gonzalez-Gil et al. [2018] and used in the context of AD modelling by Regueira et al. [2021]. Log-residuals of the calibrated model are drawn with replacement and added to the log of the calibrated model output to generate new observations. New calibration procedures with these bootstrapped observations are then performed to compute new sets of parameters. 512 bootstrapped sets of parameters were constructed for each dataset.

In order to limit the computational cost of the procedure, the new calibration procedures were restarted from the result of the first calibration procedure, with less stringent convergence criteria (tolerance in risk of 2e-4 compared to e-8 for the initial calibration procedure, and a maximum of 30 optimisation steps compared to 250). While this decreases the ability of the bootstrap procedure to construct sets of parameters far from the calibrated model, it proved necessary to control the computational cost.

This procedure was assessed only on the AM2 model, since its computational cost would have been prohibitive for ADM1 (see table 3.9).

The implementation of the bootstrap procedure is available in the project's repo ('bootstrap' in 'aduq.uncertainty' module).

3.4.5 Assessment of uncertainty quantification on parameter values

AD model parameters describe quantities which have a physical or biological interpretation and inform on properties of the AD process. As such, the uncertainty on the calibrated parameter values is an important consideration.

Table 3.3: Overview of Uncertainty Quantification routines

	Bootstrap	FIM	Beale	VarBUQ
Confidence region shape	Any	Ellipse	Any*	User chosen**
Output	Samples	Covariance	Samples	Distribution
Prerequisite	None	Statistical Model	Mean squared error calibration	Prior
Theory	None	Cramer-Rao	Beale	PAC-Bayes
Error hypothesis	None	Gaussian***	Gaussian	Bounded
Model hypothesis	None	Linear***	Linear	None

* The algorithm approximating Beale's confidence regions assumes they are connex.
** Through probability distribution class.
*** While Cramer-Rao's result holds for any statistical model, it only provides an equality for Gaussian noise and linear model.

Each UQ method is assessed through computation of p-values for tests of the hypothesis that the true set of parameters is γ^*. If the UQ method performs as it should, these p-values should be uniformly distributed between 0 and 1. Small p-values indicate that the UQ is over-confident, as the true set of parameters would be rejected, while large p-values indicate that the UQ is under confident.

For FIM UQ, the p-value is computed using the statistic $(\hat{\gamma}-\gamma^*)\hat{\Sigma}_{\text{FIM}}^{-1}(\hat{\gamma}-\gamma^*)$, which, under the null hypothesis in the linear regression setting, is distributed as a $\chi^2(d_\Gamma)$ where d_Γ is the number of parameters selected by the sensitivity analysis. Since VarBUQ considers Gaussian posteriors on the log parameters, the similar statistic $(\mu_\theta - \log(\gamma^*))\Sigma_\theta^{-1}(\mu_\theta - \log(\gamma^*))$ was used, where μ_θ and Σ_θ are the mean and covariance in log-space of $\hat{\pi}$ (with $d_\Gamma = 5$ for AM2 and 30 for ADM1).

Contrary to Beale and Bootstrap methods, FIM UQ depends on the choice of parametrisation[14] (see fig. 3.6). Notably, the confidence region provided are ellipsoids in the parametrisation considered. The choice of parametrisation can act as inductive bias; for instance, the log parametrisation used for the VarBUQ encodes the knowledge that the parameter values must be positive. We thus also tested the performance of FIM UQ performed on the log parameter[15].

For Beale's criteria, the p-values are computed using the theoretical rather than estimated confidence regions. As such, the key statistic is $\frac{N-d_\Gamma}{d_\Gamma}(R^2(\gamma^*)-R^2(\hat{\gamma}))/R^2(\hat{\gamma})$ which, under the null hypothesis, should be distributed as $F(d_\Gamma, N-d_\Gamma)$.

For the bootstrap procedure, the computation of the p-value is slightly more involved. An anomaly score function A is learned on the first half of the sample using Isolation Forest algorithm [Liu et al., 2008], then evaluated on the second half of the bootstrapped sample, resulting in a sample of anomaly scores $A(\gamma_i)_{i>n/2}$. The p-value is obtained by considering the quantile achieved by the anomaly score of the true set of parameters, $A(\hat{\gamma})$, i.e $\hat{p} = \frac{2}{n}\sum_{i>\frac{n}{2}} 1\{A(\hat{\gamma}) < A(\gamma_i)\}$. For a sufficiently large sample, the central limit theorem implies that this quantity does converge to the p-value for a family of tests assessing whether a parameter is drawn from the bootstrapped sample. Tight high confidence upper bounds for the p-value given infinite sample

[14] Note that among the assumptions for Cramer–Rao's bound, the estimator must be unbiased. This property depends on the parametrisation considered

[15] Since the uncertainty on the predictions is ascertained using linear approximation for FIM, note that FIM UQ on predictions is parametrisation independent.

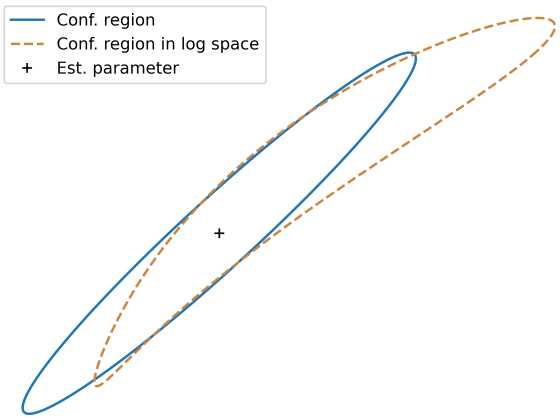


Figure 3.6: Impact of the parametrisation on the confidence regions provided through Fisher's information. Using the standard parametrisation, FIM constructs a confidence region in the shape of an ellipse (blue curve). Using a log parametrisation, FIM constructs a confidence region which is an ellipse in log space, but has a different form in the standard space (orange dotted curve).

size can be computed by noting that $A(\hat{\gamma})$ is a mean of independent Bernoulli of mean p, and that

$$\left[0, \hat{p}^+ := \sup\left\{\tilde{p} \mid \sum_{i=0}^{n\hat{p}} \binom{n}{i} \tilde{p}^i (1-\tilde{p})^{n-i} \geq \alpha\right\}\right]$$

is a confidence interval of p of level α, *i.e.* $p \leq \hat{p}^+$ with probability α.

Code for the computation of the p-values can be found in the project's repo (functions 'fim_pval', 'beale_pval', 'sample_pval'. 'sample_pval' is used for Bootstrap. Note that 'fim_pval' function can also be used to compute the p-value for VarBUQ, using as arguments the log parameter values, mean of the log parameter and covariance computed in log space).

3.4.6 Assessment of uncertainty quantification on predictions

The performance of each UQ method is also assessed on predictions using the test set (84 days). The uncertainty on the prediction is obtained by transferring the uncertainty on the parameter, through linear uncertainty transport for FIM, and through the evaluation of multiple sets of parameters for all remaining methods. Pseudo 95% **C**onfidence **I**ntervals (CIs) were then constructed for each prediction by considering quantiles, and their ability to cover the clean signal is assessed. Predictions are regrouped as gas flows ($q_{gas,M}$, $q_{gas,C}$ for AM2, q_{gas}, $q_{gas,CH4}$ for ADM1) and soluble compounds (S_1, S_2 for AM2, the four main VFAs, S_{bu}, S_{va}, S_{ac}, S_{pro} for ADM1) to assess quality.

For each group of predictions, four indicators are computed:

- The coverage of the CIs, *i.e.* the fraction of predictions inside the CIs,
- The width of the CIs,

- The prediction error of the calibrated model,
- The residual error of the CIs.

The residual error of the CI is computed by replacing the standard residuals by the distance between the ground truth and the CI. Notably, if the confidence interval completely covered the truth, the residual error of the CI would be 0.

To limit the risk of underestimating the uncertainty in the early phase of the test set, all calls to the AD models are started from day 166, 30 days before the beginning of the test set, using as initial description of the digester the result of the simulation at that day for the calibrated set of parameters (the optimised set of parameters for residual based UQ methods and the mean of the posterior for VarBUQ).

From a statistical viewpoint, the methodology used to extrapolate uncertainty from confidence region constructs confidence intervals on the predictions with matching level. However, the transfer of uncertainty guarantees that all of the clean signal should be inside the confidence intervals with probability at least α, and not that a fraction α of the clean signal should be in the confidence intervals. As such, it can not be assumed that 95% confidence regions should result in 95% coverage for a good UQ procedure - all the more so as neither the UQ methods hypothesis are met, nor is the statistical model supporting them correct. Still, since confidence regions with high confidence levels are assessed, the coverage indicator of properly working UQ methods should be high, the target coverage is set informally at 95%.

3.5 Experimental results

3.5.1 Calibration results on training set

For ADM1, the global sensitivity analysis selected from 9 to 14 parameters depending on the datasets. Only 14 parameters were at least selected once ($K_{S,\,c4+}$, $k_{m,\,c4+}$, $K_{S,\,ac}$, $k_{m,\,ac}$, $K_{S,\,pro}$, $k_{m,\,pro}$, $k_{m,\,aa}$, $k_{m,\,su}$, k_{dec}, $K_{I,\,NH3}$, $pH_{UL:LL,\,aa}$, $pH_{LL,\,aa}$, $pH_{UL:LL,\,ac}$, $pH_{LL,\,ac}$ in the original paper). Details on parameters selected for calibration for each dataset can be found in table 3.4.

The calibration algorithms were performed using the hyperparameters specified in Table 3.6. The calibrated model obtained by standard and PAC-Bayes calibration are specified in Table 3.5. For the PAC-Bayes calibration, only the mean parameter in log-space is tabulated.

Once calibrated, the models obtained train risks of about 0.09 for AM2 and 0.095 for ADM1 (see table 3.7 for detailed values). This is slightly above the contribution of the noise on the observations, and implies that the noise on the influent did increase the overall noise on observations. As expected, the optimisation-based calibration routine succeeded in finding sets of parameters achieving a lower empirical risk than the one obtained using the true set of parameters. The mean empirical risk for VarBUQ is slightly above the empirical risk of the true set of parameters for all datasets at the reference PAC-Bayes temperature (about 0.005 higher for both AM2 and ADM1, implying an absolute increase of 0.5% to the relative error). Doubling

Table 3.4: Parameter selection through Morris sensitivity analysis for ADM1.

	LN	HN	LF	HF
$k_{m,\,su}$		✓		✓
$k_{m,\,aa}$		✓		✓
$k_{m,\,c4+}$	✓	✓	✓	✓
$k_{m,\,pro}$	✓	✓	✓	✓
$k_{m,\,ac}$	✓	✓	✓	✓
k_{dec}	✓	✓	✓	✓
$K_{S,\,c4+}$	✓	✓	✓	✓
$K_{S,\,pro}$	✓	✓	✓	✓
$K_{S,\,ac}$	✓	✓	✓	✓
$K_{I,NH3}$		✓		✓
$pH_{LL\ aa}$	✓	✓	✓	✓
$pH_{UL:LL\ aa}$	✓	✓	✓	✓
$pH_{LL\ ac}$		✓		✓
$pH_{UL:LL\ ac}$		✓		

the PAC-Bayes temperature had moderate effect on the train performance of the posterior, with an increase in the risk of about 0.005. Increasing the PAC-Bayes temperature to eight time its reference value had a more noticeable effect, with a mean risk up to 0.024 higher.

3.5.2 Uncertainty on parameters values

The capacity of the UQ methods to capture the true set of parameters was assessed by computing p-values for tests indicating whether the true set of parameters belonged to the confidence regions. These p-values are tabulated in table 3.8.

Ad-hoc confidence regions constructed after standard calibration could generally not account for the large deviations between the true set of parameters and optimised set of parameters for ADM1. This results in FIM's confidence regions systematically failing to cover the true set of parameters for ADM1, where deviations are particularly noticeable for $\mathrm{k_m}$, $\mathrm{K_S}$ couples. This finding remains mostly valid in the case of AM2, since the confidence level must be chosen above the standard 95% criteria in order to cover the true set of parameters with FIM and Bootstrap confidence regions. The results of Beale's method are of particular interest. As the p-values were constructed using the theoretical criteria rather than any approximation, its failure to encompass the true set of parameters directly implies that the non-linearity in the AD models offers opportunities to reduce the noise significantly more than a linear model. Half of the p-values obtained were orders of magnitude lower than the 0.05 threshold considered, being on two occasions equal to the machine precision (2.2e-16). On the remaining datasets, only one p-value was above the threshold (0.3), while the three others were of order 1e-3.

Confidence regions constructed with Bootstrap failed to cover the true parameter for any confidence level in three of four cases. Correcting for the number of bootstrap samples generated, the 95% confidence upper bound on the p-values was above the standard 0.05 threshold for only one dataset out of four. This could be related to specific implementation choices designed to mitigate the computation time (*e.g.* restarting calibration from the previously calibrated para-

Table 3.5: Calibrated parameters

	γ^*_{N}	$\hat{\gamma}_{\mathrm{LN}}$	$\hat{\pi}_{\mathrm{LN}}$	$\hat{\gamma}_{\mathrm{HN}}$	$\hat{\pi}_{\mathrm{HN}}$	γ^*_{F}	$\hat{\gamma}_{\mathrm{LF}}$	$\hat{\pi}_{\mathrm{LF}}$	$\hat{\gamma}_{\mathrm{HF}}$	$\hat{\pi}_{\mathrm{HF}}$
k_{dis}	0.355	0.500	0.575	0.500	0.601	0.233	0.500	0.436	0.500	0.493
$k_{hyd,\ CH}$	10.0	10.0	10.5	10.0	9.90	5.19	10.0	9.74	10.0	9.34
$k_{hyd,\ PR}$	8.65	10.0	9.13	10.0	9.99	11.9	10.0	9.91	10.0	8.83
$k_{hyd,\ LI}$	23.4	10.0	9.54	10.0	10.3	2.40	10.0	10.4	10.0	9.41
$k_{m,\ su}$	37.7	30.0	32.0	2.42	30.4	47.7	30.0	32.0	107	27.5
$k_{m,\ aa}$	44.6	50.0	47.7	49.5	55.6	48.7	50.0	50.6	2.23	51.3
$k_{m,\ fa}$	5.23	6.00	5.39	6.00	6.53	4.47	6.00	7.10	6.00	5.45
$k_{m,\ c4+}$	21.3	89.6	18.1	123	18.8	17.6	36.8	17.1	44.0	16.3
$k_{m,\ pro}$	12.7	92.9	11.9	26.2	12.6	9.91	49.4	14.3	9.35	14.9
$k_{m,\ ac}$	10.7	80.0	10.3	20.7	10.6	10.2	53.1	8.39	9.29	9.75
$k_{m,\ h2}$	27.9	35.0	33.2	35.0	32.7	28.7	35.0	36.9	35.0	35.0
k_{dec}	1.57e-2	1.99e-2	1.84e-2	6.16e-3	1.66e-2	1.45e-2	1.24e-2	1.65e-2	2.77e-2	2.08e-2
$K_{S,\ IN}$	9.31e-5	1.00e-4	9.96e-5	1.00e-4	9.93e-5	9.46e-5	1.00e-4	1.01e-4	1.00e-4	9.87e-5
$K_{S,\ su}$	0.427	0.500	0.499	0.500	0.543	0.293	0.500	0.469	0.500	0.467
$K_{S,\ aa}$	0.309	0.300	0.314	0.300	0.307	0.287	0.300	0.304	0.300	0.303
$K_{S,\ fa}$	0.389	0.400	0.441	0.400	0.402	0.579	0.400	0.419	0.400	0.427
$K_{S,\ c4+}$	0.300	1.18	0.236	2.00	0.256	0.399	0.898	0.372	0.939	0.350
$K_{S,\ pro}$	0.115	0.857	0.100	0.303	0.115	6.48e-2	0.380	9.39e-2	5.07e-2	0.103
$K_{S,\ ac}$	0.147	1.21	0.129	0.210	0.133	0.243	1.50	0.153	0.227	0.194
$K_{S,\ h2}$	6.65e-6	7.00e-6	7.64e-6	7.00e-6	6.11e-6	6.12e-6	7.00e-6	7.49e-6	7.00e-6	7.33e-6
$K_{I,\ H2,\ fa}$	5.01e-6	5.00e-6	4.93e-6	5.00e-6	5.05e-6	5.60e-6	5.00e-6	5.15e-6	5.00e-6	4.99e-6
$K_{I,\ H2,\ c4+}$	1.06e-5	1.00e-5	1.00e-5	1.00e-5	9.98e-6	1.11e-5	1.00e-5	1.04e-5	1.00e-5	9.85e-6
$K_{I,\ H2,\ pro}$	3.15e-6	3.50e-6	3.49e-6	3.50e-6	3.55e-6	3.37e-6	3.50e-6	3.66e-6	3.50e-6	3.55e-6
$K_{I,NH3}$	1.88e-3	1.80e-3	1.91e-3	7.88e-4	1.75e-3	2.05e-3	1.80e-3	1.75e-3	2.91e-3	2.09e-3
$pH_{UL:LL\ aa}$	1.16	1.08	1.55	1.26	1.33	1.36	0.311	1.45	1.37	1.47
$pH_{LL\ aa}$	5.11	5.18	3.61	6.00	3.48	4.43	2.63	3.80	5.21	3.64
$pH_{UL:LL\ ac}$	1.02	1.00	1.01	0.826	0.986	1.00	1.00	0.988	1.00	1.01
$pH_{LL\ ac}$	4.53	5.00	4.98	5.33	4.90	4.76	5.00	4.97	3.80	4.93
$pH_{UL:LL\ h2}$	0.524	1.00	1.03	1.00	1.05	1.61	1.00	0.941	1.00	0.966
$pH_{LL\ h2}$	5.50	5.00	4.99	5.00	5.02	6.05	5.00	4.94	5.00	4.80

(a) ADM1

	γ^*_{N}	$\hat{\gamma}_{\mathrm{LN}}$	$\hat{\pi}_{\mathrm{LN}}$	$\hat{\gamma}_{\mathrm{HN}}$	$\hat{\pi}_{\mathrm{HN}}$	γ^*_{F}	$\hat{\gamma}_{\mathrm{LF}}$	$\hat{\pi}_{\mathrm{LF}}$	$\hat{\gamma}_{\mathrm{HF}}$	$\hat{\pi}_{\mathrm{HF}}$
$\mu_{1\mathrm{max}}$	1.23	0.55	1.17	0.91	1.18	0.99	0.86	1.02	0.91	0.86
$\mu_{2\mathrm{max}}$	1.05	0.99	0.74	0.90	0.78	1.32	2.00	0.95	1.21	1.16
K_{S_1}	7.51	5.76	7.34	6.25	7.20	15.98	15.38	15.98	15.13	14.40
K_{S_2}	11.0	11.0	9.84	10.1	9.09	4.97	5.36	4.5	4.85	4.68
K_{I2}	196.0	135	256	259	259	345	328	252	290	260

(b) AM2

γ^* stands for the true sets of parameters used to generate the data. $\hat{\gamma}$ stands for the result of the standard calibration. $\hat{\pi}\left[\gamma\right]$ stands for the mean parameter found by VarBUQ using the chosen a priori PAC-Bayes temperature in the log space, after application of an exponential transform (*i.e.* $\exp\left(\hat{\pi}\left[\log(\gamma)\right]\right)$). For ADM1, Monod half saturation constants $\mathrm{K_S}$ and inhibition constants $\mathrm{K_I}$ are in kgCOD m^{-3} except $\mathrm{K_{S,\ IN}}$ and $\mathrm{K_{I,NH3}}$ which are in kMole m^{-3}. Maximum uptake rates $\mathrm{k_m}$ are in day^{-1}. Parameters for pH upper and lower limit are comparable to pH. For AM2, the maximum growth rates μ_{max} are in day^{-1}. The Monod half saturation constant K_{S_1} is in kgCOD m^{-3}, while the inhibition constants for S_2 are in Mole m^{-3}.

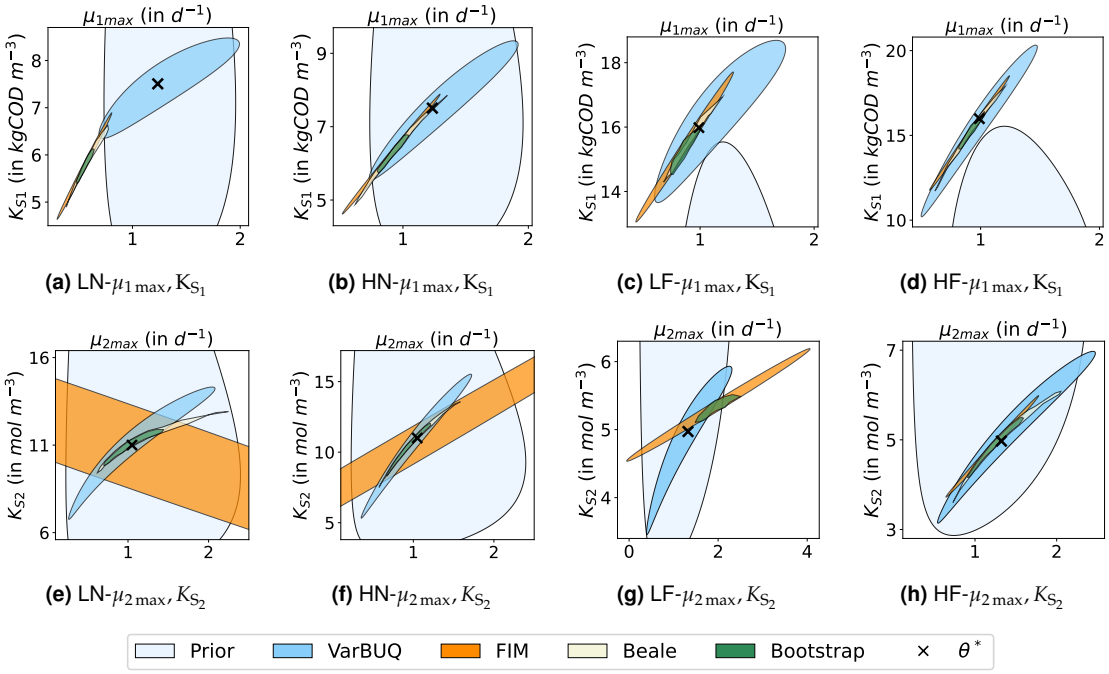
Table 3.6: Hyperparameters for calibration routines.

	step_size	PAC-Bayes. temp.	corr_eta	gen_decay	momentum	per_step	k
ADM1	0.025	0.001	0.7	0.20	0.985	160	8000
AM2	0.040	0.002	0.7	0.15	0.960	256	20000

(a) PAC-Bayes calibration

	per_step	radius_ini	radius_factor	no_change_max	keep_frac	cov_updt_speed
ADM1	96	0.2	0.7	96	0.3	0.06
AM2	480	0.1	0.7	1024	0.3	0.08

(b) Standard calibration



(a) LN-$\mu_{1\,\max}, \mathrm{K}_{\mathrm{S}_1}$ **(b)** HN-$\mu_{1\,\max}, \mathrm{K}_{\mathrm{S}_1}$ **(c)** LF-$\mu_{1\,\max}, \mathrm{K}_{\mathrm{S}_1}$ **(d)** HF-$\mu_{1\,\max}, \mathrm{K}_{\mathrm{S}_1}$

(e) LN-$\mu_{2\,\max}, K_{\mathrm{S}_2}$ **(f)** HN-$\mu_{2\,\max}, K_{\mathrm{S}_2}$ **(g)** LF-$\mu_{2\,\max}, K_{\mathrm{S}_2}$ **(h)** HF-$\mu_{2\,\max}, K_{\mathrm{S}_2}$

Figure 3.7: 95% confidence regions for AM2. The prior distribution is in light blue, the posterior in blue, FIM in orange, Beale in beige and Bootstrap in green. The true parameter is represented by a black cross. VarBUQ was the only methodology able to recapture the true sets of parameters in all settings. The methodology can benefit from the prior's inductive bias (figures 3.7a, 3.7b, 3.7g), but is also able to adapt to cases where the parameter is outside the boundary of the prior's confidence region (figures 3.7c, 3.7d). Those confidence regions are shaped as ellipses in log-space. The ellipsoidal confidence regions obtained through FIM tend to encompass those constructed through Beale's or Bootstrap methods. They suffer from some instability as exhibited in figures 3.7e, 3.7f, encompassing negative values. The confidence regions obtained through Beale's method tend to encompass those constructed through the Bootstrap method. Both methods regions with similar curvatures and an overall direction coherent with FIM's confidence regions, except for figure 3.7e. All UQ methods responded to the limited identifiability of the maximum growth rate and Monod constant (*i.e.* the fact that those parameters can compensate for one another) by constructing confidence regions which are squeezed along an axis.

Table 3.7: Summary of standard and PAC-Bayes calibration results

	$R(\gamma^*)$	$R(\hat{\gamma})$	$\hat{\pi}_\lambda[R]$	$\mathrm{KL}(\hat{\pi}_\lambda, \pi)$	$\hat{\pi}_{2\lambda}[R]$	$\mathrm{KL}(\hat{\pi}_{2\lambda}, \pi)$	$\hat{\pi}_{8\lambda}[R]$	$\mathrm{KL}(\hat{\pi}_{8\lambda}, \pi)$
AM2 LN	0.09045	0.08922	0.09401	4.423	n.a.	n.a.	n.a.	n.a.
AM2 HN	0.09297	0.09239	0.09657	4.459	n.a.	n.a.	n.a.	n.a.
AM2 LF	0.08971	0.08957	0.09296	9.152	n.a.	n.a.	n.a.	n.a.
AM2 HF	0.08640	0.08583	0.08942	9.936	n.a.	n.a.	n.a.	n.a.
ADM1 LN	0.09902	0.09711	0.10089	13.18	0.10289	11.02	0.11589	6.02
ADM1 HN	0.10106	0.09496	0.10712	11.90	0.10833	10.35	0.12426	5.78
ADM1 LF	0.09313	0.09241	0.09687	13.99	0.10112	10.47	0.11368	6.24
ADM1 HF	0.10449	0.09682	0.11031	16.50	0.11362	13.67	0.13391	6.57

$R(\gamma^*)$ is the empirical risk obtained by the true set of parameters, $R(\hat{\gamma})$ the empirical risk obtained by the set of parameters obtained by the standard calibration. $\hat{\pi}_{k\lambda}[R]$ is the average empirical risk of the posterior obtained by the PAC-Bayes calibration with PAC-Bayes temperature $k\lambda$, where λ is the PAC-Bayes temperature chosen a priori. $\mathrm{KL}(\hat{\pi}_{k\lambda}, \pi)$ is the Kullback–Leibler divergence between the posterior obtained by the PAC-Bayes calibration with PAC-Bayes temperature $k\lambda$ and the prior.

Table 3.8: Assessment of UQ methods for parameters estimation

	Bootstrap	FIM	log-FIM	Beale	VarBUQ (λ)	VarBUQ (2λ)	VarBUQ (8λ)
AM2 LN	0.0 ($<$ 1.2e-2)	0.0	0.0	5.0e-9	**0.91**	n.a.	n.a.
AM2 HN	0.0 ($<$ 1.2e-2)	3.9e-4	**0.085**	5.6e-4	**0.87**	n.a.	n.a.
AM2 LF	0.0 ($<$ 1.2e-2)	1.9e-2	**0.30**	**0.34**	**0.90**	n.a.	n.a.
AM2 HF	3.1e-2 ($<$ **5.6e-2**)	2.5e-2	**0.052**	3.8e-4	**0.91**	n.a.	n.a.
ADM1 LN	n.a.	0.0	0.0	3.3e-11	2.5e-6	5.2e-4	**0.56**
ADM1 HN	n.a.	0.0	0.0	0.0	**0.22**	**0.48**	**0.89**
ADM1 LF	n.a.	0.0	0.0	1.1e-3	0.0	7.1e-9	3.7e-3
ADM1 HF	n.a.	0.0	0.0	0.0	0.0	2.5e-8	4.1e-3

p-values in bold imply that the true set of parameters was inside the 95% confidence region. For the Bootstrap method, the upper bound given in parenthesis is valid with probability at least 0.95. Since datasets VN (resp. VF) share its training data and true parameter with dataset LN (resp. LF), the performance of the uncertainty quantification routines are identical.

meter, looser convergence tolerance). Bootstrap methods are by construction computationally intensive, requiring multiple model calibrations, which in the context of computational intensive model such as those used in AD might be prohibitive. The computational cost of the bootstrap routine could be improved by considering different calibration techniques or laxer termination criteria. However, no satisfactory trade off between performance and computational cost was found during the present study.

The change of parametrisation for Fisher's information to the log parameter remarkably improved the performance of the uncertainty quantification for AM2 datasets, with the 3 out of 4 p-values being above the 0.05 threshold. This highlights the importance of the parametrisation for Fisher's information. This log-FIM approach, which has, as far as we are aware, not been used in the field of AD modelling, provided the best UQ amongst the non Bayesian strategies for parameter estimation, and could be used as a drop in replacement for the more established FIM approach. However, the log-FIM approach still failed to obtain adequate uncertainty quantification for ADM1, with no p-value higher than the machine precision.

Of all UQ methods, VarBUQ gave the best results for parameter recovery. For ADM1, the results remained unsatisfactory. Using the reference PAC-Bayes temperature, only one p-value was above the threshold (compared to none for all remaining methods), while two of them were

equal to the machine epsilon, and the last one of order 1e-6. This was improved by doubling the PAC-Bayes temperature, which brought little train performance loss (see table 3.7), though there was still only a single p-value above the threshold. p-values were further increased by raising the PAC-Bayes temperature to eight times the reference, at the cost of noticeable decrease on train performances. In that last setting, two p-values were above the threshold, while the two remaining ones are of order 4e-3. For AM2, VarBUQ with reference PAC-Bayes temperature obtained satisfactory performance, with p-values all of order 0.9, both for L and F datasets. For the latter ones, the prior would obtain p-values of 0.05, implying that the posterior did more than inherit the induction bias.

Plots of the confidence regions constructed through each UQ method for the AM2 model (Figure 3.7) yield qualitative insight on their performances. Representations of the confidence regions are obtained by projecting it on two dimensions. For FIM and VarBUQ, exact confidence regions could be computed using 95% confidence ellipses for two dimensional Gaussians. For Bootstrap and Beale, the two dimensional confidence regions were approximated by projecting the sample and using alpha shapes to obtain a boundary [Edelsbrunner, 2011]. For the bootstrap method, 5% of the sample with highest anomaly score - as computed using Isolation Forest - are removed before constructing the boundary in order to obtain an approximative 95% confidence region. The implementations for the representation of the confidence regions can be found in the project's repo (module 'aduq.uncertainty.plot').

VarBUQ benefits from inductive bias as exhibited in figs. 3.7a, 3.7b and 3.7e to 3.7h where the confidence region constructed using the posterior remains almost entirely in the confidence region constructed using the prior. Still, VarBUQ performed satisfactorily in settings where the true set of parameters is on the boundary of the prior's confidence regions (Figures 3.7c and 3.7d). The remaining UQ methods are almost ordered, with FIM's confidence region nearly encompassing Beale's, which in turn encompasses those constructed by the Bootstrap method. While this seems incoherent with the p-values obtained in table 3.8, this could be explained by the additional approximation step required to construct Beale's confidence regions. While all UQ method indicate strong correlation between maximum growth rate and Monod constant, the exact form of the confidence regions differs. By construction, FIM's confidence regions are ellipsoidal, while VarBUQ's confidence regions are ellipsoidal in log-space. Interestingly, this second shape-constraint seems better suited to describe the relationship between the two parameters, since both Beale and Bootstrap, which outputs confidence regions with no shape constraints, obtain a somewhat similar curvature (see fig. 3.7b, 3.7c, 3.7c, 3.7d, 3.7e, 3.7f, 3.7h). This is also coherent with the improved p-values obtained for the log-FIM approach. Since FIM's confidence regions are constructed by extrapolating a local linear approximation, they can include nonphysical parameter values (*i.e.* negative values), as is the case in fig. 3.7e and 3.7f, and to a lesser degree in fig. 3.7g.

Overall, Figure 3.7 highlights two factors which contribute to VarBUQ's superior performance in comparison to the ad-hoc UQ methods for the AM2 datasets. First, the confidence region it constructed tend to be larger than those constructed using other UQ methods. Second,

VarBUQ's confidence regions are also better centered around the true parameter values, implying that the PAC-Bayes procedure offered a better calibration than the standard calibration procedure. This second feature can be attributed to the inductive bias brought by the prior: out of two sets of parameters yielding similar outputs, VarBUQ will favour the one deemed most likely by the experts (i.e., encoded in the prior), even if slightly less performant.

3.5.3 Uncertainty on prediction values

As complex AD models such as ADM1 are known to have identifiability issues, assessing the performance of the UQ on the parameter is not sufficient. Indeed, since different sets of parameters may still result in similar predictions, confidence regions centred around an incorrect set of parameters could still encapsulate the uncertainty on the predictions. Still, recovering the true set of parameters is the only way to provide full guarantees on the performance of the model on any future dataset.

Amongst all UQ methods tested, VarBUQ was best able to recapture the underlying signal. The coverage of its 95% confidence intervals is significantly higher than the other methods for both AD model (see fig. 3.8b), achieving an overall mean coverage of 69%, compared to 38% for FIM, 35% for Beale and 30% for Bootstrap. For each method, the CIs obtained higher coverage for soluble compounds than gas flows - this can be explained by the smaller sensitivity of gas flows to parameter values, as non-biological gas-related parameters such as Kp were assumed to be constant, and higher sensitivity to the input noise. The higher coverage obtained by VarBUQ is coherent with the larger width of its CIs (see fig. 3.8d). While for the gas flows, these remain too small to fully capture the signal at the target 95% level, VarBUQ's CIs can be large for soluble compounds (up to $\pm 10\%$ for AM2 and $\pm 17.5\%$ for ADM1). Notably, for AM2, VarBUQ's CIs systematically had 100% coverage, indicating under confidence in the results. Other UQ methods constructed CIs achieving above 95% coverage which were more than 3 times smaller for one dataset. Oracle symmetric CIs achieving 100% coverage could be up to 5.6 times smaller, implying that the methodology can still be improved upon. Still, no other UQ method was able to obtain consistently high coverage for soluble compounds. For ADM1, the large width of VarBUQ's CIs appears necessary to obtain the required coverage. In the single case where another UQ method covered more than 90% of the data (Beale for acetate concentration, HN), CIs' width was larger than the one obtained by VarBUQ (0.29 vs. 0.23), for lower coverage (92% vs. 95%).

VarBUQ was able to maintain a high level of coverage when the operating conditions change, for a reasonable increase in the width of the CIs. CIs constructed using Fisher's information matrix reacted drastically to changes of operating conditions. The width of the CIs reached very high levels for S_2 in the AM2 model (resp. 2.2 and 3.0 for VN and VF, implying an average factor of 8 and 20 between the lower and upper bound on the prediction). This phenomenon is also observed, though to a lesser degree, for ADM1. This could be explained by FIM using linear extrapolation of local changes in the predictions, which do not take into account saturation effects. This interpretation is corroborated by the fact that the widths of Beale's CIs do not evolve

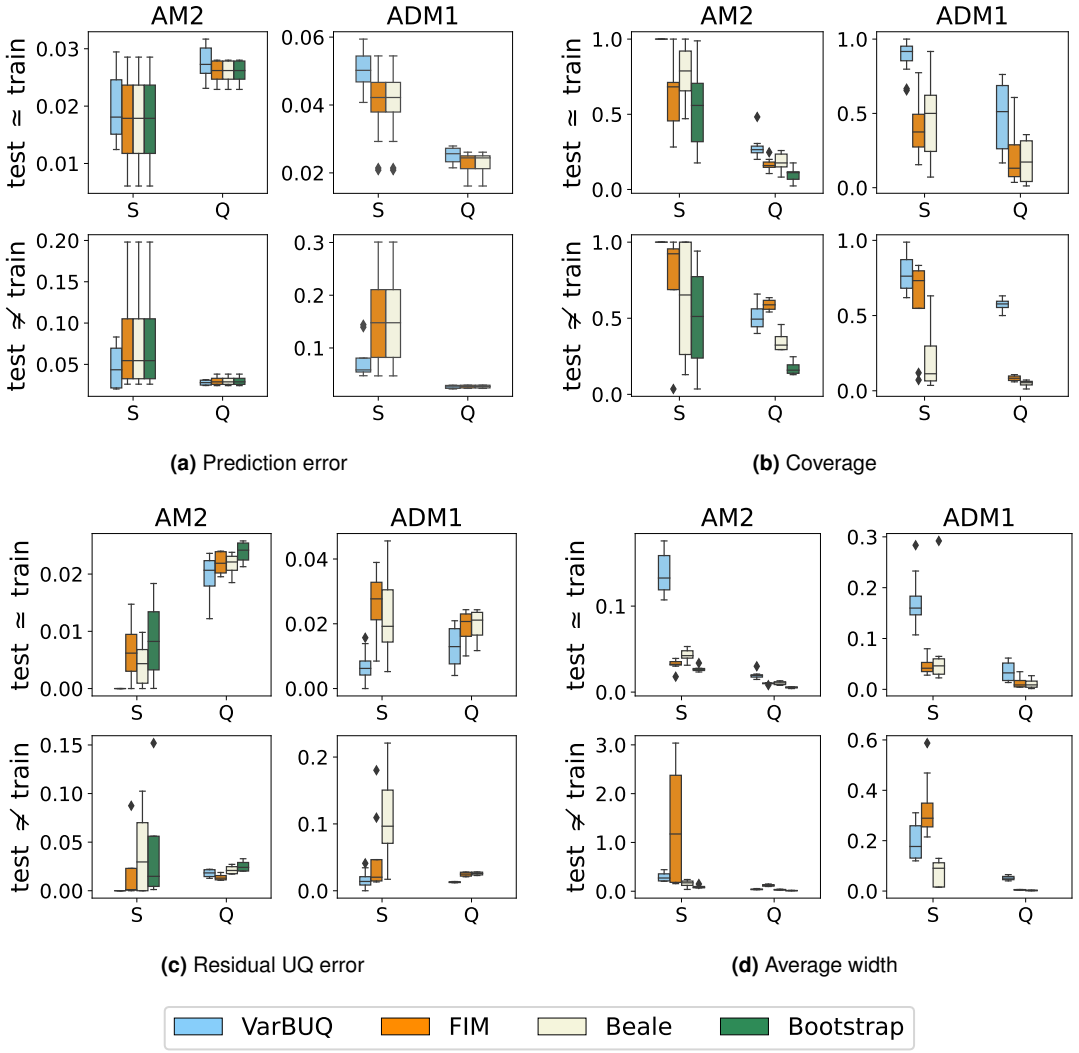


Figure 3.8: Performances of UQ methods on predictions (VarBUQ in blue, FIM in orange, Beale in beige and Bootstrap in green). Types of predictions are grouped depending on whether they are soluble compounds (S) or gas flows (Q). VarBUQ obtained slightly larger test error when the test dataset is similar to the train dataset, but noticeably lower test error when the test dataset exhibit distribution shift (3.8a). The coverage of VarBUQ's confidence intervals on the predictions was globally higher than those of the remaining methods - achieving systematically 100% coverage for soluble compounds for AM2 (3.8b). The coverage and residual error after projection on the confidence intervals are globally coherent with the width of the confidence intervals, with the notable exception of FIM's behaviour for soluble compounds under distribution shift, where the higher width of the confidence intervals does not result in higher coverage or smaller residual errors compared VarBUQ.

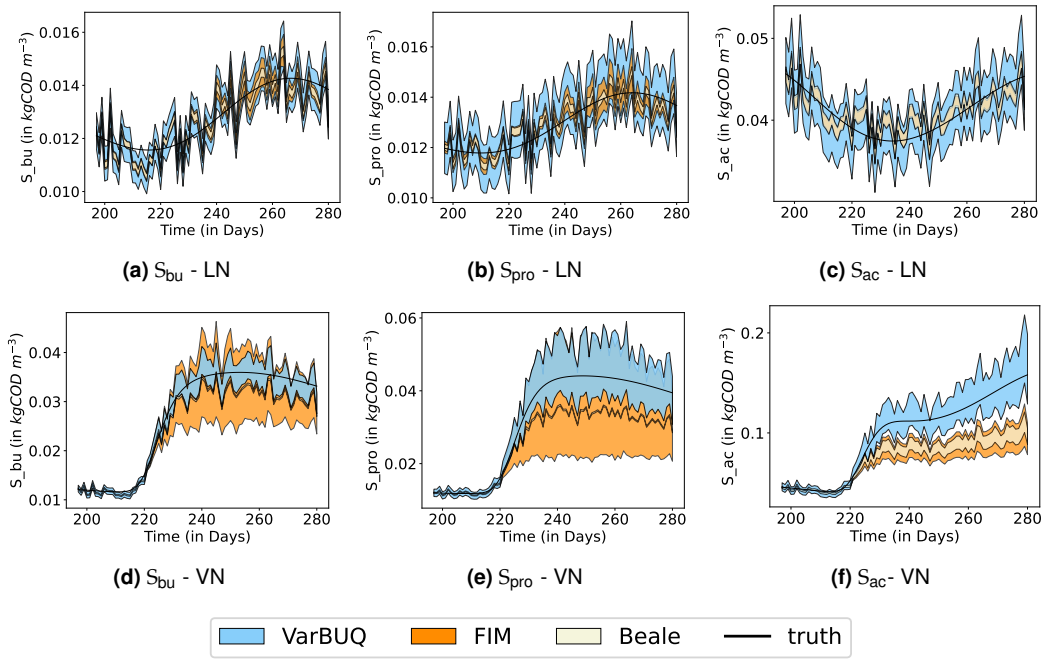


Figure 3.9: 95% Confidence intervals on predictions for the main three VFAs for LN (first row) and VN (second row) datasets (ADM1 model). The confidence intervals from all methods include high frequencies absent from the true signal, due to the noisy intrant description used. When the test set is similar to the training set, the PAC-Bayes calibration has wider confidence intervals, mitigating the impact of the intrant noise (3.9a, 3.9c). Under distribution shift, FIM's confidence intervals widen considerably, encompasses the true signal for both butyrate and propionate concentration, while Beale's confidence interval remains centred tightly around the inadequate mean prediction (3.9d, 3.9e). Both FIM and Beale's confidence intervals are unable to account for the increase of acetate production, contrary to VarBUQ's calibration (3.9f). Figures for the remaining datasets are available in the project's repository.

in such a way - resulting in a drop in coverage.

Both the standard calibration and VarBUQ obtained low errors on test sets similar to the train sets, with average prediction errors remaining lower than 0.021 for AM2, 0.04 for ADM1. While VarBUQ slightly underperformed when the test sets were similar to the training sets (obtaining test risks on average 15% higher), it exhibited stronger robustness to change of operating conditions. For those datasets, the average prediction errors of the standard optimisation were 0.041 for AM2 and 0.097 for ADM1, respectively 53% and 84% higher than the prediction errors of VarBUQ. The residual prediction error computed after projecting on the CIs is globally smaller using VarBUQ (see fig. 3.8c). Notably, it was the only method able to obtain residual prediction errors to a low level (< 0.05) for all predictions. Since the predictions are already small when the test influent is similar to the train influent, this indicator is more relevant for the datasets with distribution shift. For AM2, Bootstrap, Beale and FIM obtained their worst performance on the same variable, S_1 (for VN dataset), of respectively 0.15, 0.10 and 0.087, indicating a sizeable gap

between the signal and the CIs. For ADM1, Beale and FIM obtained non negligible residual prediction error for the concentrations of acetate (0.22 and 0.18 respectively). Beale's UQ also failed to properly account for the propionate (0.19).

Overall, VarBUQ's UQ was best able to capture the discrepancy between the predictions and the true signal. Both FIM and Beale slightly outperformed the Bootstrap method. While FIM and Beale obtained similar performances, FIM reacted better to the change in intrant characteristic, obtaining higher level of coverage.

Figure 3.9 represents the CIs for the predictions of the three main VFAs (butyrate, propionate and acetate) obtained by VarBUQ, FIM and Beale on ADM1, for LN and VN datasets. Without distribution shift (figs. 3.9a to 3.9c), the CIs constructed by VarBUQ encompass those constructed through FIM, which were on average larger than those constructed by Beale's method. All CIs exhibited high frequencies, due to the noisy intrant description. While only VarBUQ's CIs adequately covered the true signal, the remaining methods still obtained satisfactory performances as the calibrated model's predictions were suitable. Under distribution shift figs. 3.9d to 3.9f, the calibrated model's prediction diverged significantly from the truth. FIM's CIs widened sufficiently to take into account this discrepancy for butyrate and propionate concentrations, but not for acetate concentrations. The width of Beale's CIs remained noticeably too small. On the other hand, VarBUQ's CIs were centred around the true signal, and encompassed it adequately.

3.5.4 Computational cost

Computations were carried out using Microsoft Azure, on virtual machines with 32 cores, 64 Gb ram and 256 Gb of memory. Routines fully benefit from parallelisation across the 32 cores and one can assume that the number of cores have an almost linear impact on their durations. All durations are supplied in Table 3.9.

Table 3.9: Computation time for calibration and UQ routines in hours

	SA	Calib.	FIM	Beale	Bootstrap	Bayes init.	VarBUQ
AM2 LN	n.a.	0.50	0.00 (**0.50**)	0.95 (**0.45**)	50.10 (**50.60**)	n.a.	**1.70**
AM2 HN	n.a.	0.43	0.00 (**0.43**)	0.95 (**1.38**)	56.25 (**56.68**)	n.a.	**1.68**
AM2 LF	n.a.	0.47	0.00 (**0.47**)	0.6 (**1.07**)	46.14 (**46.65**)	n.a.	**1.50**
AM2 HF	n.a.	0.43	0.00 (**0.43**)	0.92 (**1.35**)	47.14 (**47.62**)	n.a.	**1.55**
ADM1 LN	0.33	2.32	0.02 (**2.67**)	1.63 (**4.29**)	n.a.	0.59	4.82 (**5.40**)
ADM1 HN	0.32	3.15	0.02 (**3.49**)	1.92 (**5.39**)	n.a.	0.59	4.80 (**5.39**)
ADM1 LF	0.32	2.29	0.02 (**2.62**)	2.22 (**4.82**)	n.a.	0.75	4.77 (**5.52**)
ADM1 HF	0.33	5.32	0.02 (**5.67**)	1.43 (**6.42**)	n.a.	0.72	6.02 (**6.74**)

SA: Sensitivity analysis, Calib.: Calibration, Bayes init.: initialisation of the posterior. For each uncertainty quantification routine, the total time, including calibration, is indicated in bold.

VarBUQ was more computationally intensive than the standard calibration, requiring an average of 1 hour 40 minutes for AM2 (resp. 30 minutes for standard calibration) and 5 hours for ADM1 (resp. 2 hours 30 minutes). This is mitigated once UQ is taken into account. While FIM method's duration is negligible, Beale's method required 50 minutes for AM2 and more than an hour and a half for ADM1, bridging a large part of the gap. The Bootstrap procedure required

prohibitive computational power. As such, this method was only assessed for AM2, with computations lasting about two days. While this computation time could be diminished, by either by reducing the number of Bootstrap procedure or relaxing convergence criteria, this would have serious consequences on the quality of the UQ.

3.5.5 Limitations of the experimental analysis

Potential bias related to calibration method

The performance of the UQ methods benchmarked are impacted by the calibration method. As such, the empirical risk minimisation approach used here should be deemed in part responsible for the obtained results. This choice of calibration method was driven by two considerations. First and foremost, it is a quite common approach in the field, and therefore the results are hopefully representative of the difficulties of obtaining proper UQ for AD models. A second point is that Beale's UQ method takes its origin in the behaviour of the minimiser of mean squared errors objective. To limit confounding factors when assessing the UQ methods, the calibration derived from Beale's method was therefore used also for FIM and Bootstrap, while VarBUQ uses the same scoring function. For FIM, such a calibration is actually ill-suited to the method's hypothesis, since the requirement that the estimator be unbiased is not met. However, it should be stressed that this hypothesis will rarely be realistic in the context of AD models, most of all for highly parametrized models. Constructing an unbiased estimator might not be feasible, even when considering a simple statistical model such as eq. (3.7) - and since the statistical models used have only limited validity, there is little point in trying. Moreover, from a statistical viewpoint, the well-known bias–fluctuation trade off indicates that biased estimator can give better performances. One important difficulty with the optimisation-based procedure used was that it could result in unrealistically high parameter values for the k_m, K_S couples. This was treated by imposing an upper bound on those values when optimising. This could explain why the optimisation procedure had poor robustness when testing on a different intrant. This hints that the calibration could benefit from penalization, in order to favour explanations remaining closer to the standard values.

Limitations of synthetic datasets

Knowledge of the true set of parameters being primordial when assessing UQ methods for parameter recovery, the benchmark was conducted using synthetic datasets. This implies some debatable modelling decisions. A first decision concerned the modelling of the noise. The signal was noised in log space. While not strictly accurate, this implies that the measurement noise will typically be better represented considering relative error. A strong hypothesis was to use the same noise level for all types of observations. This is actually a key requirement in order to use Beale's UQ technique when using different types of predictions. Adapting Beale's method when the noise levels vary is not straightforward, as a core aspect of the method consists in bypassing the estimation of the noise levels. A uniform noise structure was preferred to the

standard Gaussian noise, to test whether this slight change would give an edge to the Bootstrap procedure, specifically designed to deal with unknown noise structure. Noise on the input data resulted in prediction CIs with little smoothness. While the influent signals could have been smoothed, this could have added less detectable biases to the analysis of the results (*e.g.*, choice of the smoothing bandwidth). In practice, observations on the influent might be scarce or exhibit high frequency noise, and as such, the modelling did not seem too unrealistic.

Another consideration is that the performance of calibration and UQ using real world data will depend on the mismatch between the computational model and the physical model. Still, the experiments conducted inform on the quality of the UQ methods. A method struggling with synthetic data is unlikely to fare better with real world data. Finally, the methodology used to construct the true set of parameters might favour the PAC-Bayes framework, insofar as the prior is used. This was mitigated by assessing the performance on a set of parameters which was deemed unlikely to have been drawn from the prior (p-value of 0.05). Still, any Bayesian framework is expected to work poorly if the prior is badly constructed and is either much too large (resulting in underconfidence) or much too small (resulting in both poor calibration performances and overconfidence). Constructing adequate priors is therefore a key challenge for the use of generalized Bayes methods with real data. Thorough bibliographical work is needed to make use of numerous previous works and obtain a state of the art prior. The benchmark's results show that such work could prove valuable; although the other UQ methods can be implemented more easily as they do not require a prior, the Bayesian procedure benefitted from the prior, obtaining confidence regions better centred around the true sets of parameters and confidence intervals on predictions more robust to distribution shift. Notably, it prevents including sets of parameters which an expert would consider unrealistic.

3.5.6 VarBUQ compared to previous Bayesian routines for Anaerobic Digestion

Before the present work, Bayesian flavoured techniques had already been used in the context of AD modelling. Martin and Ayesa [2010] developed a Matlab implementation of Monte Carlo methods which could calibrate a 2-parameters AD model accurately while also assessing parameter uncertainty, adapt to non-identifiable settings, as well as construct proper and tight confidence regions for predictions Martin et al. [2011]. Couto et al. [2019] use a Bayesian framework to fit five parameters in ADM1. Pastor-Poquet et al. [2019] implemented an ad hoc Approximate Bayesian Computation (ABC) algorithm to calibrate 14 parameters on a high-solids AD model. Due to implementation choices, the actual algorithm's UQ presents characteristics between Beale and Bayesian methods. The resulting mean parameter was found to offer good predictive power for methane production, though the authors also noted discrepancies in VFA simulations which could be due to modelling issues.

These Bayesian inspired algorithms output the uncertainty in the form of a sample. Conversely, VarBUQ does not output a sample. The algorithm computes hyperparameters defining a probability distribution belonging to a user chosen parametric class (e.g., multivariate Gaus-

sian). This offers a more interpretable description of the uncertainty, able to effortlessly generate any number of samples from the posterior. This description can be furthermore stored and used for further calibration, assuming more data has been collected; as such, the algorithm can easily be used in an online learning set-up. In addition, VarBUQ considers a simplified Bayesian framework limiting the interactions between parameters to specific cases, chosen through expert knowledge. For instance, in this study, only interactions between parameters acting on the same biological reaction were allowed. This was implemented by considering Gaussian distributions with block diagonal covariance, which significantly reduce the number of hyperparameters (*e.g.*, 54 versus 465 for ADM1). This more rigid set-up limits the ability of the posterior to fit the data but reduces the number of model evaluations needed compared to learning a full covariance matrix or general distribution.

Improving VarBUQ

The PAC-Bayes paradigm showcased here can be improved both in terms of methodology and implementation. A key aspect is the construction of the prior, which could take into account observed correlations between parameters. This would help the posterior further concentrate by removing unlikely combinations of parameters. Another leverage for improvement is the procedure choosing the PAC-Bayes temperature. Informally, the choice of PAC-Bayes temperature should be guided by how much training data is used and how far the data ought to be trusted. Quantifying this confidence would be much harder in real world scenario. Selecting the PAC-Bayes temperature through validation could be a computationally costly option.

The computational cost of the procedure could be reduced. A promising option consists in building surrogate models able to approximate the error of the model for a fraction of the computational cost. For instance, increasing the maximal time step of the ordinary differential equation solvers could be a simple option. We will detail in Chapter 4 a strategy designed to construct and use principled surrogates, able to construct the posterior for a fraction of the number of simulations.

The variational class plays an important role both in terms of computational complexity and performance. The choice investigated here, Gaussian distributions with block diagonal covariance matrix, appeared a good compromise. The block covariance structure prevented the posterior from learning spurious correlations between variables, while it was still able to investigate non identifiable cases. Gaussian distributions are also easier to manipulate compared to the prevalent choice in the AD literature, where a combination of uniform and log-uniform distributions are used to construct the prior [Martin et al., 2011, Pastor-Poquet et al., 2019, Tolessa et al., 2023]. This choice is usually motivated by the lack of prior knowledge on the parameter values beyond their plausible range, hence the use of a so called uninformed prior. On the other hand, covariance plays a crucial role in bypassing AD models' identifiability issues. Gaussian distributions offer a simple way to model covariance while uniform distributions do not. To conciliate flat priors with covariance, new parametrizations of AD models could be considered. For instance, parametrizations considering the ratio of the maximum growth rate and Monod con-

stant might reduce the need for correlations. Another option could be reparametrisations where a Gaussian prior is translated into a uniform prior (using Gaussian quantiles transform)[16]. Accumulating information about the actual prior distribution of the parameters, as observed, would inform the best practical choice.

Applicability to other models

Although VarBUQ was only evaluated on AD models, it should have similar performance when applied to models involving kindred mechanisms. Most biochemical reaction network models display similar features, relying on a combination of ODEs and algebraic equations, and using similar formulas to infer reaction kinetics from the concentration of reactants. For models focusing on microbial communities (*e.g.*, AM2, ADM1, ASM2, models for dark fermentation), the network usually corresponds to Monod equations in cascade, with corrections for the impact of environmental parameters such as pH or temperature. For cell-centred models (*e.g.*, dynamic metabolic simulation), the same approach is implemented through *e.g.* Michaelis–Menten kinetics which are mathematically analogous to the Monod equation. Thus, it could be considered that all these models form a family with comparable non-linearity and differing by their complexity, that is to say the number of represented reactions and model parameters. VarBUQ should display similar advantages and limits for models belonging to this family.

3.6 Conclusion

UQ is crucial to ensure that the right level of confidence is given to future model predictions. The PAC-Bayes calibrated posterior outperformed the most commonly used UQ techniques, both regarding parameter recovery and confidence intervals on test predictions. PAC-Bayes benefits from the inductive bias encoded in the prior, which mitigates the risk of overfitting, and improves robustness compared to standard calibration.

These results vindicate the use of a PAC-Bayes calibration strategy for AD monitoring. The PAC-Bayes bound optimisation strategy introduced here, VarBUQ, succeeded in calibrating both AD models, but had to incorporate acceleration and protection mechanisms to perform adequately. For the more complex ADM1 model, the strategy still failed if started from the prior distribution, making the use of a warm start approach necessary. This questions the ability of the algorithm to perform adequately when considering more complex settings, *e.g.* for a larger number of parameters. Indeed, the calibration strategy failed to converge in reasonable time when considering SUEZ's model, ProdAD. Finally, the calibration time using VarBUQ, while still manageable, was noticeably higher than the standard calibration approach. The main computational bottleneck of VarBUQ is the time spent in simulating. We focus in the next chapter

[16] Such variational families are implemented in 'picproba' under the name of 'GaussHypercubeMap'. By tensorization of the 'ProbaMap' instances, it would be possible to construct an analogue class for block diagonal covariance Gaussian. Such variational families could also be used with the algorithm presented in Chapter 4.

on how this bottleneck can be lightened, by improving both the calibration strategy and model implementations.

Chapter 4

Nimble PAC-Bayes: PAC-Bayes learning for intensive models

In this chapter, we tackle the issue of computational efficiency for the calibration and monitoring of AD plants. Our contributions are three fold:

- We propose a new PAC-Bayes calibration framework, SuPAC, designed to reduce the number of simulations. This approach is generic and can be applied to most PAC-Bayes bounds and a large variety of physical models. We provide an implementation of SuPAC for the minimisation of Catoni's objective on exponential families. These results were obtained in collaboration with Roman Moscoviz, Vincent Schmitt and Benjamin Guedj and presented in Picard-Weibel et al. [2024b]. The implementation was integrated in the 'picpacbayes' package.
- We improve the implementations of AD models to improve computational efficiency. This resulted in a improvement of a factor 100 on the simulation time for SUEZ's AD model ProdAD.
- We assess our new calibration algorithm on SUEZ's model, using a similar methodology as in chapter 3.

Contents

The PAC-Bayes learning strategy VarBUQ developed in chapter 3 proved able to calibrate two AD models and offered state-of-the-art uncertainty quantification. However, the calibration technique considered proved to be quite computationally intensive and suffered from instabilities. This second issue was somewhat mitigated by a mechanism preventing updates from hurting the posterior performance significantly. Still, the PAC-Bayes minimisation technique requires a large number of fresh evaluations of the model to mitigate the risk of obtaining such problematic steps. Such risk queries proved to be the main computational bottleneck, and resulted in a quite long calibration time of up to 6 hours of compute parallelised on 32 cores for ADM1. The AD model on which our PAC-Bayes strategy is to be applied, ProdAD, involves more parameters than ADM1 (more than 70[1] compared to 30 for ADM1) and has wall clock time one order of magnitude larger (190 seconds vs 28 seconds per year simulated). VarBUQ proved unable to calibrate ProdAD, even after using the warm start strategy developed for ADM1.

In this section, we address the difficulty of learning a PAC-Bayes posterior for the computationally expensive ProdAD model. To shorten a journey, one can take a shorter route, or drive faster. Similarly, we pursued two distinct tasks.

[1]The number of parameters involved on ProdAD depends on the specifications of the AD plant. A baseline of 78 parameters are involved, with a minimum of 68 being calibrated, but some plants might involve many more parameters.

- First, we introduced a generic applied framework for PAC-Bayes learning designed to limit the number of calls to the model. This strategy, improving on the methodology used to benefit from previous risk queries of VarBUQ, has theoretical support for a wide range of PAC-Bayes bound and variational classes. The resulting algorithm proved able to calibrate ADM1 with forty times less model queries than VarBUQ. These results are summarized and presented in Picard-Weibel et al. [2024b].
- Second, we considered computational optimisation on the AD models to limit the computational cost of calls to the model. We obtained a speed-up of 100 for ProdAD model, which opens up such prospects as online calibration and meta-learning strategies.

The computational bottleneck being addressed, we assessed the performance of our new calibration strategy on the update implementation of ProdAD.

4.1 Surrogate PAC-Bayes learning

4.1.1 A PAC-Bayes strategy for expensive risk

As noted in chapter 1, PAC-Bayes generalisation bounds can naturally be converted into learning objective. Since the generalisation bound holds simultaneously for all probability measures, it also holds for the minima, leading to a natural way to construct the PAC-Bayes posterior as the minimiser of the generalisation bound.

We focus here on the computational aspect of this strategy. The optimisation of generalisation bounds might not always be viable for tractable or computational reasons, or both. Most PAC-Bayes bounds do not admit a close form minima formulation; moreover, such bounds involve expectations and divergence terms which in general settings can not be evaluated in closed form and thus require the use of approximation methods such as Monte-Carlo sampling (see amongst others Seldin and Tishby [2010], Dziugaite and Roy [2017], Neyshabur et al. [2017], Mhammedi et al. [2019]). Such approximation methods can prove computationally intensive, notably if the empirical risk, whose expectation is optimised in the bound, is hard to query.

Models whose predictions require solving stiff ODE or **P**artial **D**ifferential **E**quation (PDE) such as naturally occurs in physics or biology inspired problems, result in empirical risks whose query can be computationally expensive. Most PAC-Bayes bounds involve the computation of the empirical risk's average with respect to the posterior distribution. Computing this term and its gradient rely on numerous calls to the model at each step, which can be impracticable. Indeed, for the AD models studied in chapter 3, calls to the model proved to be the main computational bottleneck when optimising the PAC-Bayes bound. The computational strain was partly relieved by reusing previous model calls using weight correction based on the ratios of density and momentum. This weight correction constructed unbiased estimators of the gradient whose variance might or might not be larger than the standard gradient estimate, depending on the

fluctuation of the ratios of density[2]. To mitigate the risk of a gradient estimator providing an inadequate gradient estimation harming the performance of the newly constructed posterior, a step removal mechanism was put into place to compare the newly constructed posterior's bound to the previous objective value. Such mechanisms proved sufficient to calibrate two standard AD models using quite powerful compute resource, in moderate to high time (6 hours for ADM1).

Still, the extra layers involved in the gradient computation (weight correction, step removal) are not wholly satisfactory and do not exploit all previous information on the empirical risk; some risk evaluations are discarded, even though the predictors might be close to the mass centre of the current posterior; risk evaluations impact the gradient estimate differently depending on the distribution from which their predictor were drawn[3].

In response to these shortcomings, we propose a new principled strategy designed to mitigate the computational cost of querying the empirical risk when optimising PAC-Bayes generalisation bounds. The key idea of this learning algorithm's is to iteratively optimise a sequence of surrogate training objectives. Such surrogate objectives are obtained by replacing the hard to query empirical risk by a simpler proxy. This proxy is built as the orthogonal projection of the true empirical risk on a functional vector space of finite dimension, which can, for well chosen variational families, be queried much more efficiently than the initial risk. A key motivation is that such surrogate objectives can offer adequate approximations of the true objective valid much further away than the linear approximation offered by the gradient, and enable larger optimisation steps. This effectively decouples the complexity of querying the empirical risk and optimising PAC-Bayes objectives.

Our contributions. The three main contributions of this section span theory, algorithmic and numerical experiments.

1. We provide a generic recipe for learning via surrogate PAC-Bayes bounds, which we believe is of practical interest for machine learning tasks involving computationally intensive models with moderate dimension (e.g. physics models with less than few hundred parameters), and is of practical use when considering AD models,

2. contribute theoretical results establishing that iteratively optimising our surrogates implies the optimisation of the original generalisation bounds. This is established by Theorem 4.2 and further developed in Corollary 4.1 and Theorem 4.4,

[2]The variance should be decreased once the posterior distribution stabilizes. In the limit where the posterior has converged, the number of data increases from n to Kn where n is the number of predictors evaluated at each step, and K the number of generations aggregated. This results in a decrease of variance by a factor $1/\sqrt{K}$. On the other hand, if the fluctuations of the ratio of density are large, the variance can be increased, since the estimator depends on the variance of $\frac{\mathrm{d}\pi_t}{\mathrm{d}\pi_{t-K}}(\gamma)R(\gamma)\ell(\gamma)$. Note that this analysis does not take into account the fact that the distribution π_t depends on the values of the predictors γ drawn from π_{t-K}, which further complicates the analysis of the actual gradient estimator.

[3]Indeed, the weight correction mechanism will lead to problematic step if a predictor γ was drawn from π_{t-K} with $\frac{d\pi_t}{d\pi_{t-K}}(\gamma) \gg 1$ (high variance of the gradient estimation), even though this indicates that γ is informative about the values taken by empirical risk on predictors which the *current* posterior π_t deems likely. In other words, the mass correction mechanism penalizes cases where a past posterior drew by accident a predictor which is now deemed much more likely.

3. illustrate our approach with numerical experiments inspired by an industrial biochemical setting using an anaerobic digestion model.

4.1.2 A first step towards surrogate PAC-Bayes

A well known phenomena of Bayesian statistics is the notion of conjugate prior. Starting from a Gaussian prior $\pi_p = \mathcal{N}(\mu, \Sigma)$ and considering a statistical model $y = \gamma + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \Sigma_\varepsilon)$, it is well known that the posterior distribution is the Gaussian distribution $\mathcal{N}((\Sigma^{-1} + \Sigma_\varepsilon^{-1})^{-1}(\Sigma^{-1}\mu + \Sigma_\varepsilon^{-1}y), (\Sigma^{-1} + \Sigma_\varepsilon^{-1})^{-1})$. This conjugation of a Gaussian prior and Gaussian model resulting in a Gaussian posterior remains valid in the PAC-Bayes setting when considering Gibbs posteriors. In this case, the log likelihood is replaced by the loss, and hence the PAC-Bayes conjugation holds for a Gaussian prior in a quadratic loss setting.

Bayesian conjugation is related to the notion of exponential family. Using the natural parametrisation, an exponential family is defined as the set of probability measures whose log density with respect to a common measure is of form $\theta \cdot T - g(\theta) + \psi$, where $T, \psi \in \mathbb{R}^{\mathcal{H}}$ and $g \in \mathbb{R}^{\Theta}$. The function ψ is a measure carrier function which can be factored into the the common measure. The function g is a renormalisation function, ensuring that the integral of the density sums to 1. The exponential family considers all θ such that the density is renormalizable, hence

$$\mathcal{P}_T = \left\{ \pi_\theta, \frac{\mathrm{d}\pi_\theta}{\mathrm{d}\pi_{\mathsf{ref}}} = \exp(\theta \cdot T - g(\theta) + \psi) \mid \pi_{\mathsf{ref}}\left[\exp(\theta \cdot T + \psi)\right] < \infty \right\}. \tag{4.1}$$

For Gibbs posterior of temperature λ using a prior π_{θ_p}, the density with respect to π_{ref} is proportional to $\exp(\theta \cdot T - g(\theta) + \psi - \lambda^{-1}R)$. In particular, if $R = \theta_R + c$ and the prior belongs to the exponential family, *i.e.* $\pi_{\mathsf{p}} = \pi_{\theta_p}$, it follows that the posterior distribution belongs to the exponential family and values $\pi_{\theta_p - \lambda^{-1}\theta_R}$. This defines a generic PAC-Bayes coupling phenomena for exponential family considering Gibbs posteriors.

Exponential families define a tractable, yet flexible class of probability families, spanning from simple, fixed variance distributions to multimodal distributions [Cobb et al., 1983]. They cover most familiar distribution families such as multivariate Gaussians, Beta and Gamma [Brown, 1986]. The Gaussian with block diagonal covariance family considered in chapter 3 is also an exponential family. The form of the risk for with which they are conjugated equally vary (*e.g.* linear risk for exponential distributions, quadratic risk for Gaussian distributions, quadratic risk with constraints on the Hessian for subclass of Gaussian distributions).

Gibbs posteriors can be motivated as being minimisers of Catoni's bound (1.20) [Alquier, 2024]. Let us consider the PAC-Bayes variational framework using Catoni's bound for exponential family. In this setting, the posterior distribution minimising the bound is of form π_{θ^*}. This implies that the log ratio of density between prior and posterior is $(\theta^* - \theta_p) \cdot T - g(\theta^*) + g(\theta_p)$, where θ^* is a function of the empirical risk. Remark that the same posterior distribution would be obtained by replacing the empirical risk by $-\lambda(\theta^* - \theta_p) \cdot T + c$. This observation gives us our first insight about surrogate risks: by limiting our exploration to a parametric family, we are creating equivalent classes for the risks defined as the set of risks resulting in the same posterior

distribution. The observation above implies that for any risk, its equivalent class intersects the family $\{\theta \cdot T + c \mid \theta \in \mathbb{R}^d, c \in \mathbb{R}\}$.

Intuitively, the equivalent class should group risks which are similar to one another, that is to say that the function $-\lambda(\theta^* - \theta) \cdot T$ should provide an optimal approximation of R. A difficulty is that the mapping is invariant for the addition of constants. We show that this is (almost) the only difficulty when defining the optimal θ_R.

Theorem 4.1

(Informal[a]) Consider a risk function R and an exponential family defined by $\mathcal{E}(T, h, \pi_{\text{ref}})$ and parametrized by an open subset $\Theta \subset \mathbb{R}^{d_\Theta}$. Assume that the Catoni's bound has a minima on the exponential family and let θ^* denote the parameter defining the minimiser of Catoni's bound for temperature λ and prior π_{θ_p}, *i.e.*

$$\theta^* \in \arg\inf \pi_\theta [R] + \lambda \mathsf{KL}(\pi_\theta, \pi_{\theta_p}).$$

Then a surrogate risk inducing the same posterior, defined as $R_{\theta_R} = \theta_R \cdot T$ with

$$\theta_R := -\lambda(\theta^* - \theta_p)$$

is the best quadratic approximation of the risk R up to a constant, *i.e.*

$$\theta_R \in \arg\inf_{\theta \in \mathbb{R}^{d_\Theta}} J(\theta) := \pi_{\theta^*}\left[\left(R - \theta \cdot T - \pi_{\theta^*}[R - \theta \cdot T]\right)^2\right]. \tag{4.2}$$

[a] A rigorous version of Theorem 4.1 can be obtained as a corollary of Theorem 4.4.

Proof. First of all, the function $J(\theta)$ is convex, being the integral (with respect to γ) of convex function

$$\theta \mapsto \left(\theta \cdot \left(T(\gamma) - \pi_{\theta^*}[T]\right) - \left(R(\gamma) - \pi_{\theta^*}[R]\right)\right)^2$$

for fixed γ. This implies that any θ satisfying $\nabla J(\theta) = 0$ is the global minimiser of J.

Allowing ourselves to differentiate under the integral sign, one obtains

$$\nabla J = 2\pi_{\theta^*}\left[\left(\theta \cdot T - R\right)\left(T - \pi_{\theta^*}[T]\right)\right].$$

Taking the problem from the other end, we consider Catoni's objective

$$\begin{aligned} C(\theta) :=& \pi_\theta[R] + \lambda \pi_\theta\left[\log\left(\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\pi_{\theta_p}}\right)\right] \\ =& \pi_\theta\left[R + \lambda\left(\left(\theta - \theta_p\right) \cdot T - g(\theta) + g(\theta_p)\right)\right]. \end{aligned}$$

Taking it for granted that this objective is differentiable and can be differentiated under the integ-

ral sign, one obtains

$$\begin{aligned}
\nabla C(\theta) =& \partial_\theta \left(\pi_{\theta_p} \left[\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\pi_{\theta_p}} \left(R + \lambda \left((\theta - \theta_p) \cdot T - g(\theta) + g(\theta_p) \right) \right) \right] \right) \\
=& \pi_\theta \left[\partial_\theta \left(R + \lambda \left((\theta - \theta_p) \cdot T - g(\theta) + g(\theta_p) \right) \right) \right] \\
&+ \pi_\theta \left[\left(R + \lambda \left((\theta - \theta_p) \cdot T - g(\theta) + g(\theta_p) \right) \partial_\theta \left(\log \left(\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\pi_{\theta_p}} \right) \right) \right) \right] \\
=& \pi_\theta \left[\left(R + \lambda \left((\theta - \theta_p) \cdot T - g(\theta) + g(\theta_p) \right) \right) \left(T - \pi_\theta [T] \right) \right] \\
=& \pi_\theta \left[\left(-\lambda \left(\theta - \theta_p \right) \cdot T - R \right) \left(T - \pi_\theta [T] \right) \right],
\end{aligned}$$

where we use the well known identity $\pi_\theta [T] = \nabla g$ (see Brown [1986]) to obtain the third equality. Since $\nabla C(\theta^*) = 0$, this implies that

$$\pi_{\theta^*} \left[(R + \lambda(\theta^* - \theta_p) \cdot T)(T - \pi_{\theta^*} [T] \right] = 0,$$

which implies that $-\lambda(\theta^* - \theta_p) = \theta_R$ minimises J. This concludes the informal proof (conditions implying that the differentiation under the integral sign are possible will be stated in the main version of the theorem). □

Theorem 4.1 can at first seem of limited practical value, since the criteria used to project the risk relies on the knowledge of the posterior itself. This however naturally leads to a learning process alternating between two steps: at fixed posterior, projecting the risk, then using the risk projection, updating the posterior distribution. Then Theorem 4.1 implies that the true posterior distribution is a fixed point of this algorithm[4]. Theorem 4.1 relies both on the fact that the PAC-Bayes objective considered is Catoni's bound (in order to use the closed form expression for the posterior), and on the fact that the bound is minimised on an exponential family (in order to define the space of surrogate functions). A natural question is whether this surrogate strategy can be extended for the minimisation of generic PAC-Bayes bound, or for generic families of distributions. We show that we can simultaneously perform both these extensions.

4.1.3 A generic surrogate framework

Before establishing theoretical motivation, let us introduce the SuPAC strategy as a reasonable heuristic for the minimisation of PAC-Bayes bounds. Let us consider a generic PAC-Bayes bound PB, which is a function of the posterior distribution, the empirical risk, the prior distribution, and other factors (*e.g.* the confidence level, the PAC-Bayes temperature, the number of data) which we regroup in the notation ξ. A PAC-Bayes learning task, limited to the variational

[4] As can be inferred from the proof, any θ defining a local minima of the objective, or even any θ such that $\nabla C(\theta) = 0$, will also be a fixed point.

family $\mathcal{P} \subset \Pi$, consists in solving

$$\arg \inf_{\pi \in \mathcal{P}} \mathrm{PB}\left(\pi, R, \pi_{\mathsf{p}}, \xi\right). \tag{4.3}$$

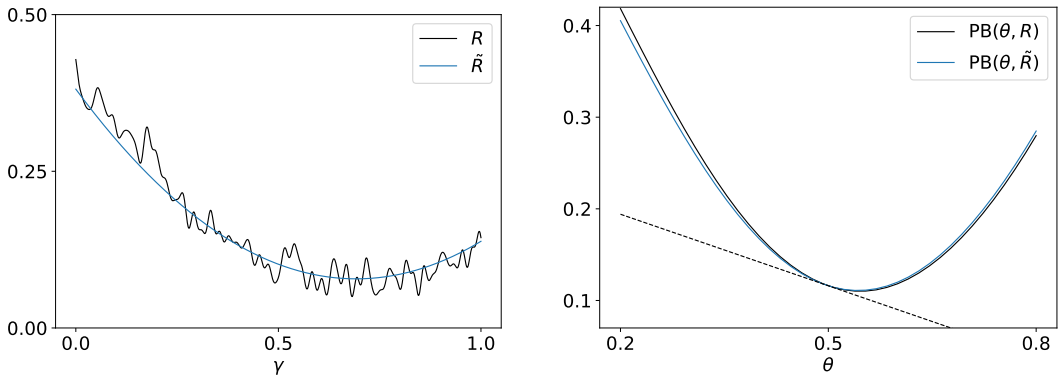
Restriction of the minimisation problem to the variational family might be justified by various considerations, including storage of the calibrated distribution, simplification of the minimisation task [Alquier et al., 2016, Dziugaite and Roy, 2017] or, as seen in chapter 3, as a way to encode expert knowledge. When $\mathcal{P}$ can be parametrized by a finite dimensional space, this effectively turns the PAC-Bayes learning task from a infinite to finite dimension minimisation problem. Even this simplified minimisation problem might prove computationally difficult for GD based algorithm. This is especially the case when evaluating the empirical risk is costly, e.g. when the prediction model involves solving stiff ODEs or PDEs. As PAC-Bayes bounds depend on the π-mean of the empirical risk, each gradient estimation rely on numerous new evaluations of the empirical risk. For ODEs $\dot{X} = F(X, t, \gamma)$ where F is very sensitive with respect to X, numerous evaluations of F are required to obtain adequate numerical solutions. These evaluations must moreover be performed iteratively, and hence can not be parallelised. Moreover, implementing the ODE solver in a way to benefit from GPU speeds up might not be practicable, since most ODE solver use a varying step size which will depend on γ. This will result in typically long model calls which cannot be massively parallelised.

To overcome this difficulty, we introduce the Surrogate PAC-Bayes bound learning framework (SuPAC), which is based on alternatively building and solving surrogate problems. Formally, we consider an approximation algorithm $\mathsf{F} : \mathcal{P} \times \mathcal{M}(\mathcal{H}) \mapsto \mathcal{M}(\mathcal{H})$ in conjunction with an approximate solving algorithm $\mathrm{Solve} : \Pi \times \mathcal{P} \times \mathcal{M}(\mathcal{H}) \mapsto \Pi$. Informally, F constructs a proxy of the empirical risk valid for the current posterior estimation π; while Solve updates the posterior estimation by solving the resulting surrogate objective.

Algorithm 3 Surrogate PAC-Bayes Learning framework (SuPAC)

Require: PB, $\pi_0 \in \mathcal{P}$, $\pi_{\mathsf{p}} \in \Pi$, $R \in \mathcal{M}(\mathcal{H})$
  $\pi \leftarrow \pi_0$
  while not converged **do**
    $f \leftarrow \mathsf{F}(\pi, R)$
    $\pi \leftarrow \mathrm{Solve}(\pi_{\mathsf{p}}, \pi, f)$
  end while

Algorithm 3 offers a lot of leeway for building surrogates (e.g., iteratively refining an ODE or PDE solver, tailored made surrogates for physical models, polynomial approximations) as well as solving the surrogate problem. For such a framework to be practicable, two conditions should apply: the construction of the surrogate and approximate solving should be faster than solving the initial problem, and the algorithm's result should tend to diminish the PAC-Bayes bound. Intuitively, the choice of the approximation mechanisms plays a critical role; indeed, the more precise the approximation, the more likely is the minima of the surrogate task to be close to the true minimiser, but the harder the approximation task and the surrogate solving task.



(a) Surrogate risk

(b) Surrogate Objective

Figure 4.1: Impact of the risk approximation on a Catoni base PAC-Bayes objective. Replacing the complex risk R by a quadratic approximation $\tilde{R}$ (Figure 4.1a) does not impact much the objective for normal distributions of variance 1 (Figure 4.1b). Here, the quadratic approximation $\tilde{R}$ is defined as the best quadratic approximation of R weighted by $\mathcal{N}(0.5, 1)$. As a result, Theorem 4.2 guarantees that the gradient of the two objectives are identical (dotted line). Still, the surrogate objective provides an adequate approximation of the true objective for values of θ where the gradient approximation does not.

4.1.4 Constructing surrogate function spaces

A core contribution of this chapter is to show that for generic PAC-Bayes bounds and generic probability families $\mathcal{P}$ of dimension d_Θ, $\mathrm{L}^2(\pi)$ orthogonal projection of the empirical risk on a functional vector space of dimension $d_\Theta + 1$ is sufficient to obtain convergence guarantees.

A few assumptions on the PAC-Bayes bounds, the risk R and the probability family $\mathcal{P}$ are required.

Assumptions. *(A_1) $\mathcal{P} = \{\pi_\theta, \theta \in \Theta\}$ is a parametric set indexed by an open subset $\Theta \subseteq \mathbb{R}^{d_\Theta}$;*

(A_2) $\forall \theta \in \Theta$, π_θ is absolutely continuous with respect to π_p and $\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\pi_\mathsf{p}}(x) = \exp(\ell(\theta, x))$ with $\theta \mapsto \ell(\theta, x)$ differentiable for all x;

(A_3) $\forall \theta \in \Theta$, $\exists N_\theta$ a neighbourhood of θ such that $x \mapsto \sup_{\theta \in N_\theta} |\partial_\theta \ell(\theta, x)| \in L^2(\pi_\theta)$;

(A_4) $R \in \cap_{\theta \in \Theta} L^2(\pi_\theta)$;

(A_5) There exists $\widetilde{\mathrm{PB}}$ such that $\mathrm{PB}(\pi_\theta, R, \pi_\mathsf{p}, \xi) = \widetilde{\mathrm{PB}}(\theta, \pi_\theta[R], \pi_\mathsf{p}, \xi)$ (i.e. the PAC-Bayes bound dependence on the empirical risk is limited to the posterior average of the empirical risk). Moreover, $\widetilde{\mathrm{PB}}$ is differentiable with respect to its two first arguments.

We emphasise that these assumptions are valid for practically all PAC-Bayes bounds, most risks, and for a wide variety of probability families, and are thus rather more technical than restrictive. Although the second assumption rules out probability distributions whose support is not included in the prior support, such distributions usually obtain unbounded PAC-Bayes

bounds due to penalisation terms, and as such are already ruled out. Most standard family of distributions, including Gaussian and Gaussian mixtures, satisfy (A_1) to (A_3) for adequate parametrizations. The fourth assumption is automatically satisfied for all bounded risks, which is a typical assumption of PAC-Bayes bounds, but also allows for unbounded risks provided that they are square integrable (*e.g.* polynomials if $\mathcal{P}$ span Gaussian would satisfy (A_4)). The last assumption is satisfied by most PAC-Bayes bound, *e.g.* McAllester [2003], Maurer [2004][5].

Since $\mathcal{P}$ is parametrised by Θ, we will abuse notations for functions of $\mathcal{P}$ and write *e.g.* $G(\theta) := G(\pi_\theta)$. For a given θ, the functional vector space generated by the gradient of the log-likelihood,

$$\mathcal{F}_\theta := \left\{ f_{\eta,C} : x \mapsto \eta \cdot \partial_\theta \ell(\theta, x) + C \mid \eta \in \mathbb{R}^{d_\Theta}, C \in \mathbb{R} \right\} \tag{4.4}$$

provides a natural approximation space of dimension $d+1$. We are now in a position to state our main approximation result

Theorem 4.2

Under assumptions (A_1) to (A_5), replacing the empirical risk R by its $L^2(\pi_\theta)$ projection on $\mathcal{F}_\theta$ defined as in eq. (4.4) (proxy risk)

$$f^{R,\theta} := \arg \inf_{f \in \mathcal{F}_\theta} \pi_\theta \left[(R - f)^2 \right]$$

leaves the gradient of the objective PB invariant, i. e.

$$\partial_1 \mathsf{PB}(\theta, R, \pi_\mathsf{p}, \xi) = \partial_1 \mathsf{PB}\left(\theta, f^{R,\theta}, \pi_\mathsf{p}, \xi\right).$$

This result also holds if the approximation space $\mathcal{F}_\theta$ is replaced by the larger approximation set $\{f + \rho \mid f \in \mathcal{F}_\theta, \rho \in \mathcal{G}\}$ for any set $\mathcal{G} \subset \mathsf{L}^2(\pi_\theta)$.

Proof. Assumptions *(A_3)* and *(A_4)* allow differentiating $\theta \mapsto \pi_\theta[R] = \pi\left[\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\pi} R\right]$ under the integral sign (see Theorem 6.28 in Klenke [2020]), yielding $\nabla \pi_\theta[R] = \pi_\theta[R \partial_\theta \ell]$. As such, the derivative of $\widetilde{\mathsf{PB}}(\theta, \pi_\theta[R], \pi_\mathsf{p}, \xi)$ with respect to θ equals

$$\partial_1 \widetilde{\mathsf{PB}}(\theta, \pi_\theta[R], \pi_\mathsf{p}, \xi) + \partial_2 \widetilde{\mathsf{PB}}(\theta, \pi_\theta[R], \pi_\mathsf{p}, \xi)\, \pi_\theta[R \partial_\theta \ell].$$

As the only dependence on the gradient with respect to R is on the value of $\pi[R]$ at which the derivative is evaluated and on the vector $\pi_\theta[R\partial_\theta \ell]$, it follows that $\partial_\theta \mathsf{PB}$ is not modified by replacing R by a function $f \in L^2(\pi_\theta)$ satisfying the following linear system:

$$\begin{cases} \pi_\theta[R \partial_\theta \ell] &= \pi_\theta[f \partial_\theta \ell], \\ \pi_\theta[R] &= \pi_\theta[f]. \end{cases} \tag{4.5}$$

[5] It is not however satisfied for the PAC-Bayes bound of Tolstikhin and Seldin [2013] which also involve the variance of the risk.

By construction of $\mathcal{F}_\theta$, the linear system (4.5) is satisfied if and only if $(f - R) \in \mathcal{F}_\theta^\perp$, where $A^\perp$ denotes the orthogonal complement of A in $\mathsf{L}^2(\pi_\theta)$. Hence for any set $\mathcal{G} \subset \mathsf{L}^2(\pi_\theta)$, the orthogonal projection of R on $\widetilde{\mathcal{F}} = \mathcal{F}_\theta + \mathcal{G}$ satisfies the linear system (4.5). Noticing that the orthogonal projection $f^{R,\theta}$ of R on space $\widetilde{\mathcal{F}}$ satisfies $f^{R,\theta} = \arg\inf_{f \in \widetilde{\mathcal{F}}} \pi_\theta \left[(R - f)^2\right]$ completes the proof. □

Informally, Theorem 4.2 guarantees that if searching for a PAC-Bayes posterior in a space of size d_Θ, adequately projecting the score on a space of dimension at most $d_\Theta + 1$ preserves the immediate surrounding of the PAC-Bayes objective. For some families of probability distributions, the approximation space dimension can be lower than $d_\Theta + 1$, if the functions $\partial_\theta \ell$ are not linearly independent. Moreover, the dimension might not be constant: for mixtures of 2 Gaussians, the dimension changes when the Gaussians coincide. If the approximation built at θ maintains near optimal performance for a large neighbourhood of θ, this surrogate task provides a valid approximation of the true task for a wide range of distributions, and offers approximate solutions $\tilde{\theta}$ much further away than the range of validity of the objective's gradient.

The extension of the result for $\mathcal{F}_\theta + \mathcal{G}$ implies that proxy score functions combining a known, simplified model with a learnt correction term can be used. For $\mathcal{G} = \{\rho\}$, it implies that the result holds if the approximation space consists of a fixed user defined proxy and a correction term. This can have direct practical implications in settings where efficient, natural proxy are available. For the AD setting for instance, such a function ρ could be obtained by using a coarser resolution for solving the ODE. Another venue for this extension could be to consider larger functional vector space, for instance polynomial approximation of the empirical risk for a Gaussian variational class which only require quadratic approximations. A key incentive for such strategy is that the learnt corrective term would presumably be smaller, and hence the approximation's validity larger. Another incentive is that this can simplify the approximation strategy, if this removes the dependency of the approximation space on the current posterior - we will show in Theorem 4.3 that this is virtually limited to the case when $\mathcal{P}$ is a subset of an exponential family..

A direct consequence of Theorem 4.2 is a fixed point characterisation of the minima of the PAC-Bayes objective for instances of Algorithm 3 using GD based surrogate solver:

Corollary 4.1

Under assumptions *(A_1)* to *(A_5)*, and defining the approximation space $\mathcal{F}_\theta$ through eq. (4.4), the minimiser θ^* of the original PAC-Bayes bound is a fixed point of any instance of Algorithm 3 such that:

- the approximation function is $\mathsf{F}(\pi_\theta, R) := \arg\inf_{f \in \mathcal{F}_\theta} \pi_\theta \left[(R - f)^2\right]$,
- the surrogate solving Solve strategy is any (corrected) gradient descent strategy starting at the current θ, using update steps of form

$$\mathsf{Updt}(\theta) = \theta - M(\pi, \theta, f, \xi)\partial_\theta \mathsf{PB}(\theta, f, \pi_\mathsf{p}, \xi),$$

where M stands for any function returning an endomorphism, for any number of steps, any convergence criteria.

As in Theorem 4.2, the approximation space $\mathcal{F}_\theta$ can be replaced by the larger approximation spaces $\mathcal{F}_\theta + \mathcal{G}$ with $\mathcal{G} \in \mathsf{L}^2(\pi_\theta)$. Moreover, $\mathcal{G}$ can be a function of the step and the current θ.

Proof. As assumptions *(A_1)* to *(A_5)* hold, Theorem 4.2 can be used. It implies that replacing R by $f^{R,\theta}$ does not change the gradient of PB. Hence, starting from $\theta = \theta^*$, since

$$\partial_1 \mathsf{PB}(\theta^*, R, \pi_\mathsf{p}, \xi) = \partial_1 \mathsf{PB}(\theta^*, f^{R,\theta^*}, \pi_\mathsf{p}, \xi) = 0,$$

the update step in the solving strategy satisfies

$$\mathsf{Updt}(\theta^*) = \theta^* - M(\pi, \theta, f, \xi) \times 0 = \theta^*.$$

Hence, by recursion, it follows that $\mathrm{Solve}\left(\pi_\mathsf{p}, \pi_{\theta^*}, f^{R,\theta^*}\right) = \pi_{\theta^*}$. Since $\mathsf{F}(\pi_\theta, R) = f^{R,\theta}$, this implies that π^{θ^*} is a fixed step of $\pi \mapsto \mathrm{Solve}\left(\pi_\mathsf{p}, \pi, \mathsf{F}(\pi, R)\right)$, and hence that the posterior is a fixed point of SuPAC for the specified F and Solve strategies, concluding the proof. □

It should be stressed that Corollary 4.1 does not imply that algorithm 3 improves on GD. Corollary 4.1 only guarantees that replacing the empirical risk by a low dimensional approximation is harmless locally. Informally, if the approximation built at θ maintains near optimal performance for a large neighbourhood of θ, this surrogate task provides a valid approximation of the objective for this wide radius, and can construct approximate solutions $\tilde{\theta}$ much further away than the range of validity of the gradient (see Figure 4.1). SuPAC decouples the variations of the bound due to the evolution of θ and $f^{\theta,R}$; such a decoupling is particularly interesting if $f^{\theta,R}$ is stable.

4.1.5 Exponential family and Catoni's bound

Closed form surrogate solution and fixed point property

From Theorem 4.2, we can revisit the minimisation of Catoni's objective on exponential family setting studied in Theorem 4.1. First of all, let us remark that the notion of exponential family naturally arises from Theorem 4.2. Theorem 4.2 involves approximation of the empirical risk through orthogonal projection on a local functional vector space $\mathcal{F}_\theta$ of dimension at most $d_\Theta + 1$. A setting of particular interest concerns families of probabilities such that the space $\mathcal{F}_\theta$ does not depend on θ. We establish the following theorem, which guarantees that if Θ is connected and the likelihood smooth, then restrictions of exponential families are the only family of distributions satisfying $\forall \theta \in \Theta, \mathcal{F}_\theta = \mathcal{F}_{\theta_0}$.

Theorem 4.3

For family of distributions satisfying the first three hypotheses of Section 4.1.4 such that, moreover:

- Θ is a connected,
- $\theta \mapsto \ell(\theta, x)$ is twice continuously differentiable for all x.

If there exists a vector space of finite dimension $\mathcal{F}$ such that $\mathcal{F}_\theta \subset \mathcal{F}$ for all $\theta \in \Theta$, then there exists an exponential family $\mathcal{P}_T$ defined on $\widetilde{\Theta}$ and a connected set $\Theta_{\mathcal{P}_T}$ such that $\mathcal{P}_T = \{\pi_\theta \mid \theta \in \Theta_{\mathcal{P}_T}\}$.

Proof. For $\mathcal{F}$ of dimension $\tilde{d}+1$, choose $T_1, \ldots, T_{\tilde{d}}$ and $T_{\tilde{d}+1} := 1$ a basis of $\mathcal{F}$. Then, for all θ, there exists a unique matrix $A(\theta) \in \mathbb{R}^{\tilde{d}+1,\tilde{d}+1}$, and a unique vector $c \in \mathbb{R}^{\tilde{d}+1,1}$ such that

$$\partial_\theta \ell = \begin{pmatrix} A(\theta) & c(\theta) \end{pmatrix} \begin{pmatrix} T_1 \\ \cdots \\ T_{\tilde{d}+1} \end{pmatrix}.$$

Assume that $A(\theta)$ and $c(\theta)$ are differentiable (this is proved afterwards). Since ℓ is twice continuously differentiable, it follows $\partial_{\theta_i}\partial_{\theta_j}\ell = \partial_{\theta_j}\partial_{\theta_i}\ell$, and therefore that $\partial_{\theta_i} A_{j,k} = \partial_{\theta_j} A_{i,k}$ and that $\partial_{\theta_j} c_i = \partial_{\theta_i} c_j$. This, in conjunction with the hypothesis that Θ is connected, implies that $A(\theta)$ is a gradient of some $\beta : \Theta \mapsto \mathbb{R}^{\tilde{d}}$ while c is the gradient of some $-g : \Theta \mapsto \mathbb{R}$ (see Lang [1999]). Hence, $\ell(\theta) = \beta(\theta) \cdot T - g(\theta) + \psi$ for h a solution of $\partial_\theta \psi = 0$. Since Θ is connected, this implies that h can not be a function of θ. Hence $\mathcal{P}$ is the restriction of an exponential family on Θ.

If Θ is not connected, $\mathcal{P}$ is an exponential family in each connected part, each sharing a common function T, but the parametrisation β as well as the carrier measure ψ might change between the different connected parts.

It remains to show that $A(\theta)$ and $c(\theta)$ are differentiable. First of all, for all finite collection of linearly independent real valued functions $(f_1, \ldots, f_n)$, there exists n points $(x_1, \ldots, x_n)$ such that $(f_i(x_j))_{i,j\leq n}$ is invertible. Indeed, this result holds for a single function, since f_1 must be non zero. Then if the result holds for $x_1, \ldots, x_k$, then the determinant $D = \left|(f_i(x_j))_{i,j\leq k}\right| \neq 0$. Consider the matrix $m(z) = (f_i(\tilde{x}_j)_{i,j\leq k+1}$ with $\tilde{x}_j = x_j$ if $j \leq k$, $\tilde{x}_{k+1} = z$. Then the determinant of matrix m is $Df_{k+1}(z) + \sum_{i\leq k} C_i f_i(z)$. Since $f_1, \ldots, f_{k+1}$ are linearly independent and since D is not zero, there must exist z such that $|m(z)| \neq 0$, which we can pick as x_{k+1}. This proves the result by recursion.

Since $T_1, \ldots T_{\tilde{d}+1}$ are linearly independent, we can therefore pick such $x_1, \ldots, x_{\tilde{d}+1}$. By

definition of $A(\theta)$ and $c(\theta)$, it follows that for all θ,

$$\begin{pmatrix} A(\theta) & c(\theta) \end{pmatrix} = \begin{pmatrix} \partial_{\theta_1}\ell(\theta, x_1) & \dots & \partial_{\theta_1}\ell(\theta, x_{\tilde{d}+1}) \\ \vdots & & \vdots \\ \partial_{\theta_k}\ell(\theta, x_1) & \dots & \partial_{\theta_k}\ell(\theta, x_{\tilde{d}+1}) \end{pmatrix} \begin{pmatrix} T_1(x_1) & \dots & T_1(x_{\tilde{d}+1}) \\ \vdots & & \vdots \\ T_{\tilde{d}+1}(x_1) & \dots & T_{\tilde{d}+1}(x_{\tilde{d}+1}) \end{pmatrix}^{-1}.$$

This implies that A and c are smooth functions of the differentiable $(\partial_\ell(\cdot, x_i))_{i\in[1,\tilde{d}+1]}$, and hence that they are differentiable. □

Theorem 4.3 states that (almost) the only case where one can perform the approximation algorithm 3 by projecting the risk on a fixed, finite dimensional vector space is when the variational class is an exponential family. As discussed in Section 4.1.2, exponential families interact nicely with Catoni's bound in the sense that they offer prior to posterior conjugation. That is to say that when considering the minimisation of Catoni's objective on exponential family, solving the surrogate risk can be performed using a closed form expression, rather than gradient descent or other heuristics. More precisely, let us consider the following two assumptions on the prior :

Assumptions. *(A_6)* $0 \in \Theta$, $\pi_{\mathsf{ref}} = \pi_0$;

(A_7) π_{ref} *is absolutely continuous with respect to* π_{p};

(A_8) $\forall\theta \in \Theta$, $\psi_p := \log\left(\frac{\mathrm{d}\pi_{\mathsf{ref}}}{\mathrm{d}\pi_{\mathsf{p}}}\right) \in L^2(\pi_\theta)$.

Assumption *(A_6)* can always be satisfied up to a shift of the parametrisation of Θ if the exponential family is not empty. Its purpose is to simplify notations. Let us remark that if $\pi_{\mathsf{p}} = \pi_{\theta_p} \in \Theta$, then both assumptions *(A_7)* and *(A_8)* are automatically fulfilled. Assumption *(A_7)* guarantees that *(A_2)* is satisfied, while *(A_8)* guarantees that $\forall\theta, \{\lambda\psi_p\} \subset \mathsf{L}^2(\pi_\theta)$.

Then let us consider as the set of approximation functions

$$\widetilde{\mathcal{F}} = \{f_{\eta,C} := \eta \cdot T + C + \lambda\psi_p \mid \eta \in \mathbb{R}^d,\ C \in \mathbb{R}\}.$$

Since $\widetilde{\mathcal{F}} = \mathcal{F}_\theta + \{\lambda\psi_p\}$, it is a set of approximation function in the sense of Theorem 4.2. For any risk in this set, the minimiser of Catoni's objective over all distributions belongs to the exponential family, and hence coincides with the minimiser of Catoni's objective on the exponential family. This implies the closed form solution $\theta^* = -\lambda^{-1}\eta$ for risk $f_{\eta,C}$ provided this defines a probability distribution (else Catoni's bound does not reach its minima on $\mathcal{P}$ nor Π). The surrogate risk $f_{\eta,C}$ does not necessarily satisfy the assumptions of the empirical risk which may (or may not) imply that an optimal posterior distribution exists. For instance, the surrogate risk might no longer be lower bounded (*e.g.* because it is estimated close to a local maxima). We stress that while Θ might be a subset of $\mathbb{R}^d$, the surrogate risk parameter $\eta \in \mathbb{R}^d$ (*e.g.* for Gaussians, normalisation implies that the log density is a quadratic form with negative Hessian, but the approximation function will consider all quadratic form). A simple example of a case where no posterior distribution minimises Catoni's objective with temperature $\lambda = 1$ can

be obtained by considering the one dimensional surrogate empirical risk $f(x) = -2x^2$ to be minimised on the variational family $\{\mathcal{N}(\mu, 1) \mid \mu \in \mathbb{R}\}$ with prior $\mathcal{N}(0, 1)$. In that case, KL is $.5x^2$, while the average empirical risk is $-2x^2 - 1$, and therefore the minima is reached at $-\infty$.

Using the exact solution of the surrogate PAC-Bayes bound could significantly speed up the learning process, avoiding multiple evaluations of the surrogate objective gradient. However, Corollary 4.1 only provides guarantees for gradient descent based algorithm, and hence, *a priori*, the exact solving step can not be used. Hence theorem 4.1 is not a direct consequence of corollary 4.1. We provide in the following lemma the missing link, which reframes the closed form solving step as a corrected gradient descent step.

Lemma 4.1

Consider an empirical risk function R, a prior distribution π_{p} and an exponential family $\mathcal{P} := \{\pi_\theta \mid \theta \in \Theta\}$ with sufficient statistic T satisfying assumptions *(A_1)* to *(A_8)*.
Noting $\widetilde{\mathcal{F}} := \{f_{\eta,C} : x \mapsto \eta \cdot T(x) + C + \lambda\psi_p \mid \eta \in \mathbb{R}^d, C \in \mathbb{R}\}$, let $f_\eta \in \widetilde{\mathcal{F}}$. Then for any prior parameter $\theta_p \in \Theta$, for any parameter θ, the mapping $\tilde{\theta}(\eta) := -\lambda^{-1}\eta$ satisfies:

$$\tilde{\theta} = -\lambda^{-1} I(\theta)^{-1} \nabla_\theta \mathsf{PB}_{\mathsf{Cat}}(\theta, \pi_{\mathsf{p}}, f_{\eta,C}, \xi) + \theta,$$

where $I(\theta)$ denotes Fisher's information matrix

$$I(\theta) = \pi_\theta \left[\partial_\theta \log\left(\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\pi_{\mathsf{ref}}}\right) \partial_\theta \log\left(\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\pi_{\mathsf{ref}}}\right)^t \right].$$

Proof. For any $f_{\eta,C} \in \widetilde{\mathcal{F}}$, the solver of Catoni's bound on all distributions is given by $\tilde{\theta} = -\lambda^{-1}\eta$. Note that the choice of $\tilde{\theta}$ is coherent with the formula given in Lemma 4.1 when the prior belongs to $\mathcal{P}_T$, since in that case $h = \theta_p \cdot T$, leading to a change of coordinate in the definition of $\mathcal{F}$.

Under the assumptions, Catoni's bound is differentiable and its gradient with respect to θ can be computed under the integral. Thus, for risk $f_{\eta,C}$,

$$\begin{aligned}
\nabla \mathsf{PB}_{\mathsf{Cat}} &= \pi_\theta \left[f_{\eta,C}(T - \nabla g(\theta)) \right] + \lambda \pi_\theta \left[(\theta \cdot T - g(\theta) - \psi_p)(T - \nabla g(\theta)) \right] \\
&= \pi_\theta \left[\left(f_{\eta,C} + \lambda\theta \cdot T - g(\theta) - \lambda\psi_p \right) (T - \nabla g) \right] \\
&= \pi_\theta \left[\left(f_{\eta,C} + \lambda\theta \cdot T - \lambda\psi_p \right) (T - \pi_\theta [T]) \right] \\
&= \pi_\theta \left[(\eta \cdot T + C)(T - \pi_\theta [T]) \right] + \lambda \mathbb{V}_{\pi_\theta} [T] \, \theta \\
&= \mathbb{V}_{\pi_\theta} [T] \, (\eta + \lambda\theta).
\end{aligned}$$

For exponential families, the derivative of the log density is $\partial_\theta \log\left(\frac{\mathrm{d}\pi_\theta}{\mathrm{d}\pi_{\mathsf{ref}}}\right) = T - \pi_\theta [T]$, which implies that the variance $\mathbb{V}_{\pi_\theta} [T]$ coincides with Fisher's information. Hence the previous equality reads $\nabla \mathsf{PB}_{\mathsf{Cat}} = \lambda I(\theta)(\theta - \tilde{\theta}(\eta))$, which implies Lemma 4.1. □

Lemma 4.1 states that the closed form solution can be interpreted as a corrected gradient descent procedure. Interestingly enough, the correction factor does not match the standard

Gauss-Newton second order correction, and remarkably does not involve the empirical risk function. Fisher's information matrix provides a natural metric on $\mathcal{P}$ [Čencov, 2000] - giving a geometrical interpretation to this correction.

A direct consequence of Lemma 4.1 is that Corollary 4.1 applies when using the exact solver for the surrogate Catoni task. Since Fisher's information is positive, it follows that the update direction $\tilde{\theta} - \theta$ always diminishes the bound locally. We summarise these results in the following theorem.

Theorem 4.4

The minimiser of Catoni's PAC-Bayes objective on an exponential family is a fixed point of Algorithm 3 with approximation function

$$\mathsf{F}(\pi_\theta, R) := \arg\inf_{f \in \mathcal{F}} \pi_\theta \left[(R - f)^2 \right],$$

and surrogate solver

$$\mathrm{Solve}\left(\pi_{\mathsf{p}}, \theta, f_\eta\right) := -\lambda^{-1}\eta = \arg\inf_{\tilde{\theta} \in \Theta} \mathsf{PB}_{\mathsf{Cat}}\left(\tilde{\theta}, \pi_{\mathsf{p}}, f_\eta, \xi\right).$$

Moreover, for all θ,

$$\nabla \mathsf{PB}_{\mathsf{Cat}} \cdot (\mathrm{Solve}(\pi_{\mathsf{p}}, \theta, \mathsf{F}(\theta, R)) - \theta) \leq 0.$$

Since the surrogate solver in theorem 4.4 does not depend on the current posterior parameter θ, it implies theorem 4.1 by considering $\theta = \theta^*$. When $\pi_{\mathsf{p}} = \pi_{\theta_p}$, one can pick $\psi_p = 0$ in the definition of $\mathcal{F}$, which will lead to the more natural parametrisation $\tilde{\eta} = \theta_p - \lambda^{-1}\eta$ for the minimiser.

As discussed above, the solution of the surrogate task must belong to Θ to define a probability distribution. There is however no guarantee that such is the case for any approximated risk. For instance, if the risk is estimated close to a local maxima by a quadratic function, the resulting surrogate task might not have a minima, and hence the resulting $\theta(\eta)$ might fail to be a probability distribution, causing the algorithm to break. Another difficulty lies in solving the approximation task. As it involves an integral of a function of the risk, finding the best approximation theoretically requires evaluations of the risk at all predictors, defeating the purpose of the surrogate approach. We show in the next section how both these issues can be solved in practice.

Framework implementation: SuPAC-CE

Following theorem 4.4, we propose an algorithm, SuPAC-CE designed to efficiently find the minimiser of Catoni's bound on Exponential families. The implementation of SuPAC-CE can be found in the package picpacbayes ('ScoreApproxPBayesSolver' class for exponential fam-

ily, 'PreExpSABS' for exponential family not using the natural parametrisation[6]. An optimiser dedicated to Gaussian families, 'GaussianSABS' is also available[7]).

Approximating the risk Let us consider the surrogate risk construction task at a fixed π,

$$\text{Find } \eta(\pi) := \arg\inf_{\eta \in \mathbb{R}^{d_\Theta}} \pi\left[(f_\eta - R - \pi\left[f_\eta - R\right])^2\right].$$

As the surrogate PAC-Bayes bound is solved using a closed form expression, the computational bottleneck of algorithm 3 is this approximation task. As f_η is a η weighted linear combination of functions, this is formally a least square weighted linear approximation problem with infinite number of observations, whose solution can be explicitly written as

$$\eta(\pi) = \mathbb{V}_\pi\left[T\right]^{-1} \pi\left[R\left(T - \pi\left[T\right]\right)\right].$$

This solution can be approximated using a finite number of function evaluations $R(\gamma_i)$, replacing the probability π by an empirical counterpart $\pi_{\text{emp}} = \sum_{i=1}^N \omega_i \delta_{\gamma_i}$.

Different choices of (γ_i, ω_i) can be considered. A first approach consists in drawing i.i.d. samples from π_θ and considering uniform weights. This guarantees that the approximated objective is unbiased. A main shortcoming of this approach, however, is that it disregards all previous risk evaluations at each step. The ratio of density weight corrections used in Chapter 3 could be used to salvage samples drawn from $\pi_{\tilde{\theta}}$, all the while guaranteeing unbiased approximated objective. This, as previously noted, has intractable impact on the variance, and thus might not be practical.

We advocate a "generation agnostic" approach for the weighing process, which treats all available risk evaluations in a like manner. Our strategy assumes that $\mathcal{H}$ is a metric space. For all predictors $(\gamma_i)_{i\in[1,N]}$ whose risk $R(\gamma_i)$ is known, target weights $\tilde{\omega}_i$ are defined as the probability given to the Voronoi cell

$$\overline{\gamma}_i = \{\gamma \in \mathcal{H} \mid i = \arg\min d(\gamma, \gamma_i)\}$$

by the measure π, *i.e.* $\omega_i = \pi\left[\mathbb{1}_{\overline{\gamma}_i}\right]$. We estimate this target weight by solving a large number of nearest neighbour search in $(\gamma_i)_{i\in[1,N]}$ for M draws $\tilde{\gamma} \sim \pi$. This result in an unbiased estimation $\hat{\omega}_i$ of the Voronoi cell weights with controlled variance for each cell $\frac{1}{M}(1-\omega_i)(\omega_i)$ and satisfying[8] $\sum_{i=1}^N \hat{\omega}_i = 1$. For further flexibility, we allow the distance d used for the Voronoi cell to depend on the distribution π. For instance, if π is a Gaussian distribution, a natural choice for d would be the Mahalanobis distance. This approach requires, if the empirical distribution

[6]The instance of the 'PreExpFamily' used must contain the implementation of methods specifying how to move from the parametrisation to the natural parametrisation and back

[7]The Voronoi partition technique described in section 4.1.5 uses Mahalanobis distance rather than the default euclidean distance.

[8]This supposes having either chosen some rule to break up distance ties, or dispatching the weights of cell boundaries to different cells (*e.g.* if $d(\gamma, \gamma_i) = d(\gamma, \gamma_j)$, contribute 0.5 to both cell i and j), or knowing that the boundary is π-almost never drawn (typically the case for continuous distributions).

$\sum \omega_i \delta_{x_i}$ is to form an adequate approximation of the distribution π, some fresh queries from to π. The stack of function evaluation is hence appended at each approximation step by evaluating samples from π_θ. As this weight computation can bring some overhead, it is only appropriate when risk queries are the main computational bottleneck.

Our surrogate risk building strategy has two limitations. First of all, the linear regression least square problem has a unique solution only if the number of observations is higher than the number of features, which translates into a condition that the number of predictors γ for which we have a risk evaluation exceeds the dimension of the variational family. Moreover, it is uncertain that the 'generation agnostic' strategy based on Voronoi partition used to define the empirical approximation of measure π will provide a satisfactory approximation in any sense if the dimension of the predictor space is too large, due to concentration of measure. In this case, the distance between two predictors drawn from the distribution π will be almost constant, and as such the notion of proximity in the predictor space no longer meaningful[9].

The linear least square approach could also be solved in a heterogeneous way, by combining the exact computation $\mathbb{V}_\pi[T]$ to an empirical $\pi[R(T - \pi[T])]$, *i.e.* defining η as

$$\mathbb{V}_\pi[T]^{-1} \sum_i \omega_i R(\gamma_i) \left(T(\gamma_i) - \sum \omega_i T(\gamma_i)\right).$$

Such an approach would however loose one of the main benefits of solving the linear least square approach for the empirical distribution $\sum \omega_i \delta_i$, namely the convergence in one step if the risk belongs to $\mathcal{F}$. Early experiments on Gaussian distributions did not exhibit any improved behaviour for the optimisation task for the heterogenous surrogate building, which was not further investigated.

Boundary issues PAC-Bayes bounds typically hold for empirical risk functions satisfying moment bounds (with respect to the data generation mechanism) or boundedness conditions (the latter being usually required for Catoni's bound). Such assumptions might no longer be met for the approximated risks. A consequence is that the minimiser of the surrogate task might not exist. For instance, a local quadratic approximation of the risk near a local maxima can induce a surrogate task whose minima is $-\inf$.

To ensure that for any risk approximation $f_{\eta,C}$, the surrogate solver always defines a probability distribution, two regularisation hyperparameters $\mathrm{kl}_{\max}$ and $\alpha_{\max}$ are introduced. $\mathrm{kl}_{\max} \in \mathbb{R}_+ \cup \{+\infty\}$ determines the maximum step size allowed between two successive posterior estimation, measured in Kullback–Leibler divergence. $\alpha_{\max} \in]0,1]$ acts as a dampening hyperparameter. The corrected update rule is changed to $\tilde{\theta}_c(\theta) = \tilde{\alpha}(\tilde{\theta}(\eta) - \theta) + \theta$ with $\tilde{\alpha}$ the highest $\alpha \leq \alpha_{\max}$ such that $\mathrm{KL}(\tilde{\theta}_c, \theta) \leq \mathrm{kl}_{\max}$. Such $\tilde{\alpha}$ can be easily obtained through a Newton scheme or dichotomy, noticing that it is defined through $f(\tilde{\alpha}) = C$ for a non decreasing function f.

[9] For instance, if one has access to the risks for N draws $(\gamma_i)_{i \leq N}$ of $\mathcal{N}(\gamma_0, \Sigma)$ as well as the evaluation of γ_0, the distance between a new draw γ and γ_i is distributed as $2\chi^2(d_\Gamma)$ while the distance between γ and γ_0 is distributed as $\chi^2(d_\Gamma)$. Since $\chi^2(d_\Gamma)/d$ concentrates around 1 with fluctuations $1/\sqrt{d_\Gamma}$, large values of d_Γ will result in large weight ω_0, implying that the other points are disregarded

We summarize the behaviour of the regularized algorithm in the following theorem.

Theorem 4.5

For $\alpha_{\max} \in]0,1]$ and $\mathrm{kl}_{\max} \in \mathbb{R}_+ \cup \{+\infty\}$, the minimiser of Catoni's PAC-Bayes objective on an exponential family is a fixed point of Algorithm 3 with approximation function

$$\mathsf{F}(\pi_\theta, R) := \arg\inf_{f \in \mathcal{F}} \pi_\theta \left[(R - f)^2 \right],$$

and surrogate solver

$$\widetilde{\mathrm{Solve}}(\pi_{\mathsf{p}}, \theta, f_\eta) := \tilde{\alpha}(-\lambda^{-1}\eta - \theta) + \theta$$

where

$$\tilde{\alpha} := \sup \left\{ \alpha \in [0, \alpha_{\max}] \mid \mathsf{KL}(\pi_{\tilde{\alpha}(-\lambda^{-1}\eta - \theta) + \theta}, \pi_\theta) \leq \mathrm{kl}_{\max} \right\}.$$

Moreover, for all θ,

$$\nabla \mathsf{PB}_{\mathsf{Cat}} \cdot (\widetilde{\mathrm{Solve}}(\pi_{\mathsf{p}}, \theta, \mathsf{F}(\theta, R)) - \theta) \leq 0.$$

Furthermore, for $R \in \mathcal{F}$,

- if $\alpha_{\max} = 1$, $\mathrm{kl}_{\max} = \infty$, the algorithm converges in one step for all initialization,
- if $\alpha_{\max} = 1$, $\mathrm{kl}_{\max} > 0$, the algorithm converges in a finite number of steps,
- if $\alpha_{\max} < 1$, $\mathrm{kl}_{\max} > 0$, the algorithm has geometric convergence.

Proof. First of all, one can remark that the modification of the solving step does not impact the fixed point property of Theorem 4.4 since if $\theta = \theta^*$, $-\lambda^{-1}\eta(\theta^*) = \theta^*$. Moreover, since by construction the solving procedure $\widetilde{\mathrm{Solve}}$ satisfies

$$\widetilde{\mathrm{Solve}}\left(\pi_{\mathsf{p}}, \theta, f_{\eta(\theta)}\right) - \theta = \tilde{\alpha}\left(\mathrm{Solve}\left(\pi_{\mathsf{p}}, \theta, f_{\eta(\theta)}\right) - \theta\right)$$

for $\tilde{\alpha} \geq 0$, it follows that the update direction remains positively correlated to the gradient direction.

If $R = f_\eta \in \mathcal{F}$, the best risk approximation is clearly the perfect approximation f_η, and as such the surrogate problem (and its solution) is identical to the original problem, implying one step convergence for the uncorrected step direction. If $\alpha_{\max} = 1$ and $\mathrm{kl}_{\max} = \infty$, then $\tilde{\alpha} = 1$ (no correction), and hence the corrected step converges in one step.

Since the uncorrected step converges in one step for all θ, this implies that the update direction is always θ^*. Hence all successive posterior estimation θ_i belong to the segment $[\theta_0, \hat{\theta}] := \{(1-t)\theta_0 + t\theta^* \mid t \in [0,1]\}$. Note $\Delta\theta = \theta^* - \theta_0$. Since the normalisation function g is strictly convex, it follows that the function $t \mapsto \Delta\theta \cdot \nabla g(\theta_0 + t\Delta\theta)$ is non decreasing, and hence, for all t,

$$\Delta\theta \cdot \nabla g(\theta_0) \leq \Delta\theta \cdot \nabla g(\theta_0 + t\Delta\theta) \leq \Delta\theta \cdot \nabla g(\hat{\theta}).$$

Using the convexity of g, this implies that for $t_1 < t_2$, $g(\theta_0 + t_1\Delta_\theta) - g(\theta_0 + t_2\Delta_\theta) \leq (t_1 -$

$t_2)\Delta_\theta \cdot \nabla g(\theta_0)$ while $(t_2 - t_1)\Delta_\theta \cdot \nabla g(\theta + t_2\Delta\theta) \leq (t_2 - t_1)\Delta_\theta \cdot \nabla g(\hat{\theta})$.

It follows that for all $t_1 < t_2$,

$$\mathrm{KL}(\theta_0 + t_2\Delta\theta, \theta_0 + t_1\Delta\theta) \leq (t_2 - t_1)\Delta\theta \cdot (\nabla g(\hat{\theta}) - \nabla g(\theta_0)).$$

This implies that for $\theta_i = \theta_0 + t_i\Delta\theta$, $\theta_{i+1} = \theta_0 + t_{i+1}\Delta\theta$, if the condition $\mathrm{KL}(\theta_{i+1}, \theta_i) \leq \mathrm{kl}_{\max}$ is active, then

$$t_{i+1} - t_i \geq \frac{\mathrm{kl}_{\max}}{\Delta\theta \cdot (\nabla g(\theta^*) - \nabla g(\theta_0))}.$$

Since $t_{i+1} - t_i \geq 0$ and for all i, $t_i \leq 1$, this implies that the condition is active at most $\frac{\Delta\theta \cdot (\nabla g(\theta^*) - \nabla g(\theta_0))}{\mathrm{kl}_{\max}}$ time, i. e. a finite number of time. In the case of $\alpha_{\max} = 1$, this implies convergence in a finite number of steps. For $0 \leq \alpha_{\max} < 1$, this implies that after some K, $t_{i+K} = (1 - \alpha_{\max})^i(1 - t_K)$, and hence geometric convergence of (θ_i) to $\hat{\theta}$. □

The convergence rates provided in theorem 4.5 do not consider the empirical risk approximation step and assume that the surrogate objective uses the exact risk surrogate. The surrogate risk approximation technique described above returns the empirical risk when it belongs to the approximation class. Indeed, this is always the case when replacing the expectation with respect to π by an empirical counterpart with a sufficient number of data points to ensure that a single minimiser exist. In that setting, the empirical risk is guaranteed to be optimal, since it belongs to the class and achieves 0 error. However, for the heterogenous case where closed form expression for the variance $\mathbb{V}_\pi[T]$ are used, the surrogate risk returned will no longer be the empirical risk for a finite number of samples γ_i, and hence the convergence rates no longer hold.

Implementation details SuPAC-CE algorithm is summarized in the pseudo code algorithm 4. A graphical description of the algorithm is provided in fig. 4.2.

Algorithm 4 Surrogate Catoni solver for exponential families (SuPAC-CE)

Require: $\lambda > 0$, $\theta_0 \in \Theta$, $\theta_p \in \Theta$, $R \in \mathcal{M}(\mathcal{H})$, $\mathsf{Ev} = (x_i, R(x_i))_{i=1}^n$, $0 < \alpha_{\max} \leq 1$, $0 < \mathrm{kl}_{\max}$

$\theta \leftarrow \theta_0$

while not converged **do**

Draw i.i.d. $x_{n+1}, \ldots, x_{n+k} \sim \pi_\theta$

$\mathsf{Ev}, n \leftarrow \mathsf{Ev} \cup ((x_{n+1}, R(x_{n+1})), \ldots, (x_{n+k}, R(x_{n+k}))), n + k$

$\omega_i \leftarrow \pi[\overline{x}_i]$ ▷ Solving nearest neighbour problems

$\eta^*, C = \arg\inf_{\eta, C} \sum_{i \leq n} \omega_i (T(x_i) - R(x_i) - C)^2$

$\delta\theta = \theta_0 - \lambda^{-1}\eta^* - \theta$

$\tilde{\alpha} \leftarrow \sup\{\alpha \mid \alpha < \alpha_{\max}, \mathrm{KL}(\theta + \alpha\delta\theta, \theta) \leq \mathrm{kl}_{\max}\}$

$\theta \leftarrow \theta + \tilde{\alpha}\delta\theta$

end while

To fit our new setting, classes for exponential families of distributions were introduced in our 'picproba' package, and implementation of the Gaussian family classes modified accordingly. A modular and generic solver class for the minimisation of Catoni's PAC-Bayes bound

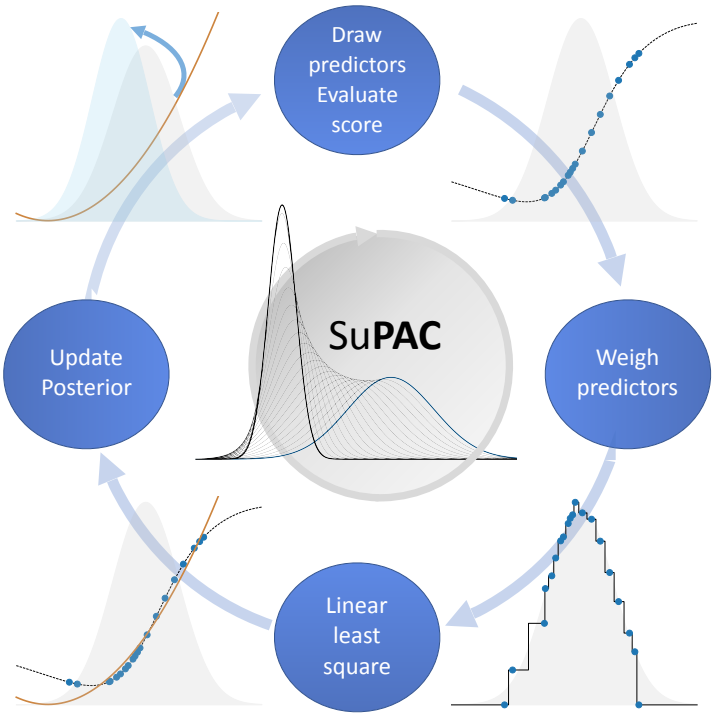


Figure 4.2: Overview of SuPAC-CE in a one dimensional setting. At each step, some new predictors are drawn from the current posterior approximation (grey area) and evaluated (top right figure). All evaluated predictors are then weighted according to the weight of their Voronoi cell (bottom right figure). These weighted evaluations are used to construct an optimal approximation of the risk through a linear least square task (bottom left figure). The approximated risk is used to update the posterior (light blue area) using a closed form expression (top left figure). This procedure is looped until convergence (center).

on exponential families was introduced, as well as more specific implementations for probability families outputting Gaussian distributions, using the Mahalanobis distance when approximating the weights. These solvers rely on closed form expressions for the Kullback–Leibler divergence and its derivative, inferred from the normalisation function and its derivatives.

The default weighing approach for the risk approximation uses exact 1-NN for a user specified number of samples ("n_estim_weights" argument), performed using Faiss library [Douze et al., 2024]. Another weight approximation method, relying on approximate k-NN solving, is also provided.

The corrected update rule parameter $\tilde{\alpha}$ is estimated by dichotomy, using the fact that for all θ, $\delta\theta$, the function $\alpha \mapsto \mathrm{KL}(\theta + \alpha\delta\theta, \theta)$ is not decreasing. The resulting $\tilde{\alpha}$ is guaranteed to result in a Kullback–Leibler step of less than $\mathrm{kl}_{\max}$.

Optimisation of Maurer-Langford-Seeger's objective

While designed to optimise Catoni's bound, SuPAC-CE can be used as a tool to optimise the tighter, but less tractable MLS bound, by alternating optimisation on a Catoni-like objective and optimisation on an intermediary temperature parameter.

Let us recall the MLS objective: for n i.i.d. observations z_i, risks $R = \frac{1}{n}\sum_{i=1}^{n} \ell_{z_i}$ with $0 \leq \ell_{z_i} \leq 1$, for any $0 < \delta < 1$, with probability at least $1-\delta$, for any posterior distribution $\pi \ll \pi_p$,

$$\mathsf{PB}_{\mathsf{MLS}}(\pi, \pi_{\mathsf{p}}, R, n, \delta) := \mathsf{kl}^{-1}\left(\frac{1}{n}\left(\mathsf{KL}(\pi, \pi_p) + \log\left(\frac{\xi(n)}{\delta}\right)\right), \pi\left[R\right]\right) \tag{4.6}$$

where $\xi(n) = 2\sqrt{n}$ and kl^{-1} is an inverse of kl the Kullback–Leibler divergence between two Bernoulli distributions defined as $\mathsf{kl}^{-1}(r,p) = \sup\{q \leq 1, \mathsf{kl}(q,p) \leq r\}$. Germain et al. [2009] notes that for $0 \leq q \leq p < 1$, $\mathsf{kl}(q,p)$ satisfies

$$\mathsf{kl}\left(q,p\right) = \max_{C \geq 0} \left\{-\log\left(1 - p\left(1 - \exp\left(-C\right)\right) - Cq\right)\right\}. \tag{4.7}$$

This (4.7) can be plugged in the MLS bound (4.6), and, after the change of variable $\lambda = C^{-1}$, this implies the following identity for MLS

$$\mathsf{PB}_{\mathsf{MLS}}(\pi, \pi_{\mathsf{p}}, R, n, \delta) = \inf_{\lambda > 0} \widetilde{\mathsf{PB}}_{\mathsf{MLS}}\left(\pi\left[R\right], \mathsf{KL}(\pi, \pi_{\mathsf{p}}), \lambda, n, \delta\right)$$

where

$$\widetilde{\mathsf{PB}}_{\mathsf{MLS}}\left(r, kl, \lambda, n, \delta\right) := \frac{1 - \exp\left(-\lambda^{-1}\left(r + \frac{\lambda}{n} kl\right) - \frac{\log(\xi(n)/\delta)}{n}\right)}{1 - \exp\left(-\lambda^{-1}\right)}.$$

At a fixed temperature $\lambda > 0$, minimising $\widetilde{\mathsf{PB}}_{\mathsf{MLS}}$ on the posterior is equivalent to minimising the objective defined by Catoni's bound. At a fixed posterior π, the right hand side is a smooth function of the temperature, and derivatives of arbitrary orders can be computed if $\mathsf{KL}(\pi, \pi_p)$ and $\pi\left[R\right]$ are known. Searching for the minimiser of MLS's bound in an exponential family leads us to Algorithm 5, relying on SuPAC-CE for optimisation at a given temperature.

We stress that in practical implementations of Algorithm 5, the hyperparameters of SuPAC-CE should be modified after the first optimisation procedure to lower the number of risk queries. Indeed, the risk queries from previous optimisation procedure (conducted at other temperature) are used in new optimisation procedures through the *generation agnostic* weighing process.

Moreover, the two level optimisation strategy pursued by Algorithm 5 relies on a greedy optimisation of the PAC-Bayes temperature. While this might lead to convergence to a local minima, it has the benefit of guaranteeing that the objective can only decrease during the optimisation procedure (as long as proper safeguards are put into place when using SuPAC-CE, *e.g.*, by adapting the step size).

More generally, SuPAC-CE can be used for the minimisation of generic Kullback–Leibler

Algorithm 5 Surrogate PAC-Bayes Learning for MLS (SuPAC-MLSE)

Require: $\theta_0 \in \Theta$, $\theta_p \in \Theta$, $R \in \mathcal{M}(\mathcal{H})$, $\lambda_0 \in \mathbb{R}^+$, $n \in \mathbb{N}$, $\delta \in [0,1]$
  $\theta \leftarrow \theta_0$
  $\pi \leftarrow \pi_0$
  $\lambda \leftarrow \lambda_0$
  $\mathsf{Ev} = ()$ ▷ No evaluations
  while not converged **do**
    $\theta, \mathsf{Ev} \leftarrow \text{SuPAC-CE}(\theta, \theta_p, R, \lambda, \mathsf{Ev})$
    $r \leftarrow \pi_\theta[R]$ ▷ Use Ev
    $kl \leftarrow \mathsf{KL}(\pi_\theta, \pi_{\theta_p})$ ▷ Closed form
    $\lambda \leftarrow \arg\inf_{\lambda>0} \widetilde{\mathsf{PB}}_{\mathsf{MLS}}(r, kl, \lambda, n, \delta)$ ▷ e.g. Newton, Householder
  end while

penalised PAC-Bayes bounds on exponential family. Indeed, assuming that the bound of interest is of form $\mathsf{PB}(\pi, R, \pi_\mathsf{p}, \xi) = F(\pi[R], \mathsf{KL}(\pi, \pi_\mathsf{p}), \xi)$ with F decreasing in its first two arguments, then it follows that the minimizer of the bound on θ must be on the Pareto front of $\theta \to (\pi_\theta[R], \mathsf{KL}(\pi_\theta, \pi_\mathsf{p}))$, which is covered by the solutions of SuPAC-CE for varying temperatures[10].

4.1.6 Experiments

On a synthetic Rosenbrock risk

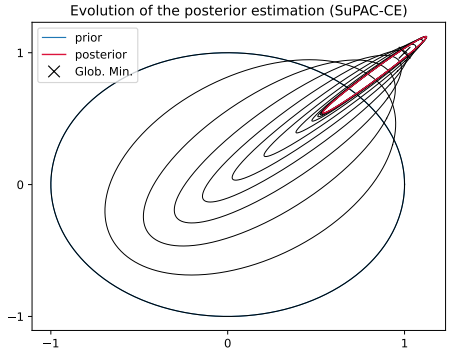
We first compare VarBUQ to SuPAC-CE on a modified two dimensional Rosenbrock function. The risk is defined as

$$R\left(\begin{pmatrix}\gamma_1 & \gamma_2\end{pmatrix}\right) = 5 \times \arctan\left(0.2 \times (\gamma_2 - 1)^2 + 10 \times (\gamma_2 - \gamma_1)^2\right), \tag{4.8}$$

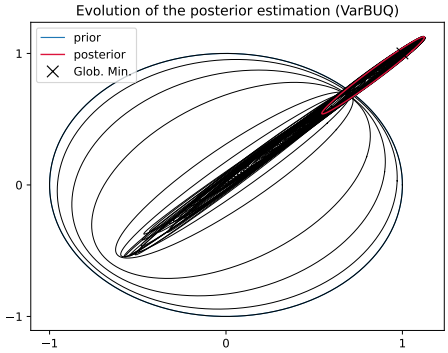
and we consider the minimisation of the resulting Catoni's objective on Gaussian distributions, with a standard Gaussian prior, for PAC-Bayes temperatures of $\lambda = 0.2$ and $\lambda = 0.05$.

The new method's training path exhibits superior stability (see Figure 4.3) and constructs a distribution with better performance (*i.e.*, the posterior distribution constructed has a smaller Catoni's bound). The improvement of the procedure is more notable for smaller PAC-Bayes temperature λ (*i.e.* when the learning rate is important). For $\lambda = 0.2$, SuPAC-CE obtained a Catoni's bound of 1.0359 ± 0.00011 while VarBUQ obtained a Catoni's bound of 0.10392 ± 0.00012, implying that SuPAC-CE obtained only a marginally better result. For $\lambda = 0.05$, SuPAC-CE obtained a bound of 0.331051 ± 0.000025 while VarBUQ obtained a bound of 0.87770 ± 0.00034. It appears VarBUQ had difficulty when the posterior distribution must exhibit high correlations to obtain low mean risk, as is the case in the test Rosenbrock function. Such high correlations are usually due to low joint identifiability of parameters. The performance decrease can be explained by the choice of parametrization used for the Gaussian distributions in 'picproba' – namely, the (matrix)

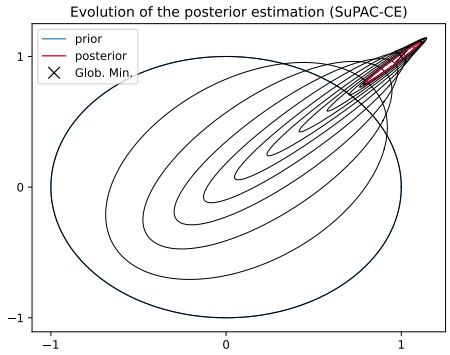
[10] The same argument implies that the minimizer on all distributions of such PAC-Bayes bounds must be a Gibbs posteriors



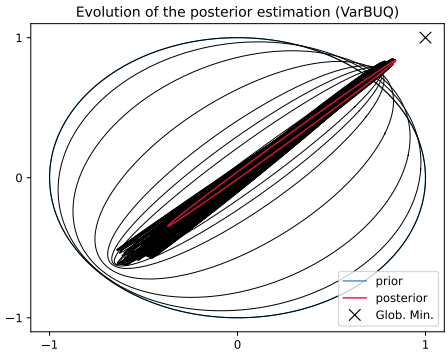
(a) SuPAC, $\lambda = 0.2$



(b) VarBUQ, $\lambda = 0.2$



(c) SuPAC, $\lambda = 0.05$



(d) VarBUQ, $\lambda = 0.05$

Figure 4.3: Posterior evolution for $\lambda = 0.2$ and $\lambda = 0.05$, using Rosenbrock risk, for algorithms SuPAC-CE and VarBUQ. Grey ellipses represent confidence regions of level 0.393 for the posterior approximation during the different steps of training. The minimiser of the risk is (1,1), and the posterior's confidence region is expected to contract around this value. For $\lambda = 0.2$, the posterior distributions obtained by both approaches are similar, although the path of SuPAC-CE is smoother, and convergence faster. For $\lambda = 0.05$, VarBUQ failed to shift its mass towards the minima.

square root of the covariance matrix – the covariance matrix having typically a large conditioning number, it can be difficult to find an adequate step size for the gradient. On the other hand, the update rule for the posterior parameter of SuPAC-CE is done in the natural parametrization of the exponential family, which is better behaved (*e.g.* it is a convex space, the Kullback-Leibler divergence is a Bregman divergence)[11].

The number of function calls was considerably smaller for SuPAC-CE (for the PAC-Bayes temperature of 0.05, convergence was achieved in 27×320 calls compared to lack of convergence after 500×160 calls for VarBUQ). Such behaviour is not wholly surprising, considering that the empirical risk is approximately quadratic. For a true quadratic functions, SuPAC-CE with no regularisation would converge in a single step.

Shortcomings of the surrogate strategy

We investigate failure modes of SuPAC-CE on toy exemples. While the minimiser of Catoni's bound is guaranteed to be a fixed point, SuPAC-CE can still fail to converge to it. Two such settings are explored here: a case where the approximated posterior sequence oscillates, and a case where the approximated posterior sequence converges to a local minima of Catoni's bound. These failure modes of SuPAC-CE are similar to those expected for Gradient Descent.

Oscillations An oscillation instability was observed using the Rosenbrock risk of eq. (4.8) in conjunction with minimisation of Catoni's bound on Gaussians with covariance matrix known up to a multiplicative factor. The family of distributions $\left\{\mathcal{N}\left(\mu, \beta\Sigma_0\right) \mid \mu \in \mathbb{R}^{d_\Gamma}, \beta \in \mathbb{R}_+\right\}$ is an exponential family. The resulting family of approximation functions are quadratic functions with Hessian collinear to Σ_0^{-1}. As a result, the empirical risk no longer approximately belong to the set of approximation function. The covariance chosen was $\Sigma_0 = \begin{bmatrix} 0.5 & 0.4 \\ 0.4 & 0.5 \end{bmatrix}$. Oscillations were observed using a dampening factor of $\alpha_{\max} = 0.9$ and a maximum Kullback–Leibler step of $\mathrm{kl}_{\max} = 0.04$ (see Figure 4.4).

The oscillation phenomena can be fixed by modifying the regulation hyperparameters (see Figure 4.4). Keeping the maximum Kullback-Leibler step size at $\mathrm{kl}_{\max} = 0.04$, the oscillation phenomena, still perceptible for a dampening parameter of $\alpha_{\max} = 0.18$, has disappeared for a dampening parameter of $\alpha_{\max} = 0.16$. Lowering the maximum Kullback-Leibler step size while maintaining the dampening parameter to $\alpha_{\max} = 0.9$ proved less efficient in removing the oscillations, which diminish in amplitude but are still visible for maximum step size as low as $\mathrm{kl}_{\max} = 0.0005$.

[11] The parametrisation of 'picproba' was designed in such a way as to ensure that almost all $\theta \in \mathbb{R}^D$ defines a valid distribution (see section 3.3.1). This design choice is effective in preventing Gradient Descent procedure from failing with a fixed step size. However, this strategy has its drawbacks. Consider a simple one dimensional case where the variance is $(0.1)^2$, parametrized by 0.1, and the step -0.5. The key message of the step is that the covariance should contract. However, after the step, the parametrisation becomes -0.4, resulting in a much higher variance of $(0.4)^2$. Such behaviour is no longer possible when considering the natural parametrisation.

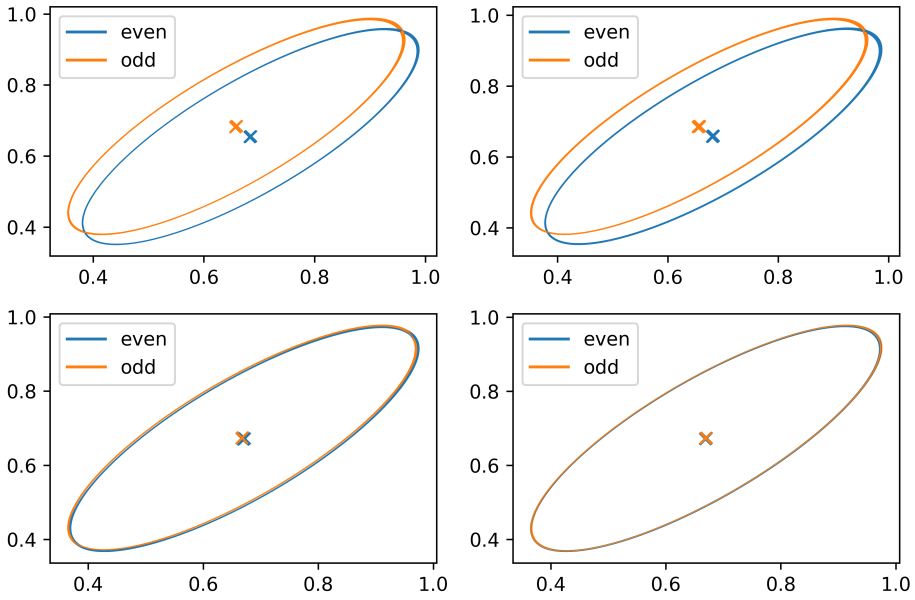


Figure 4.4: Oscillations occurring during SuPAC-CE minimisation algorithm. Ellipses represent confidence regions of the posterior distribution of identical level. From upper left, reading clockwise: $\alpha_{\max} = 0.9$, $\mathrm{kl}_{\max} = 0.04$; $\alpha_{\max} = 0.18$, $\mathrm{kl}_{\max} = 0.04$; $\alpha_{\max} = 0.16$, $\mathrm{kl}_{\max} = 0.04$, $\alpha_{\max} = 0.9$, $\mathrm{kl}_{\max} = 0.0005$. A phase transition between oscillating and converging procedures seems to occur when modifying $\alpha_{\max}$. On the other hand, $\mathrm{kl}_{\max}$ impacts the amplitude of the oscillations.

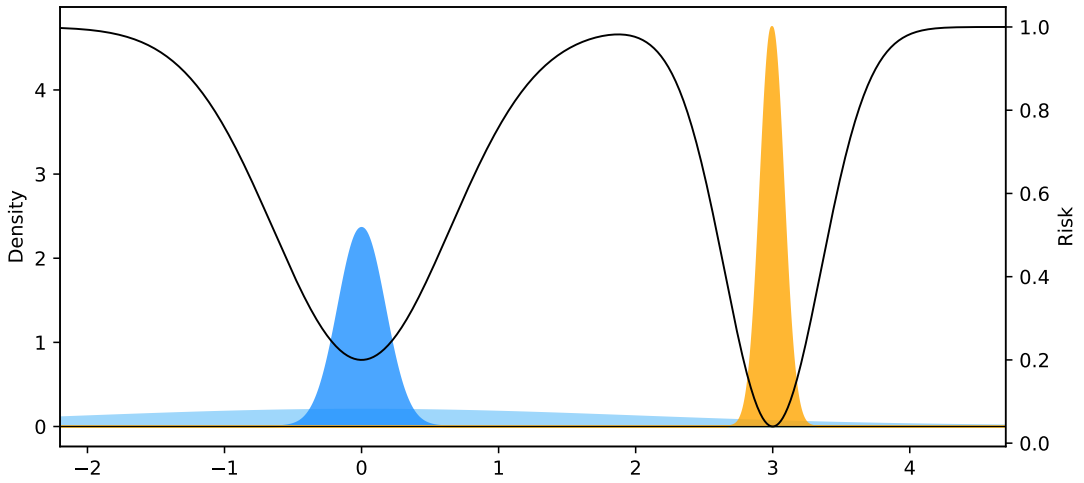


Figure 4.5: Convergence of SuPAC-CE to a local minima for multimodal risks. The one dimensional risk function is represented by the black line. The prior density is represented in light blue. The posterior returned by SuPAC-CE's density is represented in blue. The posterior minimising the learning objective is represented in orange. SuPAC-CE fails to shift the centre of mass of the posterior distribution towards the risk minimiser $\gamma = 3$, although the oracle posterior is centred around it.

Convergence to a local minima As with GD approaches, SuPAC-CE might fail to discover the global minima of Catoni's bound and remain stuck in a local minima. We illustrate this issue using a new synthetic risk function with 2 modes, in a 1-dimensional setting:

$$R(\gamma) = \frac{1.25 - 1.2\exp(-4 \times (\gamma - 3)^2) - \exp(-1.2\gamma^2)}{1.25}$$

The global minima of this function is achieved close to $\gamma = 3$, implying that for small enough PAC-Bayes temperature, the minimiser of Catoni's bound should concentrate around this value. We consider a standard Gaussian $\mathcal{N}(0, 1)$ as prior. This prior distribution is centred around a local minima of the risk. For $\lambda = 0.05$, the algorithm, starting with initial guess of posterior the prior, converges to a distribution centred around 0 ($\mu = -2.5e - 7$ with low deviations $\sigma = 0.17$, achieving a Catoni's bound of 0.325. Using the knowledge of the true global minima of the risk, the calibration procedure started from posterior guess $\mathcal{N}(3, 1)$ converges to a distribution centred around 3 ($\mu = 2.99$) with low deviations ($\sigma = 0.084$), achieving a lower Catoni's bound of 0.256 (see Figure 4.5).

The algorithm therefore can fail in multimodal settings to find the global minimiser of Catoni's bound. Two reasons could explain this failure: either an issue with exploration (*i.e.*, the algorithm was unaware of the global minima close to 3.0 because it never evaluated the risk in this area), or an intrinsic limitation of the routine (the algorithm would fail even if the risk approximation problem could be perfectly solved at each step). Further tests, where SuPAC-CE was fed information on the risk close to the global minima, showed that the first reason could be disregarded. This confirms that the routine can get stuck in a local minima by conception. Indeed, all local minima of the PAC-Bayes bound are also fixed points of both GD based approaches and SuPAC-CE.

For ADM1 calibration

SuPAC-CE was assessed on the learning task studied in Chapter 3. Catoni's PAC-Bayes objective is minimised on Gaussian distributions with block diagonal covariance in order to calibrate 30 parameters of model ADM1 [Batstone et al., 2002a]. ADM1 relies on solving a stiff ODE to predict the evolution of the states, and is therefore quite computationally intensive, at about 3 seconds per model query in our experiments[12].

The PAC-Bayes temperature was set to 0.002 (this corresponds to $2\lambda_{\text{ADM1}}$). Mean risks were assessed at test time by resampling new predictors from the posterior. Training procedures were repeated 20 times.

We compared SuPAC-CE to standard GD on the dataset LF from Chapter 3 using the same family of distributions and risk function. A maximal budget of 9 600 empirical risk queries was fixed. For SuPAC-CE, the regularisation hyperparameters were set to $\mathrm{kl}_{\max} = 1$ and $\alpha_{\max} = 0.5$,

[12] As 196 days were simulated, this is five times smaller than the average 15 seconds per model query observed in Chapter 3. This is due to improvements in the implementation of ADM1, which are described in Section 4.2. Note that the speed up is not as massive for ADM1 as it will prove to be for ProdAD.

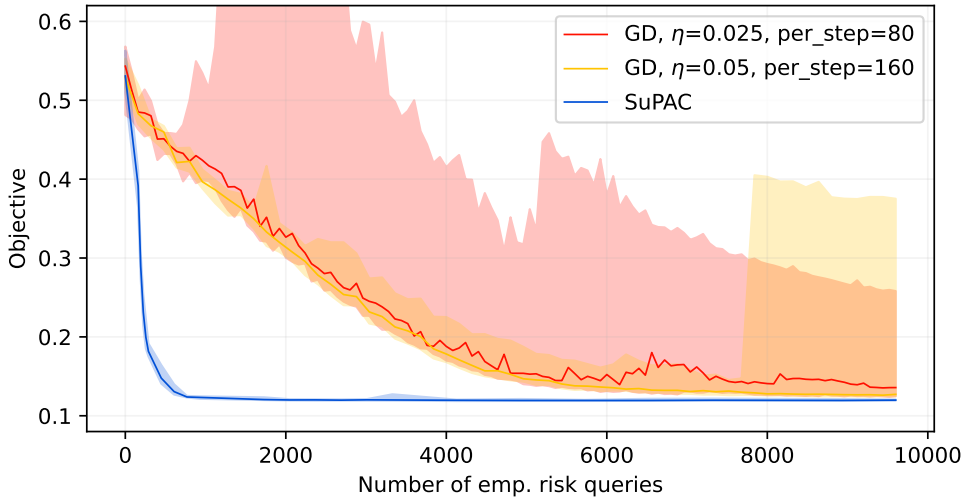


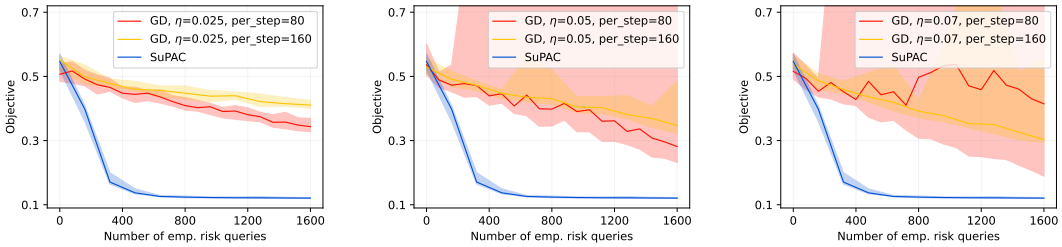
Figure 4.6: Comparison of the optimisation procedures as performed by SuPAC-CE and gradient descent (GD) for two selected sets of hyperparameters on an ADM1 calibration task. Each optimisation procedure was repeated 20 times; the median performance and 0.2 and 0.8 quantiles are represented. SuPAC-CE was performed with hyperparameters $\alpha_{\max} = 0.5$ and $\mathrm{kl}_{\max} = 1$.

while the number of samples generated to evaluate the weights was set to 40 000. The optimisation algorithm was trained on 296 steps; for the initial step, 160 risk queries were performed, while for all the remaining steps, 32 risk queries were performed. This larger number of queries for the initial step is due to the necessity of having a least more evaluations than the dimension of the family of probability.

Hyperparameters for GD were selected after assessing the grid $(\text{per\_step}, \text{step\_size}) \in \{80, 160\} \times \{0.025, 0.05, 0.07\}$ on a preliminary 1 600 risk queries budget, with 20 repeats. The larger step size 0.07 was rejected due to its erratic behaviour between repeats, obtaining both optimal and worse GD performance. This erratic behaviour was also observed for step size 0.05 when estimating gradients from 80 risk queries. On the other hand, for per_step set to 160, the step size of 0.025 clearly under-performed compared to the step size of 0.05, although this was slightly more stable. This led to the selection of the two sets of hyperparameters, (per_step=80, step_size=0.025) and (per_step=160, step_size=0.05), which had similar performances. Both were assessed, and the set of hyperparameters obtaining the lowest risk, (per_step=160, step_size=0.05), was kept for comparison (see fig. 4.6).

The performance of the sequence of posteriors were compared by aligning the number of empirical risk queries. Indeed, the main motivation of SuPAC-CE is the setting when querying the empirical risk is computationally expensive, and can be assumed to be the computational bottleneck. This is indeed the case for the AD example considered here. At equal number of risk queries, SuPAC-CE required on average an extra 3.5% processing time compared to gradient descent, mainly caused by the weighing process.

SuPAC-CE proved significantly more efficient at minimising the bound than GD (see Fig-



(a) η = 0.025 **(b)** η= 0.05 **(c)** η = 0.07

Figure 4.7: Preliminary gradient descent optimisation procedures for different choices of hyperparameters. The evaluations of each optimisation procedure was repeated 20 times; the median performance and 0.2 and 0.8 quantiles are represented. The performance of SuPAC-CE (with hyperparameters $\alpha_{\max} = 0.5$ and $\mathrm{kl}_{\max} = 1$) is given for comparison.

ure 4.6). The average performance of SuPAC-CE proved better after 1 800 queries than the best performance obtained after the full 9 600 queries for GD. The experiments also indicate that SuPAC-CE offered much higher stability compared to GD, both during training and between the training duplicates. This could be attributed to the "generation agnostic" weighing approach, which relies on all previous risk evaluations at each step and is thus more stable. On the other hand, the noisy gradients estimates have some probability of leading to problematic steps during GD, leading to sharp increase in the objective. In the experiments, 4 out of 20 GD procedures thus led to a worse performance than the one obtained by a single optimisation step of SuPAC-CE. The posterior distributions constructed through SuPAC-CE obtained an average empirical risk of 0.102 ± 0.003, similar to the 0.101 value from table 3.7. The resulting PAC-Bayes bound proved also similar (0.121 ± 0.004 vs. 0.122). Thus SuPAC-CE constructed as good a posterior as VarBUQ, but with twenty times less evaluations of the risk (convergence in approx. 2 000 evaluations of the empirical risk versus 40 000 for VarBUQ).

The impact of SuPAC-CE's hyperparameters was investigated by running further optimisation procedures with different choices of hyperparameters. A grid was assessed, with values of $\mathrm{kl}_{\max}$ in $\{0.5, 1, 2\}$ and $\alpha_{\max}$ in $\{0.25, 0.5, 0.75\}$, with each optimisation process repeated ten times (see fig. 4.8). The resulting optimisation procedures proved to all have similar performances, with only a slight decrease in speed in the early phase between the most regularized and less regularized hyperparameters which was below the noise level after the fourth optimisation step (see fig. 4.8). Two further sets of slow hyperparameters values ($(\mathrm{kl}_{\max}, \alpha_{\max}) \in \{(0.1, 0.9), (0.01, 0.9)\}$) and fast hyperparameters values ($(\mathrm{kl}_{\max} = 5, \alpha_{\max} = 0.1)$, $(\mathrm{kl}_{\max} = 10, \alpha_{\max} = 0)$) were also assessed, with 8 repeats (see fig. 4.9). The slow hyperparameters led to more stable and reproducible optimisation procedures. For the small maximum step size of $\mathrm{kl}_{\max} = 0.01$, the average performance of the optimisation process was similar (*i.e.* difference below the noise level) to the performance of the optimisation process with standard hyperparameters after 2000 risk queries. The highest maximal step size assessed of $\mathrm{kl}_{\max} = 10$ resulted in a final average PAC-Bayes bound of 0.147 ± 0.022, with a standard deviation between runs of

0.061, significantly higher than the standard deviation for the standard hyperparameters (0.0032, p-value of $1.95e-09$).

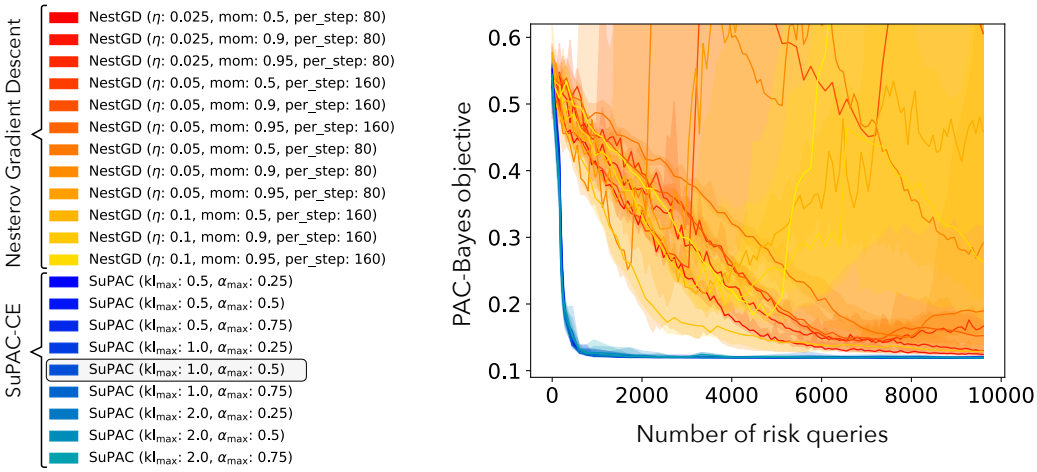


Figure 4.8: Comparison of SuPAC-CE with Nesterov accelerated gradient descent for a variety of hyperparameters choices. Each optimisation procedure was repeated 8 times; the median performance and 0.2 and 0.8 quantiles are represented. SuPAC-CE proved to be consistently more efficient for all hyperparameters values tested. The hyperparameter for SuPAC-CE assessed in fig. 4.6 is highlighted.

SuPAC-CE was further compared to Nesterov accelerated gradient descent as implemented in chapter 3. Starting from the two sets of hyperparameters preselected for GD, optimisation procedures using a momentum of 0.5, 0.9 and 0.95, and either the original step size or twice the step size were assessed. Each of these 12 new optimisation procedures was repeated 8 times, and compared to SuPAC-CE (see fig. 4.10). For no choice of hyperparameter values did Nesterov accelerated GD prove more efficient than SuPAC-CE (fig. 4.8). The increase of step size in conjunction with the moderate momentum improved the speed of the optimisation procedure, but at the cost of a higher risk of optimisation failure, leading to 3 out of 8 runs (resp. 2 out of 8 runs) for 160 simulations per step (resp. 80 simulations per step) with a final objective higher than the initial objective. Higher momentum led to major instabilities, with less than 3 runs out of 8 managing to reduce the objective below 0.2 (compared to 0.121 obtained by SuPAC-CE) for all hyperparameter combinations. For the original step size, momentum appeared to improve the stability of the procedures for all setting except moderate momentum for a per step hyperparameter of 80. Higher momentum procedures led to a speed decrease, caused by the larger number of steps necessary for momentum to build up.

Computations were performed using Azure Machine Learning compute clusters with 32 cores and Intel Xeon Platinum 8272CL processors.

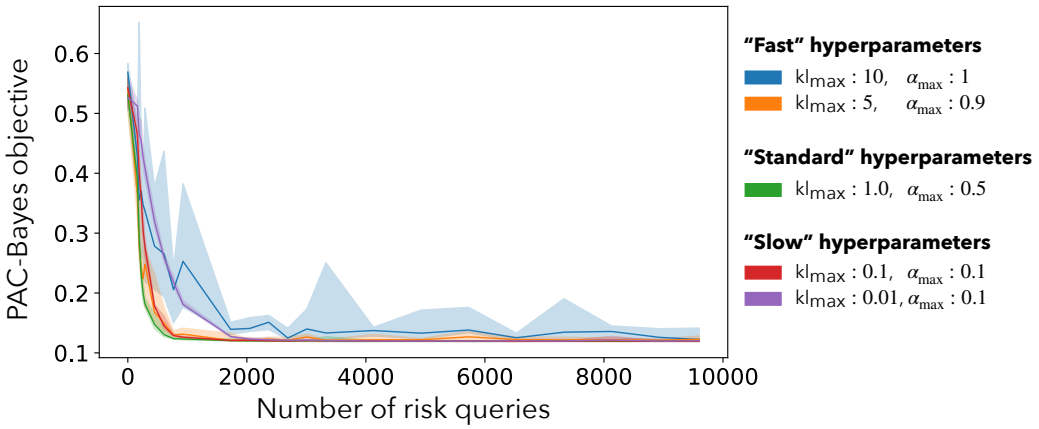


Figure 4.9: Performance of SuPAC-CE with extreme hyperparameters values. Each optimisation procedure was repeated 8 times; the median performance and 0.2 and 0.8 quantiles are represented. SuPAC-CE exhibited noticeable instabilities and speed loss for hyperparameters leading to insufficient regularization (blue curve). Too much regularisation lead to speed decrease in the early phase of the optimisation procedure (purple curve)

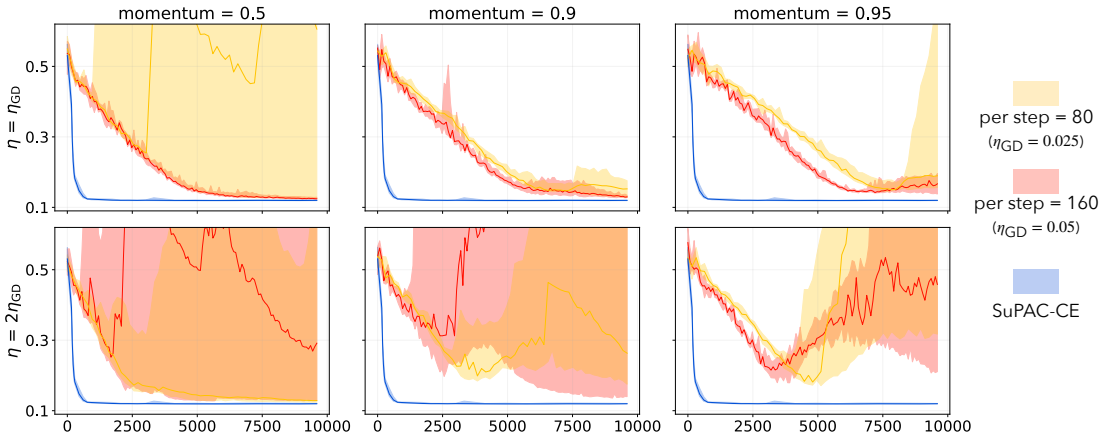


Figure 4.10: Comparison of the optimisation procedures as performed by SuPAC-CE and Nesterov accelerated gradient descent (x axis: number of empirical risk queries). Each optimisation procedure was repeated 8 times; the median performance and 0.2 and 0.8 quantiles are represented. SuPAC-CE was performed with hyperparameters $\alpha_{\max} = 0.5$ and $\mathrm{kl}_{\max} = 1$. Momentum of 0.5, 0.9 and 0.95 were assessed for Nesterov gradient descent. Both the original step size (η) parameter as well as twice the step size parameter for gradient descent comparisons were investigated. At twice the step size, all momentum accelerated procedures proved unstable. At the original step size, the momentum tended to increase the stability of the procedure at the cost of speed. All Nesterov accelerated gradient descent procedures assessed were slower than SuPAC-CE

4.1.7 Limitations and prospects

Theorem 4.2 shows that it is possible to locally decouple the complexity related to querying the empirical risk and the minimisation of a PAC-Bayes bound. A main motivation for such decoupling is that the approximated risk function define non linear surrogate objectives which might be valid for a wider range of probabilities than the linear approximations offered by the gradients. As a consequence, the surrogate bound solution can be reasonably allowed to be much further away from the current posterior estimation than is the case for GD. A key implementation difficulty remains picking the range of validity, *i.e.* how far away from the current posterior the surrogate solver can be allowed to choose a distribution. Such a choice, formalised in the selection of an adequate surrogate solving algorithm, is analogue to the choice of a step size in gradient descent procedures, and balances the stability and speed of the procedure. Automating the selection of the surrogate validity range offers an exciting prospect for the framework.

The Voronoi cell weighing approach used to solve the approximation problem is equivalent to replacing the empirical risk function by a 1-nearest neighbour trained predictor, and approximating this predictor. Variants following this two step approximation approach could be worth investigating. Notably, an interesting perspective would be to approximate the empirical risk through Gaussian processes, taking inspiration from Bayes Optimisation.

Bayes Optimisation or active learning inspired methods could also improve the sampling procedure used during the learning process. The current strategy evaluates a given number of parameters during each step, drawn at random from the current posterior approximation. This randomised approach might be inefficient. Gaussian process approximation of the empirical risk inform on where the risk value is less known. This in turn can inform on which predictor evaluations might have the largest impact on the resulting posterior distribution, and help avoid querying predictors bringing little information[13]. The uncertainty quantification on the risk function could also help choosing the number of risk queries required at a given step, making this hyperparameter more adaptative.

A key restriction is that our surrogate PAC-Bayes framework is only practicable when the dimension of the predictor space and of the probability family are small (*i.e.* less than a few hundreds). This is due to two factors; first of all, the larger the dimension of the probability family, the larger becomes the approximation space, and hence the more empirical risk evaluations are required. Notably, at least $d_\Theta + 1$ evaluations of the empirical risk are required for probability families of dimension d_Θ. The second factor is that the "generation agnostic" weighing approach described in Section 4.1.5 is unlikely to give adequate performances if $\mathcal{H}$ is high dimensional. This effectively rules out deep learning settings, which have been recently the main focus of the PAC-Bayes community. Still, we believe that PAC-Bayes learning offers meaningful prospects for a wide range of physics, biology or medical inspired problems which involve few parameters

[13] While Bayes Optimisation designs its acquisition function to favour parameters likely to improve a score, an acquisition function for SuPAC-CE should favour predictors leading to drastic changes to the update rule. For instance, let us note $\delta(\gamma, r)$ the difference between the posterior constructed with, and without assuming that $R(\gamma) = r$. We can consider the average posterior update of a predictor γ as the average displacement $\delta(\gamma, r)$ with r following the distribution of $R(\gamma)$ given by the Gaussian process. This could serve as an acquisition function for the selection of predictors γ to query.

and expensive model computations, and therefore can be efficiently trained using our framework. Concrete fields of application of SuPAC-CE include, but are not limited to, fluid dynamics simulations with dimension reduction [Callaham et al., 2021], metabolic models for microbial communities [Cerk et al., 2024] and greenhouse gas emission inverse problems [Nalini et al., 2022].

4.1.8 Conclusion

SuPAC is a generic framework for minimising PAC-Bayes bounds designed to tackle computationally intensive empirical risks for low to moderate dimensional problems such as naturally arise in physical models.

We established that SuPAC was theoretically well supported. We instantiated this framework for the optimisation of bounds on exponential family. Preliminary experiments showed that our framework could significantly reduce the number of empirical risks queries when calibrating a biochemical model, thus opening exciting new fields of applications for PAC-Bayes.

4.2 Building fast Anaerobic Digestion models

The monitoring of AD plant using a computational model requires calibration of such models. Due to the number of parameters to calibrate (more than fifty), such calibration can require thousands of evaluations of the model. Having a computationally efficient model is therefore a key consideration to develop a usable and cost efficient calibration methodology. SUEZ's in-house anaerobic digestion model, ProdAD, is developed in Python. Using the Python language simplifies the interaction with other algorithms developed by SUEZ. Moreover, ProdAD still being in development, Python's flexible nature and ease of implementation simplify testing and integrating model modifications.

Being an interpreted language, Python is however not as computationally efficient as compiled languages such as C++. This seriously impacts the duration of the Bayesian routine selected to calibrate ProdAD, as well as the cost of the calibration.

We present here some implementation modifications designed to

- obtain a computationally efficient implementation of ProdAD model;
- ensure that this implementation can smoothly integrate improvements on the model.

4.2.1 Methodology

The implementation of ProdAD model was modified in two ways:

- the code was checked for duplicate computations,
- the code was modified to accommodate Just In Time (JIT) compilation using package numba.

ProdAD model solves an ODE using a standard Euler method. The inputs are characterized daily, while the step size is set by defaults to 5 minutes. This results in simulations being processed through a double loop scheme. The implementation was modified to ensure that computations which could be cached are not unnecessarily looped.

JIT compilation of the code is performed using the package "numba". Numba provides an elegant framework to offer JIT compilation to standard python functions. At its simplest, numba turns a python function into a JIT compiled function simply by adding the decorator "@numba.jit" to functions. However, to obtain interesting speed ups, the functions follow certain guidelines enabling the so-called "nopython" compilation mode ("@numba.njit" decorator). The code was modified to follow numba's requirements. Numba being optimised for manipulation of arrays, the custom classes as well as dictionary and list objects were removed from the core model code and replaced by array input. Additional layers were added to the code to provide interfaces between the previous inputs and outputs to numba compatible inputs and outputs. As such, the inputs and outputs of the main model code are not modified, and routines built on top of ProdAD model can be used without any further changes.

Some flexibility of the previous implementation was traded off for efficiency and readability of the code. Some arguments of ProdAD model have been turned into fixed global variables. As such, the biochemical reaction network considered by ProdAD is now fixed and can no longer be changed through a configuration file (the source code must be modified). The key characteristics of these biochemical reactions were set to the default values. To ensure stable behaviour when reading the configuration dictionaries - which is necessary to ensure that parameters will be decoded appropriately-, the python version requirement was raised to 3.6.

The core model function, which computes the state a time $t + \delta t$ from time t was modified to take only arrays and standard python objects as inputs (floats, strings, booleans). The dictionaries were removed from the arguments and replaced by the values of keys accessed in the function. In a similar fashion, custom class arguments were replaced by all attributes accessed during the function call. Some inner functions were externalized. Although this is not a mandatory requirement of numba, this simplified the implementation. As required by numba, all functions called by ProdAD_model are also decorated by "numba.njit". Explicit signatures were given to all functions to check that the downstream flow of function is properly written[14].

The outputs of "ProdAD_model" were modified. The previous implementation outputted two to three dictionaries. The current implementation outputs a tuple containing floats and arrays. It should be noted that dictionary outputs are not incompatible with having JIT compiling "ProdAD_model" using numba. However, this would prevent JIT compiling the whole simulation. This causes the code to be slightly less flexible and changes to "ProdAD_model" outputs will require changes to the "run" function as well (unpacking and writing to temporary

[14]This serves both as a way to avoid unexplainable numba compilation error, but proved also useful in making slightly more efficient in two ways. First providing a signature to numba forces compilation when the function is loaded rather than first executed, which means that there is no risk of the compilation step being done in a parallelised call (which is inefficient since the compilation must be done in all cores and every time the parallel call is relaunched as the result is not stored). Moreover, providing a single signature prevents the function from having multiple signatures, which can bring some overhead when calling the function.

memory). The "run" method of the simulation class was remodelled. While the previous version of the run function directly implemented the ODE solver loops, repeatedly calling the core "ProdAD_model", the new version relies on a JIT compiled "run" function which calls the also JIT compiled core "ProdAD_model". As such, the computations no longer use temporary arrays stored as attributes of the simulation class. As of version 0.58.1, numba provides only limited support for exception handling. Exception handling was therefore simplified compared to the previous implementation. In the previous implementation, assertion failure during the core model call function (t to $t + \delta t$ computation) was caught during the simulation function (main loop) and raised again as a custom error containing details on the failure, specifically the time of failure, the state of the digester before failure and the feed information. Due to numba's inability to catch custom exception in a JIT compiled function, the custom error is directly raised, without the fine grain details provided by ProdADFailure.

A similar approach was also applied to both implementations of AM2 and ADM1. Both model relies on more sophisticated ODE solving approach than the Euler scheme used for ProdAD and relies on routines provided in 'scipy.integrate.solve_ivp'. As a result, the function computing the derivative passed to the scipy routine was JIT decorated. The implementation of ADM1 alternates solving an ODE and solving differential algebraic equations. This second part was externalized from the main function call and JIT decorated as well. For ADM1, this two step process results in a loop which is not compiled, resulting in a lower speed up.[15]

4.2.2 Comparison with previous implementation of ProdAD

The package numba is still in development and has yet to have a stable release version. Consequently, its version was pinned to 0.58.1 to ensure reproducible behaviour. The computations were performed using Python version 3.9.18 and numpy version 1.23.5.

The outputs of the new and previous version of the code were compared on a synthetic dataset of 280 days with variable intrant description. No differences were observed beyond a relative level of 1e-15, which can be accounted for by modifications in the order of some computations (*e.g.* $a \times (b + c)$ rather than $(a \times b) + (a \times c)$, leading to slight differences which can build up when solving the ODE).

Simulation time for a single simulation

The new implementation's efficiency was assessed by simulating 280 days with variable intrant description. Computations were performed on a notebook, on the same **V**irtual **M**achine (VM) (4 cores). Simulation with the previous implementation took 146 seconds. Simulation with the new code took 1.5 seconds. The modifications on the code therefore led to a speed up of a

[15] Since writing the initial version of the manuscript, I have modified the implementation of both ADM1 and AM2 so has to have the whole simulation numba decorated. This was done by implementing Runge-Kutta solvers in numba. The newest version of the AD models in the package 'anaerodig' uses this latest implementation, which led to a further speed up of approximately a factor 2 to 3 compared to the implementations relying on the 'scipy' ODE solver. Note that the actual ODE solver has been changed to 'LSODA' to Runge-Kutta (by default of order 4(5), as in the reference Matlab implementation of ADM1).

factor 100. The JIT compilation time lasted approximately 35 seconds and therefore should not be considered an issue.

Simulation time for multiple, parallelised simulations

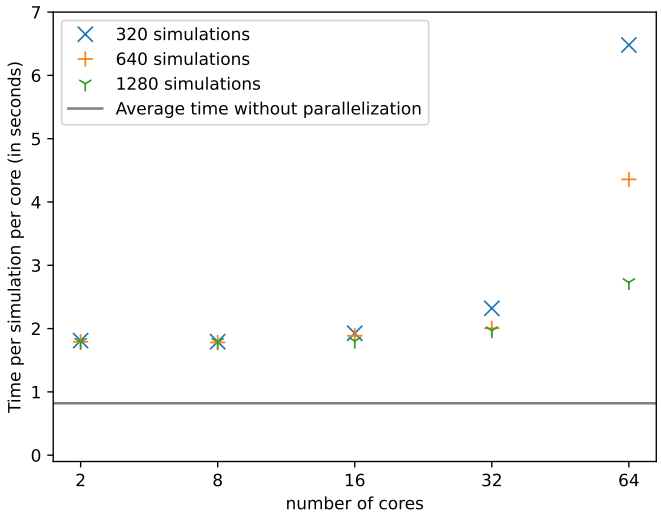


Figure 4.11: Simulation time for the simulation of 196 days using ProdAD model.

The learning algorithm SuPAC-CE benefit from parallelisation, as multiple evaluations of the model can be performed at each calibration step (so called “embarrassingly parallel” setting). The performance of the new implementation of ProdAD in a parallel context was assessed on experiments launched in Azure Machine Learning. The computations were performed on compute optimised VMs with varying number of cores to assess overhead.

The average time for the evaluation of one model on 196 days was 0.83 seconds[16] and remained stable between the VMs of different number of cores without parallelisation. Once parallelised, the average computation per core lasted between 1.78 seconds and 6.48 seconds (see fig. 4.11), from 2 to 8 times higher. Such difference was related to the initialization of computation jobs in each core, denoted overhead. Yet, it is to be noted that the total computation time was still lower using the parallelised routine. For instance, the 64 core VM was able to compute 640 model calls in 40 s (*i.e.*, about 4 s/model/core) while the same calculation lasted 580 s on a 2 core VM (*i.e.*, about 1.8 s/model/core). For a target of 640 parallelised evaluations per calibration step, a good trade off between computation time and performance decline due to overhead can be obtained using up to 32 cores VMs (about 0.4-0.45 performance ratio, *i.e.*

[16]This improves on the computation time of 1.5 seconds for 280 days reported above. The difference is due to the use of compute optimised VM for the evaluation of parallelised simulations, while the performance of a single simulation was assessed on a general purpose compute.

from 1.8 to 2 seconds per model evaluation). For 320 evaluations, the overhead involved in the 32 cores VMs for the parallelisation starts becoming too important (about 2.3 seconds per model evaluation). 64 cores VMs proved inefficient in our setting. Notably, for 320 simulations, the parallelised procedure proved significantly longer on 64 cores VMs compared to 32 cores VMs (32.4 seconds compared to 23.2 seconds)! This hints that numba's compilation for a single model might have been optimised for multiple cores VMs and therefore led to inefficient computations when they had to be performed on a single core. Such an explanation is corroborated by the computation time when parallelisation is handled by numba (performance ratio of 0.39, compared to 0.12 - 0.2 for custom parallelisation procedure).

Parallelisation of the simulations using numba is technically feasible at the price of custom exception handling, which are useful to monitor the behaviour of ProdAD model. As ProdAD is still being actively developed, such a sacrifice is not justified by the small edge numba parallelisation appeared to have on preliminary assessments (performance ratio of 0.4 - 0.49).

4.2.3 Conclusion

The modifications resulted in a code running orders of magnitude faster (100 times faster with no parallelisation, approximately 12 to 50 times faster with parallelisation) with no detectable changes in the outputs. It moreover remained in the user-friendly python language, facilitating further changed by developers familiar only with python and no background in compiled languages. While less flexible than the previous implementation, the massive speed up it brought definitely tipped the scale in its favour.

4.3 Calibrating ProdAD through SuPAC-CE

Capitalizing on our simulation efficient calibration strategy SuPAC and on the speed-up implementation of ProdAD, we assessed the performance of PAC-Bayes trained posteriors on controlled ProdAD calibration tasks. We assessed both the predictive power of the calibrated model, as well as the uncertainty quantification offered by the posterior.

4.3.1 Methodology

The evaluation of the PAC-Bayes posteriors followed the same approach as chapter 3. Fifteen synthetic datasets representing 280 days of operations of single-tank digester were calibrated using noisy descriptions of the first 70% of days. The performance of the calibrated model was measured in term of its prediction error (both on the train set and test set). The performance of the uncertainty quantification technique was measured both in terms of parameter recovery (construction of confidence region covering the true parameter) and in terms of recovery of the test observations. The fifteen synthetic datasets span three types of intrant which differ in term of average HRT (L: average HRT of 30 days, H: average HRT of 15 days, V: average HRT of 30 days on the train data and 21 days on the test data), and 5 draws of parameters from the prior

(denoted A to E). Contrary to the set up considered in chapter 3, these five draws of parameter did not represent different prior plausibility scores, but were simply random draws[17].

The modelled digesters were single tank digesters with 4300 cubic meters of liquid phase and 430 cubic meters of gas phase. The initial state of each digester was computed using steady state simulation with the initial intrant description.

A geometric grid of 4 PAC-Bayes temperatures λ between 0.002 an 0.016 was assessed. The risk function was constructed as in section 3.2.4, using the predictions of VFA, pH, NH4 concentration, gas flows and gas concentrations (for methane and carbonic gas). The softclip function was modified to cap the residuals to 4; its exact form was

$$\text{softclip}(x) = 4\left(1 - \frac{\log\left(1+\exp\left(4 - \frac{(1+\exp(-4))\log(1+\exp(4))}{4}|x|\right)\right)}{\log\left(1+\exp\left(4\right)\right)}\right).$$

The offset η was left to 1e-8 for all predictions except VFA, for which it was raised to 2 gCOD/L (*i.e.* predicted concentrations of VFA much lower than 2 gCOD/L are considered negligible).

SuPAC-CE was used with regularisation hyperparameters $\mathrm{kl}_{\max} = 0.25$, $\alpha_{\max} = 0.3$. A total of 48 optimisation steps were performed, with 640 evaluations of the risk at each step. 25 000 samples from the posterior are used to estimate the weights.

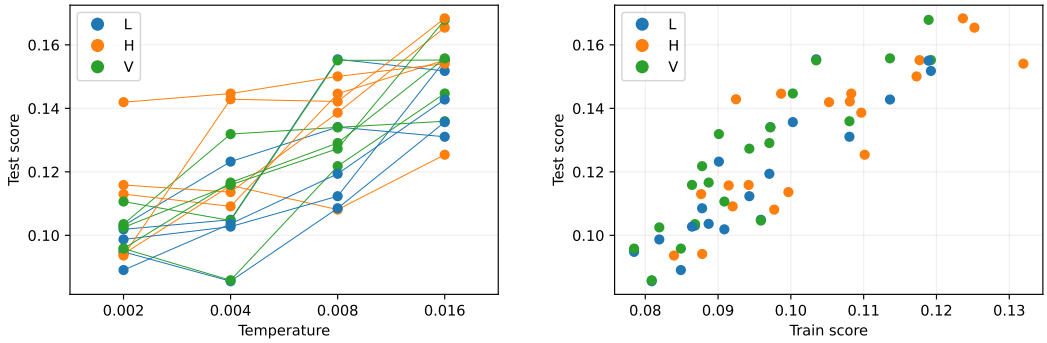
4.3.2 Calibration results

The comparison on ProdAD model is hindered by the fact that the VarBUQ failed to converge, or even to decrease the posterior average risk by any acceptable margin, even after testing multiple combinations of hyperparameters (*e.g.* initial objective after the warm start approach of 0.76, final objective of 1.13). This could be explained by the larger dimension of the problem considered, compared to those tested in Chapter 3 ($d_\Gamma = 73$ compared to 30 for ADM1, 5 for AM2). SuPAC-CE was able to substantially reduce the mean risk, and therefore can be said to have calibrated the model (prior objectives above 2.5 for all datasets, final objectives lower than 0.18 for all datasets and all temperatures).

Computational cost The computational cost was stable between all datasets. Using compute optimised VMs of 16 and 32 cores, the computation time for 48 optimisation steps with 640 risk evaluations per step (196 days modelled) and 25 000 draws for integral estimation, the experiments lasted 1h 20min $\pm$ 2min on the 16 cores VMs (1h 00min $\pm$ 2min for the 32 cores VMs).

Posterior train and test error The prediction performance of the calibration was assessed both on the train set and on the test set, for each PAC-Bayes temperature. As expected, the

[17] The plausibility indicator might not be informative in the context of ProdAD, where, depending on the feed characteristics, only part of the blocks of parameters might be considered active, *e.g.* having an impact on the simulation results. The inactive blocks of parameters would act as confounding factors when assessing prior plausibility, especially for uncertainty quantification on predictions performance indicators.



(a) Test error from λ

(b) Test error from train error

Figure 4.12: Test error as a function of the PAC-Bayes temperature (4.12a) and of the train score (4.12b). The colours represent the type of intrant (L for low intrant flow, H for high intrant flow, V for variable between training set and testing set). For 4.12a, lines link test risks for the same dataset. The test risk of the posterior tends to increase as the PAC-Bayes temperature increases. For 4.12b, a strong empirical correlation of 0.897 is observed, indicating that overfitting did not significantly harm the test performance in the range of PAC-Bayes temperatures considered.

lower the PAC-Bayes temperature, the lower the resulting mean error for the training set. This was also the case for the test set (see Figure 4.12a). Thus, the PAC-Bayes temperature which would be selected through validation considering only the test error would be the lowest, *i.e.* the highest learning rate. The test error was somewhat larger than the train error (from 0.02 to 0.04 added error), indicating mild overfitting. The test error is strongly correlated with the train error, with an empirical correlation of 0.90 (see Figure 4.12b).

4.3.3 Parameter recovery performance

As in Section 3.3.1, credible regions are constructed for posterior $\mathcal{N}(\mu, \Sigma)$ as sets $\text{CR}(\alpha, \mu, \Sigma) = \left\{\gamma \mid d(\gamma) := (\gamma - \mu)^T \Sigma^{-1} (\gamma - \mu) \leq Q_{\chi^2(d_\Gamma)}(\alpha)\right\}$, where $Q_{\chi^2(d_\Gamma)}$ is the quantile function of a $\chi^2(d_\Gamma)$ distribution. This implies that the true parameter γ^* is rejected with confidence α if the statistic $d(\gamma^*)$ is above the threshold $Q_{\chi^2(d_\Gamma)}(\alpha)$. The statistics observed for the different datasets depending on the PAC-Bayes temperature λ are specified in Figure 4.13. Note that the statistics for the datasets of intrant type "V" are not represented, as they exactly match those of the datasets of type "L".

The constructed credibility regions recovered the true parameter for reasonable confidence levels ($\leq 95\%$) for about half of the datasets tested, even for low PAC-Bayes temperature. This represents an improvement from the results reported for ADM1 model in Chapter 3 (where VarBUQ was used to calibrate), where recovery was achieved only for one out four datasets. For low PAC-Bayes temperatures (*i.e.* high learning rates) of $\lambda \in \{0.002, 0.004\}$, the true parameter was the inside the confidence region in 4 out of 10 cases. This increased to 5 for $\lambda = 0.008$ and 6 for $\lambda = 0.016$. While reducing the learning rate helps recover the true set of parameters,

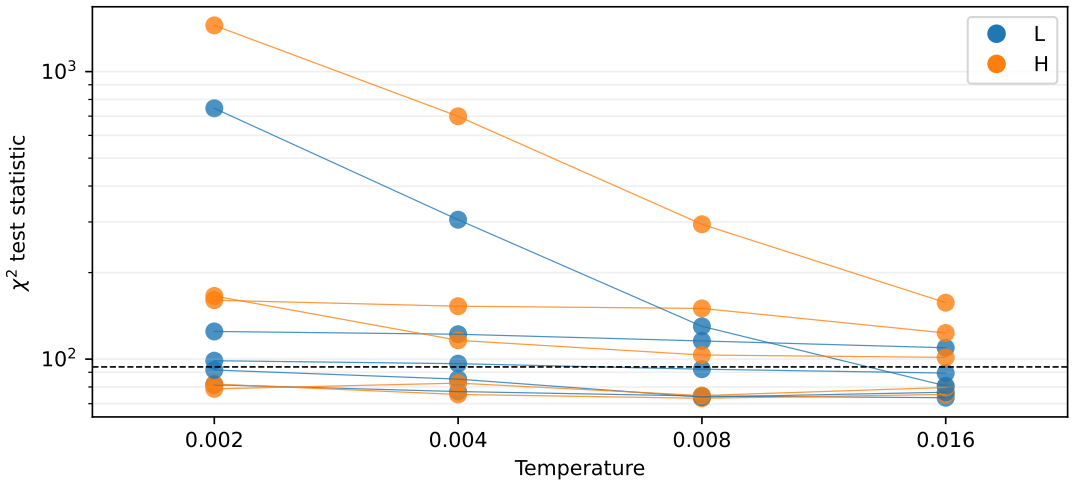


Figure 4.13: χ^2 test statistic for rejection of the true parameter by the posterior distribution. The colours represent the type of intrant (L for low intrant flow, H for high intrant flow – V datasets are not represented, as they share their statistics with L datasets). The dotted line represents the threshold for acceptance of the true parameter at 95% confidence (*i.e.*, above this threshold, the true parameter is deemed not drawn from the posterior). Lines link statistics for the same dataset.

it seems that certain datasets were more prone to suffer from overfitting than others. This is corroborated by Figure 4.13, which shows that the influence of the PAC-Bayes temperature on the test statistic can greatly differ depending on the dataset.

4.3.4 Performance of uncertainty quantification on predictions

The calibration procedure constructed posterior distributions which had on the whole satisfactory uncertainty quantification on the predictions. Confidence intervals on the test observations were assessed for two confidence levels: 50% (*i.e.* the quantiles 25% and 75%) and 95% (*i.e.* the quantiles 2.5% and 97.5%). The four performance indicators described in chapter 3 (coverage of CIs, width of CIs, median error, residual error of CIs) were computed for each type of predictions. For VFA concentrations, the performance indicators were computed using the log residuals after shifting the predictions and observations by 2 gCOD/L, to match the formula in the risk. The coverage indicator was computed as the percentage of the true signal (*i.e.* computed without any noise on the inputs and outputs) inside the confidence regions. Its average value among all predictions ranged from 52% ($\lambda = 0.002$) to 62% ($\lambda = 0.008$) for a confidence level of 50%, and from 90% ($\lambda = 0.002$) to 94% ($\lambda = 0.016$) for a confidence level of 95%, in both cases indicating adequate coverage performances. While increasing the PAC-Bayes temperature (*i.e.* reducing the learning rate) tended to increase the coverage, the coverage of the lowest PAC-Bayes temperature tested ($\lambda = 0.002$) proved satisfactory, even for the datasets whose training data and test data differ (V datasets). The coverage performance was heterogenous considering the type of predictions, with pH confidence intervals always achieving perfect cov-

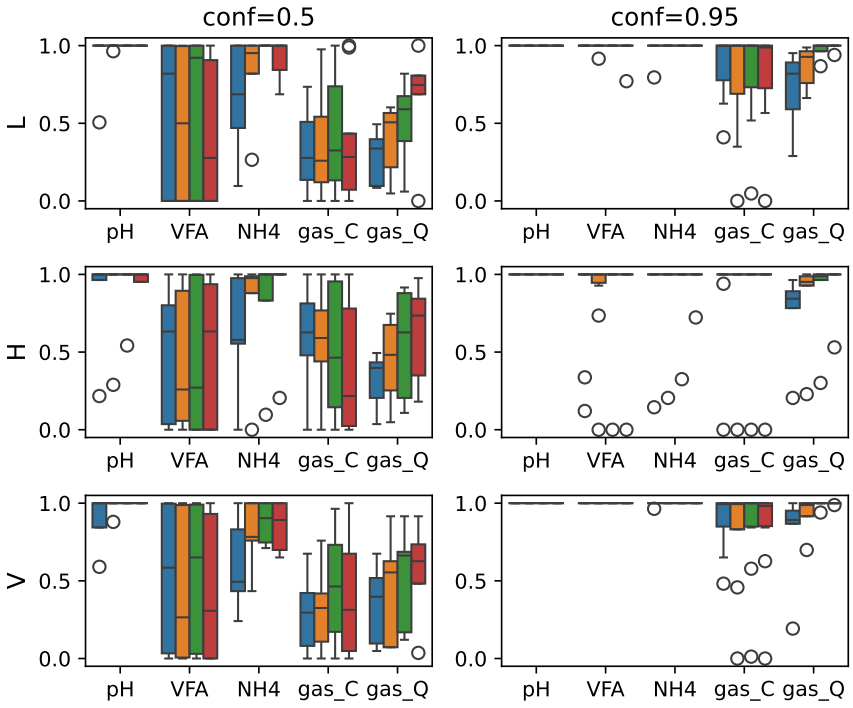


Figure 4.14: Coverage of the confidence intervals. The PAC-Bayes temperature is increasing from left to right (blue: $\lambda = 0.002$, orange: $\lambda = 0.004$, green: $\lambda = 0.008$, red: $\lambda = 0.016$). 'gas_C' describes the molar content of the gas phase in methane and carbonic gas, while 'gas_Q' describes the methane and carbonic gas flows.

erage for 95% confidence level and confidence intervals for methane production occasionally failing to cover any single predictions (although, averaging on all datasets, 85% of methane and carbonic gas points are in the 95% confidence intervals). A description of the coverage performance is given in Figure 4.14.

The CIs width, computed as root mean square of the log ratios between the upper and lower bound, proved to be reasonably small for small PAC-Bayes temperatures (average over all datasets and all predictions type except VFAs of 0.08 for a confidence level of 95% for $\lambda = 0.002$), but increased significantly with the PAC-Bayes temperature (average of 0.24 for a confidence level of 95% for $\lambda = 0.016$) (see Figure 4.15). For the largest PAC-Bayes temperature notably, confidence intervals could measure up to 0.7 for gas flows, indicating an average ratio of 2 between the upper and lower bound. As expected, the confidence intervals are much smaller for the confidence level of 0.5 (average width of 0.033 for $\lambda = 0.033$, 0.071 for $\lambda = 0.016$). Uncertainty on VFA predictions tends to be high (up to 1.7, due to a confidence interval on propionate concentration ranging from 1 gCOD/L to 10 COD/L). The confidence intervals do not appear to be significantly larger for datasets whose train and test sets do not match (V).

The performance of the calibrated model is assessed by computing the error of the median

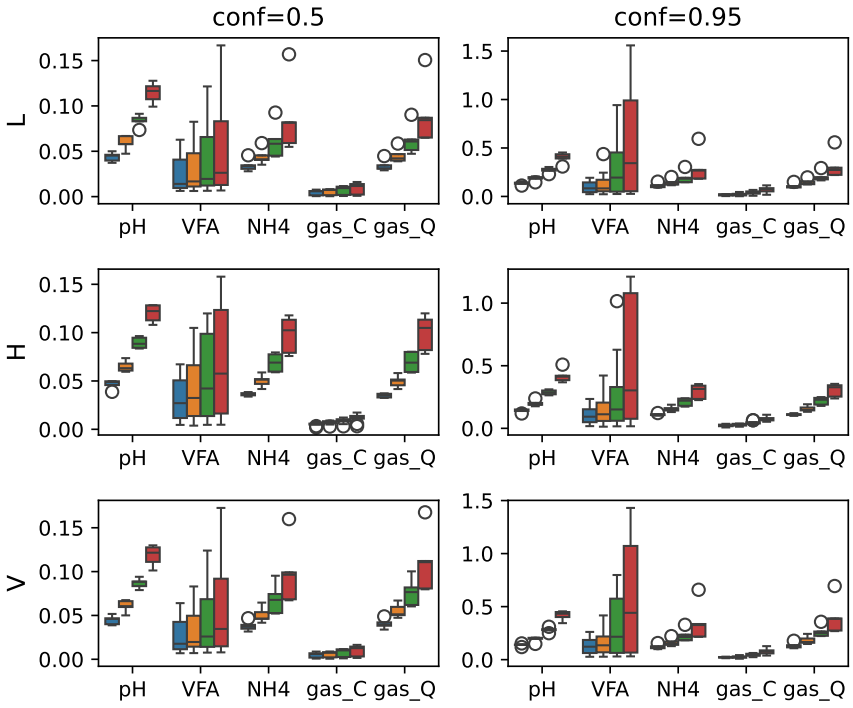


Figure 4.15: Width of confidence intervals (computed as the mean log ratio).The PAC-Bayes temperature is increasing from left to right (blue: $\lambda = 0.002$, orange: $\lambda = 0.004$, green: $\lambda = 0.008$, red: $\lambda = 0.016$). While confidence intervals for a confidence level of 0.5 have a moderate width of up to 0.17 excluding volatile fatty acids (relative uncertainty of $\pm 9\%$), confidence intervals for a confidence level of 0.95 can have extremely large width of up to 0.7 for gas (a factor of 2 between the upper and lower bound), and up to 1.5 for volatile fatty acids. The width of the confidence intervals increases with the PAC-Bayes temperature for both confidence levels.

prediction for all types of predictions (see Figure 4.16). These indicators therefore do not depend on the confidence level considered. Average errors on all datasets are well controlled for every type of predictions and every PAC-Bayes temperature, ranging from 0.004 to 0.056. For some datasets, the error can however be as high as 0.17 (about 18% relative error) (achieved for gas flows at the lowest PAC-Bayes temperature tested). While NH4 concentrations, VFA concentrations and gas flows predictions can differ significantly from the truth, the gas content and pH predictions tend to only marginally deviate from the truth (less than 5% error for gas content and less than 0.053 absolute difference for pH).

The coverage indicator does not inform on how far the data points which are not covered by the confidence intervals are from these confidence intervals. To remediate this, the residual error after projection on the confidence intervals was measured. This residual error was mostly negligible for the 95% confidence intervals (less than 0.035), except in the case of the H datasets, where it could rise for up to 0.12 for gas flows for $\lambda = 0.002$ (see Figure 4.17). Raising the PAC-Bayes temperature helps diminishing the residual error, indicating that the increase in the

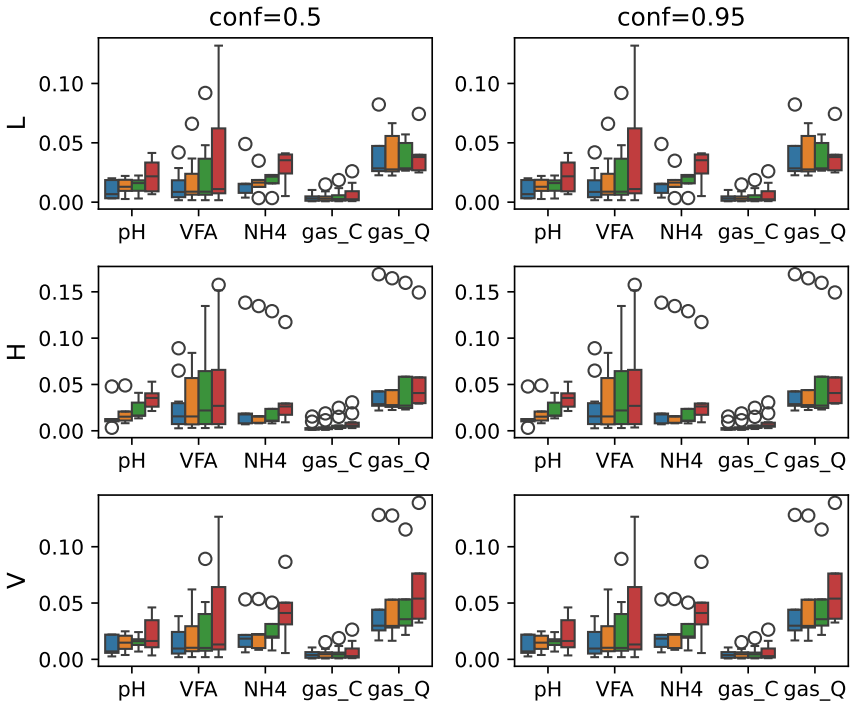


Figure 4.16: Test prediction error of the median. The PAC-Bayes temperature is increasing from left to right (blue: $\lambda = 0.002$, orange: $\lambda = 0.004$, green: $\lambda = 0.008$, red: $\lambda = 0.016$). Confidence does not play a part in this setting. Prediction errors remain lower than 0.17 (approx. 18% relative error). Decreasing the PAC-Bayes temperature decreased the prediction error for all types of predictions except gas flows.

width of the confidence intervals more than compensates for the increase of error of the median. The maximum residual error is lowered to 0.10 for $\lambda = 0.004$, 0.07 for $\lambda = 0.008$ and becomes almost negligible (0.04) for $\lambda = 0.016$, although at the cost of much larger confidence intervals. The residual error for the 50% confidence intervals follows a similar trend to the median error, but with errors on average 75% lower. However, the residual error tends to be lowered more significantly for predictions with low errors (*e.g.* pH), and in the worst case, the residual error for the 50% confidence intervals was only marginally (9%) lower than the median error (gas flow prediction for dataset DH).

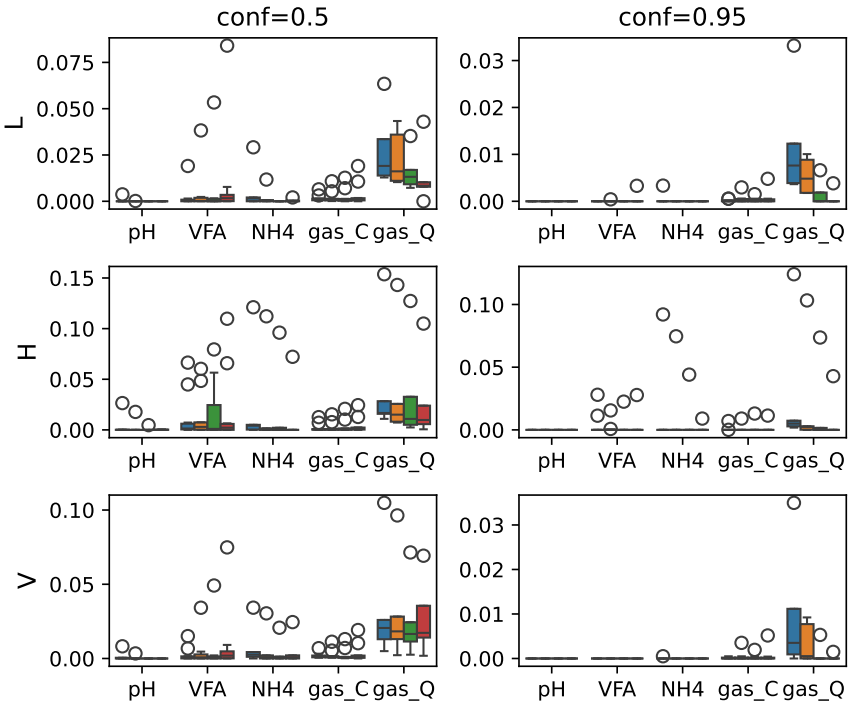


Figure 4.17: Residual error after projection on the confidence intervals. This indicator measures how much of the true signal is not captured by the confidence intervals. The confidence intervals appear to miss a more important part of the signal for datasets with short hydraulic retention time (H), with a residual error for confidence level of 0.6 of up to 0.15 for gas flows. The residual errors for 95% confidence intervals are mostly negligible, remaining under 0.04.

4.3.5 Stability of the calibration procedure

The calibration procedure used is stochastic, relying on random draws of sets of parameters. As such, some randomness is expected in the algorithm output, as would also be the case with other stochastic minimisation procedures (*e.g.* stochastic gradient descent). Whether or not this randomness is negligible depends on the termination criteria used as well as the geometry of the empirical risk. Intuitively, one expects that the more complex the risk is (*i.e.* the less it can be well approximated by quadratic functions in our setting), the higher the dimension of the problem, and the lower the PAC-Bayes temperature, the more important would be the outputs randomness.

To assess this phenomenon in the case of ProdAD, triplicate calibration procedures were performed using identical inputs, and the discrepancy between the outputs was measured. The lowest PAC-Bayes temperature of 0.002 was assessed, as well as the PAC-Bayes temperature of 0.008. For $\lambda = 0.002$, all three calibration procedures concentrated around a similar mean parameter, with natural distance induced by the prior $\mathcal{N}(\mu_{\text{prior}}, \Sigma_{\text{prior}})$ between the posterior means $d(\mu_i, \mu_j) = \frac{1}{d} (\mu_i - \mu_j)^t \Sigma_{\text{prior}}^{-1} (\mu_i - \mu_j)$ of 0.0056, 0.0074 and 0.0056. The posteriors still were

somewhat distinct, with KL divergences between the three distributions ranging from 1.5 to 2.3. For comparison, the KL divergences between the prior and posterior range from 8.6 (computed as $\mathrm{KL}(\pi_{\text{post}}, \pi_{\text{prior}})$) to 130 (computed as $\mathrm{KL}(\pi_{\text{prior}}, \pi_{\text{post}})$). For $\lambda = 0.008$, the distributions were at KL divergence between 0.8 to 1.1 from one another, while the prior induced distance between the means ranged from 0.0027 to 0.0037, indicating improved stability. Therefore, stability issues cannot be fully ignored at low PAC-Bayes temperature. This lack of stability hints that the algorithm either failed to converge to a same local minima in the given computational budget, or that it could find multiple local minima posteriors with identical centres, but differing covariance structures. This first hypothesis was partially confirmed by extending the calibration computational budget. The stability of the algorithm is also hurt by the randomness of the methodology used to estimate integrals with respect to the posterior, a key part of the calibration strategy. This randomness prevents the algorithm from stabilizing. To minimise this behaviour, more computational budget was allotted to the integral estimation by increasing the number of samples used to estimate the weights from 25 000 to 250 000. This impacts the training time (about twice as long), and might not be practicable for real world use case. After an additional 24 calibration steps, the resulting posteriors were all at KL divergence between 0.4 to 0.9 from one another for $\lambda = 0.008$. This suggests that all three procedures approximate the same local minima.

4.3.6 Spurious learning on inactive blocks

Closely related to the question of calibration stability is the presence of spurious learning of the calibration procedure. For given operational conditions, some blocks of parameters in ProdAD model are known to play no part in the predictions. Such blocks of parameters having no impact on the risk should have a posterior distribution matching the prior distribution, as this minimises the KL penalization. However, due to the limited number of risk evaluations and the presence of confounding factors (namely, the parameters which do not have impact on the risk), the algorithm spuriously learns a posterior distribution on these blocks of parameters differing from the prior distribution[18]. The importance of this phenomena was quantified by computing the contribution to the KL divergence caused by inactive blocks.

For AL dataset at $\lambda = 0.016$, the KL contributions from inactive blocks range from 0.014 to 0.025. On the other hand, the KL contributions from comparable active blocks range from $0.1$to 1.4. When the PAC-Bayes temperature is reduced, the KL contributions from inactive blocks tend to increase, ranging from 0.04 to 0.074 at a PAC-Bayes temperature of 0.002, while the comparable active blocks contributions still range from 0.1 to 1.4. This impact of the temperature was expected, since spurious learning is less penalized as the PAC-Bayes temperature is decreased.

This spurious learning indicates that the algorithm has not exactly stabilized at the local minimiser of Catoni's bound. Given sufficient computational power, the spurious learning phe-

[18]To illustrate this point, one can consider the predictive model $f(x, y) = g(y)$. For a finite sample (x_i, y_i) drawn at random, the empirical correlation between x_i and $g(y_i)$ will not be 0 (although hopefully small), and a model learnt from the finite sample may consider an influence of x.

nomena could be attenuated. To attenuate this phenomena, the posterior distribution returned by the algorithm is cleaned up. Posterior blocks whose KL to the corresponding prior block are less than a threshold $\varepsilon_{\text{KL}} = 0.1$ are put back to the prior. This is similar to the behaviour of ℓ_1 penalized regression[19]. If the resulting posterior obtains a PAC-Bayes bound less than $\varepsilon_{\text{PB}} = 0.001$ more than the PAC-Bayes bound of the initial posterior, it is accepted (else the initial posterior is used).

4.3.7 Impact of hyperparameters

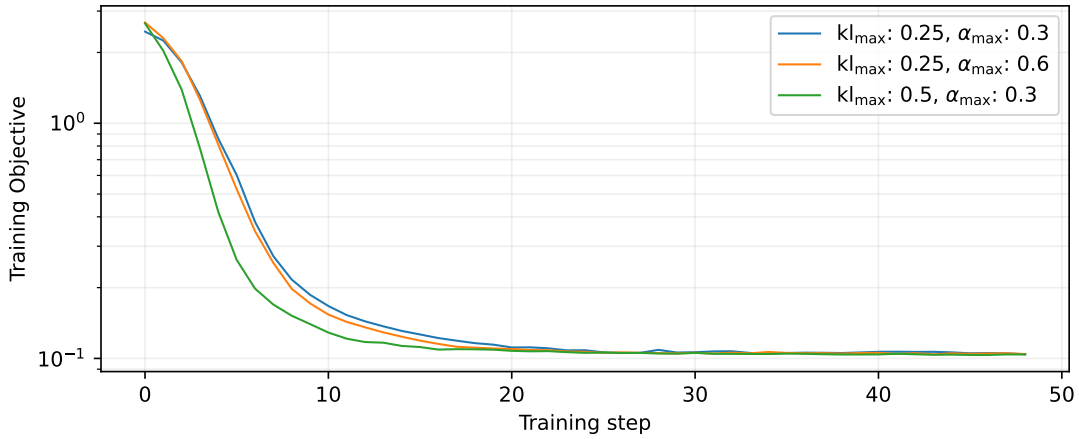


Figure 4.18: Impact of the hyperparameters on the convergence of the calibration procedure. The algorithm appears to be limited by the kl_{max} parameter

The impact of the regularisation hyperparameters were assessed through additional calibrations. While all previous calibrations routine were performed using $\text{kl}_{\text{max}} = 0.25$ and $\alpha_{\text{max}} = 0.3$, two additional calibrations routine for the dataset AL at a PAC-Bayes temperature of $\lambda = 0.0002$ were launched, one using $\text{kl}_{\text{max}} = 0.5$ and $\alpha_{\text{max}} = 0.3$, the other using $\text{kl}_{\text{max}} = 0.25$ and $\alpha_{\text{max}} = 0.6$. All three calibration procedure stabilized around a similar risk. The calibration procedure using the higher kl_{max} hyper parameter proved to converge somewhat faster than the remaining two procedures, which are to all purposes undistinguishable. This strongly hints that the training path was regularized mainly by the kl_{max} parameter, while the additional dampening was not high enough to play a part.

4.3.8 Conclusion

The PAC-Bayes calibration strategy SuPAC-CE provided efficient when assessed on a full calibration of ProdAD model. The resulting stochastic calibrated model had good predictive power

[19]Note that contrary to true ℓ_1 penalized regression, the blocks above the threshold are not modified. This could be done by solving $\text{KL}(\theta_p^B + x(\theta^B - \theta_p^B), \theta_p^B) = \max(0, \text{KL}(\theta^B, \theta_p^B) - \varepsilon_{\text{KL}})$

and adequate uncertainty quantification, both on the values of the calibrated parameters and on new predictions.

The choice of the hyperparameters used for the calibration of ProdAD model could be further refined. Due to the somewhat large dimension of the variational family considered, a large number of simulations (of order 500) is necessary for the initial step of the algorithm. While SuPAC-CE was evaluated here using this number of simulations constant during the learning process, we confidently expect that it could be reduced as the algorithm starts converging, as was done when applied to ADM1. This would further decrease the number of simulations during the calibration, which could prove useful when considering complex plants involving multiple digestion tanks and increased simulation time.

4.4 General Conclusion

To calibrate SUEZ's ProdAD model using PAC-Bayes objective in reasonable time, we tackled the main computational bottleneck - the simulation time - through two design changes. First, we proposed a PAC-Bayes bound minimisation framework, SuPAC, which is designed to limit the number of queries to the model. Second, we modified the implementations of AD models to gain a substantial speed gain, with no noticeable impact on the model predictions. Taking advantage of these improvements, we assessed the performance of PAC-Bayes trained posteriors for ProdAD model using controlled, synthetic data. Building on this, we are now in a position to tackle the further challenges brought by real-world data and online monitoring.

Chapter 5

From PAC-Bayes theory to industrial impact

In this chapter, we consider PAC-Bayes calibration of real-world, industrial AD plants. We propose a set of rules to perform the calibration process in a harmonized way on diverse AD plants, and assess the posteriors obtained on two industrial plants. We also adjust the PAC-Bayes calibration methodology for the online calibration and monitoring of AD plants.

Contents

5.1 Adjusting to real world data

The PAC-Bayes calibration strategy SuPAC-CE introduced in section 4.1 proved able to efficiently calibrate synthetic AD data using both ADM1 (section 4.1.6) and SUEZ in-house ProdAD

model (section 4.3). Real world data, however, adds new layers of difficulty, which call for some design changes and further assessment. The noise pattern of true data is presumably much more complex than the log-uniform noise signal used in the experiments. The data available for calibration also differs from the data used in the synthetic experiments in terms of available features and frequency, and differs from plant to plant depending on available sensors.

Moreover, some design choices left open in the previous section should be fixed in order to quickly deploy SuPAC-CE calibration to any new plant. This includes automating the choice of PAC-Bayes temperature for a given plant, as well as automating the choice of hyperparameters.

Fine tuning the hyperparameters of the calibration algorithm is an important consideration, as these balance the stability and the speed of the algorithm. The choice of hyperparameters used for the tests above was rather conservative, with a longer, but safer learning strategy. As the optimal choice of hyperparameters depends on the structure of the error function, thorough tests should be performed using real world data. Moreover, the algorithm's performance can benefit from using adaptative hyperparameters, evolving during the calibration.

5.1.1 Construction of a generic risk function

Depending on the plant's configuration, observation data used for calibration can change; plants involving multiple digestion units might for instance either only measure the total amount of methane produced, or individual methane production for each digester; pH measurement might be lacking or wholly inaccurate (sensor not properly calibrated); the quality of each individual sensor might vary; measurement frequency also depend on the type of observation (*e.g.* online gas measurement are more frequent than lab measurements of VFA concentrations).

On the other hand, applying our calibration strategy to a new plant should involve as few expert design choices as possible, in order to limit the deployment time and to rely on a generic approach. Constructing highly fine-tuned error functions for each individual plant would be time consuming and limit the transfer abilities of our learning approach from one plant to another. By making the learning task more distinctive, less information could be shared at the meta level.

Our approach to balancing the unique characteristics of each plant with the need for a generic risk function involves adjusting the log-RMSE error from Chapter 3 using a task-specific configuration file, along with general rules for handling missing data. The task specific configuration essentially encodes two sparse matrices, specifying how the model's outputs and the observations are to be linearly transformed in order to match one another[1]. The common rules are leveraged among all tasks and could eventually be refined[2].

[1] This approach is well suited to account for flows which might be joined in observation data, but not for ratios and multiplications such as are involved between transformations to mass or mole. In practice, the model's default outputs were extended to contain both concentrations and mole quantities for each tank. More complex feature computation schemes could also be encoded in the configuration file in the future

[2] This could clash with our meta-learning and online learning strategies, if the updated risk evaluations can not be inferred from stored data. If this happens, SuPAC-CE algorithm will need to query the risk a large number of time for each task. Such major design changes are therefore limited to major version update of the code base.

The plant specific configuration is a triplet specifying two sparse matrices and a weight vector. The two matrices define the multi-dimensional regression task for the plant,

$$M_{\text{pred}}\text{AD}(\gamma, \text{Feed}) \sim M_{\text{obs}}\text{Obs}$$

where Obs is a matrix containing the observation data, containing missing values. Using matrixes M_{pred} and M_{obs} offer some much needed flexibility when matching the model predictions to available measurements. For instance, VFA measurements can be describe either the total VFA concentrations, or the concentration of specific VFA (e.g. acetate and propionate concentration); some plants can either measure the gas flow of each individual tank, or the gas flow for a set of tanks. The risk configuration matrices M_{pred} and M_{obs} cover this wide range of practices by comparing linear combinations of predictions to linear combinations of observations; for instance, if the first observation is the joint gas flow of tank A1 (column j_1 of the prediction matrix) and tank A1 (column j_2 of the prediction matrix), the line i in the matrix M_{pred} and M_{obs} corresponding to equation $\text{Obs}_1 \sim$ Gas Flow A1 + Gas Flow A2 would be $M_{\text{pred},i} = (\delta_{j,j_1} + \delta_{j,j_2})_j$ while $M_{\text{obs},i} = (1, 0, \dots)$. The use of such matrices:

- avoids having to compute by default a large number of predictions (*plant specific predictions can be added*);
- is expressive enough (*no urgent need for more elaborate options, such as divisions, multiplications*);
- can be easily configured and does not necessitate writing any specific code for the plant (*simple communication with the model and easy to store*).

We will denote $\overline{\text{Pred}} := M_{\text{pred}}(\gamma)\text{AD}(\gamma, \text{Feed})$ and $\overline{\text{Obs}} := M_{\text{obs}}\text{Obs}$ the linearly transformed model predictions and observations. These are matrices of size (T, n_{eq}) with n_{eq} the number of features to be regressed (*number of equations*) and T the number of days simulated. Each column defines a type of observation (*e.g.* pH in the post digester, overall biogas production), denoted obs. A type of observation is unique for each plant and belong to one of ten possible families of observations (*e.g.* pH, VFA concentration, methane flow, concentration of methane in biogas) denoted f. The family to which belongs a given type of observation is specified in the configuration, and impacts the way the residual is computed. For each family f of observations is associated an offset η_f and a transform function trf_f used to compute the residuals as

$$\text{res}_{t,\ \text{obs}}(\gamma) = \text{trf}_f(\overline{\text{Obs}}_{t,\ \text{obs}} + \eta_f) - \text{trf}_f(\overline{\text{Pred}}_{t,\ \text{obs}}(\gamma) + \eta_f).$$

The transform functions are currently logarithms except for pH where the transform is the identity function (*i.e.* no transform applied). The offsets η are set to 10^{-4} for all families of observation except VFA, where it is set to 1 gCOD/L. This implies that VFA observations will only play a part whenever either the observation or the prediction is large enough.

These residuals are then used to compute a mean square error, slightly modified to account for the number of available data. Noting $\mathcal{T}_{\text{obs}}$ the set of timestamps for which an observation

of type obs is available and $|\mathcal{T}_{\text{obs}}|$ the number of observations, the mean square error for a the prediction type obs is

$$\text{MSE}_{\text{obs}}(\gamma) = \frac{1}{|\mathcal{T}_{\text{obs}}| + N_f} \sum_{t \in \mathcal{T}_{\text{obs}}} \text{softclip}\left(\text{res}_{t,\ \text{obs}}(\gamma)\right)^2$$

where the softclip function follows a similar form as in section 3.2.4, but, as in section 4.3, considering a maximum value of 4 instead of 3. If no observation is available, the observation type is disregarded for the error computation. The set of time stamps $\mathcal{T}_{\text{obs}}$ can be modified to exclude some data at the beginning to account for the approximation in the initialization (this has been implemented but not tested). Finally, the mean square errors are combined together using the weights specified in the configuration, or default weights depending on the observation family if weights are not provided. The offset N_f induces a weight correction $\frac{|\mathcal{T}_{\text{obs}}|}{|\mathcal{T}_{\text{obs}}|+N_f}$, limiting the impact of a given prediction type in the low data regime. This smooths the transition between no data (weight 0) to few data points (weight $\omega\frac{n}{n+N_f}$). The value of N_f varies from 5 to 20 and can be roughly interpreted as the minimal number of data which should be gathered before considering that observation type in the risk computation; AD modellers can recognize in this weight correction a form of Monod inhibition.

All in all, the risk is computed as

$$R(\gamma) = \sqrt{\frac{\sum_{\text{obs}} \omega_{\text{obs}} \mathbb{1}(|\mathcal{T}_{\text{obs}}| > 0) \text{MSE}_{\text{obs}}(\gamma)}{\sum_{\text{obs}} \omega_{\text{obs}} \frac{\mathbb{1}(|\mathcal{T}_{\text{obs}}|>0)|\mathcal{T}_{\text{obs}}|}{|\mathcal{T}_{\text{obs}}|+N_{f_{\text{obs}}}}}} \tag{5.1}$$

for all parameters γ for which the simulation of ProdAD succeeded (*i.e.* did not lead to negative concentrations or other physically unfeasible values). Parameters γ whose simulations failed were attributed the maximal risk value of 4[3].

5.1.2 Selecting new hyperparameters for SuPAC-CE

The most important hyperparameter of SuPAC-CE is the choice of PAC-Bayes temperature, as it directly impacts the learning objective. PAC-Bayes generalisation bounds provide guidelines on the way the temperature should be chosen. The choice of temperature for the analysis of the UQ ability of PAC-Bayesian method in section 3.3.4 was based on the analysis of the extra term $\lambda \log(\delta) + \frac{C}{8\lambda n}$ where C is an upper bound of the risk. We follow a similar course and define the temperature as $\lambda := \frac{1}{0.8n}$. This is an optimistic (*i.e.* small) value, amounting to an extra term in the generalisation bound of $0.4 - 1.25\log(\delta)/n$ *without* factoring in the fact that the data are

[3]Another option consisted in computing the score on the section of the simulation which did not fail, then using the maximal residual value of 4 on the part of the simulation which could not be carried out. This option could positively impact the learning procedure in the setting where most draws from the prior lead to simulation failure, by informing the learning algorithm on which failed simulations had most success. Unfortunately, this strategy was ruled out by the difficulty of recovering the partial predictions after simulation failure in the current, numba compliant implementation due to its limited error handling abilities. Moreover, our experiments hinted that most simulation failures happened at few specific dates, making it unclear whether much information could have been extracted by this process (similar risks).

dependent. Building on Section 1.2.3, taking into account dependency in the data should result in temperatures a factor HRT higher (typically 10 to 30 times higher depending on the plant). Still, the analysis of Section 4.3 showed that such low temperatures improved test prediction and offered adequate uncertainty quantification on the predictions.

The other hyperparameters for SuPAC-CE impact the calibration speed, and to some degree the calibration result. A first difficulty, shared by VarBUQ and SuPAC-CE, concerns the stopping criteria of the procedure. The training objective involves the posterior average of the empirical risk; as the algorithm has access to a finite number of evaluation of the risks, this estimate involves some noise, which will be typically larger than usual termination criteria. On our experiments, SuPAC-CE only met the convergence criteria on simple toy use cases where the score can adequately be approximated by a surrogate risk (*e.g.* in the modified Rosenbrock risk of section 4.1.6).

As convergence of the procedure is thus unlikely due to its stochastic nature, we rely on a fixed sequence of hyperparameters to perform the calibration. This sequence was semi arbitrarily built, guided by early feedback on the calibration procedure for real world AD task. During the learning procedure, the regularisation hyperparameters become increasingly more restrictive, while the number of new parameters evaluated is decreased. The sequence of hyperparameters is built around five phases:

- A first phase of 4 steps, with regularisation $\alpha_{\max} = 0.5$, $\mathrm{kl}_{\max} = 0.2$, 640 parameters drawn per step;
- A second phase of 4 steps, with regularisation $\alpha_{\max} = 0.4$, $\mathrm{kl}_{\max} = 0.2$, 640 parameters drawn per step;
- A third phase of 10 steps, with regularisation $\alpha_{\max} = 0.35$, $\mathrm{kl}_{\max} = 0.2$, 480 parameters drawn per step;
- A fourth phase of 10 steps, with regularisation $\alpha_{\max} = 0.35$, $\mathrm{kl}_{\max} = 0.2$, 320 parameters drawn per step;
- A fifth and final phase of 15 steps, with regularisation $\alpha_{\max} = 0.3$, $\mathrm{kl}_{\max} = 0.2$, 160 parameters drawn per step.

In total, 15 520 parameters are evaluated during the learning procedure. This is almost half the number of evaluations considered in section 4.3 (30 720 evaluations). Between these phases are intercepted learning phases relying only on existing risk evaluations. These use regularisation hyperparameters of $\alpha_{\max} = 0.1$, $\mathrm{kl}_{\max} = 0.02$ and are run for 5 steps. The learning sequence ends with the same 'no new risk evaluation' sequence, running for ten steps.

5.1.3 Calibration results

The calibration procedure was assessed on data from an industrial mesophilic AD plant (plant A). Anaerobic digestion is performed on two tanks operating in parallel (D1, D2), receiving sim-

ilar influent. These tanks are assumed to have identical parameters. The available observations are the biogas flow, VFA concentration, **T**otal **A**mmonia **N**itrogen (TAN) concentration, pH measurements, and VS fraction, all of these for each digester tank. Calibration was performed using the first 280 days of data, the remaining days being left out for test assessment (see fig. 5.1).

During the early phases, the posterior learnt which parameters do not lead to simulation failure. Simulation failure occurs when the model predicts a sharp increase of the acids concentration in the digester, leading to irreversible pH decrease (acidification), computational instabilities and eventually to negative concentrations of some compounds. As such behaviour does not occur for the digester's data, such models are grossly inadequate, and receive the highest possible risk value. The initial proportion of simulation failure for the prior was 79.4% (95% confidence interval of $[76.1\%, 82.4\%]$). This proportion decreased to 59.2% ($\in [55.3\%, 63.1\%]$) at the end of the first sequence, and further decreased to 35.5% ($\in [31.8\%, 39.3\%]$) after the "no new risk evaluation" sequence. The proportion falls to 13.6% at the end of the second sequence, then to 0.6% ($\in [0.15\%, 1.8\%]$) at the end of the third sequence. This resulted in a decrease in the objective from 3.32 to 2.57 at the end of the first sequence (2.01 after the "no new evaluation" sequence) to 0.88 at the end of the second sequence (0.63 after the "no new evaluation sequence") to 0.27 at the end of the third sequence. The objective was then further decreased to 0.191 at the end of the calibration sequence, for an average error of 0.151, or 15% more than the smallest error observed during the calibration procedure (0.1295).

Compared to the synthetic data studied in Section 4.3, the uncertainty quantification appears less effective (see fig. 5.1), with most observations points outside the confidence regions. We stress that contrary to some other uncertainty quantification methods such as conformal predictions, the uncertainty quantification provided by the posterior distribution does not try to account for the observation noise. It strives to offer uncertainty quantification for the underlying true signal, rather than the noisy observations. As such, it is not surprising that the noisy observations fail to belong to the confidence intervals for the target level, since the calibration performs de-noising. As the true signal is not observed, it is impossible to assess the coverage of the confidence intervals. However, some predictions such as VFA and VS appear to have systematic bias both on the train and test data. This hints that no likely explanation could correct the bias - at least without negatively impacting other predictions taking into account. This could be explained by a systematic bias in input measurements.

The calibration procedure was further assessed on a second industrial plant (plant B). Two digesters in succession perform AD in this setting. Measurements of VFA and TAN concentrations and pH are available for the first digester, as well as the overall gas production. The model was calibrated using the first 459 days of data (note that there are no observations for the first 179 days), and its performance assessed on the remaining data. The predictions and uncertainty quantification are given in Figure 5.2. The calibration procedure decreased the PAC-Bayes objective from 2.52 to 0.337 and appeared to have converged (average decrease of PAC-Bayes objective of 0.00025 per step in the last ten steps). The calibrated model extrapol-

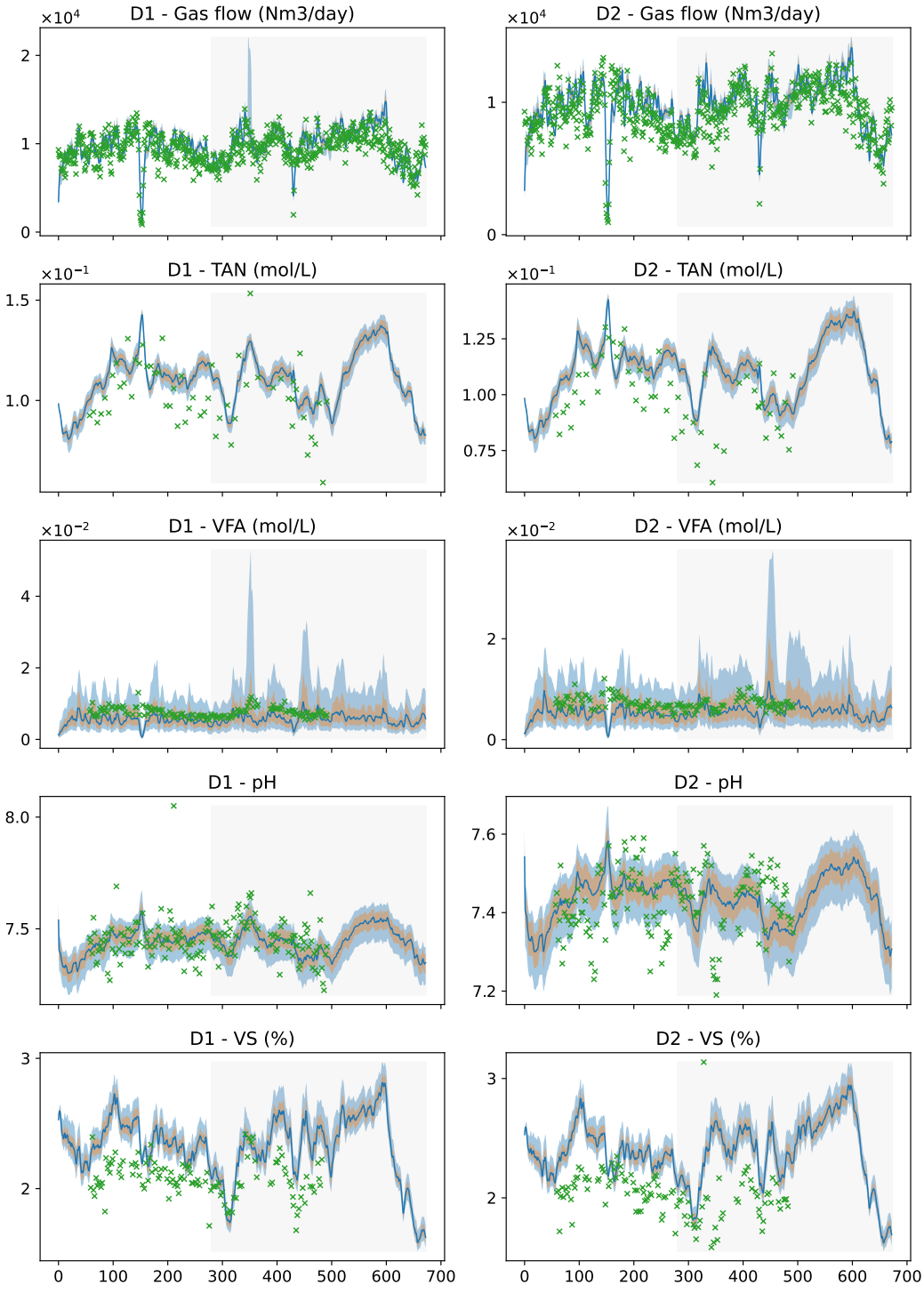


Figure 5.1: Predictions of calibrated ProdAD model on real plant data (plant A). Confidence intervals of level 66% (orange) and 95% (blue) and median predictions from the posterior (blue line) are plotted. Observations are represented by green crosses. The grey zone indicate test data.

ated well on the test data, with a smaller RMSE[4] on the test set than on train set for gas flows (2543 vs 2960 Nm3/day) and VFA concentrations (0.0067 vs 0.012 mol/L) and slighlty higher test RMSE for TAN (0.084 vs 0.076 mol/L) and pH (0.12 vs 0.11).

For Plant B, the initial proportion of parameter failure for the prior was 63.6% (95% confidence interval of $[59.8\%, 67.3\%]$). This decreased to 6.4% ($\in [4.7\%, 8.6\%]$) at the end of the second sequence. This large proportion of parameter failures for the prior in both cases questions its quality and hints that it could be improved through meta-learning, especially if parameters leading to simulation failure are similar for various plants.

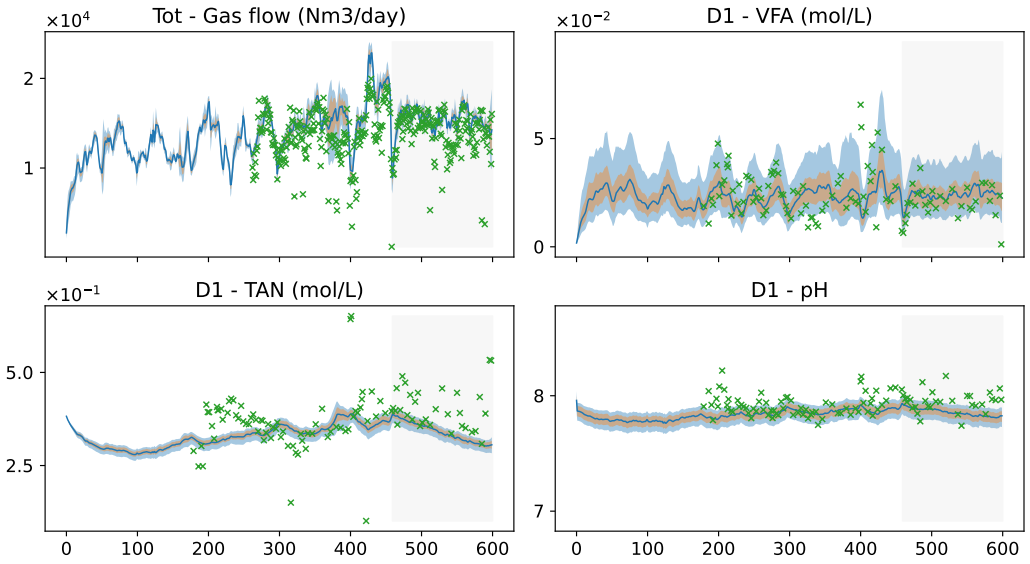


Figure 5.2: Predictions of calibrated ProdAD model on real plant data (plant B). Confidence intervals of level 66% (orange) and 95% (blue) and median predictions from the posterior (blue line) are plotted. Observations are represented by green crosses. The grey zone indicate test data.

The sequence of training hyperparameters gave adequate results on two distinct AD plants. Tests on further digesters would be necessary to assess the robustness of the procedure. Moreover, the arbitrary nature of the choice of hyperparameters is not wholly satisfactory. Unfortunately, experimental assessments of an optimal sequence of hyperparameters would be complicated as the dimension of the design choices is large (for a given number of iterates n and k changes of hyperparameters and considering a grid of p options for the main three hyperparameters, the number of hyperparameters to investigate is $\binom{n}{k}p^3$). Bayesian optimisation methods could be applied once a sufficient number of tasks are collected, in order to find a set of hyperparameters which give adequate performance for all digesters. However, we expect the optimal hyperparameters to vary depending on the task at hand. Future prospects include auto adjustment of the hyperparameters during the learning task, or theoretically motivated choice of hyperparameters schedule.

[4]The RMSE is computed using the median prediction.

5.2 Online monitoring of Anaerobic digestion plants

AD plants accumulate data in an online fashion, both from online sensors and for lab analysis. This data describes both the description of the intrant, as well as further observation data. Online learning strives to incorporate these observation data into the calibrated model as they appear. In the context of AD modelling, this is of particular interest since further data can explore new operational condition and help identifiability of some parameters. For instance, early signal of acidification could help reduce the parameter uncertainty and help operators make better informed decisions.

The calibration strategy SuPAC-CE relies on an increasingly large stack of risk evaluations, continuously increasing the knowledge of the risk landscape close to the current posterior mass centre. For generic risk functions, this is an issue in an online setting, since a new stack of risk evaluations must be reconstructed as the dataset is updated. For additive risks however, of form $R(\gamma) := \frac{1}{t}\sum_{i=1}^{t} \ell(\gamma, z_i)$, SuPAC-CE can fit easily in an online setting, since the stack of previous risk evaluations can be updated with little computational cost by evaluating $\ell(\gamma_i, z_{t+1})$. While the risk function defined in Section 5.1.1 does not exactly fit this simple pattern, we will show that this strategy can still be pursued with few adjustments.

A more delicate issue concerns the validity of the data used at a given time to perform simulations and calibration. Some data may be uploaded with unavoidable delays (*e.g.* laboratory analysis), leading to a discrepancy between the risk function used for the risk evaluations in memory (with partial observations) and the risk function used for fresh risk evaluations (with all observations). Such discrepancy could potentially harm the calibration process if the posterior has already concentrated, by adding perturbations to the perceived risk.

A final question concerns the regularity and computational budget which should be allocated to the online calibration process. The number of risk queries evaluated per year as well as the number of risk evaluations kept in memory and requiring regular update will impact the cost of the calibration procedure. Numerical experiments for various online calibration design are performed to assess an adequate balance between calibration precision and computational cost.

5.2.1 Online SuPAC calibration for ProdAD risk

SuPAC-CE relies on a growing population of predictors γs and corresponding errors $R(\gamma)$. For the calibration of ProdAD, these errors are computed by confronting the simulation with observation as described in section 5.1.1. To obtain uncertainty quantification on the predictions, the simulations $\mathsf{State}(t, \gamma)$ were saved during the calibration procedure, and quantiles were obtained by weighing each γ according to the posterior distribution.

ProdAD models the behaviour of an AD Plant through the ODE

$$\frac{\mathrm{d}\mathsf{State}}{\mathrm{d}t}(t) = \mathsf{ProdAD}\left(\mathsf{State}(t), \gamma, \mathsf{Feed}(t)\right). \tag{5.2}$$

This differential equation is solved using Euler's scheme with a fixed time step. Consequently, simulating between t_1 and $t_2 > t_1$ for a given parameter only requires the knowledge of State(t_1) and Feed(t) for $t \in]t_1, t_2]$.

During the calibration procedure, the last states of the simulations are also saved (these contain the state of the digester at midnight for each day, while the simulations contain averages of the states during the day[5]). After the feed data has been updated, new simulations can be restarted from the last day in the predictions to the last day in the feed data, and these new simulations can be aggregated using the parameter weights. To be able to perform online calibration with few new simulations, it is necessary to track the evolution of the errors for each simulation parameter in memory. If for each predictor, the simulation was stored in memory, it would then be possible to assess the risk efficiently by concatenating the previous and new simulation, and confronting it to the full observation. Storing all simulations performed during calibration would require large memory usage and was ruled out[6] On the other hand, it is not possible to reconstruct the error for a simulation between time t_0 and time t_2 from the error between the simulation error between t_0, t_1 and the simulation error between t_1, t_2, since the total error is a combination of multiple errors for different types of observations. For each of this individual error however, it is possible to combine the error values from disjoint time intervals, using

$$\begin{aligned}&\mathsf{MSE}_{\mathsf{obs}}(\gamma, [t_0, t_2]) \\ =&\frac{\left(\left|\mathcal{T}_{\mathsf{obs}}^{[t_0,t_1]}\right| + N_f\right) \mathsf{MSE}_{\mathsf{obs}}(\gamma, [t_0, t_1]) + \left(\left|\mathcal{T}_{\mathsf{obs}}^{]t_1,t_2]}\right| + N_f\right) \mathsf{MSE}_{\mathsf{obs}}(\gamma,]t_1, t_2])}{\left|\mathcal{T}_{\mathsf{obs}}^{[t_0,t_2]}\right| + N_f}.\end{aligned}$$

Hence, the number of observations and pseudo square error are also saved during the calibration procedure, and the above formula is used to update the errors when the feed is updated. The recalibration is performed using SuPAC-CE, with an initial population of γs and $R(\gamma)$s consisting of all simulations in memory. The algorithm is run with hyperparameters specific to the online setting (see below for selection procedure). This recalibration procedure is run when more than a specified number of days have appeared since the last calibration. During the recalibration procedure, new simulations are evaluated; these are merged with the previous simulations by selecting a fixed number of simulations (N_{memory}); the simulations receiving lowest weight by the posterior distribution are disregarded (see Figure 5.3).

The update of existing simulations (triggered by the feed update) and the recalibration process are decoupled. To avoid storing too large amount of data on the database, individual predictions for each simulation are not stored. Only aggregated predictions containing the quantiles

[5] While the difference is small, using the average state rather than the last state has non negligible impact on the predictions, and should therefore be ruled out

[6] SuPAC-CE proved quite memory intensive and require careful handling for simulation storages during batch calibration. This has a noticeable impact (up to 20%) on the calibration procedure time due to the necessity of accessing simulations stored on the disk when computing quantiles. We expect this could be improved, as the amount of data, if quite large (tensors of $\sim$ 500 predictions for $\sim$ 500 days for ~ 5000 sets of parameters), remains smaller than the amount of data involved in deep learning settings.

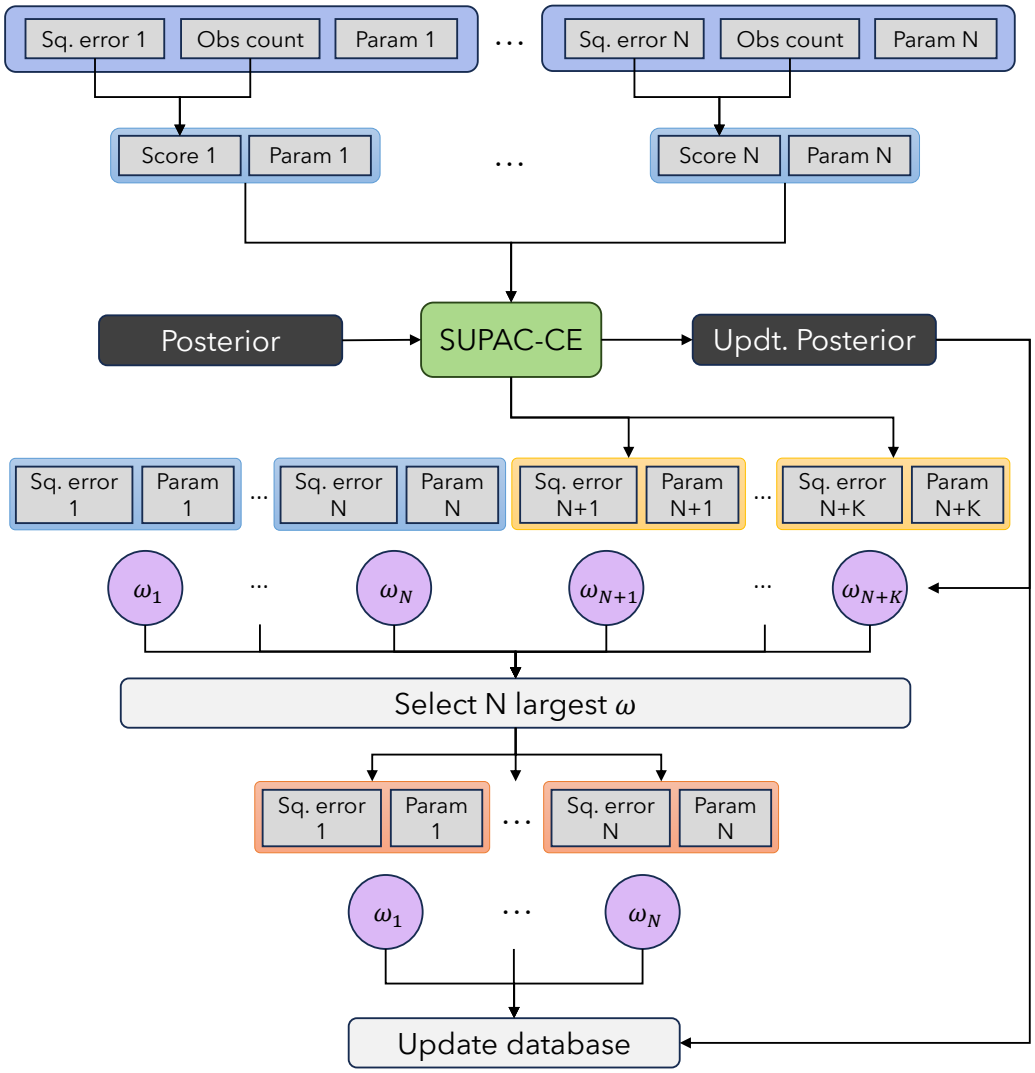


Figure 5.3: Online calibration scheme. The tracked square errors associated to each parameter are used to infer error scores. This initial population of parameters and scores (in blue) are used to update the posterior using SuPAC-CE algorithm. K new parameters are evaluated during the online calibration (new evaluations in yellow). Using the posterior to define weights, the fraction of parameter with highest mass are conserved to define the new tracked simulations.

are stored in the base. The predictions quantiles are computed during the update of existing simulations, and are not updated backward (*i.e.* if the feed is updated to contain new data between t_1 and t_2, the prediction update script will write the prediction quantiles between t_1 and t_2, while prediction quantiles before t_1 will not be changed). Thus, the historic prediction quantiles displayed in the interface are not modified by the change of posterior caused by the online recalibration. Only future predictions will be affected. This delay in updating the posterior should not be an issue if data is regularly uploaded to the database but could lead to performance decrease if the feed data is updated at long intervals.

5.2.2 Handling data validity

In practice, the feed and observation data might be uploaded with a delay. This is notably the case for offline analyses which require a few days or even a few weeks to be completed. To be able to run simulations, the missing feed values will be completed by filling missing values by the latest available analysis. In these cases, the resulting feed data might be backward corrected, *i.e.*, for $t < t_1 < t_2$, noting Feed^{t_1} (resp. Feed^{t_2}) the feed downloaded at time t_1 (resp. t_2), it could be that $\text{Feed}^{t_1}(t) \neq \text{Feed}^{t_2}(t)$. A similar difficulty holds for the observation data; values which were considered missing at time t_1 might become available at time t_2, which breaks the decomposition of the pseudo square error and number of observations resulting in the error decomposition. To overcome this issue, a back up state, which corresponds to the state of the digester 60 days before the end of the simulation, was saved during the calibration, and the score was changed so that it no longer considers the last 60 days of the simulation. These last 60 days of data are considered not to be consolidated, while on the contrary, the observation and feed data which are 60 days older than the last feed data are considered to be consolidated. As long as the consolidated data has not changed, the online procedure can run smoothly. If the consolidated data has been modified, a new batch calibration is required. If the unconsolidated data has been modified, simulations are automatically re-run from the back up (see Figure 5.4).

5.2.3 Assessing online calibration hyperparameters

Hyperparameter description

The resulting framework requires the selection of 3 hyperparameters:

- The period of recalibration ($\Delta_{\text{t, recal}}$)
- The number of simulations kept in memory (N_{memory})
- The hyperparameters of SuPAC-CE routine, and in particular the weekly number of simulations ($n_{\text{simu/week}}$)

The impact of these hyperparameters were investigated by simulating the performance of the online procedure on 280 days (forty weeks) on a digester initially calibrated on 280 days on a

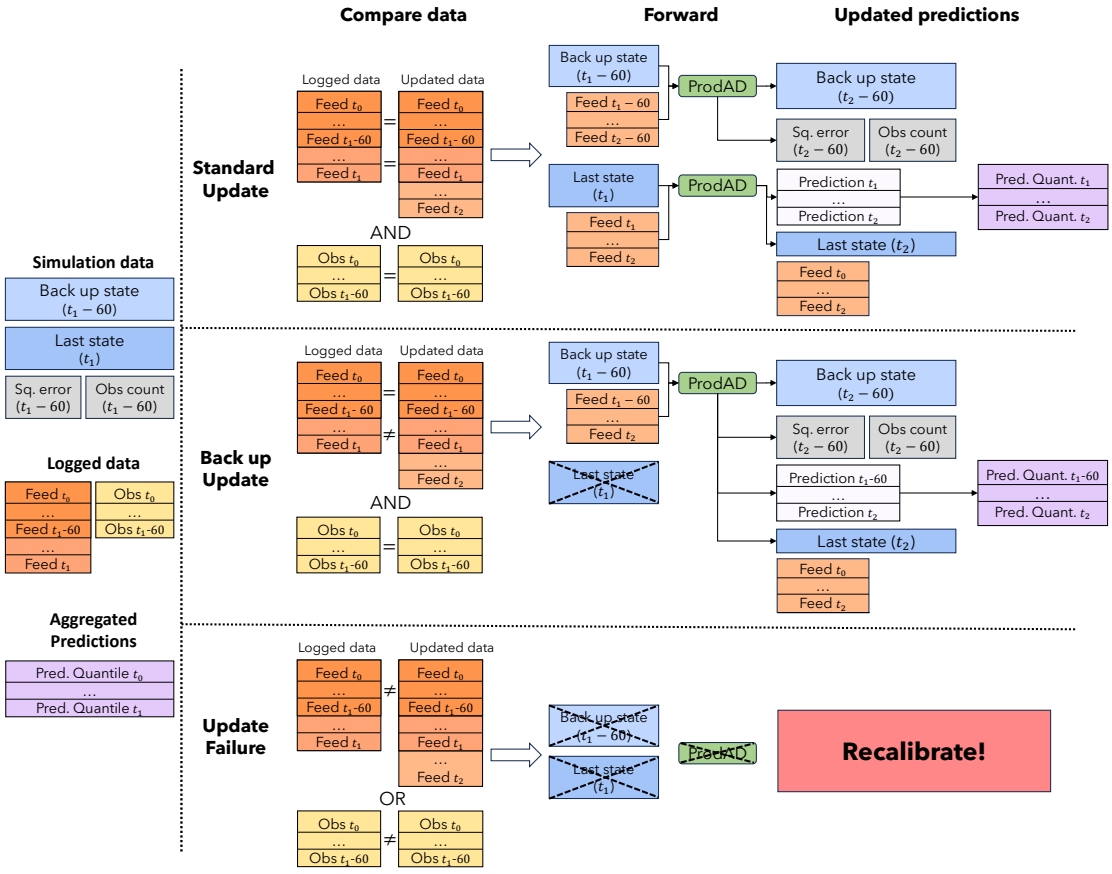


Figure 5.4: Online prediction updates scheme. Depending on whether the consolidated data and unconsolidated data was changed, the update scheme follows one of three options (top graph: both unconsolidated and consolidated data were preserved; middle graph: consolidated data was preserved but unconsolidated data changed; bottom script: the consolidated data was changed). After ProdAD simulations for all stored parameters have been completed, predictions quantiles are obtained and logged to the base, while individual simulations are disregarded.

grid of hyperparameters. Real digester data was used, corresponding to plant A. Values investigated for $\Delta_{\text{t, recal}}$ were 7, 14, 28 and 56 days. Values investigated for N_{memory} were 2500, 5000 and 10000. Values investigated for $n_{\text{simu/week}}$ were 32, 64 and 128. The theoretical computation cost of the procedure, counting only the simulation time, depends on the hyperparameters in the form of

$$\begin{aligned}\text{Cost} =& N_{\text{memory}} \times \left(T_{\text{end onl.}} - T_{\text{beg. onl.}}\right) \\ &+ \left(n_{\text{simu/week}}/7\right) \times \left(\left(T_{\text{beg. onl.}} + T_{\text{end onl.}}\right)/2 - T_0\right) \times \left(T_{\text{end onl.}} - T_{\text{beg. onl.}}\right),\end{aligned} \tag{5.3}$$

where $T_{\text{beg. onl.}}$ is the start date of the online procedure in days (here 280), $T_{\text{end. onl.}}$ is the end date of the online procedure in days (here 560), T_0 is the start date of the simulation (here 0). Noting that simulations run in approximately 15 seconds per year of simulation[7], the longest incompressible compute time amounts to 3.5 hours on a 16 core compute. This should be compared to the annual 9.3 hours spent in daily opening the compute to check if updates are required.

Other design choices for SuPAC-CE algorithm are as follow. Per optimisation step, 32 new simulations are performed. The number of optimisation step is adjusted in order to guarantee that $n_{\text{simu/week}}$ simulations are performed per week (*e.g.* if $\Delta_{\text{t, recal}} = 14$ and $n_{\text{simu/week}} = 128$, 8 optimisation steps are performed every two weeks of new data). Parameters impacting the speed and stability of the calibration procedure are chosen so that the procedure favours stability over speed ($\alpha_{\max} = 0.1$, $\mathrm{kl}_{\max} = 0.1$, number of samples for weights estimation of 100 000). The remaining parameters are left to their default values.

Analysis and results

Procedure time The procedure time varied between 124 minutes ($n_{\text{simu/week}} = 32$, $\Delta_{\text{t, recal}} = 56$, $N_{\text{memory}} = 2500$) and 523 minutes ($n_{\text{simu/week}} = 128$, $\Delta_{\text{t, recal}} = 7$, $N_{\text{memory}} = 1000$) on 16 cores CPU. This largest time was more than twice the theoretical time of 3h30. This performance loss could be explained by the extensive use of parallelisation for moderately short tasks ($\sim$ 0.3 s to simulate 7 days). In this setting, parallelisation overhead becomes apparent. This explanation was corroborated by the longer time reported for procedure with smaller $\Delta_{\text{t, recal}}$, which splits the simulations in shorter periods. In practice, this extra overhead time would depend on the regularity of upload of feed data rather than on $\Delta_{\text{t, recal}}$.

Impact of the hyperparameters on the end performance All online procedures resulted in smaller objectives compared to the one obtained by the calibration process launched on all 560 days of data at once (resulting objective of 0.1678). This vindicates the performance of the online PAC-Bayesian approaches. Indeed, the extra optimisation steps involved in the online approach appear to outweigh the downsides of calibrating online. This also hints that the batch

[7] Contrary to the use case explored in Section 4.2, the simulations also involve complex inlet flow resolutions as well as multiple digester tanks.

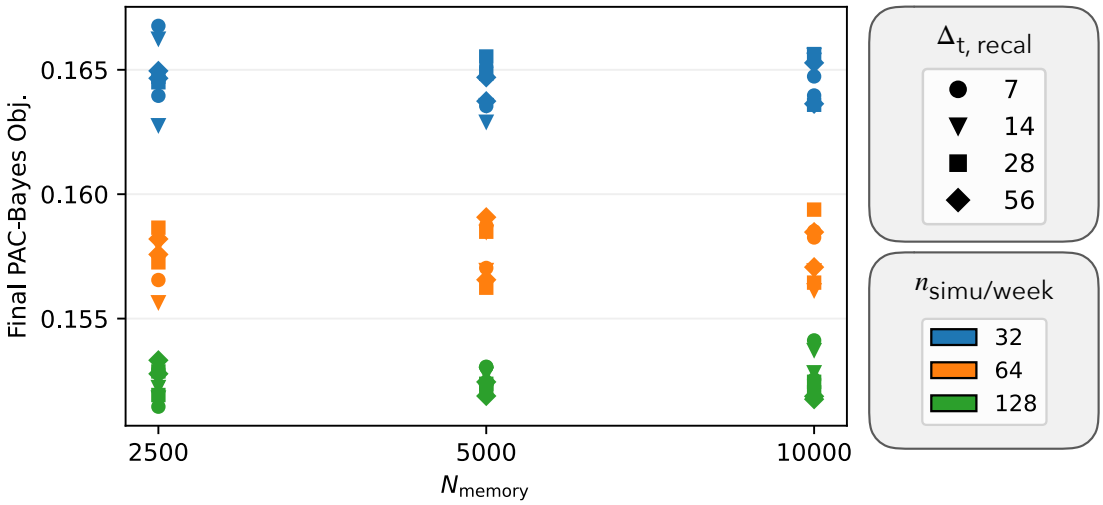


Figure 5.5: PAC-Bayes objective at the end of the online procedure as a function of the hyperparameters.

calibration budget should be increased. Interactions between the different hyperparameters on the final PAC-Bayes objective were found to be statistically insignificant. As a result, the simplified model

$$\text{Bound} = f_0 + f_1(n_{\text{simu/week}}) + f_2(\Delta_{\text{t, recal}}) + f_3(N_{\text{memory}}) + \varepsilon \tag{5.4}$$

was analysed. This model, with 8 degrees of freedom, was fitted on the 72 simulation results (adjusted R^2 of 0.966). Analysis of variance showed that only $n_{\text{simu/week}}$ had a statistically significant impact on the final score (p-value of $4.5e-49$), with $f_1(64) \in [-0.008, -0.006]$ and $f_1(128) \in [-0.013, -0.012]$ (the reference being $f_1(32) = 0$). The hypotheses that $f_2 = 0$ and $f_3 = 0$ could not be rejected (p-values of 0.58 and 0.74 respectively). The impact of the hyperparameters on the final performance of the posterior is summarized in Figure 5.5.

Stability of the end performance The simulations were performed in duplicates to assess the impact of SuPAC-CE's inherent randomness on the results. A standard deviation of 0.0094 for the PAC-Bayesian objective at the end of the online procedure was measured between the duplicates. While $\Delta_{\text{t, recal}}$ and N_{memory} had no measurable impact on the stability of the algorithm, $n_{\text{simu/week}}$ had, resulting in a significantly lower standard deviation between duplicates for $n_{\text{simu/week}} = 128$ compared to $n_{\text{simu/week}} = 32,\ 64$ (p-values of 1.02 e-4 and 3.3 e-3 for F-tests).

Impact of the hyperparameters on the procedure resilience The dynamic evolution of the PAC-Bayes objective during the online procedure was also investigated (Figure 5.6). A large fraction of the simulations constructed while computing the initial posterior distribution on the

first 280 days of data failed between day 340 and day 360. These failed simulations result in the maximum error value of 4, forcing the posterior to shift is weight away from these simulations. The impact of the hyperparameters on the resilience of the online procedure was qualitatively assessed. $n_{\text{simu/week}}$ proved to have the largest impact on the highest PAC-Bayes objective reached during the event, with the larger values leading to more resilient models. $\Delta_{\text{t, recal}}$ also impacted on the resilience, with larger values leading to smaller peak objective. Such procedures spend their optimisation budget to once sufficient data is accumulated, resulting in smaller optimisation bounds, but at the cost of being less responsive. The number of simulations kept in memory proved not to have marked impact.

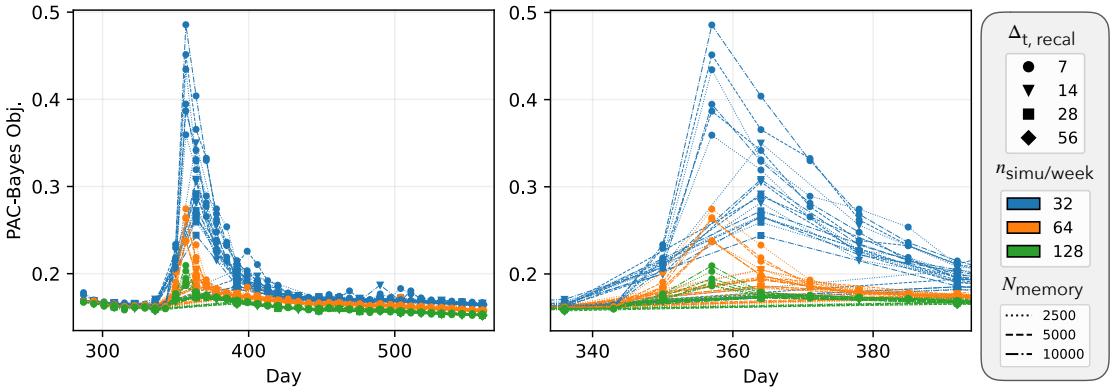


Figure 5.6: Evolution of the PAC-Bayes objective during the online calibration process for different choices of hyperparameters. The figure on the left represents the evolution of the PAC-Bayes objective throughout the experiment, while the figure on the right focus on the evolution between days 335 and 395, where a temporary increase of the objective occurs.

5.2.4 Final design of the online procedure

All three analyses support choosing a large value of $n_{\text{simu/week}}$. $\Delta_{\text{t, recal}}$ impacts the trade-off between the responsivity of the procedure and the performance. As N_{memory} has little impact on the performance, a moderate value should be selected to limit the computational cost of the procedure.

The hyperparameters of the online procedure were set to $n_{\text{simu/week}} = 128$, $\Delta_{\text{t, recal}} = 14$ or 28 and $N_{\text{memory}} = 5000$. While the experiments suggested choosing the smaller $N_{\text{memory}} = 2500$, the current choice is more conservative and was preferred pending evaluations on more plants. Indeed, the procedure's hyperparameters should be reassessed on other plants to check if the findings above can be generalized. Larger values for $n_{\text{simu/week}}$ as well as smaller values for N_{memory} could also be investigated.

Other designs for SuPAC-CE's hyperparameters during recalibration could be investigated. Notably, adding optimisation steps with no new simulations while lowering the optimisation speed could result in improved results. Such a strategy was used to advantage in the meta-learning experiments of Chapter 6.

The difference between the measured online procedure times and the theoretical counterpart showed that the procedure could benefit from lowering the parallelisation overhead. Similarly, a large time of the compute time of the online procedure as currently implemented has been identified to be importing the code module when launching the scripts. To mitigate these issues, the online calibration scripts are executed on smaller 8 cores computes.

5.2.5 Conclusion

We introduced an online calibration and monitoring procedure, working in synergy with the batch calibration procedure. The procedure is designed to tackle asynchronous upload of data. The online procedure was assessed for various hyperparameters using real digester data. This guided the choice of the hyperparameters for the online procedure currently used for monitoring two industrial AD plant operated by SUEZ. This preliminary design should be further refined after assessing more AD plants.

5.3 General conclusion

The calibration strategy SuPAC-CE, assessed previously on synthetic AD tasks, proved adequate on two real-world, industrial AD plants. We introduced common rules to construct comparable risk functions on diverse plants, and proposed an initial sequence of hyperparameters for the calibration process obtained satisfactory results. An online calibration and monitoring routine was constructed, and is now employed to monitor these plants. Due to the limited number of AD plants configured to be modelled with SUEZ's ProdAD model, these results should at this stage be considered preparatory, and the current routines design should be reassessed and updated when new plants data are available.

Chapter 6

Meta Anaerobic Digestion modelling

This chapter tackles the use of meta-learning strategies to AD modelling. It consists of two, mostly independent sections:

- in Section 6.1, the prediction of a key performance indicator of AD is performed through a data-pooling, non PAC-Bayes, approach. This represents a first, rudimentary way to benefit from multiple plants data, and proved adequate when applied to the modelling of a single feature of oversized digester processes. Results from this section were obtained in collaboration with Mathieu Haddad, Danielle Trap, Damien Batstone and Roman Moscoviz, and are currently under review at the Journal of Water Process Engineering.
- in Section 6.2, an optimisation based meta-learning approach using the PAC-Bayes learning algorithm of Chapter 4 is introduced, and assessed on synthetic toy tasks and AD calibration tasks. This section extends on the PAC-Bayes meta-learning strategy and preliminary experiments published in Picard-Weibel et al. [2024b].

Contents

A first approach towards learning from multiple plants consist in simply pooling the different plants data when calibrating. Such an approach should provide adequate results when the different datasets considered can be described by a shared model, *e.g.* when the relationship between input and output data is identical between the different plants (same parameter), but the input data varies between the plants (change of operational conditions); training on the pooled data should result in a more robust model. This approach was applied to infer VSR of sewage sludge AD in WWTPs, using a non-PAC Bayesian learning strategy.

Modelling sludge AD, which involves oversized digesters and a limited variety of feed flows, for a specific target is well suited to a data pooling approach. The calibration of generic AD process is a more complex setting. Reports from the literature indicate wide range of plausible values for the parameters to be calibrated. Rosén and Jeppsson [2006] reports uncertainty of 300% for some parameters. Hence the data pooling approach is unlikely to give satisfactory results, since no single set of parameters can give adequate performance on a wide range of AD process. The two-level learning strategy pursued by meta-learning appears better suited to tackle the task of AD modelling. At the meta-learning level, the common structure of AD process can be learnt, improving the inner-learning calibration strategy used for each AD process by feeding it inductive bias.

In the context of the PAC-Bayes calibration strategy used to calibrate AD process, the inductive bias which can be learnt at the meta-level is the prior. Our second contribution in this chapter is presenting how the PAC-Bayes learning strategy SuPAC-CE can efficiently be extended to perform meta calibration without too drastic increase in the number of risk queries.

6.1 Learning from multiple plants: volatile solid reduction use case

This section is based on work done conjointly with Mathieu Haddad, Danielle Trap, Damien Batstone and Roman Moscoviz. I contributed to the training approach, the benchmark methodology and the actual implementation.

6.1.1 Introduction

VSR is a key performance indicator of anaerobic digestion processes. VSR is defined as a mass ratio of VS consumed during anaerobic digestion and VS entering the digester. This can be computed using mass balance during a period of time (*e.g.* 30 days), *i.e.*

$$\mathrm{VSR} = (\mathrm{VS_{in}} + \mathrm{VS_{accu}} - \mathrm{VS_{out}})/\mathrm{VS_{in}} \tag{6.1}$$

where $\mathrm{VS_{in}}$ is the mass of VS which entered the digester during the considered time span, $\mathrm{VS_{out}}$ the mass of VS which left the digester during the considered time span, and $\mathrm{VS_{accu}}$ the accumulation of VS during the time span (the difference between the amount of VS at the end and the beginning, which might be positive or negative). Building AD systems with high VSR has two positive implications for the process. First, the larger the VSR, the more organic matter has been consumed and converted to valuable biogas. Second, the larger the VSR, the smaller the total solids content of the digestate, implying less weight in the solid fraction of the digestate which, for wastewater treatment plant, has to be further processed. Indeed, the historic motivation for anaerobic digestion for wastewater plant was reducing the total solids content of the sludges rather than biogas production.

Knowledge of a process's VSR can be combined with estimation of the methane yield, defined as the amount of methane produced per unit of mass of VS consumed, and methane content (fraction of methane in the biogas) to estimate the production of biogas of a plant, given a feedstock VS content [Banks et al., 2011]. Measures of methane yield can be indirectly obtained from the COD to VS ratio (COD/VS), which can be measured for a given feedstock, and from the theoretical production of 350 NL of methane per kg of removed COD [Jensen et al., 2022], while Moscoviz and Jimenez [2021] constructed a procedure to predict biogas content with high precision from usually measured characteristics of a feedstock. The prediction of the VSR is a less studied problem. [Liptak, 1974] proposed estimating the VSR from the **S**olid **R**etention **T**ime (SRT) in the anaerobic digester as

$$\mathrm{VSR} = 13.7 \ln(\mathrm{SRT}_d) + 18.9, \tag{6.2}$$

where VSR is the volatile solids reduction, in %, and SRT_d is the design solids retention time in the anaerobic digester in days.

Equation (6.2) is still cited in reference works [Appels et al., 2008, Turovskiy and Mathai,

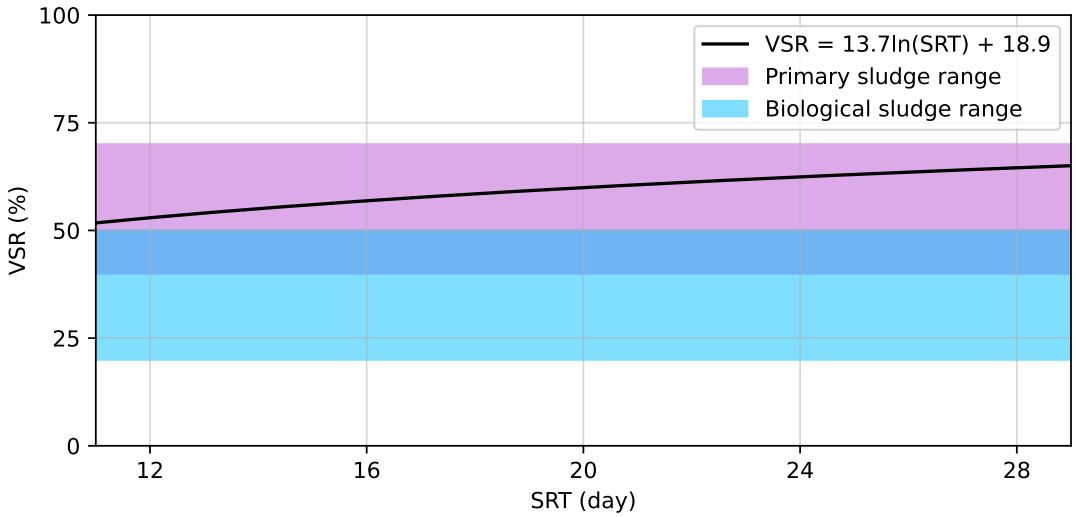


Figure 6.1: Volatile Solids Reduction (VSR) prediction from Solids Retention Time (SRT), using Equation (6.2). The typical range of VSR for primary sludge (resp. biological sludge) are represented in blue (resp. purple). The predicted VSR is contained in a much narrower range of values compared to what is reported in the literature.

2006]. However, literature reports for the typical values of VSR for wastewater sludges appear at odds with the model. Typical VSR values for primary sludge span from 40% to 70%, while typical VSR values for secondary activated sludge span from 20% to 50% [Albertson et al., 1987, Bolzonella et al., 2005, Arnaiz et al., 2006, Tezel et al., 2011]. For typical SRT of 15 to 20 days, Equation (6.2) predicts a much narrower range of VSR values, spanning from 56% and 60% for any type of sludge (see Figure 6.1). AD models, such as ADM1, provide another way to predict a process VSR. Calibrating ADM1 is necessary to adequately reproduce a digester's behaviour. Using default parameters, simulation outputs can result in an inaccurate biogas production estimation - Baquerizo et al. [2021] reported a 36% underestimation of biogas production - which is translated into inadequate VSR prediction. Indeed, for feedstock with low total solid contents such as occur for wastewater treatment, hydrolysis limits the process. Values for the hydrolysis rate have been reported to vary on 3 order of magnitude (0.001 to 2.2 day^{-1}, Mo et al. [2023]), making the default value of 0.5 day^{-1} little more than an educated guess. Assuming that hydrolysis follows a first order rate [Vavilin et al., 2008], the parameters which require calibration are the degradation extend, the fraction of substrate that can be converted and the first order hydrolysis rate coefficient (k_{hyd}).

When it comes to predicting AD performances for WWTP that are not yet equipped with an AD system, calibration of the model is no longer possible. Due to the limited predictive ability of mechanistic AD models using default parameters, an empirical model that accurately predicts the VSR achieved with different types of municipal sludge is thus critical. For this new model to be as widely applicable as possible, it must include more than the common AD operating

parameters, which are the HRT, temperature, pH and alkalinity, organic loading rate, nutrients, and inhibitors [Tezel et al., 2011]. For instance, even if the methane production is directly linked to the HRT of the anaerobic digester, more parameters were shown to have an impact on the VSR [Jensen et al., 2022]. Bolzonella et al. [2005] demonstrated a 61% decrease in the specific biogas production of waste activated sludge when increasing the sludge age in the aeration tanks from 8 to 35 days. Authors were able to establish empirical relationship between the biogas yield and the sludge age achieved in the activated sludge process. These results are in line with the work of Chen et al. [2020] that showed an exponential decrease of VSR of secondary activated sludge when the sludge age during wastewater treatment increased from 5 to 40 days. Authors demonstrated that an increase in molecular weight and stability of the extracellular polymeric substances produced lead to the decline in sludge biodegradability during AD. It is thus to be expected that the primary to secondary activated sludge ratio, their respective VS content as well as the environmental conditions in which they were produced, such as the effluent temperature, will impact the VSR obtained when digesting them jointly or separately. Therefore, a universal empirical VSR model for municipal sludge digestion would need to include most of the parameters linked to the above mentioned feed sludge quality and not solely focus on the AD operating conditions.

This works aims at establishing a predictive empirical model for VSR of municipal sludge digestion operating under mesophilic conditions. This was done using over 32 years of daily data from 6 different industrial plants and by considering both the wastewater and sludge treatment line for each plant. Performances were compared to those obtained using the fitted and un-fitted mechanistic ADM1 model, used as a reference.

6.1.2 Methodology

Plants description

The main characteristics of each wastewater treatment plant considered in this study are described in Table 6.1. All WWTPs included a conventional activated sludge or a high rate activated sludge process with or without primary settling tanks upstream. The temperature of all digesters remained in a similar range of $37 \pm 2^{\circ}$C, with extreme observations of 33 to 42°C.

Variable definitions

The datasets had been previously cleaned to remove outliers and infer part of the missing data.

Modelling was carried out based on directly measured and indirectly measured variables. Measured variables included wastewater temperature, **T**otal **S**olids (TS) concentration and VS fraction of both **P**rimary **S**ludge (PS) and **W**aste-**A**ctivated **S**ludge (WAS) as well as the PS to WAS ratio (VS-based) at the digestion system inlet. Indirectly measured variables included the sludge age of the activated sludge biomass, the HRT of the AD system and the VSR obtained in the AD system.

N°		1	2	3	4	5	6
Plant Location		Oceania	Middle East	South America	Central America	Western Europe	Western Europe
Period (days)		1810	4696	1811	2415	762	192
Sludge type processed by AD		WAS	PS + WAS	PS + WAS	WAS	PS	WAS
Wastewater temperature (℃)	**min**	20.70	15.25	15.10	24.61	13.69	13.95
	avg	25.75	18.95	23.06	28.83	17.93	15.79
	max	30.16	21.95	28.87	30.66	22.18	18.46
Activated Sludge age (days)	**min**	8.27	1.94	6.20	1.89	-	16.82
	avg	13.29	3.25	9.58	5.71	-	19.05
	max	23.65	5.44	21.19	16.78	-	23.20
Digester HRT (days)	**min**	21.98	22.26	16.56	16.65	14.74	53.62
	avg	33.67	26.04	29.98	28.03	20.66	66.60
	max	43.68	31.52	56.51	53.00	35.18	81.97
PS/ (PS + WAS) (VS-basis)	**min**	0%	52.45%	50.29%	0%	100%	0%
	avg	0%	63.30%	61.06%	0%	100%	0%
	max	0%	75.35%	74.56%	0%	100%	0%
VS content of PS (%TS)	**min**	-	63.64%	53.98%	-	65.76%	-
	avg	-	74.66%	70.14%	-	78.06%	-
	max	-	80.50%	77.03%	-	82.11%	-

Table 6.1: Wastewater Treatment Plant and sludge characteristics used to establish and test the predictive Volatile Solids Reduction model. PS: Primary sludge; WAS: Waste activate sludge; VS: Volatile Solids; TS: Total Solids.

Empirical Volatile Solids reduction modelling

To predict VSR, a linear model with feature engineering was considered. The initial five features $(X_i)_{i\in[1,5]}$ considered for VSR predictions were:

- HRT, the digester Hydraulic Retention Time, in days
- SRT_{AS}, the combined aerobic and anoxic activated sludge age, in days
- T_{ww}, the temperature of the waste water, in degrees Celsius
- VS_{PS}, the primary sludge Volatile Solids (VS) content, in $\%_{TS}$
- m_{PS}, the mass-based fraction of primary sludge in the mixed sludge fed to the AD system, in $\%_{VS}$

To these original features were added the following engineered features:

- Powers, X^k, for $k \in \{-2, -1, 2\}$ and X any of the original features (*e.g.* HRT^2, T_{ww}^{-1}),
- Multiplications, $X_i \times X_j$, for X_i and X_j any two original features (*e.g.* $HRT \times T_{ww}$ but not $HRT^2 \times T_{ww}$)
- Divisions, $\frac{X_i}{X_j}$, for X_i and X_j any two original features (*e.g.* HRT/T_{ww}).

This results in a total of 50 features. However, whenever an engineered feature could not be computed for all data points (*e.g.* X_j^{-1} if X_j contains 0, for instance for m_{PS}), it was removed.

The resulting features were then renormalised to define the feature matrix $\mathbf{X}$, with columns designating features, and rows observations. The goal is to construct a coefficient vector w such that VSR $\simeq Xw$.

Weights were assigned to each data point in the following fashion. Denoting n_p the number of data points for a plant p, a data point for plant p is given the weight $\frac{1}{\sqrt{n_p}}$. This implies that the total weight of a plant is $\sqrt{n_p}$, striking a balance between giving identical weight to each data point and giving identical weight to each plant. The resulting weight vector is denoted ω.

The model was learned by minimising an ℓ_1-penalized least squares objective (Lasso),

$$\min_{\mathbf{w}} \frac{\|\omega \times (y - \mathbf{X}\mathbf{w})\|_2^2}{2n_{\text{samples}}\|\omega\|_1^2} + \alpha\|\mathbf{w}\|_1 \tag{6.3}$$

where

- y is the target Volatile Solids Reduction (VSR), in $\%_{\text{VS}}$, determined earlier,
- $\mathbf{X}$ is the feature matrix described above,
- w is the coefficient vector to be estimated,
- ω is the weights vector,
- α is the hyperparameter that controls the level of regularisation ($\alpha \geq 0$).

ℓ_1 penalisation favours linear models with few non zero coefficients (*i.e.* sparse), and hence integrates variable selection into the learning process [Tibshirani, 1996]. The higher the hyperparameter α, the fewer features will be used in the model. The hyperparameter α is selected through a modified leave one out validation approach, where a whole plant is left out (**L**eave **O**ne **P**lant **O**ut (LOPO), see Figure 6.2). For each plant, the model is trained on all remaining plants and assessed on the plant by computing the RMSE of the residual. Averaging the results obtained for all plant gives a score to the hyperparameter; the hyperparameter obtaining the lowest score is selected to train the model on all plants. The hyperparameter $\alpha = 0.1701$ was selected after assessing a geometric grid of 40 values between 0.01 and 100.

VSR prediction using ADM1

The VSR for each plant was predicted using a calibrated and non-calibrated version of the ADM1 model for each plant (see Figure 6.2). All parameters, except those calibrated, were set to the values given by Rosén and Jeppsson [2006]. Values for the initial state of the digester were taken from the same source.

Three key parameters of ADM1 were considered for calibration: the BMP of primary sludge (B_0^{prim}, default to 450 NL/kgVS), the BMP of biological sludge (B_0^{bio}, default to 250 NL/kgVS) and the hydrolysis coefficient (k_{hyd}, default to 0.5 day^{-1}). The calibration was performed using a grid search algorithm, with values between 200 and 500 NmL/gVS investigated for B_0^{prim} (grid step size of 10), values between 100 and 350 NmL/gVS investigated for B_0^{bio} (grid step size of

10) and values between 0.1 and 1.0 day^{-1} investigated for k_{hyd} (step of 0.05). For each plant, the set of parameters achieving the lowest RMSE on the first 75% part of the data (train data) was selected. When the dataset consisted of two non contiguous periods of time, the train data consisted in the union of the first 75% of each period. The calibrated model is then assessed on the remaining part of the data (test data).

The intrant description, based on VS content, was changed into ADM1's COD based intrant description using the following hypotheses: the fractions of lipids, proteins, and carbohydrates in the degradable fraction of the VS were set to 0.8, 0.1, 0.1 for the primary sludge (resp. 0.1, 0.5, 0.4 for biological sludge). To compute the degradable fraction of the VS, the COD per VS ratio of the primary sludge was assumed to be 1.8 gCOD/gVS (resp. 1.5 for biological sludge). All fractions other than lipids, proteins, carbohydrates and inerts were deemed negligible and set to zero.

VSR was inferred from the output of ADM1 assuming gCOD per gVS ratios of 1.03 for carbohydrates, 1.5 for proteins, 2.0 for lipids and 1.5 for inerts. After computing the VS mass in the intrant and the digester, the VSR was computed using a balance on 30 days using Equation (6.1).

Model assessment

The performance of the models was assessed through the RMSE and bias. These are computed on the last 25% part of the data for each plant, corresponding the test datasets for ADM1's fitting procedure (see Figure 6.2). To ensure that no contamination between train and test data takes place when evaluating the empirical methodology, the following two-level LOPO procedure was performed. For a plant P, LOPO validation is used on the remaining plants to select the hyperparameter. The model is then trained on all the remaining plants. As such, the model constructed is wholly independent from the original plant P, since neither the hyperparameter selection procedure nor the learning procedure used data from P. The performance of this model on P assesses the test performance of the LOPO hyperparameter selection methodology. These model performances were measured in terms of RMSE and bias and are reported in Figure 6.4.

Empirical approach performance and benchmarking

Intermediate empirical models for VSR prediction were developed for the digestion system installed on each of the 6 WWTPs. Indeed, different empirical models were obtained for each plant as the result of the "Leave one plant out" cross validation approach. Performances of these models were then compared to the ones obtained with both fitted and un-fitted mechanistic ADM1 models using the same test datasets (see Figure 6.3).

First, the unfitted ADM1 (ADM1-default) model was used for VSR prediction as a baseline control. Indeed, it corresponds to a straightforward and readily available method which could be used to predict the performances of a new AD plant (*e.g.*, to be built). Thus, the minimal objective of the proposed empirical model would be to provide better performances than the

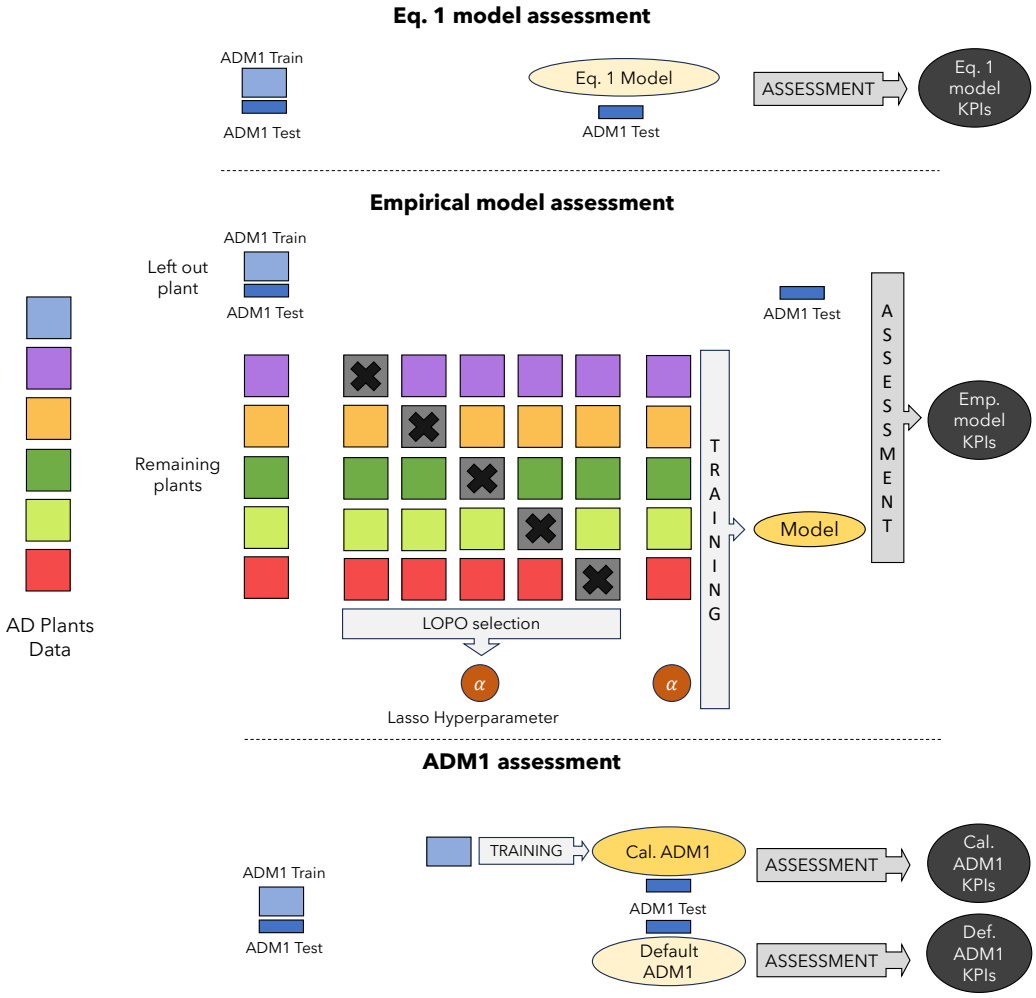


Figure 6.2: Assessment methodology for the models. All trained models were assessed on the data corresponding to the test data for the ADM1 procedure. A two level Leave One Plant Out (LOPO) approach was used the evaluation of the empirical methodology performance. For each plant, the LOPO procedure is used on the remaining plants to select the best hyperparameter value and train a model. This model is then evaluated on the last part of the original plant's data, coinciding with the test sets for the calibrated ADM1 models.

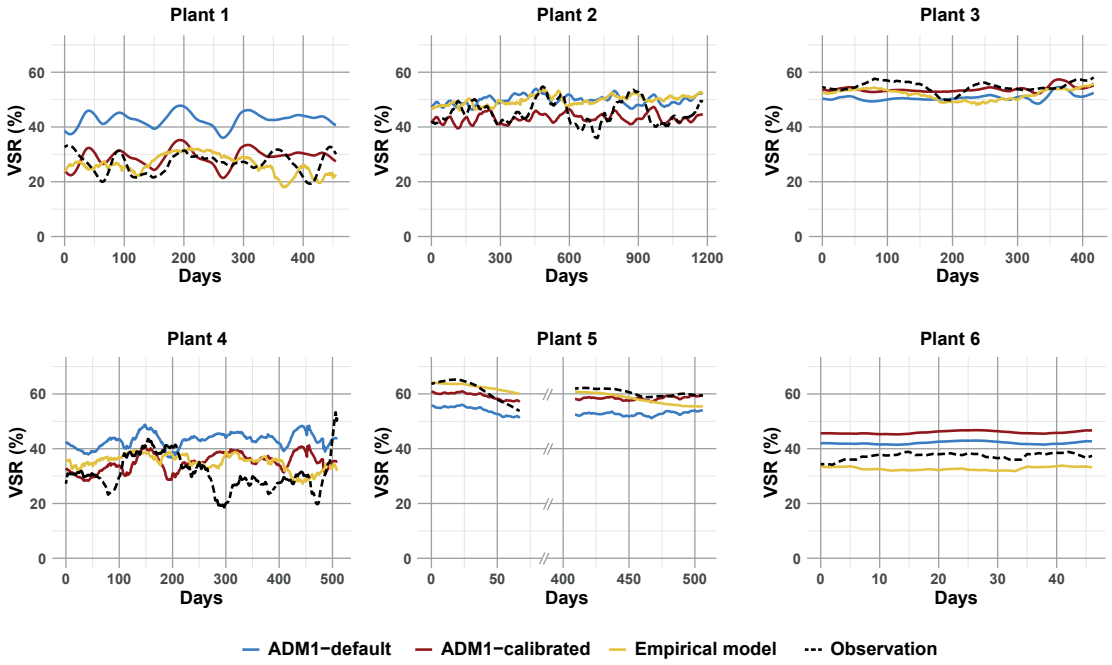


Figure 6.3: Time series of the Volatile Solids Reduction measured and predicted with the calibrated ADM1 model, the ADM1 model with default parameters and the empirical model. Only the test dataset used for prediction performance assessment is represented (25% of observations).

unfitted ADM1. On average, this approach performed poorly (average RMSE and bias of 9.0 and 8.1%VSR, respectively, see Figure 6.4) and provided the lowest performances of all models, highlighting that generic average feedstock characteristics cannot represent the diversity of actual sludge treated in industrial plants. It is to be noted that the unfitted ADM1 approach still provides a better insight compared to previously proposed empirical model (Equation (6.2)) which would lead to an average RMSE of 24.5 %VSR.

Applying a standard fitting procedure to ADM1 (ADM1-calibrated) could significantly improve the prediction performances (average RMSE and bias of 5.4 and 3.3 %VSR, respectively) despite the limited number of fitting parameters (BMPs and hydrolysis constant). However, this procedure would only be feasible for an existing AD plant with available historical data. As such, the fitted ADM1 strategy was applied as a positive control for VSR prediction when actual data from the plant is available and was expected to provide the best results of the benchmark. Surprisingly, without having access to any data from the modelled plant, the empirical models developed in this study still performed slightly better on average (average RMSE and bias of 4.7 and 2.8 %VSR, respectively) than fitted ADM1. Part of this improved performance might be related to the ability of the empirical model to integrate seasonal variability (*e.g.*, through the wastewater temperature), while ADM1 (as currently implemented) assumes constant feedstock characteristics. Those differences will be discussed in the next sections. Finally, it is worth noting that ADM1-calibrated displayed poorer performances than ADM1-default for plant

6. Such behaviour highlights the risks of ADM1 overfitting, especially when data is scarce (*e.g.*, only 144 days available for plant 6 training), while the empirical model can typically benefit from increasing data from several plants to offset such risks.

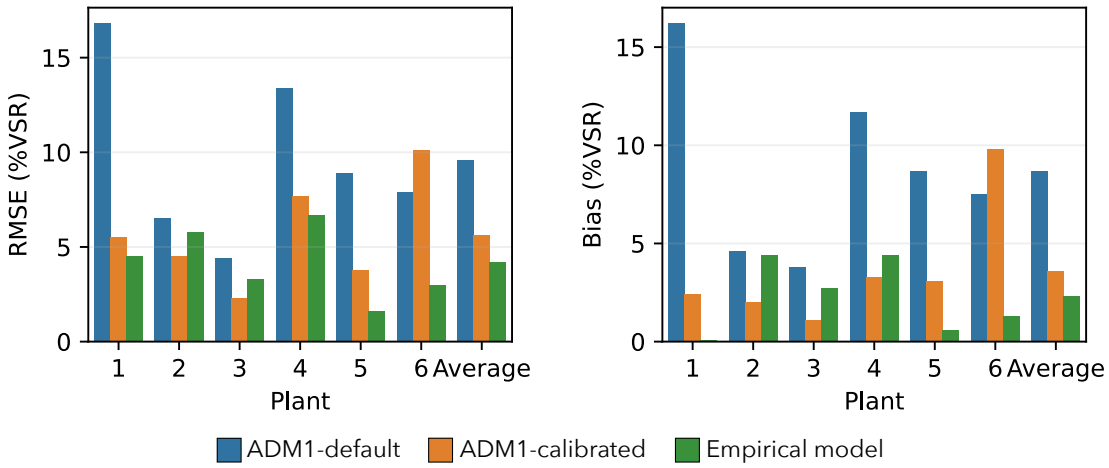


Figure 6.4: Prediction performances achieved for the Volatile Solid Reductions obtained with the calibrated ADM1 model, the ADM1 model with default parameters and the empirical model.

A robust empirical model for Volatile Solids Reduction prediction

Once the empirical model learning methodology was validated through LOPO, a single linear model was obtained from all available data. This model aims at predicting the VSR achievable on a single-stage mesophilic AD system processing primary and/or secondary municipal sludge. The equation for this empirical model is as follow:

$$\begin{aligned} \text{VSR} &= 47.5 + 12.59\text{VS}_{\text{PS}} \times \text{m}_{\text{PS}} + 8.05\text{VS}_{\text{PS}}^2 + 0.21\text{m}_{\text{PS}} - 0.014\text{T}_{\text{ww}} \times \text{SRT}_{\text{AS}} \\ &\quad - 0.16\text{T}_{\text{ww}} - 0.0014\text{T}_{\text{ww}}^2 + 54.90\frac{\text{VS}_{\text{PS}}}{\text{T}_{\text{ww}}} + 32.42\frac{\text{VS}_{\text{PS}}}{\text{HRT}} - 15.03\frac{\text{SRT}_{\text{AS}}}{\text{HRT}} \\ &\quad - 3.66\frac{\text{T}_{\text{ww}}}{\text{HRT}}. \end{aligned} \tag{6.4}$$

Although fully empirical, the model parameters selected during the training procedure were found consistent from a process perspective. First, all terms featuring HRT were selected by the automated feature engineering as the inverse of HRT. This means that for high HRT value, digestion performances would converge towards a maximum value. Thus, Equation (6.4) can be interpreted as a first part independent from HRT and corresponding to a maximum VSR (VSR_{max}), and a second part proportional to HRT^{-1} describing the interaction with the digestion process. Such inverse relationship between VSR and HRT can also be found in the theoretical equation obtained by considering a steady state in a continuous stirred-tank reactor and assuming that hydrolysis is the process-limiting step following a first order reaction of constant

k_{hyd} [Vavilin et al., 2008]:

$$VSR = VSR_{max}\left(1 - \frac{1}{1 + k_{hyd}HRT}\right). \tag{6.5}$$

As the HRT in our datasets varied between 14.7 to 81.8 days, this formula can be well approximated as $VSR = VSR_{max} - C/HRT$ provided the hydrolysis constant is not too low.

Thus, such inverse relationship between VSR and HRT is likely robust. The other part of Equation (6.4) related to VSR_{max} can also be physically interpreted or related to earlier scientific publications. Overall, this part of Equation (6.4) independent from HRT is positively correlated with the ratio of primary sludge over total sludge (m_{PS}) as well as the VS content of the primary sludge (VS_{PS}). Such behaviour is in line with primary sludge being generally more biodegradable than biological sludge [Albertson et al., 1987, Arnaiz et al., 2006, Bolzonella et al., 2005, Tezel et al., 2011]. Regarding the VS content of the primary sludge, Liu and Smith [2022] demonstrated that the biogas yield (*i.e.*, itself correlated to the VSR), increased significantly with the overall amount of VS contained in digester feed sludge. A higher VS content was correlated with an increase with the fat and cellulose contents in raw sludge, consistent with the high digestibility of both substrates [Liu and Smith, 2022]. On the other hand, an increase of wastewater temperature (T_{ww}) and/or the biological sludge age (SRT_{AS}) is predicted by the model to reduce the VSR. Higher temperatures generally enhance microbial activity, leading to increased rates of substrate utilization in aeration tanks [Wiesmann et al., 2007]. Conversely, lower temperatures slow down microbial processes, resulting in decreased degradation rates and potential accumulation of particulate organic matter. Temperature also plays a crucial role in the decay of activated sludge. Higher temperatures generally accelerate the decay rate of microbial biomass, leading to increased endogenous respiration and cell lysis [Sayigh and Malina Jr, 1978]. This results in a higher rate of sludge reduction and oxidation and consequently, a lower biodegradability. Moreover, a rise in the temperature might also accelerate the microbial activity in the sewer network, upstream of the WWTP, leading to a loss of methane potential of the primary sludge. Regarding the sludge age, Ismail et al. [2024] demonstrated that an increase of the latter decreased the methane production potential as the available biodegradable material decreased. Authors reported a biodegradability decrease from 50% to 32% as the sludge age increased from 5 to 60 days. As for the wastewater temperature, such loss in biodegradability could be attributed to a more complete oxidation of wastewater particulate pollutants as well as increased endogenous respiration.

Overall, Equation (6.4) can be considered consistent with earlier observations from the literature. It was built based on over 32 years on data obtained from 6 WWTPs located across 4 different climates, which correspond to the human climate niche described by Klinger and Ryan [2022]: moist continental mid-latitude, moist subtropical mid latitude, dry and tropical climates. Moreover, data originating from both High-Rate Activated Sludge process and Conventional Activated Sludge process, operating with and without a primary settling stage positioned upstream, were used to train the model. Therefore, this allows covering the spectrum of typical wastewater treatment lines that can be encountered, making the model already suitable for predicting the

digestion performances achievable on most plants. Yet, such empirical model can readily be improved in the future for better accuracy by increase the size of the training data (*i.e.*, including data from a wider diversity of industrial plants).

Implications for AD systems design

The present work demonstrates that the lack of accurate wastewater substrate characterization can further exacerbate the discrepancies between model predictions and actual system behaviour. At design phase, this can lead to suboptimal sizing of the ancillary equipment, potentially causing inefficiencies and increased operational costs. When applying the ADM1 model with default parameters for secondary and primary sludge for a greenfield project, an average bias of over 8 %VSR was obtained compared to the actual VSR measured on site. Comparatively, the empirical approach was able to provide a bias of less than 3 %VSR. Considering an average VSR of 45% for municipal mixed sludge [Shrestha et al., 2020], an 8 %VSR uncertainty would lead to under sizing or oversizing the biogas system and related equipment by around 18%. This is to be compared to approximately 6% with the empirical model.

While oversizing might not seem problematic from a biogas piping perspective, it can be impactful when it comes to equipment operation. Indeed, the typical design approach already factors in safety margins [Pronto et al., 2012]. This buffer capacity helps in maintaining stable operation during periods of high organic load or when the feedstock composition changes. Yet, significant oversizing results in greater initial cost to purchase equipment, lower operating efficiency in converting biogas to electricity (in the case of a Combined Heat and Power unit), and higher maintenance costs. As per the later, operating at partial load can cause increased wear and tear on the equipment.On the other hand, undersizing the biogas network and related equipment can have several significant impacts on the efficiency, reliability, and overall performance of the biogas system. First, undersized piping can lead to higher pressure drops in the biogas network. This can reduce the flow rate of biogas and eventually leading it to being vented through the pressure relief valves of the digester or flared during production peaks, which is both wasteful and environmentally harmful [Kapoor et al., 2019]. Similarly, smaller downstream equipment might not be able to process the entire flow of biogas, leading to the same negative consequences.

Implication for AD system operation

Treatment trains used in conventional WWTPs, such as the activated sludge process with extended aeration, sequential nitrification and denitrification, sequencing batch reactor, and oxidation ditch typically require high energy inputs. In a conventional activated sludge process, aeration of the biological tank alone represents nearly 60% of the total energy consumption of the treatment plant, with additional energy expended on sludge pumping, processing, and disposal [Tsalas et al., 2024]. With an upcoming directive from the European Union aiming to achieve energy neutrality in WWTPs, plant-wide optimisation becomes paramount.

The focus, so far, has been on reducing energy consumption while maintaining the same wastewater effluent standards. This is typically done through aeration control strategies, which ranges between 0.25 to over 1 kWh.m^{-3} [Maktabifard et al., 2018]. Yet, another step towards energy neutrality of WWTPs would be increasing the on-site energy production. While sludge pre-treatment technologies such as thermal hydrolysis have been widely implemented [Devos et al., 2020], the present work offers new leverage for enhancing biogas production: optimising the activated sludge age. Indeed, the empirical model developed in this work clearly demonstrates the impact of the activated sludge age on the VSR and consequently on energy production. For instance, consider an activated sludge age produced without primary settling at a temperature of 14 degree Celsius. According to Equation (6.4), the achieved VSR drops from 33% to 24% when increasing the activated sludge age from 10 to 20 days respectively. This translates into a 30% drop in biogas production when digesting the waste activated sludge. Activated sludge age thus becomes an additional decision-making factor when implementing control strategies. Minimising this factor allows a positive impact on both water and sludge treatment trains. The former is done by enhancing oxygen transfer through the reduction the concentration of biomass in the bioreactor, while later is carried out through an increase the AD performance. Moreover, as it is linked to the wastewater temperature, the present model allows for a dynamic control strategy, which takes into account seasonal variability of the wastewater.

6.1.3 Conclusion

The proposed empirical model for VSR prediction, built over 32 years of data from six different industrial plants, represents a significant advancement in the design and operation of AD systems. By providing accurate and reliable predictions, it supports the development of more efficient and sustainable wastewater treatment processes, contributing to the broader goal of energy neutrality in wastewater treatment plants. As more data becomes available, the model can be further refined to enhance its accuracy and applicability. Future research could focus on expanding the dataset to include a wider variety of industrial plants and operational conditions, thereby improving the model's robustness and reliability.

To build our VSR model, we assumed that a common equation could be used to describe all plants simultaneously. We assume that this universal formula has access to all the factors which can affect VSR, and that further details on the digester is unnecessary. As WWTPs involve stable feedstock and are operated in mild conditions (low TS content, high HRT), and we are only considering the prediction of single feature, this approximation is acceptable. To model complex AD processes however, constructing a single shared predictor is no longer possible. The data pooling approach considered here is no longer adequate. Still, the different AD processes follow the same biochemical reactions and therefore have some affinity with one another. This motivates the use of two level learning strategy such as optimisation based meta-learning.

In the following section, we propose a PAC-Bayes meta-learning objective which can be used in synergy with SuPAC-CE, which we assess for a toy experiment and two AD experiments.

6.2 Meta PAC-Bayes learning

6.2.1 Surrogate PAC-Bayes in a Meta-learning framework

Both the Bayes and PAC-Bayes framework offer a natural connection with Meta-Learning, as both involve a natural inductive bias in the form of the prior. The aim of PAC-Bayes Meta-Learning is the construction, from a sample of independent train tasks, of a prior yielding optimal generalisation bounds on new unknown test tasks. Such optimisation of the prior brings two benefits: tighter generalisation bounds (smaller penalisation); and simplified PAC-Bayes learning task (better initial guess).

As discussed in section 1.3.4, previous work which studied Meta-Learning for PAC-Bayes mostly focused on the two fold PAC-Bayes strategy, which applies a PAC-Bayes learning algorithm both as the meta-learning algorithm and as the inner learning algorithm. As a result, the meta PAC-Bayes community constructs not an optimal prior distribution, but rather a randomised prior distribution (hyper posterior). As in classic PAC-Bayes, this hyper posterior has generalisation guarantees in the form of a high probability upper bound on the performance of the learning procedure on the average future test task.

Such generalisation guarantees involve a penalisation term between the hyper prior and hyper posterior, distributions on families of distribution. Typically, the hyper posterior is estimated in a variational family of distributions, such as Gaussian distributions on the hyperparameters of Gaussian distributions. The large number of dimensions involved can lead to large or rapidly increasing Kullback–Leibler divergence between hyper prior and hyper posterior, resulting in vacuous generalisation guarantees. For instance, Amit and Meir [2018] considered an hyper prior $\mathcal{N}(0, \sigma_0^2 \mathrm{Id})$ and constructed an hyper posterior of form $\mathcal{N}(\theta, \sigma_p^2 \mathrm{Id})$ for fixed σ_0, σ_p values; this design choice results in a meta PAC-Bayes guarantee affected by the curse of dimensionality (regardless of the meta posterior chosen), leading to vacuous meta generalisation guarantees[1]. For more adequate choice of hyper distributions, this issue could be mitigated. When the hyper prior equals the hyper posterior, the curse of dimensionality is broken, and the trained objective must improve on the objective evaluated at the hyper prior. As discussed in Section 2.2, the meta generalisation guarantees depends on the quality of the hyper prior, *i.e.* how likely it is to draw a high performing prior.

The inner PAC-Bayes algorithm already provides a guarantee on the generalisation ability of a predictor trained on a test task. These guarantees differ in nature: the meta PAC-Bayes approach provides generalisation guarantees on an average test task (average test guarantee on test task), while the inner PAC-Bayes algorithm provides generalisation guarantees on a given test task (point wise test guarantee on test task). These test task specific bounds are arguably more informative than the "mean" bounds provided through the meta PAC-Bayes approach; taking the AD modelling setting as an exemple, a generalisation guarantee for the calibrated AD

[1] The authors chose $\sigma_0^2 = 2e3$ and $\sigma_p^2 = 1e-3$, which result in a minimum KL term of $\frac{1}{2}(d_\Gamma \log(2e6) + d_\Gamma)$. The objective is lower bounded by $\sqrt{(\mathrm{KL} + \log(2N/\delta))/N} \geq \sqrt{7d_\Gamma/N}$ where N is the number of tasks (5 or 10). As the dimension d_Γ far exceeds the number of tasks, the resulting generalisation guarantees are vacuous.

model on a given plant is easier to interpret and use than a (vacuous) generalisation guarantee valid only for the average plant. Moreover, the two level PAC-Bayes approach involves two layers of uncertainty to track; having to rely a random prior for the inner learning algorithm mitigates the claim that meta-learning would accelerate the calibration process for a given task.

We introduce a PAC-Bayes meta-learning strategy learning a deterministic prior using a MAML like strategy of empirical risk minimisation at the meta level. The computational bottleneck of optimisation based meta-learning strategy is usually the differentiation of the meta parameter to trained estimator map. For the typical neural network setting studied in meta-learning, this requires computation of second order derivative of the neural network, which have larger memory footprint and computational cost. As discussed in Section 1.3.2, practitioners have shown that MAML like algorithm can be trained using first order [Finn et al., 2017], or even rougher approximation of this gradient term [Nichol et al., 2018], with no performance loss. This empirical finding is theoretically grounded when using exact Gibbs posterior as inner learning algorithm. In this setting, the three methods coincide when computing gradients in an appropriate space.

Remark 6.1

Using Gibbs posterior as inner learning algorithm, the prior to posterior map can be interpreted as a linear function on the un-normalized log density. Indeed, considering π_{ref} a reference measure, and priors such that $\pi_{\text{p}} \ll \pi_{\text{ref}}$, the Gibbs posterior satisfy

$$\log\left(\frac{\mathrm{d}\hat{\pi}}{\mathrm{d}\pi_{\text{ref}}}\right) = \log\left(\frac{\mathrm{d}\pi_{\text{p}}}{\mathrm{d}\pi_{\text{ref}}}\right) - \lambda^{-1}R + C$$

with C the renormalisation constant enforcing $\pi_{\text{ref}}\left[\frac{\mathrm{d}\hat{\pi}}{\mathrm{d}\pi_{\text{ref}}}\right] = 1$.
Let us consider the set $\mathcal{F}$ of bounded measurable functions. The quotient space $\tilde{\mathcal{F}}$ of $\mathcal{F}$ with the set of constant function is in bijection with the set of measures such that $\frac{\mathrm{d}\hat{\pi}}{\mathrm{d}\pi_{\text{ref}}}$ is bounded and bounded away from 0. Indeed, for $\tilde{f} \in \tilde{\mathcal{F}}$, one can construct a prior by choosing a member f of the equivalent class of $\tilde{f}$ in $\mathcal{F}$, then defining π_{p} through $\frac{\mathrm{d}\pi_{\text{p}}}{\mathrm{d}\pi_{\text{ref}}} = \exp f - \log\left(\pi_{\text{ref}}\left[f\right]\right)$; the inverse of this map is obtained by considering the equivalent class of $\log\left(\frac{\mathrm{d}\pi_{\text{p}}}{\mathrm{d}\pi_{\text{ref}}}\right)$ in $\tilde{\mathcal{F}}$, noting that the 'bounded' assumption in the definition of $\mathcal{F}$ exactly translates into upper bounded and bounded away from 0 on the distribution space.
Using the vector space $\tilde{\mathcal{F}}$ as a representation of probability measures, the prior to Gibbs posterior map is an affine map between the prior representation and posterior representation, *i.e.*

$$\tilde{f}^{\text{post}} = \tilde{f}^{\text{prior}} - \lambda^{-1}\tilde{R}.$$

This implies that the differentiation of the prior to Gibbs posterior map is the identity for the un-normalized log-density representation. As such, first order MAML, Reptile [Nichol

et al., 2018] and exact MAML coincide when meta-learning is performed on this space.

This analysis only holds when using exact Gibbs posterior. Moreover, the mapping between the posterior representation in log density and the posterior involves the computation of a renormalisation term (for exact access to the density) or sampling methods which may require large number of evaluations of the risk. Rather than perform meta-learning through MAML on this function space, we introduce a new meta-learning objective whose gradient can be computed *without* estimating the gradients of the prior to posterior map, for any PAC-Bayes bound, which can be efficiently trained using SuPAC.

The true objective in optimisation based meta-learning strategy (see Equation (1.30)) has empirical counterpart

$$M(\phi) = \frac{1}{N}\sum_{i=1}^{N} R_i^{\text{test}}(\mathcal{A}(\phi, R_i^{\text{train}})) \tag{6.6}$$

where $\mathcal{A}$ is the inner learning algorithm, ϕ the meta parameter to train and $(R_i^{\text{train}}, R_i^{\text{test}})_{i\in[1,N]}$ are the train and test risk functions summarizing the tasks[2]. Using PAC-Bayes learning at the inner level, the PAC-Bayes bound provides a natural proxy for the test error of the calibrated model. The PAC-Bayes translation of Equation (6.6) becomes

$$M(\pi_{\mathsf{p}}) = \frac{1}{N}\sum_{i=1}^{N} \inf_{\pi\in\Pi} \mathrm{PB}(\pi, R_i, \pi_{\mathsf{p}}, \eta_i). \tag{6.7}$$

Contrary to the classic Meta-Learning setting, this objective does not require access to test data, since it (optimistically) replaces the test performance by the PAC-Bayes high probability bound. We stress that, much like empirical risk minimisation, this objective brings no theoretical guarantees (the PAC-Bayes bound holds for untrained priors, and the sum of high probability bounds would require a union bound argument turning the confidence level from δ to $N\delta$). On the other hand, the fact that the objective trained at the inner level is used as a proxy of the test performance at the meta level leads to a major simplification of the meta gradient. To establish this, we consider the case where the prior distribution π_{p} is optimised in a parametric family of distributions Π^{prior} parametrized by $\phi \in \Phi$. We suppose that assumptions *(A_1)*, *(A_4)* are satisfied, and we require the following assumptions:

Assumptions. *(A_9)* $\forall \pi_\phi \in \Pi^{\text{prior}}$*, assumptions* ($A_2$), ($A_3$), ($A_5$) *are satisfied.*

(A_{10}) The PAC-Bayes PB *is differentiable with respect to* ϕ*.*

(A_{11}) The map $\hat{\theta}_i : \phi \mapsto \inf_{\theta\in\Theta} \mathrm{PB}(\theta, R_i, \phi, \eta_i)$ *is differentiable.*

While the assumptions *(A_9)*, *(A_{10})* will typically be satisfied for adequate choices of family of distributions and parametrisations, assumption *(A_{11})* is much less likely to be satisfied in the general case (*e.g.*, the objective's global minima might change from one local minima to

[2] In practice, optimisation is usually implemented using mini batches of the train data, which require access to the whole dataset rather than the "compressed" information brought by the total training risk.

another as the prior evolves, leading to discontinuity in the infimum map). The gradient of the meta objective with respect to the meta prior parameter simplifies to

$$\begin{aligned}\nabla M(\phi) =& \frac{1}{N}\sum_{i=1}^{N} \partial_\phi \mathsf{PB}(\hat{\theta}_i(\phi), R_i, \phi, \eta_i) \\ =& \frac{1}{N}\sum_{i=1}^{N} \partial_1 \mathsf{PB}(\hat{\theta}_i(\phi), R_i, \phi, \eta_i)\partial_\phi \hat{\theta}_i + \partial_3 \mathsf{PB}(\hat{\theta}_i(\theta_p), R_i, \phi, \eta_i) \\ =& \frac{1}{N}\sum_{i=1}^{N} \partial_3 \mathsf{PB}(\hat{\theta}_i(\phi), R_i, \phi, \eta_i).\end{aligned}$$

The computation of $\partial_{\pi_\mathsf{p}} \hat{\theta}_i$ is not required since, as $\hat{\theta}_i$ minimises the PAC-Bayes objective, $\partial_1 \mathsf{PB} = 0$. A key consequence is that training the meta-learning algorithm is as hard as cycling all the Bayesian optimisation tasks. Meta-learning is as hard as re optimising the PAC-Bayes bound for a new prior.

The efficiency of the meta-learning procedure is thus a direct consequence of the efficiency of the PAC-Bayes calibration process. In the computationally costly risk query setting of AD, this justify the use of risk query efficient strategy such as SuPAC for the inner algorithm. Using SuPAC-CE brings additional benefits. First, the "generation agnostic" weighing approach enables re use of risk queries from calibration performed on previous meta-learning steps. As a consequence, re optimisation of a PAC-Bayes bound for a new prior can conceivably be performed with few risk queries, bringing an additional speed-up. Moreover, when optimising the prior on the same family of distributions as used for the posterior ($\Pi^{\text{prior}} = \Pi$, $metaparam = \theta_p$), the meta gradient benefits from the analytical expression

$$\nabla M(\theta_p) = \frac{1}{N}\sum_{i=1}^{N} \lambda_i (\nabla g(\hat{\theta}_i) - \nabla g(\theta_p)). \tag{6.8}$$

This can be efficiently evaluated if expressions of the gradient of g are available. Such analytical expressions can also be obtained in other specific cases (*e.g.* optimising Gaussian prior with diagonal covariance for generic Gaussian posteriors).

Remark 6.2

Using MLS objective instead of Catoni's PAC-Bayes bound, one can adjust the gradient expression of eq. (6.8). Reusing the link between Catoni's objective and MLS mentioned in section 4.1.5, noting $\hat{\lambda}_i$ the optimal temperature for task i, n_i the number of data in task i and $\widehat{\mathsf{PB}}_{\mathsf{MLS},i}(\theta_p)$ the bound obtained for task i, one obtains

$$\nabla M_{\mathsf{MLS}}(\theta_p) = \frac{1}{N}\sum_{i=1}^{N} \frac{(\nabla g(\hat{\theta}_i) - \nabla g(\theta_p)) \left(1 - \widehat{\mathsf{PB}}_{\mathsf{MLS},i}(\theta_p) \left(1 - \exp(-\lambda_i^{-1})\right)\right)}{n_i \left(1 - \exp\left(-\lambda_i^{-1}\right)\right)}. \tag{6.9}$$

The PAC-Bayes meta-learning strategy consisting in minimising the objective (6.7) on priors, for Catoni's PAC-Bayes objective (1.20) and using the meta gradient expression of Equation (6.8) was implemented for three GD strategies: vanilla GD, SGD and Adam[3] [Kingma and Ba, 2015]. The implementations search for prior and posterior distribution on the same family of probability measures. If this family is an exponential family, gradient steps are computed and executed in the natural parametrisation if the implementation of the gradient of the renormalisation function g is provided. Similar to SuPAC-CE's implementation, a regularization hyperparameter $\mathrm{kl}_{\max}^{\mathrm{meta}}$ is introduced to enforce small KL steps (*i.e.* $\mathrm{KL}(\pi_{\mathsf{p}}^{t+1}, \pi_{\mathsf{p}}^{t}) \leq \mathrm{kl}_{\max}^{\mathrm{meta}}$).

6.2.2 Numerical experiments

We assessed the meta-learning strategy on three different use cases of increasing difficulty. The first experiment studies wholly synthetic learning tasks whose minimum values nearly belong to a affine space of low dimension (ambient dimension of $d_\Gamma = 8$, optimal solutions close to an affine space of dimension 2). A second experiment considers meta-learning a prior for AM2 model, with synthetic tasks generated from a known prior. The third experiment is similar, but considers the more complex ADM1 model instead.

Meta-learning for dimension reduction

We first assess the performance of the meta-learning objective given by Equation (6.7) using synthetic tasks whose risk minima belong to an affine subset of the ambient predictor space. The risk functions considered were bounded, smooth functions of $\mathbb{R}^8$, achieving their global minima at $x^* \sim \mathbb{N}(\tilde{x}^*, \Sigma^*)$. $\tilde{x}^*$ was chosen so that $\|\tilde{x}^*\| = 2$, and Σ^* such that only two of its eigenvalues are higher than 0.05^2 (drawn at random between $\exp(-1)$ and $\exp(1)$). Formally, the risk functions were of form

$$R_{\omega,A,x^*} : x \mapsto \tanh(f(\omega \|A(x - x^*)\|^2)/10) \tag{6.10}$$

with $f(x) = \cos(x) + x$. x_0 is the only global minima of R_{ω,A,x_0}, while all xs such that $\omega\|A(x - x_0)\|^2 = \pi/2 + 2k\pi$ have 0 gradient. The global minima x^* is drawn from $\mathcal{N}(\tilde{x}^*, \Sigma^*)$, the frequency coefficient ω from $\mathcal{U}(\frac{3}{2}\pi, \frac{5}{2}\pi)$ and the matrix parameters $A_{i,j} \sim \mathcal{N}(\delta_{i,j}, \sigma^2 = 0.05^2)$. All risk parameters are drawn independently. The mean parameter $\tilde{x}^*$ was chosen at random on the sphere of radius 2, while the covariance Σ^* was drawn at random as

$$\Sigma^* = O \times \begin{pmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_8^2 \end{pmatrix} \times O^t,$$

[3] Respectively methods 'meta_learn', 'meta_learn_batch' of 'picmeta.MetaLearningEnv' and 'meta_learn_batch' of 'picmeta.MetaLearningEnvAdam'.

with $\sigma_1, \ldots, \sigma_6 = 0.05$, $\sigma_7, \sigma_8 \sim \exp(\mathcal{U}(-0.5, 0.5))$ and O drawn at random amongst orthonormal matrices.

Such choices ensure that the original prior distribution, $\mathcal{N}(0, \mathrm{Id})$, can be improved upon both by shifting its mass centre and adjusting its covariance. Two PAC-Bayes temperatures, $\lambda = 0.1$ and $\lambda = 0.01$, were assessed for the inner learning algorithm. Meta training was performed on 100 training tasks for 120 epochs (here an epoch indicates that the meta-learning algorithm has calibrated all task once; here, each task is calibrated 120 times). Adam Stochastic Gradient Descent was used with hyperparameters $\alpha = 0.8$ (with $\alpha = 1.0$ for the first epoch), $\beta_1 = 0.95$, $\beta_2 = 0.99$ and $\epsilon = 1e-4$ (using the notations of Kingma and Ba [2015]). The maximum KL meta-step size between successive priors was set to $\mathrm{kl}_{\max}^{\mathrm{meta}} = 0.05$. The size of the mini batch and SuPAC-CE's hyperparameters evolved during the meta-learning procedure in the following way:

- The initial calibration phase for each task was performed in 15 steps, with 100 score queries for the first five steps and 50 score queries for the remaining steps. The regularisation hyperparameters were set to $\mathrm{kl}_{\max} = 0.5$, $\alpha_{\max} = 0.3$, and 10^4 samples were used to estimate weights. This initial epoch used a mini batch size of 1 (*i.e.* the prior is updated after each task calibration).
- After all tasks have been trained once, the hyperparameters for SuPAC-CE were modified: the calibration process performs 4 iterations, with 20 evaluations of the risks for the first and third step, and none for the second and fourth. Reducing the number of risk queries is motivated by the expectation that the posterior distribution updates should be small after the initial calibration. The mini batch size remains 1. These hyperparameters are used for 49 epochs.
- For epochs 50 to 75, the mini batch size is increased to 10.
- From epochs 75 onward, the hyperparameters of SuPAC-CE are further modified, to 4 iterations with 2 risk queries per iteration. The size of the mini batch is increased, with 35 epochs with a mini batch size of 25.
- Finally, ten further meta epochs are performed using a mini batch size of 50 (half the tasks).

This sequence was designed in a semi-arbitrary way. The increasing size of the minibatch is designed to improve the precision of the gradients estimate as the posterior approximation starts to converge. The decreasing number of risk queries account for the increasingly large number of evaluations from the previous calibration processes.

The performance of sequence of priors was assessed in the following way. Using the same task generating procedure, 40 test tasks were drawn. For each prior, a full independent calibration was performed on each task, the same hyperparameters as were used to perform the initial calibration for the training tasks. The resulting posterior performance is assessed by computing the PAC-Bayes objective using 10^4 fresh evaluations of the risk. The mean of these PAC-Bayes

objectives over the test tasks defines the *meta-test score*. The dispersion of these test performance between different test task is assessed by computing the quantiles 0.1 and 0.9 of the test performances at a given prior. This procedure being quite computationally intensive, only the first ten priors constructed and afterwards one prior out of ten were assessed. All computations were performed using Azure Machine Learning compute clusters with 32 cores and Intel Xeon Platinum 8272CL processors.

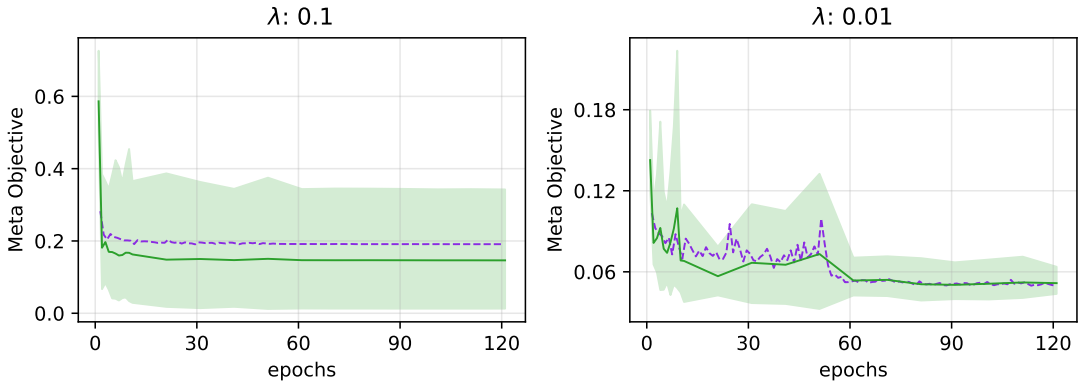


Figure 6.5: Experimental investigation of the train and test performance of the meta-learning approach of Section 6.2.1. The average test performance (green line), as well as quantiles 0.1 and 0.9, of the sequence of priors are assessed on 40 tasks and compared to the train performance (dotted violet line). The increase of the batch size after step 50 had a clear impact on the stability of the learning algorithm for the lower temperature $\lambda = 0.01$.

The meta-learning algorithm was able to satisfactorily reduce the training objective (see violet curve in Figure 6.5), from an initial average generalisation bound of 0.59 (resp. 0.14) to 0.19 (resp. 0.050) after 120 gradient steps for $\lambda = 0.1$ (resp. $\lambda = 0.01$). Most of the meta-objective reduction takes place during the early phase of training. For $\lambda = 0.1$, the objective has decreased to 0.204 by the third epoch, accounting for more than 95% of the objective's decrease. For $\lambda = 0.01$, the optimisation path is rougher during the first 50 meta epochs (while the mini batch size is 1). Still, the meta objective has been decreased by 60 % by the third epoch - and by 95% by the 60th epoch. For both temperatures tested, the average performance on the test tasks followed the objective decrease throughout training. For $\lambda = 0.1$, the meta-test score is even consistently lower than the training meta objective by 0.045, with a final test average performance of 0.146. We interpret this counter intuitive fact by the different methodology used to compute the prior performance score in the meta objective (on train tasks) and for the meta-test score. The test tasks prior evaluations benefit from a full calibration, starting from the current prior, while the meta-objective is computed using few calibration steps from the previous posterior. This might not allow the new posterior to differ enough from the previous one (smaller inner re-training budget). Moreover, posteriors might get stuck in a different local minima in the meta-training objective computation (different initialisation). Finally, the average score is computed in the meta-objective using the weighing procedure of SuPAC-CE. This may

overestimate weights given to risk queries from early steps, notably if few recent risk queries are available, resulting in overestimating the average risk. For $\lambda = 0.01$, train and test performances remained similar throughout the learning process, with the meta-learnt prior obtaining an average performance of 0.052 ± 0.0014.

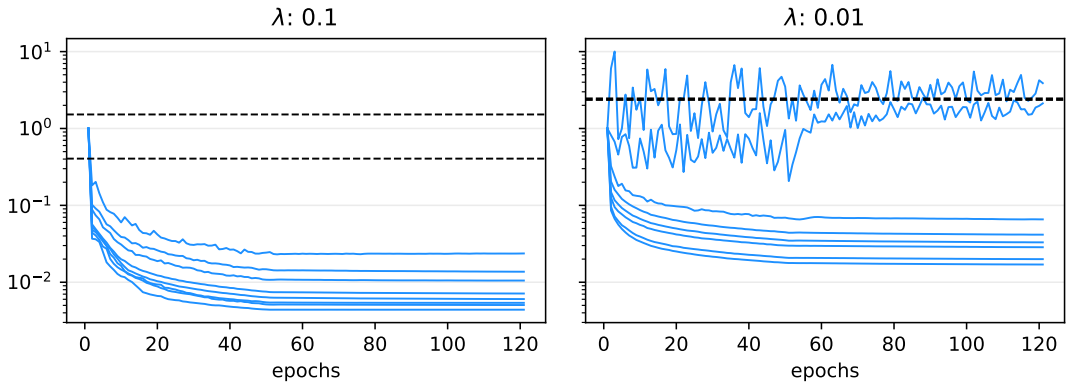


Figure 6.6: Evolution of the prior's eigenvalues during the meta-learning process. Dotted black lines indicate the two largest eigenvalues of the covariance from which the task minimisers were drawn (remaining eigenvalues of 0.0025). For large temperature $\lambda = 0.1$, all eigenvalues decrease, indicating a "frozen" prior (little learning possibilities at the inner level). For the smaller temperature $\lambda = 0.01$, the two largest eigenvalues stabilize close to the oracle eigenvalues, while the remaining eigenvalues decrease, indicating that the learning procedure recovered the true dimension of the meta-learning tasks.

The meta-learning's output differed in nature between the two learning temperatures. For the highest temperature $\lambda = 0.1$, all eigenvalues of the prior's distribution covariance decreased during the learning process, resulting in low learning capabilities at the inner level (see Figure 6.6). The constructed meta-learnt prior can be thought of as an approximate Dirac distribution on the average of the training tasks minimiser. For the lowest temperature $\lambda = 0.01$, two of the eigenvalues of the covariance stabilized around values similar to the eigenvalues of the covariance used to draw the task minimisers, while all the others decreased. This indicates that the meta-learning sequence adequately learnt the number of dimensions of the space on which live the task minimisers. The impact of the temperature on the returned prior can be interpreted in the following way. For high temperatures, the inner algorithm's learning ability is too restricted. The posteriors distribution will always remain too close to the initial posterior, and as such all randomness in the prior will result in an increase of score variance. Hence the meta-learning strategy concentrates on a single predictor which works adequately for all tasks simultaneously, without having to adapt - this is equivalent to pooling all the task's data. For the smaller temperature, learning is possible for the inner algorithm. Hence the optimal prior distribution should be large enough to encompass all the task's minimisers, and put little weight on the rest of the space. As the tasks minimisers are by design close to an affine space of dimension 2, the meta prior adequately concentrates its mass on this space, with only two 'large' eigenvalues in the

covariance.

We recapture in Remark 6.3 the transition between a frozen optimal prior for finite high temperature and a non trivial prior for a simple PAC-Bayes meta-learning setup for low temperature.

Remark 6.3

Consider the following meta-learning setup. Tasks risk are defined from γ_0 through $\tilde{R}(\gamma) = \|\gamma - \gamma_T\|^2$, with γ_T being distributed as $\mathcal{N}(0, \mathrm{Id})$. The inner learning algorithm is the minimisation of Catoni's PAC-Bayes bound, and the meta-learning objective is Equation (6.7) in the limit where an infinite number of tasks are available (the average is replaced by an expectation). The prior is optimised amongst Gaussian priors.
Then the optimal prior is $\mathcal{N}(0, \max(0, 1 - \lambda/2)\mathrm{Id})$. Notably, the optimal Gaussian prior is a Dirac distribution whenever the PAC-Bayes temperature λ is higher than 2.
It is unclear whether the Dirac distribution is the optimal prior distribution when $\lambda \geq 2$ for generic priors (*i.e.* not necessarily Gaussian). For such low inner learning rate, the previous result implies that Gaussian are not conservative enough, in the sense that they always put too much mass on inadequate predictors - which can not be ruled out by the posterior due to the low learning rate. Other distributions, such as uniform distribution on balls, would not suffer from such disadvantages, and might improve on the Dirac prior.

Proof. For clarity's sake, we first prove the result when optimising on homoscedastic Gaussian distributions $\mathcal{N}(\gamma_p, \alpha^2\mathrm{Id})$. Catoni's PAC-Bayes bound obtained for such prior on a task γ_T is

$$\begin{aligned}
\ell(\alpha, \gamma_p, \gamma_T) :=& -\lambda \log\left(\mathbb{E}_\epsilon\left[\exp(-\lambda^{-1}\left\|\alpha\epsilon + \gamma_p - \gamma_T\right\|^2\right]\right) \\
=& -\lambda \log\left(\mathbb{E}_\epsilon\left[\exp\left(-\lambda^{-1}\alpha^2\left\|\epsilon - \frac{\gamma_0 - \gamma_p}{\alpha}\right\|^2\right)\right]\right) \\
=& -\lambda \log\left(\frac{\exp\left(\frac{-\lambda^{-1}\left\|\gamma_p - \gamma_0\right\|^2}{1+2\lambda^{-1}\alpha^2}\right)}{(1+2\lambda^{-1}\alpha^2)^{d/2}}\right) \\
=& \frac{\left\|\gamma_p - \gamma_0\right\|^2}{1+2\lambda^{-1}\alpha^2} + \frac{\lambda d}{2}\log(1+2\lambda^{-1}\alpha^2).
\end{aligned}$$

The third equality is obtained by recognising in the expected value the moment generating function of a non central χ^2 distribution. As such, the meta-objective of prior $\mathcal{N}(\gamma_p, \alpha\mathrm{Id})$ is

$$\begin{aligned}
\mathbb{E}_{\gamma_T}\left[\ell\left(\alpha, \gamma_p, \gamma_T\right)\right] =& \mathbb{E}_{\gamma_0}\left[\frac{\left\|\gamma_p - \gamma_T\right\|^2}{1+2\lambda^{-1}\alpha^2}\right] + \frac{\lambda d}{2}\log\left(1+2\lambda^{-1}\alpha^2\right) \\
=& \frac{d + \left\|\gamma_p\right\|^2}{1+2\lambda^{-1}\alpha^2} + \frac{\lambda d}{2}\log\left(1+2\lambda^{-1}\alpha^2\right).
\end{aligned}$$

As a function of γ_p, this is minimised for $\gamma_p = 0$. The resulting function $\alpha \mapsto \frac{d}{1+2\lambda^{-1}\alpha^2} + \frac{\lambda d}{2}\log\left(1+2\lambda^{-1}\alpha^2\right)$ achieves its minima at $\alpha^2 = \max(0, 1-\frac{\lambda}{2})$. This concludes the proof for the homoscedastic prior.

Considering heteroscedastic priors $\mathcal{N}(\gamma_p, \Sigma)$, the performance of the prior on task γ_T values

$$\begin{aligned}
\ell(\Sigma, \gamma_p, \gamma_T) :=& -\lambda \log\left(\mathbb{E}_\epsilon\left[\exp(-\lambda^{-1}\left\|\sqrt{\Sigma}\epsilon + \gamma_p - \gamma_T\right\|^2\right]\right) \\
=& -\lambda \log\left(\mathbb{E}_\epsilon\left[\exp\left(-\lambda^{-1}\sum_{i=1}^{d}(\sigma_i\epsilon_i - (\gamma_T-\gamma_p)_i)^2\right)\right]\right) \\
=& -\lambda \sum_{i=1}^{d}\log\left(\mathbb{E}_{\epsilon_i}\left[\exp\left(-\lambda^{-1}\sigma_i^2\left(\epsilon_i - \frac{(\gamma_T-\gamma_p)_i}{\sigma_i}\right)^2\right)\right]\right) \\
=& -\lambda \sum_{i=1}^{d}\log\left(\frac{\exp\left(\frac{-\lambda^{-1}\left((\gamma_T-\gamma_p)_i\right)^2}{1+2\lambda^{-1}\sigma_i^2}\right)}{(1+2\lambda^{-1}\sigma_i^2)^{\frac{1}{2}}}\right) \\
=& \sum_{i=1}^{d}\frac{(\gamma_{T,i}-\gamma_{p,i})^2}{1+2\lambda^{-1}\sigma_i^2} + \frac{\lambda}{2}\log\left(1+2\lambda^{-1}\sigma_i^2\right).
\end{aligned}$$

where $\sqrt{\Sigma}$ is the matrix square root[4] of Σ, σ_i the eigenvalues of $\sqrt{\Sigma}$, and the notation γ_i indicates the i-th component of vector γ using as basis eigenvectors of Σ. Noting that $\gamma_{T,i}$ is distributed as $\mathcal{N}(0,1)$, it follows that the expected prior performance on the task is

$$\mathbb{E}_{\gamma_T}\left[\ell(\Sigma, \gamma_p, \gamma_T)\right] = \sum_{i=1}^{d}\frac{1+\gamma_{p,i}^2}{1+2\lambda^{-1}\sigma_i^2} + \frac{\lambda}{2}\log\left(1+2\lambda^{-1}\sigma_i^2\right).$$

Hence the meta-objective is the sum of d independent contributions depending only on $\gamma_{p,i}$, σ_i, whose minima are obtained for $\gamma_{p,i} = 0$, $\sigma_i^2 = \max(0, 1-\lambda/2)$, finishing the proof. □

We assessed whether the meta priors sequence $\pi_p^t = \mathcal{N}(x_t, \Sigma_t)$ tended to shift the centre of mass towards the centre of the tasks minimisers by evaluating the norm of $\|x_t - x^*\|_2$. Similarly, the behaviour of the prior covariance matrix was assessed by evaluating the ratio of Frobenius norm (matrix 2-norm) of the difference with the Frobenius norm of the 'oracle' covariance Σ^*, *i.e* $\frac{\|\Sigma_t - \Sigma^*\|_F}{\|\Sigma^*\|_F}$. These are represented in Figure 6.7. For $\lambda = 0.1$, the centre of mass is adequately approximated, with the difference norm decreasing from the initial 2 to 0.075. The covariance, on the other hand, is not adequately approximated; coherent with the observation that the prior's covariance to all purpose goes to 0, the ratio of Frobenius norm converges to 1 (last value of 0.99) - coherently with the frozen prior phenomena discussed above. For temperature $\lambda = 0.01$, both the mean and covariance are learnt, with the covariance norm ratio decreasing to 0.46, and the difference of mean norm decreasing to 0.55. Compared to the larger temperature, the training path is much less stable, and the approximated mean further away from the oracle mean.

[4] The positive definite matrix satisfying $\sqrt{\Sigma}^2 = \Sigma$

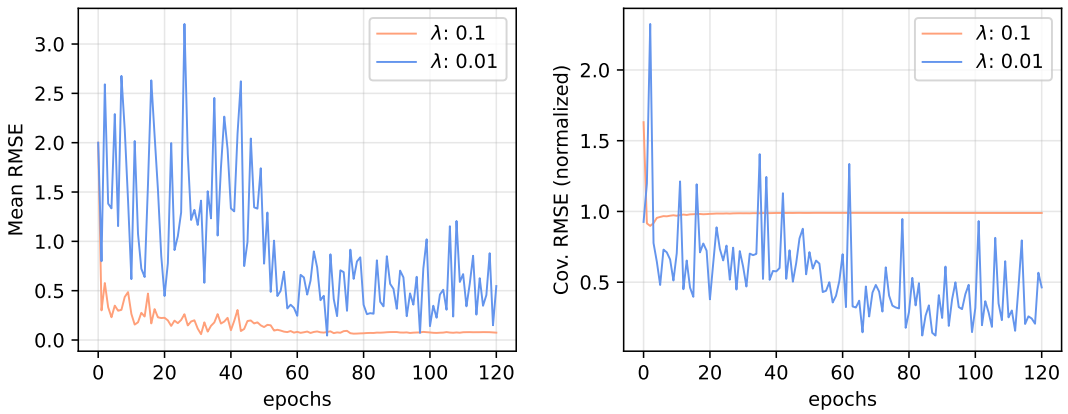


Figure 6.7: Assessment in the inference of the oracle mean and covariance during the meta-learning process on synthetic tasks for two different PAC-Bayes temperatures. The higher temperature obtained a better estimation of the oracle mean, while the lower temperature obtained a better approximation of the covariance.

We interpret this as a consequence of the greater learning ability of the lower temperature inner algorithm, which makes precise centring of the prior distribution less critical - the posterior is still able to shift its mass towards the task's minimiser. A potential work-around this lack of precision in the meta-learnt prior localisation could be to consider a decreasing sequence of learning temperature during training (similar to simulated annealing), or to combine meta-learnt priors learnt for different temperatures.

Meta-learning for AM2

We consider our PAC-Bayes prior meta-learning framework for the construction of an optimal prior for AM2. As in Chapter 3, we will consider synthetic tasks, where the observations are noisy versions of AM2's outputs using known parameters. These parameters are drawn from a known distribution, which differs from the prior; the goal of meta-learning procedure is to construct a new prior distribution, adjusted to these tasks. We will consider the case where the distribution used to generate the true parameter values (the oracle prior) belongs to the class of probability measures on which the prior is searched (Gaussian distributions in the log-space with block diagonal covariance). This represents an optimistic setting, which might not be representative of real-world tasks, for which the 'oracle prior' could be for instance multi modal.

All tasks share the same input data and initial state, which corresponds to the dataset LN from Chapter 3. To lower the computation time, the influent description is limited to the first 140 days. For a parameter γ, a task is constructed by evaluating the model's predictions, adding uniform noise $\mathcal{U}(-\sigma,\sigma)$ in log-space, with noise level $\sigma = 0.15$, and defining the score as in Section 3.2.4.

The parameters γ are drawn from the probability measure π^* constructed in a similar way as

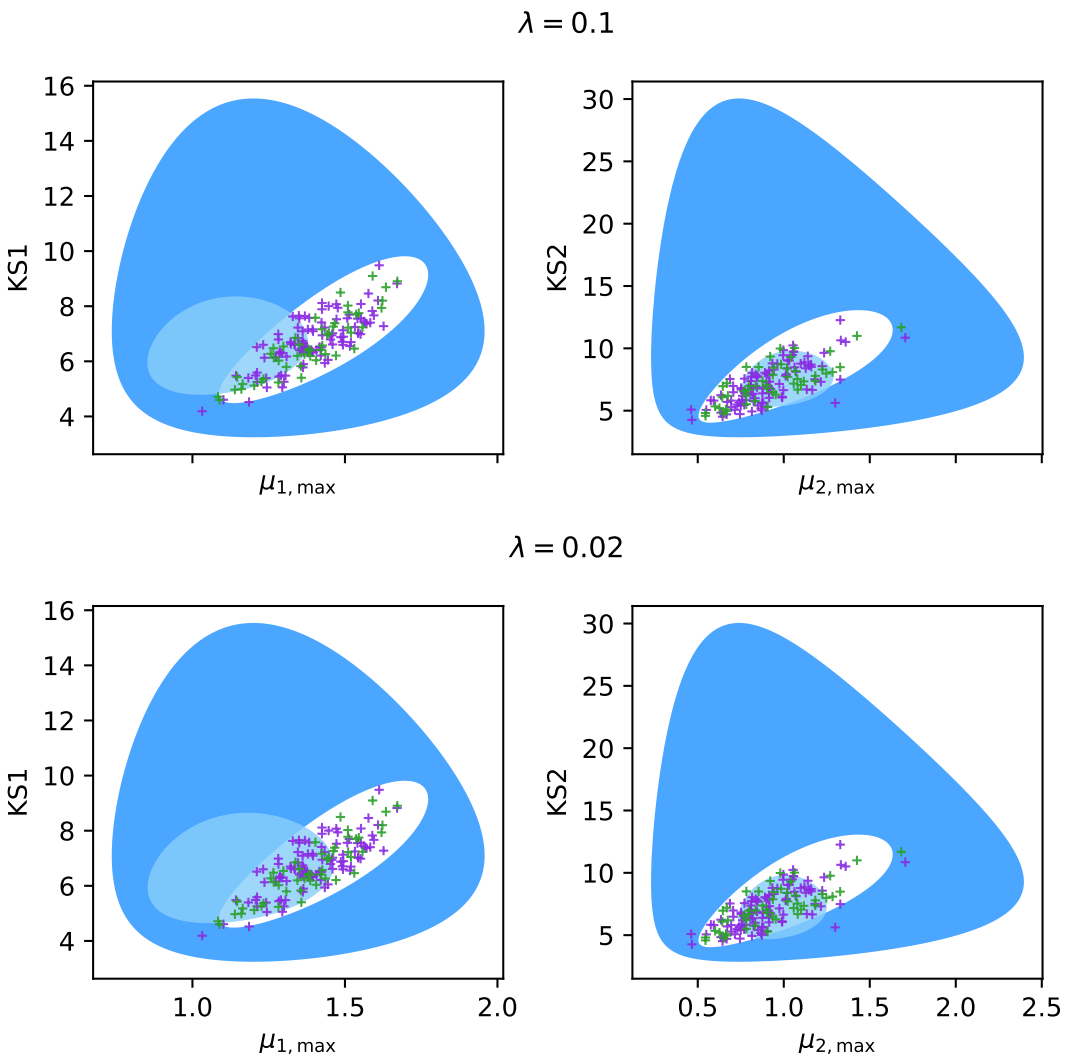


Figure 6.8: Initial prior distribution (deep blue), meta-learnt prior distribution (light blue) and oracle prior distribution (white) for AM2 Meta-Learning, using two different temperatures ($\lambda \in [0.1, 0.02]$). The log-Gaussian distributions are represented using 95% confidence ellipses region in log space. Violet crosses (resp. green) indicates the training tasks (resp. test tasks) oracle parameters. The meta-learnt prior for both temperatures are similar, and differ from the oracle distribution. The meta-learnt prior for $(\mathrm{K}_{\mathrm{S}_1}, \mu_{1,\max})$ is shifted away from the oracle prior, while the meta-learnt prior for $(\mathrm{K}_{\mathrm{S}_2}, \mu_{2,\max})$ is mostly encompassed in the oracle prior.

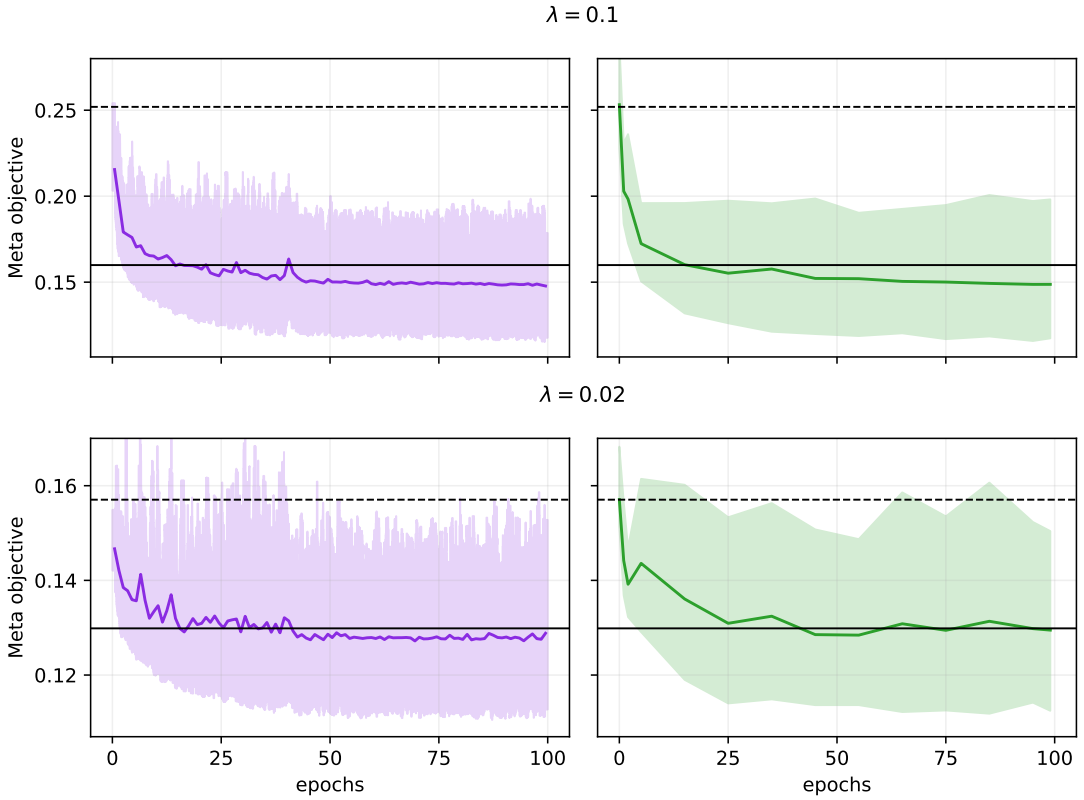


Figure 6.9: Evolution of the meta-learning objectives for AM2, for two learning temperatures ($\lambda \in [0.1, 0.02]$). The dotted black line indicate the objective obtained by the initial prior distribution, while the solid black line indicate the objective obtained by the oracle prior distribution used to generate the tasks. The evolution of the meta-objective during the training process is given by the violet line - quantiles 0.1 and 0.9 of the PAC-Bayes bounds obtained for the individual training tasks are indicated by the light violet area. The performance of the sequence of meta priors is indicated in green, with quantiles 0.1 and 0.9 of the PAC-Bayes bounds obtained for individual test tasks indicated by the light green area. For both temperatures, the test performance are better, or similar to the test performance of the oracle prior, indicating successful meta-learning. Test performance remains similar to the train performance throughout the learning task, supporting the meta-learning objective.

the prior, with the following differences: 1. The standard deviations from Table 3.2b are a factor 2 smaller; 2. the maximum growth rates $\mu_{\max}$ and Monod constants $\mathrm{K_S}$ are correlated (correlation drawn at random between 0.7 and 0.85), for both reactions considered. 100 training tasks, and 50 test tasks are constructed from this prior. Two learning temperatures, $\lambda \in [0.1, 0.02]$ are evaluated for the inner learning SuPAC-CE algorithm. Both of these temperatures are assessed on the same training and test tasks.

Meta training was performed using Adam Optimiser with a mini-batch size of 1 and $\mathrm{kl}_{\max}^{\mathrm{meta}} = 0.05$ throughout the process. The hyperparameters of Adam were set to $\beta_1 = 0.95$, $\beta_2 = 0.995$, $\epsilon = 0.0002$. The initial meta step size α was set to 0.7 for $\lambda = 0.1$, and 0.35 for $\lambda = 0.02$. The sequence of hyperparameters used was as follow:

- The initial calibration was performed through 27 training steps using a total of 512 risk queries (first 160 queries, then 64 three times, then 32 three times, with 2 no query steps following each step with risk queries). The regularisation hyperparameters are set to $\alpha_{\max} = 0.3$, $\mathrm{kl}_{\max} = 0.05$.
- In the four next epochs, PAC-Bayes calibration was performed through 18 training steps using a total 224 risk queries (a sequence of 64 risk queries followed by two "no query" steps, repeated twice followed by a sequence of 32 risk queries followed by three "no query" steps, repeated three times). The other hyperparameters are not modified.
- In the next thirty five epochs, PAC-Bayes calibration was performed through 6 training steps using a total 64 risk queries (a sequence of two "no query" steps followed by 32 risk queries, repeated twice). The other hyperparameters are not modified.
- Finally, sixty further training epochs are performed using a meta step size twice smaller, and a calibration procedure of 4 steps using 32 risk queries (32 risk queries followed by three "no query" steps). The regularisation hyperparameters of the calibration algorithm are strengthened to $\alpha_{\max} = 0.1$, $\mathrm{kl}_{\max} = 0.025$.

The initial prior and meta-learnt prior are compared to the oracle prior π^* in Figure 6.8. Both the meta-learnt prior and oracle prior are encompassed in the original prior distribution for the Monod Constant and maximum growth rate parameter. The meta-learnt prior for $\mathrm{K_{I2}}$, not represented, concentrates on values one order of magnitude smaller than the oracle prior (posterior average of $\mathrm{K_{I2}}$ of 11.9 for $\lambda = 0.02$ (resp. 17.9 for $\lambda = 0.1$) with a standard deviation in log of 0.025, compared to 215 for the oracle distribution, with a standard deviation in log of 0.04). The meta-learnt prior appears to be little sensitive to the temperature chosen, and to differ from the oracle prior distribution. Its meta-test performance is of the same order for $\lambda = 0.02$ (0.1295 for the meta-learnt prior versus 0.1299 for the oracle prior), and notably better for $\lambda = 0.1$ (0.1487 vs 0.1600). This improved performance of the meta-learnt prior and the fact that similar meta-learnt prior are found for both temperatures indicate that the meta-learning strategy has worked adequately and resulted in a distribution which is in nature different from the oracle prior. While the meta-learnt prior improved on the test performance of the oracle prior, we stress that

the latter also holds valuable theoretical information which appears to be lost during the meta-learning procedure.

Meta-learning for ADM1

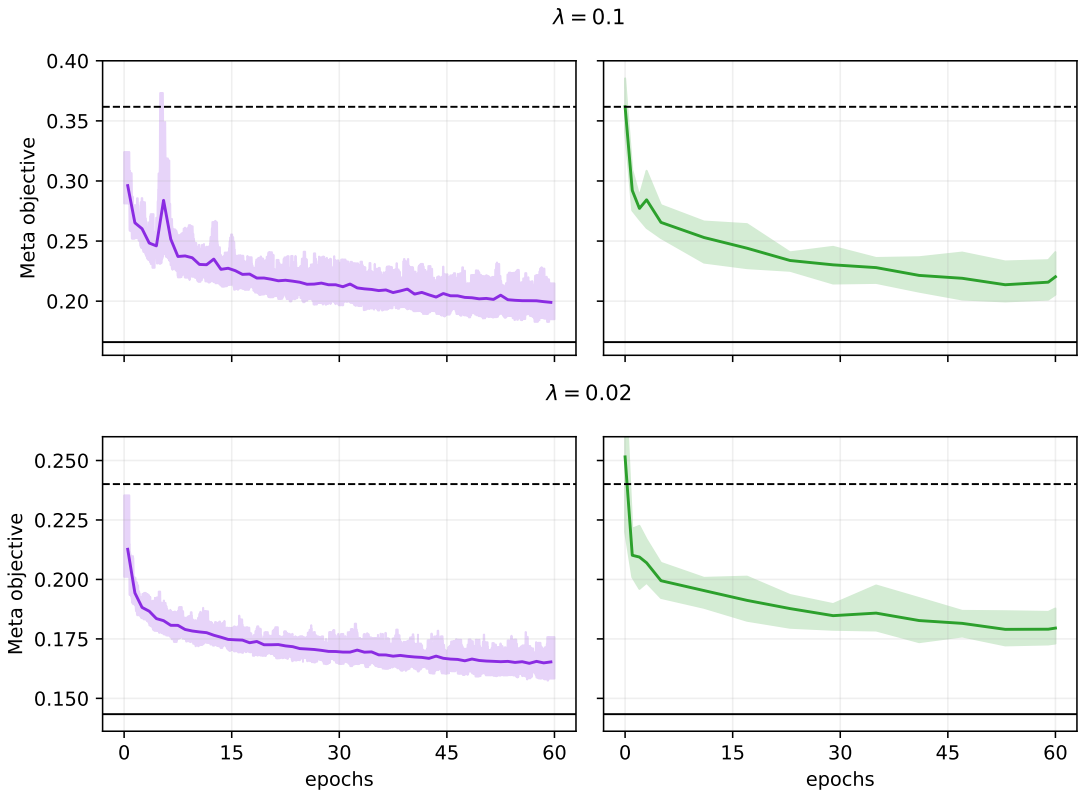


Figure 6.10: Evolution of the meta-learning objectives for ADM1, for two learning temperatures ($\lambda \in [0.1, 0.02]$). The dotted black line indicate the objective obtained by the initial prior distribution, while the solid black line indicate the objective obtained by the oracle prior distribution used to generate the tasks. The evolution of the meta-objective during the training process is given by the violet line - quantiles 0.1 and 0.9 of the PAC-Bayes bounds obtained for the individual training tasks are indicated by the light violet area. The performance of the sequence of meta priors is indicated in green, with quantiles 0.1 and 0.9 of the PAC-Bayes bounds obtained for individual test tasks indicated by the light green area. For both temperatures, the test performance of the meta-learnt prior improves on the original prior, but fail to match the performance of the oracle prior after 60 training epochs. The test performance is also consistently higher than the training objective throughout the meta-learning process.

We also assess our meta-learning strategy for the construction of an optimal prior for the more complex ADM1 model. We generate train and test task using the same strategy as in Section 6.2.2. The feed data and initial state, kept constant throughout the tasks, are those of dataset LF from Chapter 3. The oracle prior is constructed in the same way, adding correlation

factors between the maximum uptake rates $\mathrm{k_m}$ and Monod half saturation constants $\mathrm{K_S}$ for all reactions. To limit computation time, the number of tasks considered was lowered to 40 train tasks and 20 test tasks. As the dimension of the parameter space is larger than for AM2 (30 vs 5), the risk of meta overfitting is therefore greater. Indeed, the number of training tasks is lower than the dimension of the family of distributions on which the prior is looked for (84 vs 40). The same two PAC-Bayes temperatures were assessed.

Similarly to AM2, meta training was performed using Adam Optimiser with a mini-batch size of 1 and $\mathrm{kl}_{\max}^{\mathrm{meta}} = 0.05$ throughout the process. The hyperparameters of Adam were set to $\beta_1 = 0.95$, $\beta_2 = 0.995$, $\epsilon = 0.0001$. The initial meta step size α was set to 0.1 for $\lambda = 0.1$, and 0.05 for $\lambda = 0.02$. The sequence of hyperparameters is identical to the sequence of hyperparameters used for AM2, except that the meta step size is first divided by $\sqrt{2}$ after the 20th epoch, and then again divided by $\sqrt{2}$ after the 40th epoch (compared to divided by 2 after the 40th epoch for AM2).

For both temperature, the meta-learnt prior test performance (0.220 ± 0.0029 for $\lambda = 0.1$; 0.180 ± 0.0014 for $\lambda = 0.02$) improved on the original prior test performance (0.362 ± 0.0047 for $\lambda = 0.1$; 0.240 ± 0.0041 for $\lambda = 0.02$), but remained significantly higher than the oracle prior's performance (0.166 ± 0.0044 for $\lambda = 0.1$; 0.143 ± 0.0026 for $\lambda = 0.02$). Moreover, the test performance remains consistently higher than the meta-learning objective during the training process (see Figure 6.10), with final objectives of 0.199 for $\lambda = 0.1$ and 0.165 for $\lambda = 0.02$. Nonetheless, both curves tended to decrease during the sixty epochs, indicating that overfitting did not harm the constructed meta-learnt prior.

6.2.3 Perspectives

Beyond average task error meta-learning

Our meta-learning objective is built as a PAC-Bayes analogue of the classic meta-learning average task error (1.30). This objective is a tractable, smooth function of all the individual task contribution. On the other hand, focusing on average error might not be the most relevant approach, notably if one is concerned with risk management. An alternative objective could be to control quantiles of the task error distribution, less sensitive to the behaviour of a single task. While this is still easy to compute, the quantile function is not a smooth function, and hence minimisation of the empirical quantile is not as tractable. If such a minimisation task proves impracticable, other approaches involving smooth function of the risk assessment could be investigated, such as

- variance penalized objectives. Noting $L_i(\phi)$ the performance of meta parameter ϕ on task i and $\overline{L}(\phi)$ the average performance, the meta objective becomes

$$\overline{L}(\phi) + \alpha \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (L_i(\phi) - \overline{L})^2};$$

- exponentially weighted aggregates of the performance, *i.e.* meta objectives of form

$$\frac{\sum_i L_i(\phi) \exp(\beta L_i(\phi))}{\sum_i \exp(\beta L_i(\phi))}.$$

As in Section 6.2.1, these meta-learning objectives can be translated into PAC-Bayes meta-learning objectives by using the PAC-Bayes bound as the performance function L. The gradient simplification of the meta-objective still occurs for these new objective functions.

Conditional PAC-Bayes meta-learning

Conditional meta-learning strives to take advantage of task specific side information to design task specific meta parameters. Denevi et al. [2020] showed the potential benefits of such an approach when no single meta parameter is satisfactory for all tasks.

AD plants typically come with additional side information such as the type of substrates they process (*e.g.* sludge, industrial, agricultural) or the range of temperature at which they are operated (mesophilic, thermophilic). This informs on the similarity between different digesters, and hence could help theoretically help design more targeted inductive bias.

For our PAC-Bayes meta-learning framework, conditional meta-learning approaches translate into the construction of task specific priors from side information. Rather than construct a single meta-prior distribution, a meta-predictor taking as input the side information and outputting a prior is trained at the meta-level. To use conditional PAC-Bayes meta-learning, such a space of meta-predictor must be designed. Either a fully learning based approach, or expert-based design can be considered.

As conditional meta-learning is more expressive than classic meta-learning, it comes with greater risk of meta-overfitting. This is particularly true when the number of learning tasks is limited. In the context of AD prior learning, the amount of training task is expected to remain moderate - there can after all be no more tasks than there are properly monitored AD plants, which in France is of order 1500, a fraction of which only are operated by SUEZ. This encourages the use of expert-based design for AD meta-learning approach.

A potential way to implement conditional PAC-Bayes meta-learning for AD could be to take advantage of the block-diagonal structure of the variational family. The prior distribution consists of a combination on smaller priors on blocks. For a given block b, expert insight could define a fixed function f_b clustering the plants in n_b clusters from the side information. To each of this n_b cluster would be associated in a mini prior $\pi_{b,j}$. A task with side information s would then be attributed a prior $\pi_{1,f_1(s)} \times \dots \pi_{B,f_B(s)}$. Hence, while the exact prior used for any training task might be unique, the block prior $\pi_{b,j}$ is still trained from multiple tasks - namely all the tasks such that $f_b(s) = j$. This multiple clustering approach could provide a principled way to build task specific prior while avoiding training priors on too few tasks.

6.2.4 Conclusion

We introduced a PAC-Bayes meta-learning strategy which combines empirical risk minimisation at the meta-level and PAC-Bayes learning at the inner level. Contrary to the usual PAC-Bayes meta-learning framework, our objective renounce to the (usually vacuous) meta-generalisation guarantees, in exchange for a simpler formalism. It still benefits from the classic PAC-Bayes guarantees for test tasks.

The objective's gradient benefits from a remarkable simplification, paving the way for GD-type algorithms for its optimisation. The meta-learning strategy moreover fully benefits from SuPAC-CE's "generation agnostic" reuse of previous simulations, limiting the number of risk queries per task necessary after the first few meta-learning steps.

We assessed our meta-learning framework on a controlled, fully synthetic environment, and on its ability to learn a suitable prior on AD calibration tasks. The synthetic experiments highlighted the importance of the choice of PAC-Bayes temperature - which acts as a learning rate for the inner algorithm - on the meta-learnt prior, with a phase transition occurring between a frozen, Dirac prior for high temperature and expressive priors for low temperature. The preliminary experiments on AD supported our PAC-Bayes objective - the meta-test performance was adequately reduced - but also brought attention to an unexpected feature of the meta-learning procedure: the meta-learnt prior might in essence differ from the distribution from which the tasks are generated. The experiments on ADM1 model also pointed out that attention should be paid to the risk of meta-overfitting for more complex models, when few learning tasks are available.

Finally, we stress that the AD meta-learning experiments were conducted under the optimistic assumption that the oracle prior belongs to the family of measures on which the objective is minimised. As this assumption is unlikely to be fulfilled for real world data, the PAC-Bayes meta-learning algorithm should be further assessed on a misspecified oracle prior, *e.g.* a multimodal prior.

6.3 General conclusion

We investigated in this chapter the possibility of learning from multiple AD plants. We showcased first a data pooling approach applied to the prediction of a key performance indicator of sludge AD. This simple approach proved effective due to the strong affinity of the tasks considered and the study of oversized AD processes. For extensive modelling of more complex AD plants, we propose a PAC-Bayes meta-learning procedure which takes advantage of our PAC-Bayes learning algorithm's idiosyncrasies - namely, its reliance on previous simulations. This PAC-Bayes meta learning procedure was assessed on controlled settings for two AD models, AM2 and ADM1. It succeeded in improving the prior distribution, leading to smaller PAC-Bayes objectives on test tasks. These experiments brought two key findings. First, the optimal prior in the sense of meta-learning can differ from the natural prior (the distribution of the parameters). This questions the interpretability of the meta-learnt prior, and could also impact its robustness.

Further assessment of the meta-learning procedure on a wider range of tasks is necessary to address this potential issue. Second, using meta-learning for complex models such as ADM1 with few tasks could lead to meta-overfitting.

As ProdAD involves a larger number of hyperparameters and is currently used to monitor few AD plants, it is therefore too early to employ our PAC-Bayes meta-learning strategy. In the foreseeable future, variants where a simplified form of prior is learnt -*e.g.* diagonal Gaussian, or simply a multiplicative factor in front of the current covariance- could be assessed once a few more plants data become available. We also expect the distribution of real-world AD tasks to differ greatly from the simplified distributions assessed in this chapter, bringing further challenges. Presumably, these could be partly addressed using a conditional meta-learning framework relying on metadata to build a plant specific priors. However, the added flexibility of conditional meta-learning comes at the cost of increased risk of meta-overfitting, and as such we deem it prudent to exclude such complex options before a large number of plants data are available (*e.g.* a few hundreds).

Conclusion and perspectives

A last few words

The motivation of this thesis was the design of learning tools able to model AD processes, both at the individual AD process level and on a global level. These learning tools should be adapted to the specificities of AD, yet be generic enough to be useful in as many fields as possible. Early in the thesis, we planned to couple PAC-Bayes generalisation bounds as training objective for task specific learning, and meta-learning techniques to obtain a global model. Indeed, PAC-Bayes brought two major promises. First, being a Bayesian inspired technique, it automatically provides a form of uncertainty quantification. This measure on the confidence of the model's predictions is crucial in the context of AD monitoring, where identifiability issue could cause a single trained deterministic model not to foresee process failure. Second, PAC-Bayes theory revolves upon guarantees on the test performance of the calibrated model. These guarantees control the risk of overfitting, which was also an identified issue for AD models involving a large number of parameters (for a physical model). By coupling Meta-learning with PAC-Bayes, the objective was to construct task-specific priors. Benefiting from other digester's insights, these could optimistically be used to construct robust predictors even when few, or even no data would be available. Outputting prior beliefs, the results of the meta-learning process could be moreover readily interpreted.

Of these original goals, some have been fulfilled, others had to be put aside. The construction of an AD compliant PAC-Bayes bound was abandoned. The first part of Chapter 2 was carried out to extend PAC-Bayes analysis to more generic settings in the hope of eventually constructing such a bound. In the course of time, we realised that this goal was nor realistic, nor, from an industrial perspective, a priority. Indeed, lacking an accepted statistical model for the representation of AD's data generation, simplified models would remain questionable, resulting in dubious, and possibly vacuous generalisation guarantees. PAC-Bayes bounds still provided a reasonable learning objective for our stochastic models, whose qualities and shortcomings could be empirically assessed. In Chapter 3, we did empirically show that while not perfect, the uncertainty quantification brought by PAC-Bayes posteriors improved on other UQ techniques, whose validity are equally limited. We also exhibited that the prior knowledge incorporated in the PAC-Bayes learning process improved the robustness of the model, something which vindicated both the use of Bayesian inspired techniques and the necessity to construct appropriate

priors. While these findings are not theoretical guarantees and depend on the type of model considered, the type of PAC-Bayes bound considered and the learning rate (PAC-Bayes temperature), they were encouraging enough to consider the application of PAC-Bayes to monitor real-world industrial plants. The performance of the PAC-Bayes learning algorithm we initially designed did however showed its limits when coupled to SUEZ's AD model, ProdAD. The increase in the predictor space dimensionality, coupled with a much steeper initial computation time, made the learning's process unusable. The impact of the computational cost of AD models on the learning process was a difficulty whose importance was not fully measured initially. To overcome this stumbling block, a PAC-Bayes calibration framework, SuPAC, was designed to tackle settings where the simulation time (or, from a learning perspective, the risk query) is the main computational bottleneck (see Chapter 4). We believe that this framework, which is generic to most PAC-Bayes objectives, is one of this thesis's major contribution, and is of interest outside the sole scope of AD. Combined with changes in the AD models' implementations, the calibration process could be performed in reasonable time and cost. As detailed in Chapter 5, common rules were defined to calibrate real-world digesters in a unified fashion, and a strategy was devised to perform online monitoring of AD plants while managing the asynchronous uploads of data. The calibration and online monitoring strategies are now used by SUEZ to monitor two industrial digesters, with plans to extend to a broader implementation. These new plants data will offer an opportunity to assess whether the current design is robust or requires further improvements.

Whether and to which extent AD modelling can benefit from meta-learning remains an open question. The construction of a VSR prediction model through a data-pooling approach described in Chapter 6 shows that there is potential for leveraging information between different AD plants. This is however already partially done in our PAC-Bayes framework, where the prior was constructed using insights from many AD processes. The potential benefits of PAC-Bayes meta-learning will depend on how wildly off the optima this initial prior distribution is, and our ability to construct targeted priors. While, due to both lack of time and having access to too few digesters' data, this could not be measured, we were able to construct a PAC-Bayes meta-learning procedure which should prove efficient. Preliminary experiments on synthetic AD meta-learning settings showed its potential for building efficient priors, but also revealed that such meta-learnt priors could have limited interpretation, and hinted that more complex AD models might suffer from meta-overfitting. This questions whether task specific priors, or even a single good prior, could reasonably be constructed for SUEZ's ProdAD from a reasonable number of digestion units.

While meta-learning was thus limited to synthetic experiments, this thesis both investigated the benefits of a PAC-Bayes calibration approach for modelling AD, and contributed a practical implementation to perform this calibration. This implementation is efficient enough to be used for monitoring purposes, is currently used in an online fashion on 2 plants, with plans to extend to new plants in the short term. This demonstrates the relevance of PAC-Bayes methods even for computationally intensive physicochemical and biological models. The main theoretical limitation of our algorithm is its restriction to low dimensional parameter space, a condition which

is usually fulfilled for scientific models. As such, we believe that SuPAC-CE can benefit the modellers far beyond AD community. Similarly, our PAC-Bayes meta-learning framework is not limited to AD applications, but could be used to efficiently refine expert beliefs for other scientific modelling applications.

Anaerobic Digestion model meta-learning

Meta-learning strives to leverage information from different tasks to construct a robust model, which is then refined in future tasks. To this end, we considered in this thesis the goal of learning an optimal inductive bias from various AD units, which is then used as a reference for the calibration of the same, already fixed, AD model. An appealing feature of this meta-learning design is that it already incorporates strong expert knowledge through the choice of the AD model. This "modelling" based approach, contrary to pure "data-driven" approaches, inherently respects physicochemical laws, steering the learning process away from grossly inexact guess. As a result, the learnt models should prove more robust. Such an approach proved useful for AD considering the limited amount of data, its correlated nature, and its limited reliability[5].

Still, AD models rely on approximations (*e.g.* the choice of the biochemical reaction network) and empirical formulas whose exact form is arguable (*e.g.* Monod inhibition). These formulas were constructed from, and then validated on, experimental evidence. This experimental evidence is designed to target a specific phenomena (*e.g.* a specific factor on inhibition), and is as such less tangled than the evidence gathered from a complex AD process. On the other hand, this evidence might be more limited in terms of amount of data (*e.g.* a single experiment during a few months versus thousands of AD units operated all year round) and be less diverse (*e.g.* single, well controlled substrate versus multiple complex substrates).

Combining the targeted experimental data to real-world operation data could be a promising way to create a new generation of AD models. To this end, many practical issues would need to be addressed. First, access to trustworthy experimental data is essential. This supposes willingness from experimenters to share raw data, but also defining a shared standard. Second, all experimental data is not created equal, due to different experimental set up, measure precision, potential manipulation issue. Hence information on data quality (such as specification of the measurement tools) should be shared whenever possible. Moreover, the learning process should be as robust as possible to incorrect data. Third, the resulting model should be as interpretable as possible. While fully data-driven learning algorithms may provide state-of-the-art performance, interpretable models are key to designing experiments further refining the model. To construct interpretable meta-model, the meta-learning framework would require careful crafting to strike a balance between expressivity of the meta-models and expert based design constraints. Fourth, the meta-model should, if possible, be computationally efficient for each task. Building a meta-model valid for all AD process simultaneously should lead to a wide range of

[5] Prior to calibrating the models, plants operators and AD experts had to decide which sensors should be used, and exclude given timespans on which they were deemed too unreliable.

mechanisms being taken into account. For instance, both thermophilic and mesophilic micro-organisms needs to be encoded in the meta-model, while only half of these micro-organisms would typically be present for a given plant. This could lead to a sharp increase in the computational time, without any benefit in the model performance. Our hope is that task specific model can be easily inferred from the global meta-model to keep the computational cost down. Last but not least, an efficient meta-learning strategy should be devised. A potential strategy to define an interpretable meta-model could be to integrate symbolic regression (see *e.g.* Udrescu and Tegmark [2020], Kamienny et al. [2022]) to an expert designed backbone. For instance, the form of inhibition functions could be learnt at the meta-level through symbolic regression, the parameter values learnt for each task through optimisation, while the main structure of the model (*e.g.* a biochemical reaction network solved through an ODE) is designed through expert knowledge. Compared to the inductive bias learning strategy pursued in optimisation based meta-learning, learning generic form of model equation should prove an harder task. Notably, the optimisation space is no longer continuous, but a set of parametrised formulas. On the other hand, such coupling of symbolic regression and meta-learning appears a virtually unchartered field with potential groundbreaking possibility for automatic scientific discovery, and could hence be of broader interest than AD.

Beyond risk-centred PAC-Bayes

The PAC-Bayes framework attacks the problem of generalisation through a change of measure approach. The key idea is that a test bound, evaluated on a data-independent prior aggregating all predictors, is transferred simultaneously to all randomised estimators through change of measure. As a result of this framework, PAC-Bayes bounds define learning objective tantamount to penalised empirical risk minimisation.

PAC-Bayes objectives have proved their capacity to obtain generalisation guarantees even for complex models. Still, the analysis of Section 2.2 showed that most PAC-Bayes objectives inherently do not distinguish between predictors achieving the same risk. A PAC-Bayes bound will therefore shift mass from high risk predictors to low risk predictors, no matter how well each generalise. The one famous counter-example is the bound by Tolstikhin and Seldin [2013], which considers the average expected posterior empirical variance of the loss in the training objective. The empirical variance of the loss can be interpreted as an indication of the generalisation ability of a given predictor. Indeed, test bounds obtained from Bernstein's concentration inequalities decrease with the loss variance, resulting, for the same empirical risk, in tighter generalisation guarantees for predictors with a smaller variance. In short, the empirical variance informs on the generalisability of each predictor - and integrating this notion into the PAC-Bayes objective can prove beneficial[6].

In the popular classification setting however, the empirical variance indicator can prove too crude. Considering the classification loss, the empirical variance becomes a function of the em-

[6] Tolstikhin and Seldin [2013] showed that their bound is not directly comparable to MLS, and can improve generalisation guarantees in certain settings.

pirical risk. While Totstikhin's bound can still benefit from better analysis of the generalisation ability of each predictor (*i.e.* it has better incentive to shift mass towards low risk predictors because it moreover expects that those generalise better), it still constructs weight corrections between prior to posterior solely guided by the empirical risk. An interesting prospect could be to integrate other proxies for generalisation ability of the predictors into PAC-Bayes objectives. Zhou et al. [2019], Lotfi et al. [2022] considered compressibility of predictors and used it to design a prior distribution. While this strategy proved remarkably efficient, it splits the PAC-Bayes framework into a prior construction stage and a generic PAC-Bayes bound exploitation stage; similar to the sparsity inducing priors considered by Dalalyan and Tsybakov [2012a]. We take the view that such heuristics on the generalisation ability of predictors should be directly incorporated into the learning objective. We argue that this would serve three purposes. First, we expect that carefully designed, such bounds could offer tighter generalisation guarantees. Second, such an approach would prove more generic, and could be applied to settings where a natural prior is available. Last but not least, constructing such bounds require insights on how the heuristic translates into generalisation, resulting in better theoretical understanding and better interpretability of the bound. Compression based bounds might prove harder to analyse and be too specific to the neural-network settings. For smooth loss functions, or for models whose loss is a function of a smooth prediction function, incorporating the loss's derivatives could prove a more achievable option, taking advantage of the "flat-minima" view of generalisation.

Meta-learning uncertainty

The PAC-Bayes meta-learning strategy introduced in Section 6.2 is designed to construct a prior distribution optimal in terms of the generalisation guarantees it offers. The numerical experiments of Section 6.2.2 showed that this optimal prior can differ from the oracle prior, defined as the distribution of the tasks' risk minimiser. Moreover, the toy framework defined in Remark 6.1 showed that for inadequate PAC-Bayes learning rates, this optimal prior distribution can collapse to a single predictor. From a learning perspective, such results are not problematic; the use of stochastic predictors is often seen more as a bothersome technicality, and the important fact is that the predictor obtains optimal, or near optimal performance. From an uncertainty quantification perspective however, this behaviour is disturbing; such frozen priors prevent any meaningful uncertainty quantification on interpretable parameter values. While the "frozen" prior phenomena disappears for lower temperatures, the posterior distributions tend in their turn to freeze towards the empirical risk minimisers as the temperature decreases, eventually harming uncertainty quantification.

Risk minimisers being the learning analogue of true parameters in statistics, the oracle prior may be said to define the natural prior in terms of uncertainty quantification. A non meta-learning strategy to estimate this distribution could be to consider the empirical risk minimisers as a noisy sample from this oracle distribution, and to use *e.g.* density estimation techniques to infer this distribution. This could be successful whenever the noise structure is tractable *e.g.* for

convex losses. In the general case, the empirical risk minimiser and true risk minimiser may be arbitrarily far apart; for multimodal risk functions, the impact of sampling may move the minima from one risk to another. As such, the empirical risk minimisers may not be representative of the oracle distribution.

The fundamental motivation of meta-learning is the idea that some common information is shared between tasks. Treating tasks independently does not benefit from such a premise. We conjecture that meta-learning strategies may be devised to approximate the oracle prior, using the full empirical risk to build more robust approximations of each task risk minimiser (*e.g.* through uncertainty quantification) under yet to be determined assumptions on the tasks.

Bibliography

T. Abbasi, S. M. Tauseef, and S. A. Abbasi. *A Brief History of Anaerobic Digestion and "Biogas"*, page 11–23. Springer New York, Sept. 2011. ISBN 9781461410409. doi: 10.1007/978-1-4614-1040-9_2. URL http://dx.doi.org/10.1007/978-1-4614-1040-9_2. 8

C. Aceves-Lara, E. Aguilar-Garnica, V. Alcaraz-Gonzalez, O. Gonzalez-Reynoso, J.-P. Steyer, J. Dominguez-Beltran, and V. Gonzalez-Alvarez. Kinetic parameters estimation in an anaerobic digestion process using successive quadratic programming. *Water Science and Technology*, 52(1-2):419–426, 2005. 87

S. A. Alavi-Borazjani, I. Capela, and L. A. Tarelho. Over-acidification control strategies for enhanced biogas production from anaerobic digestion: A review. *Biomass and Bioenergy*, 143:105833, Dec. 2020. ISSN 0961-9534. doi: 10.1016/j.biombioe.2020.105833. URL http://dx.doi.org/10.1016/j.biombioe.2020.105833. 10

O. Albertson, B. Burris, S. Reed, J. Semon, J. Smith, and A. Wallace. Design manual: dewatering municipal wastewater sludges. Technical report, US EPA, Cincinnati, Ohio, 1987. 195, 203

P. Alquier. User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2):174–303, 2024. ISSN 1935-8237. doi: 10.1561/2200000100. URL http://dx.doi.org/10.1561/2200000100. 14, 23, 26, 27, 30, 132

P. Alquier and B. Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107 (5):887–902, 2018. URL https://doi.org/10.1007/s10994-017-5690-0. 29, 30, 44, 45, 46, 61, 63

P. Alquier, J. Ridgway, and N. Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research (JMLR)*, 17(236):1–41, 12 2016. 61, 135

A. Ambroladze, E. Parrado-Hernández, and J. Shawe-Taylor. Tighter PAC-Bayes bounds. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper_files/paper/2006/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. 69

R. Amit and R. Meir. Meta-learning by adjusting priors based on extended PAC-Bayes theory. In *International Conference on Machine Learning*, pages 205–214. PMLR, 2018. 40, 41, 206

L. Appels, J. Baeyens, J. Degrève, and R. Dewil. Principles and potential of the anaerobic digestion of waste-activated sludge. *Progress in Energy and Combustion Science*, 34(6): 755–781, Dec. 2008. ISSN 0360-1285. doi: 10.1016/j.pecs.2008.06.002. URL http://dx.doi.org/10.1016/j.pecs.2008.06.002. 194

C. Arnaiz, J. Gutierrez, and J. Lebrato. Biomass stabilization in the anaerobic digestion of wastewater sludges. *Bioresource Technology*, 97(10):1179–1184, July 2006. ISSN 0960-8524. doi: 10.1016/j.biortech.2005.05.010. URL http://dx.doi.org/10.1016/j.biortech.2005.05.010. 195, 203

J. A. Arzate, M. Kirstein, F. C. Ertem, E. Kielhorn, H. Ramirez Malule, P. Neubauer, M. N. Cruz-Bournazou, and S. Junne. Anaerobic digestion model (AM2) for the description of biogas processes at dynamic feedstock loading rates. *Chemie Ingenieur Technik*, 89(5):686–695, 2017. doi: 10.1002/cite.201600176. 12

S. Babel, K. Fukushi, and B. Sitanrassamee. Effect of acid speciation on solid waste liquefaction in an anaerobic acid digester. *Water Research*, 38(9):2417–2423, May 2004. ISSN 0043-1354. doi: 10.1016/j.watres.2004.02.005. URL http://dx.doi.org/10.1016/j.watres.2004.02.005. 10

M.-F. Balcan, M. Khodak, and A. Talwalkar. Provable guarantees for gradient-based meta-learning. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 424–433. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/balcan19a.html. 41

C. J. Banks, M. Chesshire, S. Heaven, and R. Arnold. Anaerobic digestion of source-segregated domestic food waste: Performance assessment by mass and energy balance. *Bioresource Technology*, 102(2):612–620, Jan. 2011. ISSN 0960-8524. doi: 10.1016/j.biortech.2010.08.005. URL http://dx.doi.org/10.1016/j.biortech.2010.08.005. 194

G. Baquerizo, J. Fiat, P. Buffiere, R. Girault, and S. Gillot. Modelling the dynamic long-term performance of a full-scale digester treating sludge from an urban WRRF using an extended version of ADM1. *Chemical Engineering Journal*, 423:128870, Nov. 2021. ISSN 1385-8947. doi: 10.1016/j.cej.2021.128870. URL http://dx.doi.org/10.1016/j.cej.2021.128870. 11, 13, 195

V. Barbu and T. Precupanu. *Convexity and Optimization in Banach Spaces*. Springer Netherlands, 2012. ISBN 9789400722477. doi: 10.1007/978-94-007-2247-7. URL http://dx.doi.org/10.1007/978-94-007-2247-7. 24

M. Barjenbruch, H. Hoffmann, O. Kopplow, and J. Tränckner. Minimizing of foaming in digesters by pre-treatment of the surplus-sludge. *Water Science and Technology*, 42(9):235–241, Nov. 2000. ISSN 1996-9732. doi: 10.2166/wst.2000.0215. URL http://dx.doi.org/10.2166/wst.2000.0215. 10

P. Bartlett, V. Maiorov, and R. Meir. Almost linear VC dimension bounds for piecewise polynomial networks. In M. Kearns, S. Solla, and D. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11. MIT Press, 1998. URL https://proceedings.neurips.cc/paper_files/paper/1998/file/bc7316929fe1545bf0b98d114ee3ecb8-Paper.pdf. 68

D. Batstone and J. Keller. Industrial applications of the IWA anaerobic digestion model no. 1 (ADM1). *Water Science and Technology*, 47(12):199–206, June 2003. ISSN 1996-9732. doi: 10.2166/wst.2003.0647. URL http://dx.doi.org/10.2166/wst.2003.0647. 2, 11

D. Batstone, S. Tait, and D. Starrenburg. Estimation of hydrolysis parameters in full-scale anerobic digesters. *Biotechnology and Bioengineering*, 102(5):1513–1520, Nov. 2008. ISSN 1097-0290. doi: 10.1002/bit.22163. URL http://dx.doi.org/10.1002/bit.22163. 87, 88

D. J. Batstone, J. Keller, I. Angelidaki, S. V. Kalyuzhnyi, S. G. Pavlostathis, A. Rozzi, W. T. M. Sanders, H. Siegrist, and V. A. Vavilin. The IWA Anaerobic Digestion Model No 1 (ADM1). *Water Science and Technology*, 45(10):65–73, May 2002a. URL https://doi.org/10.2166/wst.2002.0292. 2, 11, 86, 90, 100, 154, 255

D. J. Batstone, J. Keller, I. Angelidaki, S. V. Kalyuzhnyi, S. G. Pavlostathis, A. Rozzi, W. T. M. Sanders, H. Siegrist, and V. A. Vavilin. Anaerobic digestion model no. 1. Technical Report 13, International Water Association, 2002b. http://dx.doi.org/10.2166/9781780403052. 255, 257, 258, 259, 260, 261

D. J. Batstone, P. Pind, and I. Angelidaki. Kinetics of thermophilic, anaerobic oxidation of straight and branched chain butyrate and valerate. *Biotechnology and Bioengineering*, 84(2):195–204, 2003. URL https://doi.org/10.1002/bit.10753. 87, 88

M. Bauer, M. Rojas-Carulla, J. B. Świątkowski, B. Schölkopf, and R. E. Turner. Discriminative k-shot learning using probabilistic models, 2017. URL https://arxiv.org/abs/1706.00326. 40

M. J. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, University of London, University College London (United Kingdom), 2003. URL https://discovery.ucl.ac.uk/id/eprint/10101435/. 32, 90

E. M. L. Beale. Confidence regions in non-linear estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 22(1):41–76, Jan. 1960. URL https://doi.org/10.1111/j.2517-6161.1960.tb00353.x. 88, 108

L. Bégin, P. Germain, F. Laviolette, and J.-F. Roy. PAC-Bayesian bounds based on the Rényi divergence. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 435–444. PMLR, 5 2016. URL https://proceedings.mlr.press/v51/begin16.html. 23, 28, 29, 44, 45, 59, 61, 63, 72

O. Bernard, Z. Hadj-Sadok, D. Dochain, A. Genovesi, and J.-P. Steyer. Dynamical model development and parameter identification for an anaerobic wastewater treatment process. *Biotechnology and Bioengineering*, 75(4):424–438, 2001. URL https://doi.org/10.1002/bit.10036. 12, 86, 90, 92, 100, 106, 253, 255

P. Biernacki, S. Steinigeweg, A. Borchert, and F. Uhlenhut. Application of anaerobic digestion model no. 1 for describing anaerobic digestion of grass, maize, green weed silage, and industrial glycerine. *Bioresource Technology*, 127:188–194, 2013. URL https://doi.org/10.1016/j.biortech.2012.09.128. 87

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 78(5):1103–1130, 2016. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/44682909. 14

F. Blumensaat and J. Keller. Modelling of two-stage anaerobic digestion using the IWA anaerobic digestion model no. 1 (ADM1). *Water Research*, 39(1):171–183, 2005. URL https://doi.org/10.1016/j.watres.2004.07.024. 11, 13, 87

D. Bolzonella, P. Pavan, P. Battistoni, and F. Cecchi. Mesophilic anaerobic digestion of waste activated sludge: influence of the solid retention time in the wastewater treatment process. *Process Biochemistry*, 40(3–4):1453–1460, Mar. 2005. ISSN 1359-5113. doi: 10.1016/j.

procbio.2004.06.036. URL http://dx.doi.org/10.1016/j.procbio.2004.06.036. 195, 196, 203

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013. 23, 53

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004. ISBN 0521833787. URL https://stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf. 24

L. D. Brown. Fundamentals of statistical exponential families: with applications in statistical decision theory. In *Statistics*. Ims, 1986. ISBN 0940600102. 132, 134

S. Bubeck and M. Sellke. A universal law of robustness via isoperimetry. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 28811–28822. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f197002b9a0853eca5e046d9ca4663d5-Paper.pdf. 68

H. Caillet, A. Bastide, and L. Adelard. Advances in computational fluid dynamics modeling of anaerobic digestion process for renewable energy production: A review. *Cleaner Waste Systems*, 6:100124, Dec. 2023. ISSN 2772-9125. doi: 10.1016/j.clwas.2023.100124. URL http://dx.doi.org/10.1016/j.clwas.2023.100124. 11

J. L. Callaham, J.-C. Loiseau, G. Rigas, and S. L. Brunton. Nonlinear stochastic modelling with Langevin regression. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 477(2250), June 2021. ISSN 1471-2946. doi: 10.1098/rspa.2021.0092. URL http://dx.doi.org/10.1098/rspa.2021.0092. 160

A. Campos-Rodríguez, M. A. Zárate-Navarro, E. Aguilar-Garnica, V. Alcaraz-González, and J. P. García-Sandoval. Study of the behavior of alkalinities predicted by the AM2 model. *Water*, 14(10):1634, May 2022. ISSN 2073-4441. doi: 10.3390/w14101634. URL http://dx.doi.org/10.3390/w14101634. 12

M. Carballa, L. Regueiro, and J. M. Lema. Microbial management of anaerobic digestion: exploiting the microbiome-functionality nexus. *Current Opinion in Biotechnology*, 33:103–111, June 2015. ISSN 0958-1669. doi: 10.1016/j.copbio.2015.01.008. URL http://dx.doi.org/10.1016/j.copbio.2015.01.008. 1

G. Casella, C. P. Robert, and M. T. Wells. Generalized Accept-Reject sampling schemes. In *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, Lecture notes-monograph series, pages 342–347. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2004. 31

O. Catoni. *Statistical Learning Theory and Stochastic Optimization*. Lecture Notes in Mathematics: Saint-Flour Summer School on Probability Theory XXXI 2001. Jean Picard, 2004. doi: 10.1007/b99352. 15, 23, 26, 30, 47

O. Catoni. PAC-Bayesian supervised classification: the thermodynamics of statistical learning. *IMS Lecture Notes Monograph Series*, 56, 2007. URL https://doi.org/10.1214/074921707000000391. 61

N. Čencov. *Statistical Decision Rules and Optimal Inference*. American Mathematical Society, 4 2000. ISBN 9781470444679. doi: 10.1090/mmono/053. URL http://dx.doi.org/10.1090/mmono/053. 143

K. Cerk, P. Ugalde-Salas, C. G. Nedjad, M. Lecomte, C. Muller, D. J. Sherman, F. Hildebrand, S. Labarthe, and C. Frioux. Community-scale models of microbiomes: Articulating metabolic modelling and metagenome sequencing. *Microbial Biotechnology*, 17(1), Jan. 2024. ISSN 1751-7915. doi: 10.1111/1751-7915.14396. URL http://dx.doi.org/10.1111/1751-7915.14396. 160

S. Chen, D. Yang, B. Dong, N. Li, and X. Dai. Sludge age impacted the distribution, occurrence state and structure of organic compounds in activated sludge and affected the anaerobic degradability. *Chemical Engineering Journal*, 384:123261, Mar. 2020. ISSN 1385-8947. doi: 10.1016/j.cej.2019.123261. URL http://dx.doi.org/10.1016/j.cej.2019.123261. 196

L. Cobb, P. Koppstein, and N. H. Chen. Estimation and moment recursion relations for multimodal distributions of the exponential family. *Journal of the American Statistical Association*, 78(381):124–130, 1983. 132

P. Couto, M. Brustello, R. Albanez, J. Rodrigues, M. Zaiat, and R. Ribeiro. Calibration of ADM1 using the Monte Carlo Markov Chain for modeling of anaerobic biodigestion of sugarcane vinasse in an AnSBBR. *Chemical Engineering Research and Design*, 141:425–435, 2019. URL https://doi.org/10.1016/j.cherd.2018.11.014. 2, 11, 14, 89, 124

I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, pages 146–158, 1975. 25, 46, 61

S. Dabiri, P. Kumar, and W. Rauch. Integrating biokinetics with computational fluid dynamics for energy performance analysis in anaerobic digestion. *Bioresource Technology*, 373:128728, Apr. 2023. ISSN 0960-8524. doi: 10.1016/j.biortech.2023.128728. URL http://dx.doi.org/10.1016/j.biortech.2023.128728. 11

A. S. Dalalyan and A. B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 18(3), Aug. 2012a. ISSN 1350-7265. doi: 10.3150/11-bej361. URL http://dx.doi.org/10.3150/11-BEJ361. 229

A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.*, 78(5):1423–1443, 2012b. doi: 10.1016/j.jcss.2011.12.023. URL http://hal.archives-ouvertes.fr/docs/00/45/68/06/PDF/HAL_EWA_LMC.pdf. 31

J. Daniel-Gromke, N. Rensberg, V. Denysenko, W. Stinner, T. Schmalfuß, M. Scheftelowitz, M. Nelles, and J. Liebetrau. Current developments in production and utilization of biogas and biomethane in germany. *Chemie Ingenieur Technik*, 90(1–2):17–35, Dec. 2017. ISSN 1522-2640. doi: 10.1002/cite.201700077. URL http://dx.doi.org/10.1002/cite.201700077. 8

M. de Gracia, L. Sancho, J. García-Heras, P. Vanrolleghem, and E. Ayesa. Mass and charge conservation check in dynamic models: application to the new ADM1 model. *Water Science and Technology*, 53(1):225–240, Jan. 2006. ISSN 1996-9732. doi: 10.2166/wst.2006.025. URL http://dx.doi.org/10.2166/wst.2006.025. 12

J. De Vrieze, M. E. Christiaens, and W. Verstraete. The microbiome as engineering tool: Manufacturing and trading between microorganisms. *New Biotechnology*, 39:206–214, Oct. 2017. ISSN 1871-6784. doi: 10.1016/j.nbt.2017.07.001. URL http://dx.doi.org/10.1016/j.nbt.2017.07.001. 1

G. Denevi, M. Pontil, and C. Ciliberto. The advantage of conditional meta-learning for biased regularization and fine-tuning. *Neural Information Processing Systems (NeurIPS) 2020*, 2020. 38, 222

G. Denevi, M. pontil, and C. Ciliberto. Conditional meta-learning of linear representations. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=0Uejkm1GB1U. 38

K. Derbal, M. Bencheikh-Lehocine, F. Cecchi, A.-H. Meniai, and P. Pavan. Application of the IWA ADM1 model to simulate anaerobic co-digestion of organic waste with waste activated sludge in mesophilic condition. *Bioresource technology*, 100(4):1539–1543, 2009. URL https://doi.org/10.1016/j.biortech.2008.07.064. 2, 11, 13, 87

N. Deveci and G. Çiftçi. A mathematical model for the anaerobic treatment of baker's yeast effluents. *Waste Management*, 21(1):99–103, 2001. URL https://doi.org/10.1016/S0956-053X(00)00072-6. 87

P. Devos, M. Haddad, and H. Carrère. Thermal hydrolysis of municipal sludge: Finding the temperature sweet spot: A review. *Waste and Biomass Valorization*, 12(5):2187–2205, June 2020. ISSN 1877-265X. doi: 10.1007/s12649-020-01130-1. URL http://dx.doi.org/10.1007/s12649-020-01130-1. 205

N. Ding, X. Chen, T. Levinboim, S. Goodman, and R. Soricut. Bridging the gap between practice and PAC-bayes theory in few-shot meta-learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713845393. 40, 41

C. Dittmer, J. Krümpel, and A. Lemmer. Modeling and simulation of biogas production in full scale with time series analysis. *Microorganisms*, 9(2):324, Feb. 2021. ISSN 2076-2607. doi: 10.3390/microorganisms9020324. URL http://dx.doi.org/10.3390/microorganisms9020324. 2

D. Dochain and P. Vanrolleghem. *Dynamical Modelling and Estimation in Wastewater Treatment Processes*. IWA Publishing, 2001. URL https://doi.org/10.2166/9781780403045. 13, 87, 108

A. Donoso-Bravo, J. Mailier, C. Martin, J. Rodríguez, C. Aceves-Lara, and A. Wouwer. Model selection, identification and validation in anaerobic digestion: a review. *Water research*, 45 (17):5347–5364, 2011. URL https://doi.org/10.1016/j.watres.2011.08.059. 13, 87

A. Donoso-Bravo, J. Mailier, G. Ruiz-Filippi, and A. Vande Wouwer. Identification in an anaerobic batch system: global sensitivity analysis, multi-start strategy and optimization criterion selection. *Bioprocess and biosystems engineering*, 36(1):35–43, 2013. URL https://doi.org/10.1007/s00449-012-0758-5. 87, 88

A. Donoso-Bravo, D. Olivares, Y. Lesty, and H. V. Bossche. Exploitation of the ADM1 in a XXI century wastewater resource recovery facility (WRRF): The case of codigestion and thermal hydrolysis. *Water Research*, 175:115654, May 2020. ISSN 0043-1354. doi: 10.1016/j.watres.2020.115654. URL http://dx.doi.org/10.1016/j.watres.2020.115654. 2, 11

M. D. Donsker and S. R. S. Varadhan. Large deviations for Markov processes and the asymptotic evaluation of certain Markov process expectations for large times. In *Probabilistic Methods in Differential Equations*, pages 82–88. Springer, 1975. 25, 46, 61

M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou. The Faiss library, 2024. URL https://arxiv.org/abs/2401.08281. 148

G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proc. Conf. Uncertainty in Artif. Intell. (UAI)*, Sydney, Australia, 8 2017. 30, 32, 69, 98, 130, 135

G. K. Dziugaite and D. M. Roy. Data-dependent PAC-Bayes priors via differential privacy. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/9a0ee0a9e7a42d2d69b8f86b3a0756b1-Paper.pdf. 30, 31

G. K. Dziugaite, K. Hsu, W. Gharbieh, G. Arpino, and D. Roy. On the role of data in PAC-Bayes bounds. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 604–612. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/karolina-dziugaite21a.html. 30, 69

H. Edelsbrunner. Alpha shapes-a survey. In *Tessellations in the Sciences: Virtues, Techniques and Applications of Geometric Tilings*. Springer, 2011. URL http://graphics.stanford.edu/courses/cs268-14-fall/Handouts/AlphaShapes/2010-B-01-AlphaShapes.pdf. 118

D. Edwards. On the Kantorovich–Rubinstein theorem. *Expositiones Mathematicae*, 29(4): 387–398, 2011. ISSN 0723-0869. doi: 10.1016/j.exmath.2011.06.005. URL http://dx.doi.org/10.1016/j.exmath.2011.06.005. 28

A. Farid and A. Majumdar. Generalization bounds for meta-learning via PAC-bayes and uniform stability. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=RloMRU3keo3. 41

Z. Fatolahi, G. Arab, and V. Razaviarani. Calibration of the anaerobic digestion model no. 1 for anaerobic digestion of organic fraction of municipal solid waste under mesophilic condition. *Biomass and Bioenergy*, 139:105661, 2020. URL https://doi.org/10.1016/j.biombioe.2020.105661. 87

L. Fe-Fei, Fergus, and Perona. A Bayesian approach to unsupervised one-shot learning of object categories. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1134–1141 vol.2, 2003. doi: 10.1109/ICCV.2003.1238476. 39

H. Feldman, X. Flores-Alsina, P. Ramin, K. Kjellberg, U. Jeppsson, D. Batstone, and K. Gernaey. Modelling an industrial anaerobic granular reactor using a multi-scale approach. *Water research*, 126:488–500, 2017. URL https://doi.org/10.1016/j.watres.2017.09.033. 87

W. Feller. *An Introduction to Probability Theory and Its Applications*. Number vol. 1 à 2 in An Introduction to Probability Theory and Its Applications. Wiley, 1957. ISBN 9780471257097. URL https://books.google.fr/books?id=K7kdAQAAMAAJ. 23

B. Fezzani and R. Ben Cheikh. Extension of the anaerobic digestion model no. 1 (ADM1) to include phenol compounds biodegradation processes for simulating the anaerobic co-digestion of olive mill wastes at mesophilic temperature. *Journal of Hazardous Materials*, 172(2–3):1430–1438, Dec. 2009. ISSN 0304-3894. doi: 10.1016/j.jhazmat.2009.08.017. URL http://dx.doi.org/10.1016/j.jhazmat.2009.08.017. 13

B. Fezzani and R. B. Cheikh. Implementation of IWA anaerobic digestion model no. 1 (ADM1) for simulating the thermophilic anaerobic co-digestion of olive mill wastewater with olive mill solid waste in a semi-continuous tubular digester. *Chemical Engineering Journal*, 141(1–3): 75–88, July 2008. ISSN 1385-8947. doi: 10.1016/j.cej.2007.10.024. URL http://dx.doi.org/10.1016/j.cej.2007.10.024. 13

G. Fichtenholz and L. Kantorovitch. Sur les opérations linéaires dans l'espace des fonctions bornées. *Studia Mathematica*, 5(1):69–98, 1934. ISSN 1730-6337. doi: 10.4064/sm-5-1-69-98. URL http://dx.doi.org/10.4064/sm-5-1-69-98. 46

C. Finn and S. Levine. Meta-learning and universality: Deep representations and gradient descent can approximate any learning algorithm. In *Proceedings of ICLR 2018*, 2018. doi: 10.48550/ARXIV.1710.11622. URL https://arxiv.org/abs/1710.11622. 37

C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/finn17a.html. 2, 37, 207

C. Finn, K. Xu, and S. Levine. Probabilistic model-agnostic meta-learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9537–9548, Red Hook, NY, USA, 2018. Curran Associates Inc. 39

E. Fix and J. L. Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review / Revue Internationale de Statistique*, 57(3):238, Dec. 1989. ISSN 0306-7734. doi: 10.2307/1403797. URL http://dx.doi.org/10.2307/1403797. 35

X. Flotats, J. Palatsi, B. Ahring, and I. Angelidaki. Identifiability study of the proteins degradation model, based on ADM1, using simultaneous batch experiments. *Water Science and Technology*, 54(4):31–39, Aug. 2006. ISSN 1996-9732. doi: 10.2166/wst.2006.523. URL http://dx.doi.org/10.2166/wst.2006.523. 2, 11, 88

A. Y. K. Foong, W. P. Bruinsma, D. R. Burt, and R. E. Turner. How tight can PAC-Bayes be in the small data regime? In *Advances in Neural Information Processing Systems*, volume 35, 2021. 28

N. Ganidi, S. Tyrrel, and E. Cartmell. Anaerobic digestion foaming causes – a review. *Bioresource Technology*, 100(23):5546–5554, Dec. 2009. ISSN 0960-8524. doi: 10.1016/j.biortech.2009.06.024. URL http://dx.doi.org/10.1016/j.biortech.2009.06.024. 10

F. Garcıa-Ochoa, V. Santos, L. Naval, E. Guardiola, and B. Lopez. Kinetic model for anaerobic digestion of livestock manure. *Enzyme and microbial technology*, 25(1-2):55–60, 1999. URL https://doi.org/10.1016/S0141-0229(99)00014-9. 87

M. Gerber. Les données de la méthanisation en france, 2022. URL https://chambres-agriculture.fr/fileadmin/user_upload/225_chambre_dagriculture_france/Actu/actu/2022/Methanisation-france-chiffres-cles-2022.pdf. Access date: 5 december 2024. 9

P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. PAC-Bayesian learning of linear classifiers. In *Proc. Int. Conf. Mach. Learning (ICML)*, Montreal, Canada, 06 2009. doi: 10.1145/1553374.1553419. 28, 29, 32, 41, 47, 149

P. Germain, F. Bach, A. Lacoste, and S. Lacoste-Julien. PAC-Bayesian theory meets Bayesian inference. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1884–1892. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6569-pac-bayesian-theory-meets-bayesian-inference.pdf. 3

A. Goldenshluger and O. Lepski. On adaptive minimax density estimation on $\mathbb{R}^d$. *Probability Theory and Related Fields*, 159(3-4):479–543, July 2013. URL https://doi.org/10.1007/s00440-013-0512-1. 89

L. Gonzalez-Gil, M. Mauricio-Iglesias, M. Carballa, and J. M. Lema. Why are organic micropollutants not fully biotransformed? a mechanistic modelling approach to anaerobic systems. *Water Research*, 142:115–128, Oct. 2018. ISSN 0043-1354. URL http://dx.doi.org/10.1016/j.watres.2018.05.032. 110

J. Gordon, J. Bronskill, M. Bauer, S. Nowozin, and R. Turner. Meta-learning probabilistic inference for prediction. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HkxStoC5F7. 40

E. Grant, C. Finn, S. Levine, T. Darrell, and T. Griffiths. Recasting gradient-based meta-learning as hierarchical Bayes. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJ_UL-k0b. 39

P. Grünwald. Safe learning: bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity. In S. M. Kakade and U. von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 397–420, Budapest, Hungary, 6 2011. PMLR. URL https://proceedings.mlr.press/v19/grunwald11a.html. 14

P. Grünwald. *The Safe Bayesian: Learning the Learning Rate via the Mixability Gap*, page 169–183. Springer Berlin Heidelberg, 2012. ISBN 9783642341069. doi: 10.1007/978-3-642-34106-9_16. URL http://dx.doi.org/10.1007/978-3-642-34106-9_16. 20, 21

P. Grünwald and T. van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4), Dec. 2017. ISSN 1936-0975. doi: 10.1214/17-ba1085. URL http://dx.doi.org/10.1214/17-BA1085. 20

P. D. Grünwald and N. A. Mehta. Fast rates for general unbounded loss functions: From ERM to generalized bayes. *Journal of Machine Learning Research*, 21(56):1–80, 2020. URL http://jmlr.org/papers/v21/18-488.html. 20

J. Guan and Z. Lu. Fast-rate PAC-Bayesian generalization bounds for meta-learning. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7930–7948. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/guan22b.html. 41

B. Guedj. A primer on PAC-Bayesian learning. In *Proceedings of the second congress of the French Mathematical Society*, volume 33, pages 391–414, 2019. URL https://arxiv.org/abs/1901.05353. 14, 23

B. Guedj and P. Alquier. PAC-Bayesian estimation and prediction in sparse additive models. *Electronic Journal of Statistics*, 7(none):264 – 291, 2013. doi: 10.1214/13-EJS771. URL https://doi.org/10.1214/13-EJS771. 31

B. Guedj and S. Robbiano. PAC-Bayesian high dimensional bipartite ranking. *Journal of Statistical Planning and Inference*, 196:70–86, Aug. 2018. ISSN 0378-3758. doi: 10.1016/j.jspi.2017.10.010. URL http://dx.doi.org/10.1016/j.jspi.2017.10.010. 31

A. Guisasola, K. Sharma, J. Keller, and Z. Yuan. Development of a model for assessing methane formation in rising main sewers. *Water Research*, 43(11):2874–2884, 2009. URL https://doi.org/10.1016/j.watres.2009.03.040. 87, 88

C. G. Gunnerson, D. C. Stuckey, M. Greeley, R. T. Skrinde, and R. F. Ward. Anaerobic digestion - principles and practices for biogas systems. Technical Report 49, The World Bank, 181 H Street, N.W., Washington, D.C., U.S.A., 4 1986. https://documents1.worldbank.org/curated/pt/980401468740176249/pdf/multi-page.pdf. 8

J. Haag, A. Wouwer, and I. Queinnec. Macroscopic modelling and identification of an anaerobic waste treatment process. *Chemical Engineering Science*, 58(19):4307–4316, 2003. URL https://doi.org/10.1016/S0009-2509(03)00272-0. 87, 88

M. Haddouche and B. Guedj. Online PAC-Bayes learning. *Advances in Neural Information Processing Systems*, 35:25725–25738, 2022. URL https://doi.org/10.48550/arXiv.2206.00024. 15

M. Haddouche, B. Guedj, O. Rivasplata, and J. Shawe-Taylor. PAC-Bayes unleashed: generalisation bounds with unbounded losses. *Entropy*, 23(10):1330, 2021. 28, 29, 30, 33

A. Hajji, Y. Louartassi, M. Garoum, N. Laaroussi, and M. Rhachi. Modification and extension of the anaerobic model n°2 (AM2) for the simulation of anaerobic digestion of municipal solid waste. *International Journal of Renewable Energy Development*, 12(5):913–922, 2023. doi: 10.14710/ijred.2023.52798. URL https://ijred.cbiore.id/index.php/ijred/article/view/52798. 12

N. Hansen. The CMA evolution strategy: A tutorial. *arXiv preprint arXiv:1604.00772*, 2016. URL https://doi.org/10.48550/arXiv.1604.00772. 4, 106

O. Harker-Sanchez, A. A. Jaramillo, and D. M. Arias. Method to obtain parameters k2, k3 for dilution rate observer in AM2 model of the anaerobic digestion process in a batch reactor. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 46(1):3110–3123, Feb. 2024. ISSN 1556-7230. doi: 10.1080/15567036.2024.2311326. URL http://dx.doi.org/10.1080/15567036.2024.2311326. 12

S. Hassam, E. Ficara, A. Leva, and J. Harmand. A generic and systematic procedure to derive a simplified model from the anaerobic digestion model no. 1 (ADM1). *Biochemical Engineering Journal*, 99:193–203, July 2015. URL https://doi.org/10.1016/j.bej.2015.03.007. 12, 92, 93

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, Apr. 1970. ISSN 0006-3444. doi: 10.1093/biomet/57.1.97. URL http://dx.doi.org/10.1093/biomet/57.1.97. 31

T. H. Hildebrandt. On bounded linear functional operations. *Transactions of the American Mathematical Society*, 36(4):868–875, 1934. ISSN 1088-6850. doi: 10.1090/s0002-9947-1934-1501772-9. URL http://dx.doi.org/10.1090/S0002-9947-1934-1501772-9. 46

G. E. Hinton and D. van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13, 1993. URL https://doi.org/10.1145/168304.168306. 90

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, Mar. 1963. ISSN 1537-274X. doi: 10.1080/01621459.1963.10500830. URL http://dx.doi.org/10.1080/01621459.1963.10500830. 266

T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey. Meta-Learning in Neural Networks: A Survey . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 44(09):5149–5169, Sept. 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3079209. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2021.3079209. 33, 34

J. G. Ibrahim and M.-H. Chen. Power prior distributions for regression models. *Statistical Science*, 15:46–60, 2000. URL https://api.semanticscholar.org/CorpusID:121710229. 20

A. Ismail, E. Elbeshbishy, and G. Nakhla. Thermal hydrolysis pretreatment of wastewater biosolids: Modelling the impact of the aerobic sludge age. *Journal of Water Process Engineering*, 60:105114, Apr. 2024. ISSN 2214-7144. doi: 10.1016/j.jwpe.2024.105114. URL http://dx.doi.org/10.1016/j.jwpe.2024.105114. 203

P. Jaiswal, V. Rao, and H. Honnappa. Asymptotic consistency of α-Rényi-approximate posteriors. *Journal of Machine Learning Research*, 21(156):1–42, 2020. URL http://jmlr.org/papers/v21/19-161.html. 32

P. D. Jensen, S. Astals, X. Bai, L. Nieradzik, P. Wardrop, D. J. Batstone, and W. P. Clarke. *Established full-scale applications for energy recovery from water: anaerobic digestion*, page 99–139. IWA Publishing, Feb. 2022. ISBN 9781780409566. doi: 10.2166/9781780409566_0099. URL http://dx.doi.org/10.2166/9781780409566_0099. 194, 196

H.-S. Jeong, C.-W. Suh, J.-L. Lim, S.-H. Lee, and H.-S. Shin. Analysis and application of ADM1 for anaerobic methane production. *Bioprocess and biosystems engineering*, 27(2):81–89, 2005. URL https://doi.org/10.1007/s00449-004-0370-4. 11, 87

G. Jerfel, E. Grant, T. Griffiths, and K. A. Heller. Reconciling meta-learning and continual learning with online mixtures of tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/7a9a322cbe0d06a98667fdc5160dc6f8-Paper.pdf. 38

H. Kalfas, I. V. Skiadas, H. N. Gavala, K. Stamatelatou, and G. Lyberatos. Application of ADM1 for the simulation of anaerobic digestion of olive pulp under mesophilic and thermophilic conditions. *Water Science and Technology*, 54(4):149–156, 2006. URL https://doi.org/10.2166/wst.2006.536. 2, 11, 87

P.-A. Kamienny, S. d'Ascoli, G. Lample, and F. Charton. End-to-end symbolic regression with transformers. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=GoOuIrDHG_Y. 228

R. Kapoor, P. Ghosh, M. Kumar, and V. K. Vijay. Evaluation of biogas upgrading technologies and future perspectives: a review. *Environmental Science and Pollution Research*, 26(12):11631–11661, Mar. 2019. ISSN 1614-7499. doi: 10.1007/s11356-019-04767-1. URL http://dx.doi.org/10.1007/s11356-019-04767-1. 204

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *ICLR (Poster)*, 2015. URL http://dblp.uni-trier.de/db/conf/iclr/iclr2015.html#KingmaB14. 210, 211

A. Klenke. *Probability theory: a comprehensive course*. Springer, 2020. 101, 137

B. A. Klinger and S. J. Ryan. Population distribution within the human climate niche. *PLOS Climate*, 1(11):e0000086, Nov. 2022. ISSN 2767-3200. doi: 10.1371/journal.pclm.0000086. URL http://dx.doi.org/10.1371/journal.pclm.0000086. 203

J. Knoblauch, J. Jewson, and T. Damoulas. An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23 (132):1–109, 2022. URL http://jmlr.org/papers/v23/19-1047.html. 14, 32

G. R. Koch. Siamese neural networks for one-shot image recognition. In *Proceedings of the 32 nd International Conference on Machine Learning*. PMLR, 2015. URL https://api.semanticscholar.org/CorpusID:13874643. 36

K. Koch, M. Lübken, T. Gehring, M. Wichern, and H. Horn. Biogas from grass silage – measurements and modeling with ADM1. *Bioresource Technology*, 101(21):8158–8165, Nov. 2010. ISSN 0960-8524. doi: 10.1016/j.biortech.2010.06.009. URL http://dx.doi.org/10.1016/j.biortech.2010.06.009. 13

A. Kovalovszki, M. Alvarado-Morales, I. A. Fotidis, and I. Angelidaki. A systematic methodology to extend the applicability of a bioconversion model for the simulation of various co-digestion scenarios. *Bioresource technology*, 235:157–166, 2017. URL https://doi.org/10.1016/j.biortech.2017.03.101. 87

B. M. Lake, R. R. Salakhutdinov, and J. Tenenbaum. One-shot learning by inverting a compositional causal process. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/52292e0c763fd027c6eba6b8f494d2eb-Paper.pdf. 39

B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. doi: 10.1126/science.aab3050. URL https://www.science.org/doi/abs/10.1126/science.aab3050. 36, 39

S. Lang. *Fundamentals of Differential Geometry*. Springer New York, 1999. ISBN 9781461205418. doi: 10.1007/978-1-4612-0541-8. URL http://dx.doi.org/10.1007/978-1-4612-0541-8. 140

J. Langford and M. Seeger. Bounds for averaging classifiers. http://www.cs.cmu.edu/ jcl/papers/averaging/averaging_tech.pdf, 2001. 27, 47, 61

J. Lee, J. Sohl-dickstein, J. Pennington, R. Novak, S. Schoenholz, and Y. Bahri. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1EA-M-0Z. 80

E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer New York, NY, 2nd edition, Aug. 1998. ISBN 0387985026. URL https://doi.org/10.1007/b98854. 16, 88, 106

G. Letarte, P. Germain, B. Guedj, and F. Laviolette. Dichotomize and generalize: PAC-Bayesian binary activated deep neural networks. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems [NeurIPS] 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 6869–6879, 2019. URL https://papers.nips.cc/paper/8911-dichotomize-and-generalize-pac-bayesian-binary-activated-deep-neural-networks. 33

A. Leurent and R. Moscoviz. Modeling a propionate-oxidizing syntrophic coculture using thermodynamic principles. *Biotechnology and Bioengineering*, 119(9):2423–2436, June 2022. URL https://doi.org/10.1002/bit.28156. 104

D. Li, I. Lee, and H. Kim. Application of the linearized ADM1 (LADM) to lab-scale anaerobic digestion system. *Journal of Environmental Chemical Engineering*, 9(3):105193, 2021. URL https://doi.org/10.1016/j.jece.2021.105193. 2, 11, 13, 87

Y. Li and R. E. Turner. Rényi divergence variational inference. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/7750ca3559e5b8e1f44210283368fc16-Paper.pdf. 32

B. G. Liptak. Environmental engineering handbook. *Chilton Book Co., Radnor, PA*, 1974. 194

F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*. IEEE, Dec. 2008. URL https://doi.org/10.1109/icdm.2008.17. 111

J. Liu and S. R. Smith. The link between organic matter composition and the biogas yield of full-scale sewage sludge anaerobic digestion. *Water Science and Technology*, 85(5):1658–1672, Feb. 2022. ISSN 1996-9732. doi: 10.2166/wst.2022.058. URL http://dx.doi.org/10.2166/wst.2022.058. 203

Q. Liu and D. Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/b3ba8f1bee1238a2f37603d90b58898d-Paper.pdf. 39

T. Liu, J. Lu, Z. Yan, and G. Zhang. PAC-Bayes bounds for meta-learning with data-dependent prior, 2021. URL https://arxiv.org/abs/2102.03748. 40, 41

L. Lokshina, V. Vavilin, R. Kettunen, J. R., C. Holliger, and A. Nozhevnikova. Evaluation of kinetic coefficients using integrated Monod and Haldane models for low-temperature acetoclastic methanogenesis. *Water Research*, 35(12):2913–2922, 2001. URL https://doi.org/10.1016/S0043-1354(00)00595-9. 87, 88

I. López and L. Borzacconi. Modelling of slaughterhouse solid waste anaerobic digestion: Determination of parameters and continuous reactor simulation. *Waste management*, 30(10):1813–1821, 2010. URL https://doi.org/10.1016/j.wasman.2010.02.034. 88

S. Lotfi, M. Finzi, S. Kapoor, A. Potapczynski, M. Goldblum, and A. G. Wilson. PAC-Bayes compression bounds so tight that they can explain generalization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in*

Neural Information Processing Systems, volume 35, pages 31459–31473. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/cbeec55c50c3367024bafab2438a021b-Paper-Conference.pdf. 81, 82, 83, 229

F. Mairet, O. Bernard, M. Ras, L. Lardon, and J.-P. Steyer. Modeling anaerobic digestion of microalgae using ADM1. *Bioresource Technology*, 102(13):6823–6829, 2011. URL https://doi.org/10.1016/j.biortech.2011.04.015. 11, 13, 87

M. Maktabifard, E. Zaborowska, and J. Makinia. Achieving energy neutrality in wastewater treatment plants through energy savings and enhancing renewable energy production. *Reviews in Environmental Science and Bio/Technology*, 17(4):655–689, Oct. 2018. ISSN 1572-9826. doi: 10.1007/s11157-018-9478-x. URL http://dx.doi.org/10.1007/s11157-018-9478-x. 205

G. Marsh. Rise of the anaerobic digester. *Renewable Energy Focus*, 9(6):28–34, Nov. 2008. ISSN 1755-0084. doi: 10.1016/s1755-0084(08)70063-2. URL http://dx.doi.org/10.1016/S1755-0084(08)70063-2. 8

C. Martin and E. Ayesa. An integrated Monte Carlo methodology for the calibration of water quality models. *Ecological Modelling*, 221(22):2656–2667, 2010. ISSN 0304-3800. URL https://doi.org/10.1016/j.ecolmodel.2010.08.008. 14, 124

C. Martin, C. de Gracia, and E. Ayesa. Bayesian calibration of the disintegration process in WWTP sludge digesters. *8th IWA Sumposiym on Systems Analysis and Integrated Assessment (WaterMatex)*, 2011. 89, 124, 125

A. Maurer. A note on the PAC Bayesian theorem, 2004. URL https://arxiv.org/abs/cs/0411099. 27, 47, 137

D. A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37:355–363, 1999a. URL https://doi.org/10.1023/A:1007618624809. 27

D. A. McAllester. PAC-Bayesian model averaging. *COLT*, 1999b. 27, 65

D. A. McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003. 20, 61, 137

D. McDonald. Minimax Density Estimation for Growing Dimension. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 194–203. PMLR, 4 2017. URL https://proceedings.mlr.press/v54/mcdonald17a.html. 89

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953. ISSN 1089-7690. doi: 10.1063/1.1699114. URL http://dx.doi.org/10.1063/1.1699114. 31

Z. Mhammedi, P. Grünwald, and B. Guedj. PAC-Bayes un-expected Bernstein inequality. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, volume 32, Vancouver, Canada, 12 2019. 15, 30, 130

Ministère de la Transition Écologique et Solidaire. Stratégie nationale bas carbone révisée complète relative au décret n°2020-457 du 21 avril 2020 relatif aux budgets carbone nationaux et à la stratégie nationale bas-carbone, 4 2020. URL https://www.ecologie.gouv.fr/sites/default/files/documents/2020-03-25_MTES_SNBC2.pdf. 1

T. P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 362–369. Morgan Kaufmann Publishers Inc., 2001. ISBN 1558608001. 32

R. Mo, W. Guo, D. Batstone, J. Makinia, and Y. Li. Modifications to the anaerobic digestion model no. 1 (ADM1) for enhanced understanding and application of the anaerobic treatment processes – a comprehensive review. *Water Research*, 244:120504, Oct. 2023. ISSN 0043-1354. doi: 10.1016/j.watres.2023.120504. URL http://dx.doi.org/10.1016/j.watres.2023.120504. 11, 195

R. Moletta. *La méthanisation (3e éd.)*. Lavoisier, 2015. ISBN 978-2-7340-1991-4. 7, 8, 9

V. Monje, H. Junicke, D. J. Batstone, K. Kjellberg, K. V Gernaey, and X. Flores-Alsina. Prediction of mass and volumetric flows in a full-scale industrial waste treatment plant. *Chemical Engineering Journal*, 445:136774, Oct. 2022. ISSN 1385-8947. doi: 10.1016/j.cej.2022.136774. URL http://dx.doi.org/10.1016/j.cej.2022.136774. 13

M. D. Morris. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174, May 1991. URL https://doi.org/10.1080/00401706.1991.10484804. 105

M. Mösche and H.-J. Jördening. Comparison of different models of substrate and product inhibition in anaerobic digestion. *Water Research*, 33(11):2545–2554, 1999. URL https://doi.org/10.1016/S0043-1354(98)00490-4. 87

R. Moscoviz and J. Jimenez. Improving anaerobic digestion mass balance calculations through stoichiometry and usual substrate characterization. *Bioresource Technology*, 337:125402, Oct. 2021. ISSN 0960-8524. doi: 10.1016/j.biortech.2021.125402. URL http://dx.doi.org/10.1016/j.biortech.2021.125402. 194

K. Nalini, T. Lauvaux, C. Abdallah, J. Lian, P. Ciais, H. Utard, O. Laurent, and M. Ramonet. High-resolution Lagrangian inverse modeling of CO2 emissions over the Paris region during the first 2020 lockdown period. *Journal of Geophysical Research: Atmospheres*, 127(14), July 2022. ISSN 2169-8996. doi: 10.1029/2021jd036032. URL http://dx.doi.org/10.1029/2021JD036032. 160

B. Neyshabur, S. Bhojanapalli, D. A. McAllester, and N. Srebro. Exploring generalization in deep learning. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, 12 2017. 130

A. Nichol, J. Achiam, and J. Schulman. On first-order meta-learning algorithms, 2018. URL https://arxiv.org/abs/1803.02999. 37, 207

M. Odriozola, E. Abraham, M. Lousada-Ferreira, H. Spanjers, and J. Van Lier. Identification of the methanogenesis inhibition mechanism using comparative analysis of mathematical models. *Frontiers in bioengineering and biotechnology*, 7:93, 2019. URL https://doi.org/10.3389/fbioe.2019.00093. 87, 88

Y. Ohnishi and J. Honorio. Novel change of measure inequalities with applications to PAC-Bayesian bounds and Monte Carlo estimation. In A. Banerjee and K. Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 1711–1719. PMLR, 2021. URL http://proceedings.mlr.press/v130/ohnishi21a.html. 44, 61

M. Opper and O. Winther. Gaussian processes for classification: Mean-field algorithms. *Neural Computation*, 12(11):2655–2684, 2000. doi: 10.1162/089976600300014881. 32

S. Panigrahi and B. K. Dubey. A critical review on operating parameters and strategies to improve the biogas yield from anaerobic digestion of organic fraction of municipal solid waste. *Renewable Energy*, 143:779–797, Dec. 2019. ISSN 0960-1481. doi: 10.1016/j.renene.2019.05.040. URL http://dx.doi.org/10.1016/j.renene.2019.05.040. 10

W. Parker. Application of the ADM1 model to advanced anaerobic digestion. *Bioresource Technology*, 96(16):1832–1842, Nov. 2005. ISSN 0960-8524. doi: 10.1016/j.biortech.2005.01.022. URL http://dx.doi.org/10.1016/j.biortech.2005.01.022. 12

E. Parrado-Hernández, A. Ambroladze, J. Shawe-Taylor, and S. Sun. PAC-Bayes bounds with data dependent priors. *Journal of Machine Learning Research*, 13(112):3507–3531, 2012. URL http://jmlr.org/papers/v13/parrado12a.html. 30, 69

V. Pastor-Poquet, S. Papirio, J. Harmand, J.-P. Steyer, E. Trably, R. Escudiá, and G. Esposito. Assessing practical identifiability during calibration and cross-validation of a structured model for high-solids anaerobic digestion. *Water Research*, 164:114932, 2019. ISSN 0043-1354. URL https://doi.org/10.1016/j.watres.2019.114932. 14, 88, 89, 124, 125

M. Patacchiola, J. Turner, E. J. Crowley, M. O' Boyle, and A. J. Storkey. Bayesian meta-learning for the few-shot setting via deep kernels. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16108–16118. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/b9cfe8b6042cf759dc4c0cccb27a6737-Paper.pdf. 40

M. Patón and J. Rodríguez. Integration of bioenergetics in the ADM1 and its impact on model predictions. *Water Science and Technology*, 80(2):339–346, July 2019. ISSN 1996-9732. doi: 10.2166/wst.2019.279. URL http://dx.doi.org/10.2166/wst.2019.279. 13

A. Pentina and C. Lampert. A PAC-Bayesian bound for lifelong learning. In *International Conference on Machine Learning*, pages 991–999. PMLR, 2014. 2, 40, 41

M. Pérez-Ortiz, O. Rivasplata, B. Guedj, M. Gleeson, J. Zhang, J. Shawe-Taylor, M. Bober, and J. Kittler. Learning PAC-Bayes priors for probabilistic neural networks. arXiv preprint, 2021a. URL https://arxiv.org/abs/2109.10304. 30, 71, 81, 82

M. Pérez-Ortiz, O. Rivasplata, J. Shawe-Taylor, and C. Szepesvári. Tighter risk certificates for neural networks. *Journal of Machine Learning Research (JMLR)*, 22(227):1–40, 2021b. 28, 30, 32, 69, 71, 81, 82, 98

A. Picard-Weibel and B. Guedj. On change of measure inequalities for f-divergences, 2022. URL https://arxiv.org/abs/2202.05568. 4, 44

A. Picard-Weibel, G. Capson-Tojo, B. Guedj, and R. Moscoviz. Bayesian uncertainty quantification for anaerobic digestion models. *Bioresource Technology*, 394:130147, Feb. 2024a. ISSN 0960-8524. doi: 10.1016/j.biortech.2023.130147. URL http://dx.doi.org/10.1016/j.biortech.2023.130147. 4, 84

A. Picard-Weibel, R. Moscoviz, and B. Guedj. Learning via surrogate PAC-Bayes. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems [NeurIPS] 2024, 8-15 December 2024, Vancouver, BC, Canada*, 2024b. URL https://neurips.cc/virtual/2024/poster/95781. 4, 128, 130, 192

A. Picard-Weibel, E. Clerico, R. Moscoviz, and B. Guedj. How good is PAC-Bayes at explaining generalisation?, 2025. URL https://arxiv.org/abs/2503.08231. 4, 44

D. Poggio, M. Walker, W. Nimmo, L. Ma, and M. Pourkashanian. Modelling the anaerobic digestion of solid organic waste–substrate characterisation method for ADM1 using a combined biochemical and kinetic parameter estimation approach. *Waste management*, 53:40–54, 2016. URL https://doi.org/10.1016/j.wasman.2016.04.024. 87, 88

J. Pronto, C. Gooch, and N. J. Brown. Practical considerations and implementation of anaerobic digester systems, 2012. 204

J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset shift in machine learning*. Mit Press, 2008. URL https://mitpress.mit.edu/9780262545877/dataset-shift-in-machine-learning/. 88

L. Ralaivola, M. Szafranski, and G. Stempfel. Chromatic PAC-Bayes bounds for non-IID data: Applications to ranking and stationary β-mixing processes. *Journal of Machine Learning Research*, 11(65):1927–1956, 2010. URL http://jmlr.org/papers/v11/ralaivola10a.html. 29, 44, 45

I. Ramirez, A. Mottet, H. Carrère, S. Déléris, F. Vedrenne, and J.-P. Steyer. Modified ADM1 disintegration/hydrolysis structures for modeling batch thermophilic anaerobic digestion of thermally pretreated waste activated sludge. *Water research*, 43(14):3479–3492, 2009. URL https://doi.org/10.1016/j.watres.2009.05.023. 11, 13, 87

A. Regueira, R. Bevilacqua, M. Mauricio-Iglesias, M. Carballa, and J. Lema. Kinetic and stoichiometric model for the computer-aided design of protein fermentation into volatile fatty acids. *Chemical Engineering Journal*, 406:126835, 2021. URL https://doi.org/10.1016/j.cej.2020.126835. 87, 88, 104, 110

A. Rezazadeh. A unified view on PAC-Bayes bounds for meta-learning. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18576–18595. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/rezazadeh22a.html. 40, 41

L. Rieger, S. Gillot, G. Langergraber, T. Ohtsuki, A. Shaw, I. Takacs, and S. Winkler. *Guidelines for using activated sludge models*. IWA publishing, 2012. URL https://www.iwapublishing.com/books/9781843391746/guidelines-using-activated-sludge-models. 87, 88

C. P. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*, volume 2. Springer, 2007. 17

G. O. Roberts. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and Computing in Applied Probability*, 4(4):337–357, 2002. ISSN 1387-5841. doi: 10.1023/a:1023562417138. URL http://dx.doi.org/10.1023/A:1023562417138. 31

G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341 – 363, 1996. 31

C. Rosén and U. Jeppsson. Aspects on ADM1 implementation within the BSM2 framework. *Department of Industrial Electrical Engineering and Automation, Lund University, Lund, Sweden*, pages 1–35, 2006. 2, 10, 12, 86, 100, 106, 193, 198, 256, 257, 258, 259, 260, 261

J. Rothfuss, V. Fortuin, and A. Krause. PACOH: Bayes-optimal meta-learning with PAC-guarantees. In *4th Lifelong Machine Learning Workshop at ICML 2020*, 2020. URL https://openreview.net/forum?id=a5rImUm5rZR. 41

J. Rothfuss, C. Koenig, A. Rupenyan, and A. Krause. Meta-learning priors for safe Bayesian optimization. In *6th Annual Conference on Robot Learning*, 2022. URL https://openreview.net/forum?id=zNB_UVj5oKQ. 41

J. Rothfuss, M. Josifoski, V. Fortuin, and A. Krause. Scalable PAC-Bayesian meta-learning via the PAC-optimal hyper-posterior: From theory to practice. *Journal of Machine Learning Research*, 24:1–62, 2023. 40, 41

J. Rousseau, J.-B. Salomond, and C. Scricciolo. On some aspects of the asymptotic properties of Bayesian approaches in nonparametric and semiparametric models. *ESAIM: Proceedings*, 44:159–171, Jan. 2014. ISSN 1270-900X. doi: 10.1051/proc/201444010. URL http://dx.doi.org/10.1051/proc/201444010. 18

W. Rudin. *Real and Complex Analysis*. McGraw-Hill Science/Engineering/Math, 5 1986. ISBN 0070542341. URL https://dl.acm.org/doi/10.5555/26851. 21

S. Ruel, Y. Comeau, P. Ginestet, and A. Heduit. Modeling acidogenic and sulfate-reducing processes for the determination of fermentable fractions in wastewater. *Biotechnology and Bioengineering*, 80(5):525–536, 2002. URL https://doi.org/10.1002/bit.10410. 87, 88

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, Oct. 1986. 33

A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJgklhAcK7. 38

P. Sadrimajd, P. Mannion, E. Howley, and P. Lens. PyADM1: a python implementation of anaerobic digestion model no. 1. *bioRxiv*, 2021. URL https://doi.org/10.1101/2021.03.03.433746. 4, 92

M. Sales-Cruz and R. Gani. Aspects of modelling and model identification for bioprocesses through a computer-aided modelling system. In *Computer Aided Chemical Engineering*, volume 18, pages 1123–1128. Elsevier, 2004. URL https://doi.org/10.1016/S1570-7946(04)80253-0. 87

T. Sari and B. Benyahia. The operating diagram for a two-step anaerobic digestion model. *Nonlinear Dynamics*, 105(3):2711–2737, July 2021. ISSN 1573-269X. doi: 10.1007/s11071-021-06722-7. URL http://dx.doi.org/10.1007/s11071-021-06722-7. 12

B. A. Sayigh and J. F. Malina Jr. Temperature effects on the activated sludge process. *Water Pollution Control Federation*, pages 678–687, 1978. 203

R. L. Schilling. *Measures, Integrals and Martingales*. Cambridge University Press, 2005. 51

M. Schoen, D. Sperl, M. Gadermaier, M. Goberna, I. Franke-Whittle, H. Insam, J. Ablinger, and B. Wett. Population dynamics at digester overload conditions. *Bioresource technology*, 100 (23):5648–5655, 2009. URL https://doi.org/10.1016/j.biortech.2009.06.033. 87

M. Seeger. PAC-Bayesian generalisation error bounds for gaussian process classification. *Journal of machine learning research*, 3(Oct):233–269, 2002. URL https://www.jmlr.org/papers/volume3/seeger02a/seeger02a.pdf. 27, 30

Y. Seldin and N. Tishby. PAC-Bayesian analysis of co-clustering and beyond. *Journal of Machine Learning Research (JMLR)*, 11(117):3595–3646, 12 2010. 130

Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012. URL https://doi.org/10.1109/TIT.2012.2211334. 44

J. Shawe-Taylor and R. C. Williamson. A PAC analysis of a Bayesian estimator. In *Proc. Conf. Learn. Theory (COLT)*, 7 1997. doi: 10.1145/267460.267466. 27

B. Shrestha, R. Hernandez, D. L. B. Fortela, W. Sharp, A. Chistoserdov, D. Gang, E. Revellame, W. Holmes, and M. E. Zappi. A review of pretreatment methods to enhance solids reduction during anaerobic digestion of municipal wastewater sludges and the resulting digester performance: Implications to future urban biorefineries. *Applied Sciences*, 10(24):9141, Dec. 2020. ISSN 2076-3417. doi: 10.3390/app10249141. URL http://dx.doi.org/10.3390/app10249141. 204

I. Siegert and C. Banks. The effect of volatile fatty acid additions on the anaerobic digestion of cellulose and glucose in batch reactors. *Process Biochemistry*, 40(11):3412–3418, Nov. 2005. ISSN 1359-5113. doi: 10.1016/j.procbio.2005.01.025. URL http://dx.doi.org/10.1016/j.procbio.2005.01.025. 10

A. Spyridonidis, T. Skamagkis, L. Lambropoulos, and K. Stamatelatou. Modeling of anaerobic digestion of slaughterhouse wastes after thermal treatment using ADM1. *Journal of environmental management*, 224:49–57, 2018. URL https://doi.org/10.1016/j.jenvman.2018.07.001. 11, 13, 88

J. B. Tenenbaum. Bayesian modeling of human concept learning. In *Proceedings of the 11th International Conference on Neural Information Processing Systems*, NIPS'98, page 59–65, Cambridge, MA, USA, 1998. MIT Press. 39

J. B. Tenenbaum. *A Bayesian framework for concept learning*. PhD thesis, Massachusetts Institute of Technology, 1999. 39

U. Tezel, M. Tandukar, and S. Pavlostathis. *Anaerobic Biotreatment of Municipal Sewage Sludge*, page 447–461. Elsevier, 2011. ISBN 9780080885049. doi: 10.1016/b978-0-08-088504-9.00329-9. URL http://dx.doi.org/10.1016/B978-0-08-088504-9.00329-9. 195, 196, 203

J. Thual, S. Martin, and J. Mousset. La méthanisation. Technical report, ADEME, 2023. 9

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, Jan. 1996. ISSN 1467-9868. doi: 10.1111/j.2517-6161.1996.tb02080.x. URL http://dx.doi.org/10.1111/j.2517-6161.1996.tb02080.x. 100, 198

Y. M. Tobo, U. Rehman, J. Bartacek, and I. Nopens. Partial integration of adm1 into cfd: understanding the impact of diffusion on anaerobic digestion mixing. *Water Science and Technology*, 81(8):1658–1667, Feb. 2020. ISSN 1996-9732. doi: 10.2166/wst.2020.076. URL http://dx.doi.org/10.2166/wst.2020.076. 11

A. Tolessa, N. J. Goosen, and T. M. Louw. Probabilistic simulation of biogas production from anaerobic co-digestion using Anaerobic Digestion Model No. 1: A case study on agricultural residue. *Biochemical Engineering Journal*, 192:108810, Mar. 2023. URL https://doi.org/10.1016/j.bej.2023.108810. 125

I. O. Tolstikhin and Y. Seldin. PAC-Bayes-empirical-Bernstein inequality. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/a97da629b098b75c294dffdc3e463904-Paper.pdf. 72, 83, 137, 228

N. Tsalas, S. K. Golfinopoulos, S. Samios, G. Katsouras, and K. Peroulis. Optimization of energy consumption in a wastewater treatment plant: An overview. *Energies*, 17(12):2808, June 2024. ISSN 1996-1073. doi: 10.3390/en17122808. URL http://dx.doi.org/10.3390/en17122808. 204

A. Tsigkinopoulou, S. M. Baker, and R. Breitling. Respectful modeling: Addressing uncertainty in dynamic system models for molecular biology. *Trends in Biotechnology*, 35(6):518–529, June 2017. URL https://doi.org/10.1016/j.tibtech.2016.12.008. 100

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer New York, 2009. ISBN 9780387790527. doi: 10.1007/b13794. URL http://dx.doi.org/10.1007/b13794. 16

I. S. Turovskiy and P. Mathai. *Wastewater sludge processing*. John Wiley & Sons, 2006. 194

S.-M. Udrescu and M. Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16), Apr. 2020. ISSN 2375-2548. doi: 10.1126/sciadv.aay2631. URL http://dx.doi.org/10.1126/sciadv.aay2631. 228

A. van der Vaart. The statistical work of Lucien Le Cam. *The Annals of Statistics*, 30(3):631 – 682, 2002. doi: 10.1214/aos/1028674836. URL https://doi.org/10.1214/aos/1028674836. 18

M. C. M. van Loosdrecht, P. H. Nielsen, C. M. Lopez-Vazquez, and D. Brdjanovic. *Experimental methods in wastewater treatment*. IWA Publishing, London, England, 5 2016. 13, 88

J. Vanschoren. *Meta-Learning*, pages 35–61. Springer International Publishing, Cham, 2019. ISBN 978-3-030-05318-5. doi: 10.1007/978-3-030-05318-5_2. URL https://doi.org/10.1007/978-3-030-05318-5_2. 33

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer New York, 2000. ISBN 9781475732641. doi: 10.1007/978-1-4757-3264-1. URL http://dx.doi.org/10.1007/978-1-4757-3264-1. 68

V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, Jan. 1971. ISSN 1095-7219. doi: 10.1137/1116025. URL http://dx.doi.org/10.1137/1116025. 67

V. Vavilin, B. Fernandez, J. Palatsi, and X. Flotats. Hydrolysis kinetics in anaerobic degradation of particulate organic material: An overview. *Waste Management*, 28(6):939–951, 2008. ISSN 0956-053X. doi: 10.1016/j.wasman.2007.03.028. URL http://dx.doi.org/10.1016/j.wasman.2007.03.028. 195, 203

P. Viallard, M. Haddouche, U. Simsekli, and B. Guedj. Learning via Wasserstein-based high probability generalisation bounds. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=3Wrolscjbx. 28, 30, 69

O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra. Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/90e1357833654983612fb05e3ec9148c-Paper.pdf. 36

V. Vovk. Aggregating strategies. In M. Fulk and J. Case, editors, *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 371–383. Morgan Kaufmann, 1990. 20

R. Vuorio, S.-H. Sun, H. Hu, and J. J. Lim. Multimodal model-agnostic meta-learning via task-aware modulation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/e4da3b7fbbce2345d7772b0674a318d5-Paper.pdf. 38

R. Wang, Y. Demiris, and C. Ciliberto. Structured prediction for conditional meta-learning. *Neural Information Processing Systems (NeurIPS) 2020*, 2020. 38

S. Wang, C. Xu, L. Song, and J. Zhang. Anaerobic digestion of food waste and its microbial consortia: A historical review and future perspectives. *International Journal of Environmental Research and Public Health*, 19(15):9519, Aug. 2022. ISSN 1660-4601. doi: 10.3390/ijerph19159519. URL http://dx.doi.org/10.3390/ijerph19159519. 8

S. Weinrich, E. Mauky, T. Schmidt, C. Krebs, J. Liebetrau, and M. Nelles. Systematic simplification of the Anaerobic Digestion Model No. 1 (ADM1)–laboratory experiments and model application. *Bioresource Technology*, 333:125104, 2021. URL https://doi.org/10.1016/j.biortech.2021.125104. 11, 13, 87, 88, 89

M. Wichern, T. Gehring, K. Fischer, D. Andrade, M. Lübken, K. Koch, A. Gronauer, and H. Horn. Monofermentation of grass silage under mesophilic conditions: measurements and mathematical modeling with ADM1. *Bioresource technology*, 100(4):1675–1681, 2009. URL https://doi.org/10.1016/j.biortech.2008.09.030. 13, 87, 96

U. Wiesmann, I. S. Choi, and E.-M. Dombrowski. *Fundamentals of biological wastewater treatment*. John Wiley & Sons, 2007. 203

S. Xue, J. Song, X. Wang, Z. Shang, C. Sheng, C. Li, Y. Zhu, and J. Liu. A systematic comparison of biogas development and related policies between China and Europe and corresponding insights. *Renewable and Sustainable Energy Reviews*, 117:109474, Jan. 2020. ISSN 1364-0321. doi: 10.1016/j.rser.2019.109474. URL http://dx.doi.org/10.1016/j.rser.2019.109474. 8

J. Yoon, T. Kim, O. Dia, S. Kim, Y. Bengio, and S. Ahn. Bayesian model-agnostic meta-learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/e1021d43911ca2c1845910d84f40aeae-Paper.pdf. 39

H. Zakerinia, A. Behjati, and C. H. Lampert. More flexible PAC-Bayesian meta-learning by learning learning algorithms. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2024. 40

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx. 68

H. Zhang, S. Lee, and A. Tzempelikos. Bayesian meta-learning for personalized thermal comfort modeling. *Building and Environment*, 249:111129, Feb. 2024. ISSN 0360-1323. doi: 10.1016/j.buildenv.2023.111129. URL http://dx.doi.org/10.1016/j.buildenv.2023.111129. 40

Q. Zhang, J. Fang, Z. Meng, S. Liang, and E. Yilmaz. Variational continual Bayesian meta-learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24556–24568. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/cdd0500dc0ef6682fa6ec6d2e6b577c4-Paper.pdf. 40

X. Zhang, D. Meng, H. Gouk, and T. Hospedales. Shallow Bayesian Meta Learning for Real-World Few-Shot Recognition . In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 631–640, Los Alamitos, CA, USA, Oct. 2021b. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.00069. URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00069. 40

X. Zhao, L. Li, D. Wu, T. Xiao, Y. Ma, and X. Peng. Modified Anaerobic Digestion Model No. 1 for modeling methane production from food waste in batch and semi-continuous anaerobic digestions. *Bioresource technology*, 271:109–117, 2019. URL https://doi.org/10.1016/j.biortech.2018.09.091. 87, 88

H. Zhou, Z. Ying, Z. Cao, Z. Liu, Z. Zhang, and W. Liu. Feeding control of anaerobic co-digestion of waste activated sludge and corn silage performed by rule-based PID control with ADM1. *Waste Management*, 103:22–31, 2020. URL https://doi.org/10.1016/j.wasman.2019.12.021. 11, 13, 87

W. Zhou, V. Veitch, M. Austern, R. P. Adams, and P. Orbanz. Non-vacuous generalization bounds at the ImageNet scale: a PAC-Bayesian compression approach. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=BJgqqsAct7. 81, 229

Appendix A

Anaerobic Digestion models

A.1 Details on Anaerobic Digestion model 2

The Anaerobic Digestion model 2 (AM2) was introduced in Bernard et al. [2001]. The description of the model presented here closely follows the description provided in the original article.

A.1.1 State variables

AM2 considers six state variables:

- X_1, the concentration of acidogenic bacteria,
- X_2, the concentrataion of methanogenic bacteria,
- S_1, the concentration of organic substrate,
- S_2, the concentration of volatile fatty acids,
- Z, the total alkalinity,
- C, the concentration of inorganic carbon (bicarbonate and dissolved CO_2)

The overall state vector is noted $\xi = (X_1, X_2, S_1, S_2, Z, X)$.

A.1.2 Biochemical reactions

AM2 considers two reactions catalysed by micro organisms:

- The organic substrate (noted S_1) is transformed into volatile fatty acids (noted S_2) and carbonic gas by the acidogenic bacteria (noted X_1) in the reaction R_1 with stoichiometry $k_1 S_1 \rightarrow X_1 + k_2 S_2 + k_4 CO_2$;

Table A.1: Stoichiometry of AM2 reactions

Parameter	Unit	Value
k_1	g/g	42.14
k_2	mmol/g	116.5
k_3	mmol/g	268.0
k_4	mmol/g	50.6
k_5	mmol/g	343.6
k_6	mmol/g	453.0

- The volatile fatty acids are transformed by the methanogenic bacteria (noted X_2) into carbonic gas and methane in the reaction R_2 with stoichiometry $k_3 S_2 \rightarrow X_2 + k_5 \mathrm{CO}_2 + k_6 \mathrm{CH}_4$.

The stoichiometry of the reactions are considered known and fixed to the values given in table A.1.

Kinetics

The rates of the reactions are respectively noted r_i (with $i = 1, 2$) and can be computed from the growth rate μ_i through $r_i = \mu_i X_i$. The growth rates are obtained by combining the maximum growth rate ($\mu_{i,\max}$) and an inhibition factor - Monod for r_1, Haldane for r_2 -, resulting in

$$\mu_1 = \mu_{1,\max} \frac{S_1}{S_1 + K_{S1}},$$
$$\mu_2 = \mu_{2,\max} \frac{S_2}{S_2 + K_{S2} + \frac{S_2^2}{K_{I2}}}.$$

In all, five parameters impact the rates of the biochemical reactions involved in AM2: the maximum growth rate $\mu_{1,\max}$ and $\mu_{2,\max}$, and the inhibition parameters K_{S1}, K_{S2} and K_{I2}.

Beyond the biochemical reactions, the concentration of the different compounds are impacted by the dilution rate D (volume of the digester which is replaced by fresh intrant during a given unit of time), and by the characteristics of the intrant. The intrant is assumed to contain a fraction of organic substrate (concentration S_{1in}), a fraction of volatile fatty acids (concentration S_{2in}), a fraction of alkaline content (concentration Z_{in}), and a fraction of inorganic carbon (concentration C_{in}).

A.1.3 Model equations

The evolution of the state variables for the biomass, substrate, VFA and alkalinity are governed by the following equations :

$$\begin{cases} \frac{dX_1}{dt} &= (\mu_1(\xi) - \alpha D)\, X_1 \\ \frac{dX_2}{dt} &= (\mu_2(\xi) - \alpha D)\, X_2 \\ \frac{dS_1}{dt} &= D\,(S_{1in} - S_1) - k_1\mu_1(\xi)X_1 \\ \frac{dS_2}{dt} &= D\,(S_{2in} - S_2) + k_2\mu_1(\xi)X_1 - k_3\mu_2(\xi)X_2 \\ \frac{dZ}{dt} &= D\,(Z_{in} - Z) \end{cases}$$

For the inorganic carbon, the equation is slightly more involved, due to the necessity to estimate the flow of dissolved carbonic gas into gas. Noting q_C the flow of carbonic gas, the evolution of inorganic carbon follows the equation

$$\frac{dC}{dt} = D\,(C_{in} - C) - q_C(\xi) + k_4\mu_1(\xi)X_1 + k_5\mu_2(\xi)X_2.$$

Gas flows

The gas flows are computed using the following fashion. For methane, it is assumed that no methane is dissolved in the liquid phase, resulting in $q_{\mathrm{M}} = k_6\mu_2 X_2$. For carbonic gas, the following approach is used:

- First, the speciation of the inorganic carbon between carbonic gas and bicarbonate is obtained from the pH and the equilibrium constant K_b, resulting in $\mathrm{CO_2} = \frac{C}{1+10^{\mathrm{pH}-\mathrm{pK}_b}}$ (see equation 3 in Bernard et al. [2001]);
- Then, we compute the intermediary value φ following equation 16 of Bernard et al. [2001], $\phi = \mathrm{CO_2} + \mathrm{K_H} \times \mathrm{P}_T + \frac{q_M}{\mathrm{k_L a}}$ where P_T is the atmospheric pressure, $\mathrm{K_H}$ is Henry's constant and $\mathrm{k_L a}$ is the mass transfer coefficient (per day);
- From ϕ, we deduce the partial pressure of $\mathrm{CO_2}$, $P_C = \frac{\phi - \sqrt{\phi^2 - 4\mathrm{K_H}P_T(C+S_2-Z)}}{2\mathrm{K_H}}$ (where P_T is the atmospheric pressure);
- Finally, the gas flow q_{C} is obtained from the partial pressure using Henry's law, $q_{\mathrm{C}} = k_{La}(\mathrm{CO_2} - \mathrm{K_H}P_C)$.

Note that this strategy assumes that reliable pH measurements are available, and differs from the original description in Bernard et al. [2001].

A.2 Details on Anaerobic Digestion Model 1

The Anaerobic Digestion Model 1 (ADM1) was introduced in Batstone et al. [2002a]. The description of the model presented here is based on the technical report Batstone et al. [2002b]

as well as on the modifications proposed by Rosén and Jeppsson [2006], which forms the base of most implementations of ADM1.

A.2.1 State variables

ADM1 involves 28 state variables[1]:

- the concentration of composite substrate, X_c
- the concentration of carbohydrates, proteins and lipids (resp. X_{ch}, X_{pr}, X_{li}),
- the concentration of particulate inerts X_I,
- the concentration of soluble inerts S_I,
- the concentration of sugars, amino acids, and long chain fatty acids (S_{su}, S_{aa}, S_{fa})
- Volatile fatty acids, i.e. valeriate, butyrate, propionate, acetate (resp S_{va}, S_{bu}, S_{pro}, S_{ac})
- Hydrogen S_{h2}
- Inorganic carbon S_{IC}
- Inorganic nitrogen S_{IN}
- Biomass feeding on sugars, amino acids, long chain fatty acids, volatile fatty acids, hydrogen (X_{su}, X_{aa}, X_{fa}, X_{c4}, X_{pro}, X_{ac}, X_{h2}),
- concentration of other cation S_{cat}
- concentration of other anion S_{an}
- Gas phase component for hydrogen, methane and carbonic gas ($S_{gas,\,h2}$, $S_{gas,\,ch4}$, $S_{gas,\,co2}$)[2].

Concentrations are given in a COD basis (i.e. kgCOD/m3), except for the concentration of inorganic carbon, inorganic nitrogen, cations and anions. The concentration of inorganic carbon is in kmole of carbon per m3, the concentration of inorganic nitrogen in kmole of azote per m3, and cations and anions in kmole per m3.

Note that some implementations of ADM1 consider other state variables, such as the concentration of ionic form of valeriate, butyrate, propionate, acetate, bicarbonate (S_{va^-}, S_{bu^-}, S_{pro^-}, S_{ac^-}, S_{hco3^-}) and the concentration of free ammonia S_{nh3}. Here, these values will be inferred from the total concentration and the pH, assuming that the acid/base equilibrium is instantaneous (see appendix A.2.5 for pH computation).

[1] Note that contrary to AM2, the notation X is used for particular content (rather than reserved for biomass) and S is used for soluble content (rather than substrate)

[2] These last 3 state variable were added in Rosén and Jeppsson [2006]

A.2.2 Biochemical process

Reactions

The biochemical process described in ADM1 follows 5 stages:

- In the disintegration stage (reaction 1), the composite substrate X_{c} is decomposed into carbohydrates, proteins, lipids, as well as inert particulate (X_{I}) and soluble particulate (S_{I});
- In the hydrolysis stage, carbohydrates are decomposed into sugars (reaction 2), proteins into amino acids (reaction 3), and lipids into sugars and long chain fatty acids (reaction 4);
- In the acidogenesis stage, micro organisms transform sugars (reaction 5), amino acids (reaction 6) and long chain fatty acids (reaction 7) into volatile fatty acids and hydrogen. Three reactions are considered, each catalyzed by a specific community of micro organism (X_{su}, X_{aa}, X_{fa}).
- In the acetogenesis step, micro organisms transform valerate (reaction 8), butyrate (reaction 9) and propionate (reaction 10) into smaller chain volaile fatty acids, hydrogen and inorganic nitrogen. The same micro organism community (X_{c4}) is responsible for the decomposition of valerate and butyrate, while the decomposition of propionate is performed by a specific microbial community (X_{pro}).
- In the methanogenesis step, acetate (reaction 11) and hydrogen (reaction 12) are transformed into methane, inorganic carbon, inorganic nitrogen by two distinct microbial community (X_{ac} and X_{h2}).

Finally, the biomass for the different microbial community decay; the decayed part of the biomass is turned into the composite substrate (reactions 13 to 19). Overall, the stoichiometry for the reactions is given in table 3.2 from Batstone et al. [2002b]. Note that contrary to AM2, the reactions catalysed by micro organisms have a stoichiometry normalised on the substrate and not on the biomass, i.e. they are of form $S \mapsto YX + \cdots$ rather than $kS \mapsto X + \cdots$. The stoichiometry coefficient Y is called the yield (in g COD of biomass per g COD of substrate), and can be calibrated.

Kinetics

The kinetic rates of the reactions (noted ρ) are described in section 3.5.1 of Rosén and Jeppsson [2006], and follow the generic form $\rho_x = k_x \times I_x \times X_x$. k_x is the maximum rate of the reaction (in d^{-1}). Note that our implementation of ADM1 considers a single rate for decay reactions, k_{dec}, while the description of Rosén and Jeppsson [2006] considers 7 distinct decay rates. I_x is the total inhibition, and X is the particulate component - either the biomass catalysing the reaction (e.g. X_{pro}, X_{h2}, X_{ac}) or the particular compound which is being decomposed (e.g. X_{c}, X_{ch}, X_{pr}). The inhibition factor is omitted (equivalent to $I = 1$) for the disintegration, hydrolysis and decay steps.

Inhibition factors

The following inhibition factors are considered.

Monod inhibition Monod inhibition $I_{\text{Monod}} = \frac{S}{S+K_S}$ is a form of substrate inhibition. The parameter K_S is called the half saturation constant, as at $S = K_S$, $I_{\text{Monod}} = \frac{1}{2}$. Monod inhibition is considered for all biochemical reactions catalysed by a microbial community, i.e. reactions 5 to 12. Note that for reactions 8 and 9 (degradation of valerate and butyrate), which are catalysed by the same microbial community X_{c4}, the Monod half saturation constant is shared.

pH inhibition The pH impact the reactions rates through pH inhibition. The exact form of pH inhibition varies between implementations (see discussion in [Rosén and Jeppsson, 2006], section 3.5.2). The implementation of ADM1 in anaerodig uses the same form of pH inhibition as described in Rosén and Jeppsson [2006], i.e.

$$I_{\text{pH}} = \frac{K_{\text{pH}}^n}{S_{\text{h}^+}^n + K_{\text{pH}}^n} \tag{A.1}$$

with n and K_{pH} obtained from the lower pH limit and upper pH limit parameters pH_{LL}, pH_{UL} as

$$\begin{cases} K_{\text{pH}} = 10^{-\frac{\text{pH}_{\text{LL}}+\text{pH}_{\text{UL}}}{2}} \\ n = \frac{3}{\text{pH}_{\text{UL}}-\text{pH}_{\text{LL}}} \end{cases} \tag{A.2}$$

Three sources of pH inhibition are considered: amino acids ($I_{\text{pH, aa}}$), acetate ($I_{\text{pH, ac}}$) and hydrogen ($I_{\text{pH, h2}}$), each with 2 distinct parameters pH_{LL}, pH_{UL}. These pH inhibitions affect various reactions. Amino acids pH inhibition affects reactions 5 to 10; acetate pH inhibition affects reaction 11; hydrogen inhibition affects reaction 12.

Inorganic nitrogen inhibition The concentration of inorganic nitrogen inhibits reactions 5 to 12 in a Monod like way, $I_{\text{IN, lim}} = \frac{S_{\text{IN}}}{K_{\text{IN}}+S_{\text{IN}}}$. This prevents reactions from going forward in total absence of inorganic nitrogen. Note that the same half saturation constant is considered for all reactions.

Free ammonia inhibition The acetoclastic methanogenesis reaction (reaction 11) is inhibited when the free ammonia (NH3) concentration is too high. The inhibition function takes the form $I_{\text{nh3}} = \frac{K_{\text{I, nh3}}}{K_{\text{I,nh3}}+S_{\text{nh3}}}$.

Hydrogen inhibition Due to thermodynamic constraints, acetogenesis and hydrogenotrophic methanogenesis can only occur in a narrow range of hydrogen concentration (see Batstone et al. [2002b], section 3.5). This is taken into account in ADM1 through an hydrogen inhibition

factor. This affects the acetogenic reactions (reactions 7, 8, 9 and 10), through

$$I_{\mathsf{h2},\,x} = \frac{K_{\mathsf{I,\,h2},\,x}}{K_{\mathsf{I,\,h2},\,x} + S_{\mathsf{h2}}} \tag{A.3}$$

with $x \in \{\mathsf{fa}, \mathsf{c4}, \mathsf{pro}\}$[3].

Competition between valerate and butyrate The degradation of valerate (reaction 8) and butyrate (reaction 9) are catalysed by the same microbial community X_{c4}. A competitive uptake competition factor is considered for both reactions, with

$$\begin{cases} I_{\mathsf{compet,\,va}} &= \frac{S_{\mathsf{va}}}{S_{\mathsf{va}}+S_{\mathsf{bu}}} \\ I_{\mathsf{compet,\,bu}} &= \frac{S_{\mathsf{bu}}}{S_{\mathsf{va}}+S_{\mathsf{bu}}} \end{cases} \tag{A.4}$$

A.2.3 Liquid gas flow

ADM1 considers distinct liquid and gas phases for hydrogen, methane and carbonic gas. The gas transfer rates are

$$\begin{aligned} \rho_{T,8} &= \mathsf{k_L a}(S_{\mathsf{h2}} - 16K_{\mathsf{H,h2}}p_{\mathsf{gas,\,h2}}) \\ \rho_{T,9} &= \mathsf{k_L a}(S_{\mathsf{ch4}} - 64K_{\mathsf{H,ch4}}p_{\mathsf{gas,\,ch4}}) \\ \rho_{T,10} &= \mathsf{k_L a}(S_{\mathsf{co2}} - K_{\mathsf{H,co2}}p_{\mathsf{gas,\,co2}}). \end{aligned}$$

Henry's constants K_{H} depend on the temperatures through

$$\begin{aligned} K_{\mathsf{H,\,co2}} &= 10^{-1.46}\exp\left(\frac{194.1}{R}\left(T_{base}^{-1} - T^{-1}\right)\right) \\ K_{\mathsf{H,\,ch4}} &= 10^{-1.46}\exp\left(\frac{142.4}{R}\left(T_{base}^{-1} - T^{-1}\right)\right) \\ K_{\mathsf{H,\,h2}} &= 10^{-3.11}\exp\left(\frac{41.8}{R}\left(T_{base}^{-1} - T^{-1}\right)\right) \end{aligned}$$

The concentration of carbonic gas can be obtained from the concentration of inorganic carbon S_{IC} through $S_{\mathsf{co2}} = S_{\mathsf{IC}} - S_{\mathsf{hco3^-}}$. The partial pressure of each gas is obtained from the ideal gas law (see section 3.5.4 in Rosén and Jeppsson [2006]), resulting in

$$\begin{aligned} p_{\mathsf{gas,\,h2}} &= S_{\mathsf{gas,\,h2}}\frac{RT}{16} \\ p_{\mathsf{gas,\,ch4}} &= S_{\mathsf{gas,\,ch4}}\frac{RT}{64} \\ p_{\mathsf{gas,\,co2}} &= S_{\mathsf{gas,\,co2}}RT. \end{aligned}$$

[3]This follows the description of Rosén and Jeppsson [2006]; the original description of Batstone et al. [2002b] considers a shared hydrogen inhibition term.

A.2.4 Acid Base equilibrium

ADM1 considers that the acid base reactions occur at a much faster rate than the rest of the reactions. The initial description of ADM1 of Batstone et al. [2002b] proposed to use a reaction rate at least ten times larger than the remaining rates. Rosén and Jeppsson [2006] proposed to consider the acid base equilibrium as instantaneously reached. As such, the concentration of the ionic form of VFAs, bicarbonate and free ammonia concentration are computed as

$$
\begin{aligned}
S_{\mathsf{va}^-} &= \frac{K_{\mathsf{a,\,va}} S_{\mathsf{va}}}{K_{\mathsf{a,\,va}} + S_{\mathsf{h}^+}} \\
S_{\mathsf{bu}^-} &= \frac{K_{\mathsf{a,\,bu}} S_{\mathsf{bu}}}{K_{\mathsf{a,\,bu}} + S_{\mathsf{h}^+}} \\
S_{\mathsf{pro}^-} &= \frac{K_{\mathsf{a,\,pro}} S_{\mathsf{va}}}{K_{\mathsf{a,\,pro}} + S_{\mathsf{h}^+}} \\
S_{\mathsf{ac}^-} &= \frac{K_{\mathsf{a,\,ac}} S_{\mathsf{ac}}}{K_{\mathsf{a,\,ac}} + S_{\mathsf{h}^+}} \\
S_{\mathsf{va}^-} &= \frac{K_{\mathsf{a,\,va}} S_{\mathsf{va}}}{K_{\mathsf{a,\,va}} + S_{\mathsf{h}^+}} \\
S_{\mathsf{hco3}^{-1}} &= \frac{K_{\mathsf{a,\,co2}} S_{\mathsf{IC}}}{K_{\mathsf{a,\,co2}} + S_{\mathsf{h}^+}} \\
S_{\mathsf{nh3}} &= \frac{K_{\mathsf{a,\,IN}} S_{\mathsf{IN}}}{K_{\mathsf{a,\,IN}} + S_{\mathsf{h}^+}}
\end{aligned}
$$

The equilibrium constants K_a are considered constant for VFAs, and to depend on the temperature for inorganic nitrogen and carbonic gas through equations

$$
\begin{aligned}
K_{\mathsf{a,\,co2}} &= 10^{-6.35} \exp\left(\frac{76.46}{R}\left(T_{base}^{-1} - T^{-1}\right)\right) \\
K_{\mathsf{a,\,IN}} &= 10^{-9.25} \exp\left(\frac{519.65}{R}\left(T_{base}^{-1} - T^{-1}\right)\right).
\end{aligned}
$$

A.2.5 Hydrogen and pH estimation

The computation of inhibition factor requires the knowledge of hydrogen concentration and pH concentration. The dynamics for hydrogen and pH concentration are much faster than the dynamics of other reactions; thus modelling their evolution through an Ordinary Differential Equation system would require small time steps, resulting in a sharp increase in computational complexity (see discussion in Rosén and Jeppsson [2006]). An alternative approach consists in estimating the hydrogen concentration and pH value through a DAE system.

The evolution of the concentration of proton S_{h^+} is obtained by considering the charge bal-

ance equations,

$$E(S_{\mathsf{h}^+}) = S_{\mathsf{cat}} + S_{\mathsf{nh4}^+} + S_{\mathsf{h}^+} - S_{\mathsf{hco3}^-} - \frac{S_{\mathsf{ac}^-}}{64} - \frac{S_{\mathsf{pro}^-}}{112} - \frac{S_{\mathsf{bu}^-}}{160} - \frac{S_{\mathsf{va}^-}}{208} - \frac{K_W}{S_{\mathsf{h}^+}} - S_{\mathsf{an}}. \quad \text{(A.5)}$$

This charge balance remains constant at 0. The pH is deduced from the concentration of proton, obtained by solving $E(S_{\mathsf{h}^+}) = 0$. The fraction of compounds which are in ionic form is obtained through acid base equilibrium, and is therefore a function of the pH. As a result, the equation does not admit a closed form solution, and hence the solution must be estimated using *e.g.* a Newton Solver.

Similarly, Rosén and Jeppsson [2006] proposes to use a DAE solver for hydrogen concentration, this time considering the mass balance constraint:

$$\begin{aligned} E(S_{\mathsf{h2}}) =& \frac{Q}{V_{\mathsf{liq}}}(S_{\mathsf{h2,\,in}} - S_{\mathsf{h2}}) + (1 - Y_{\mathsf{su}}) f_{\mathsf{h2,\,su}}\, \rho_5 + (1 - Y_{\mathsf{aa}}) f_{\mathsf{h2,\,aa}}\, \rho_6 + (1 - Y_{\mathsf{fa}}) 0.3\, \rho_7 \\ &+ (1 - Y_{\mathsf{c4}}) 0.15\, \rho_8 + (1 - Y_{\mathsf{c4}}) 0.2\, \rho_9 + (1 - Y_{\mathsf{pro}}) 0.43\, \rho_{10} - \rho_{12} - \rho_{T,8} \end{aligned} \quad \text{(A.6)}$$

This quantity depends on S_{h2} both directly and indirectly through the hydrogen inhibition factor in the reaction rates.

A.2.6 Differential equations

Liquid phase

For particular or soluble compounds A, the concentration of the intrant is denoted A_{in}. The total flow of intrant is denoted Q, in cubic meter per day.

We denote M the stoichiometry matrix defined in Batstone et al. [2002b], table 3.2 (where columns denote compounds and rows equation). Assuming that the liquid phase volume remains unchanged, the evolution of the i-th compound A_i is given as

$$\frac{\mathsf{d}A_i}{\mathsf{d}t} = \frac{Q}{V}(A_{i,\mathsf{in}} - A_i) + \sum M_{j,i}\rho_j. \quad \text{(A.7)}$$

Note that the stoichiometry matrix M is sparse, with the evolution of compound A_i typically being impacted by a few reactions only.

Gas phase

The flow of gas leaving the digester is obtained through the difference between the gas pressure and atmospheric pressure as

$$Q_{\mathsf{gas}} = k_p(p_{\mathsf{gas}} - p_{\mathsf{atm}})\frac{p_{\mathsf{gas}}}{p_{\mathsf{atm}}} \quad \text{(A.8)}$$

where the gas pressure p_{gas} can be computed as the sum of partial pressure from methane, carbonic gas, hydrogen, and water, and k_p is a parameter (in m3/day/bar). The partial pressure of water, $p_{\text{gas, h2o}}$, is modelled as a known function of the temperature T (in Kelvin), $p_{\text{gas, h2o}} = 0.0313 \exp\left(5290\left(T_{base}^{-1} - T^{-1}\right)\right)$.

Overall, the evolution of the gas content is modelled as

$$\begin{aligned}
\frac{\mathrm{d}S_{\text{gas, h2}}}{\mathrm{d}t} &= -\frac{S_{\text{gas, h2}}Q_{\text{gas}}}{V_{\text{gas}}} + \rho_{T,8}\frac{V_{\text{liq}}}{V_{\text{gas}}} \\
\frac{\mathrm{d}S_{\text{gas, ch4}}}{\mathrm{d}t} &= -\frac{S_{\text{gas, ch4}}Q_{\text{gas}}}{V_{\text{gas}}} + \rho_{T,9}\frac{V_{\text{liq}}}{V_{\text{gas}}} \\
\frac{\mathrm{d}S_{\text{gas, co2}}}{\mathrm{d}t} &= -\frac{S_{\text{gas, co2}}Q_{\text{gas}}}{V_{\text{gas}}} + \rho_{T,10}\frac{V_{\text{liq}}}{V_{\text{gas}}}.
\end{aligned}$$

Appendix B

Technical complements on PAC-Bayes

B.1 Proofs of change of measures inequalities

The proof for each bound follows the same pattern: for each f-divergence, compute f^*, check whether $(1/f'')$ is concave, then apply accordingly either eq. (2.16) or eq. (2.15) to obtain:

$$\pi\left[D\right] \leq c + \lambda \mathsf{B}(\lambda^{-1}(D-c)) + \lambda \mathcal{D}_f(\pi, \pi_{\mathsf{p}}).$$

Then optimise on λ and c whenever feasible. We therefore sum up the proofs in Table B.1 which details the form of the f^* as well as the optimal value of λ when computable.

The Kullback–Leibler, power divergence for $1 < p \leq 2$ and Pearson χ^2 satisfy $1/f''$ concave, and we therefore use Equation (2.14). All the other bounds use Equation (2.10). For the total variation, it is simple to see that to get non trivial bounds, we need to pick $c \leq \frac{1}{2} - D_{\max}$. Diminishing c to $c - \delta c$ decreases the integral by at most δc (the threshold can only dampen the decrease), while the other term increases by δc. This implies that $c^* = \frac{1}{2} - D_{\max}$.

For the Vincze–Le Cam divergence, we are in the situation described in Remark 2.11. The upper bound obtained through $\tilde{f}$ is much more tractable than the one obtained through f, and in particular, it can be optimised on the scale parameter λ. It is this bound through $\tilde{f}$ which we use to obtain the final bound. Using the Legendre transform of $f_{|\mathbb{R}_+}$ yields this tighter, though less tractable, inequality for all $c \geq 0$, $\lambda > 0$

$$\begin{aligned}\pi\left[D\right] \leq c + D_{\max} + \pi_{\mathsf{p}}\left[\mathbb{1}\left[4\lambda - c \geq D^-\right]\left(4\lambda - 4\sqrt{\lambda(c+D^-)} + (c+D^-)\right)\right] \\ + \lambda\left(\mathsf{VC}(\pi, \pi_{\mathsf{p}}) - 2\right).\end{aligned}$$

where $D^- = D_{\max} - D$.

Table B.1: f^*, λ^* optimising change of measure inequalities

f-div	$f^*(t) = \dots$	λ^*
KL	$\exp(t-1)$	
Power-p, $1 < p \leq 2$	$\frac{p^{1-q}}{q}\max(t,0)^q + 1$	$p\left(\frac{\mathcal{D}_{f_p}(\pi,\pi_{\mathsf{p}})}{\pi_{\mathsf{p}}\left[D_+^q\right]-\pi\left[D_+^{\frac{q}{p}}\right]^p}\right)^{\frac{1}{q}}$
Power-p, $2 < p$	$\frac{p^{1-q}}{q}\max(t,0)^q + 1$	$p\left(\frac{1+\mathcal{D}_{f_p}(\pi,\pi_{\mathsf{p}})}{\pi_{\mathsf{p}}[h^q]}\right)^{\frac{1}{q}}$
Pearson χ^2	$\sqrt{2}\max(t,0)^2 + 1$	$2\left(\frac{\mathcal{D}_{f_2}(\pi,\pi_{\mathsf{p}})}{\mathbf{Var}_\pi[D_+]}\right)^{\frac{1}{2}}$
Power-p, $0 < p < 1$	$\begin{cases} -1 + \frac{p^{1-q}}{-q}(-t)^q & t < 0 \\ +\infty & \text{else} \end{cases}$	$p\left(\frac{1-\mathcal{D}_{f_p}(\pi,\pi_{\mathsf{p}})}{\pi_{\mathsf{p}}[(-D)^q]}\right)^{\frac{1}{q}}$
Squared Hellinger	$\begin{cases} -1 + \frac{1}{4}(-t)^{-1} & t < 0 \\ +\infty & t \geq 0 \end{cases}$	$\frac{\pi\left[(-D)^{-1}\right]}{2-2H^2(\pi,\pi_{\mathsf{p}})}$
Power-p, $p < 0$	$\begin{cases} 1 - \frac{(-p)^{1-q}}{q}(-t)^q & t \leq 0 \\ +\infty & \text{else} \end{cases}$	$(-p)\left(\frac{1+\mathcal{D}_{f_p}(\pi,\pi_{\mathsf{p}})}{\pi_{\mathsf{p}}[(-D)^q]}\right)^{\frac{1}{q}}$
Reverse Pearson	$\begin{cases} 1 - 2(-t)^{1/2} & t \leq 0 \\ +\infty & \text{else} \end{cases}$	$\left(\frac{1+\chi^2(\pi_{\mathsf{p}},\pi)}{\pi\left[\sqrt{(-D)}\right]}\right)^2$
Total Variation	$\begin{cases} -\frac{1}{2} & t < -\frac{1}{2} \\ t & -\frac{1}{2} \leq t \leq \frac{1}{2} \\ +\infty & t > \frac{1}{2} \end{cases}$	
Reverse KL	$\begin{cases} -(1 + \log(-t) & t < 0 \\ +\infty & \text{else} \end{cases}$	$\exp\left(\mathsf{KL}(\pi,\nu) - \pi\left[\log(-D)\right]\right)$
Lin's measure, $0 < \theta < 1$	$(1-\theta)\log\left(\frac{1-\theta}{1-\theta\exp\left(t\theta^{-1}\right)}\right)$	
Jensen–Shannon	$-\frac{1}{2}\log\left(2 - e^{2t}\right)$	
Vincze–Le Cam	$\begin{cases} -2 & t \leq -4 \\ -4\sqrt{-t} - t + 2 & -4 \leq t \leq 0 \\ +\infty & \text{else} \end{cases}$	$\left(\frac{1+\frac{1}{2}\mathsf{VC}(\pi,\pi_{\mathsf{p}})}{\pi_{\mathsf{p}}[\sqrt{-D}]}\right)^2$
$e^{t-1} - 1$	$\begin{cases} -1/e + 1 & t \leq \frac{1}{e} \\ t\log(t) + 1 & \geq \frac{1}{e} \end{cases}$	

Legendre transform of Pearson χ^2 divergence

Theorem B.1

For a generalised generalisation gap D satisfying $\pi_{\mathsf{p}}[|D|] < \infty$, the Legendre transform of Pearson χ^2 divergence can be obtained from Equation (2.11), *i.e.*

$$\overline{\chi^2}^*(D) = \frac{1}{2}\pi_{\mathsf{p}}\left[(D-c)_+^2 + 1\right] + c. \tag{B.1}$$

Moreover, if D satisfies $D > 0$, $\pi_{\mathsf{p}}[D] \leq 1$, the Legendre transform of Pearson χ^2 divergence is

$$\overline{\chi^2}^*(D) = \frac{1}{2}\mathbb{V}_{\pi_{\mathsf{p}}}[D] + \pi_{\mathsf{p}}[D]. \tag{B.2}$$

Proof. Let us prove the first statement of the theorem. The function $f^*(x) = \frac{x_+^2+1}{2}$ is differentiable, with derivative $f^{*\prime}(x) = x_+$. Since $\pi_{\mathsf{p}}[|D|] < \infty$, this implies that $\pi_{\mathsf{p}}\left[f^{*\prime}(D)\right] < \infty$ and, considering the form of $f^{*\prime}$, that $1 \leq \pi_{\mathsf{p}}\left[f^{*\prime}(D+1)\right] < \infty$. Hence the second part of Theorem 2.1 holds.

Then if $D > 0$ and $\pi_{\mathsf{p}}[D] \leq 1$, it follows that $c^* = \pi_{\mathsf{p}}[D] - 1 \leq 0$ is such that $D - c^* > 0$ and hence $\pi_{\mathsf{p}}\left[f^{*\prime}(D-c^*)\right] = 1$. Hence c^* minimises the bound. Evaluating the bound for c^* finishes the proof. □

B.2 Test generalisation bound

A test bound is a confidence interval for the mean of the risk constructed from a sample of risk values $r_1, \cdots, r_n$. We assume that the risk values are evaluated on n-i.i.d. new datapoints, independent from the data used to train the predictor. Test bounds can be inferred from some concentration inequalities in the following way.

Lemma B.1

Let $\mathcal{P}$ denote a subset of probability measures on $\mathbb{R}_+$. Consider a probability $\mathbb{P} \in \mathcal{P}$, denote $\bar{r}$ its mean. Consider n -i.i.d. draws $r_1, \ldots, r_n$ from $\mathbb{P}$ and denote $\bar{r}$ the empirical average $\bar{r} := \frac{1}{n}\sum_{i=1}^n r_i$. Consider a concentration inequality valid for all $\mathbb{P} \in \mathcal{P}^{\otimes n}$ stating that

$$\mathbb{P}[\bar{r}_n \leq \bar{r} - t] \leq \gamma(t, \bar{r}, n, \mathcal{P}).$$

Define $\hat{r}_+(\delta, n) = C_\gamma(\bar{r}_n, n, \delta, \mathcal{P}) := \sup\{m, \gamma(m - \bar{r}_n, m, n, \mathcal{P}) \geq \delta\}$. Then $[0, \hat{r}_+(\delta, n)]$ is a confidence interval of level at least $1 - \delta$.

Proof. Consider the function

$$F(p, n, \delta, \mathcal{P}) = \max C(p, n, \delta) := \{\bar{r} \mid \exists \mathbb{P} \in \mathcal{P} \text{ s.t. } \mathbb{P}[\mathsf{Id}] = \bar{r},\ \ \mathbb{P}^n[\bar{r} \leq p] \geq \delta\}.$$

By definition of F, $[0, F(\overline{r}_n, n, \delta, \mathcal{P})$ is a confidence interval of level $1-\delta$. Hence if $F(p, n, \delta, \mathcal{P}) \leq C_\gamma(p, n, \delta, \mathcal{P})$, then $[0, C_\gamma(\overline{X}, n, \alpha)]$ is a confidence interval of level at least $1-\delta$.

To prove the inequality, consider a sequence $(m_i)_{i\in\mathbb{N}}$ such that

- $p \leq m_i$,
- $\forall i,\ m_i \in C(p, n, \delta)$,
- $m_i \to F(p, n, \delta, \mathcal{P})$.

Since $F(p, n, \delta, \mathcal{P}) \geq p$, such a sequence always exists. Then for all i, there exists $\mu_i \in \mathcal{P}$ with mean m_i such that $\mathbb{P}^n[\overline{r}_n \leq m_i + p - m_i] \geq \delta$. Moreover, using the concentration inequality with $t = p - m_i$, it follows that $\gamma(m_i - p, m_i, n, \mathcal{P}) \geq \mathbb{P}^n[\overline{r}_n \leq m_i + p - m_i]$. Hence $\gamma(m_i - p, m_i, n, \mathcal{P}) \geq \delta$ for all i. By definition of $C_\gamma(r, b, \delta, \mathcal{P})$, this implies the inequality. □

The quantity $F(p, n, \delta, \mathcal{P})$ introduced in the bound is the tightest upper bound on the average which can be inferred using the empirical average; the proof consists of showing that any concentration inequality results in a looser bound. The best concentration inequality of the form studied here should achieve this bound.

We will now consider the setting of bounded risks ($\mathcal{P}$ are probability measures on $[0, 1]$), and use an extended version of Hoeffding's lemma to obtain a test bound. This extended version was in fact also established in Hoeffding's original paper [Hoeffding, 1963], and improves on the bound by considering the variance. This has a large impact when the empirical mean is small, since bounded random variables with average close to the bounds must have small variance. Notably, this results in test bounds with rate $1/n$ when the empirical mean is 0 (compared to $1/\sqrt{n}$ for the classic Hoeffding's bound).

Lemma B.2: Lemma 1 in Hoeffding [1963])

Let X be a random variable distributed from $\mathbb{P}$ taking value between a and $b > a$. Then $\forall \lambda$,

$$\mathbb{E}\left[\exp(\lambda X)\right] \leq \Psi_\lambda\left(a + (b-a)\mathcal{B}\left(\frac{\mathbb{E}\left(X\right) - a}{b-a}\right)\right) \tag{B.3}$$

where Ψ_λ is the moment generating function evaluated at λ, i.e.

$$\Psi_\lambda\left(a + (b-a)\mathcal{B}\left(\frac{\mathbb{E}\left(X\right) - a}{b-a}\right)\right) = \frac{\mathbb{E}[X] - a}{b-a}\exp(\lambda a) + \frac{b - \mathbb{E}[X]}{b-a}\exp(\lambda b) \tag{B.4}$$

Using the Chernov bound strategy used in the original Hoeffding's theorem and replacing Hoeffding's lemma with this improved version yields the following concentration inequality.

Corollary B.1

For $\mathbb{P}$ a measure on $[0,1]$ of mean $\bar{r}$, noting $\bar{r}_n$ the empirical mean for n i.i.d. observations, it follows that

$$\begin{aligned}\mathbb{P}^n\left[\bar{r}_n \leq \bar{r} - t\right] &\leq \inf_{\lambda} \exp\left(-n\lambda t + n\Psi_\lambda\left(\mathcal{B}\left(\bar{r}\right)\right)\right) \\ &= \exp\left(-n\left(\left(\bar{r}-t\right)\log\left(\frac{\bar{r}-t}{\bar{r}}\right) + \left(1-\bar{r}+t\right)\log\left(\frac{1-\bar{r}+t}{1-\bar{r}}\right)\right)\right).\end{aligned}$$

The test bounds proposed in table Table 2.2 in Section 2.2 are computed using the test bounds derived from the concentration inequality of Corollary B.1 using the concentration to test bound algorithm described in Lemma B.1. The inversion is performed using dichotomy. We remark that the slighter tighter test bounds could be obtained in the classification setting by limiting the space $\tilde{P}$ to Binomial distributions and explicitly computing the tightest confidence interval achievable in that setting. This second strategy is pursued to compute upper bounds on the p-value for the bootstrap process in Chapter 3.

Acronyms

AD	**A**naerobic **D**igestion
ADM1	**A**naerobic **D**igestion **M**odel 1
AM2	**A**naerobic Digestion **M**odel 2
BMP	**B**iochemical **M**ethane **Potential**
CI	**C**onfidence **I**nterval
COD	**C**hemical **O**xygen **D**emand
CMA-ES	**C**ovariance **M**atrix **A**daptation **E**volution **S**trategy
DAE	**D**ifferential **A**lgebraic **E**quation
ERM	**E**mpirical **R**isk **M**inimisation
FIM	**F**isher's **I**nformation **M**atrix
GD	**G**radient **D**escent
HRT	**H**ydraulic **R**etention **T**ime
JIT	Just In Time
KL	**K**ullback–**L**eibler divergence
LOPO	**L**eave **O**ne **P**lant **O**ut
MAE	**M**ean **A**bsolute **E**rror
MAML	**M**odel **A**daptative **M**eta **L**earning
MLS	**M**aureer-**L**angford-**S**eeger's bound
ODE	**O**rdinary **D**ifferential **E**quation
PAC	**P**robably **A**pproximately **C**orrect
PDE	**P**artial **D**ifferential **E**quation
PS	**P**rimary **S**ludge
RMSE	**R**oot **M**ean **S**quare **E**rror

SGD	**S**tochastic **G**radient **D**escent
SRT	**S**olid **R**etention **T**ime
TAN	**T**otal **A**mmonia **N**itrogen
TS	**T**otal **S**olids
UQ	**U**ncertainty **Q**uantification
VarBUQ	**Var**iational PAC-Bayes **U**ncertainty **Q**uantification routine
VC	**V**apnik–**C**hervonenkis
VM	**V**irtual **M**achine
VS	**V**olatile **S**olids
VSR	**V**olatile **S**olids **R**eduction
VFA	**V**olatil **F**atty **A**cid
WAS	**W**aste-**A**ctivated **S**ludge
WWTP	**W**aste**W**ater **T**reatment **P**lant

Symbols and notations

$\mathbb{R}$	The set of real numbers
λ^{Leb}	The Lebesgue measure
π_{p}	The prior distribution
π	A posterior distribution
$\hat{\pi}$	The posterior distribution minimizing a PAC-Bayes bound
PB	A PAC-Bayes bound
R	The empirical risk function
$\tilde{R}$	The risk function
KL	The Kullback–Leibler divergence
kl	The Kullback–Leibler divergence between two Bernoulli
$\mathcal{D}_f$	The f-divergence associated to f
θ	A probability measure parameter
Θ	The space of the probability measure parameters
h	A predictor (or *hypothesis*)
$\mathcal{H}$	The space of predictors
γ	A predictor parameter
Γ	The space of predictor parameters
D	A generalized generalisation gap
λ	A PAC-Bayes temperature
ϕ	A meta-parameter
Φ	The space of meta-parameters
AD	An anaerobic digestion model function
$\|\cdot\|$	Norm (euclidean norm by default)
$\|\cdot\|_1$	L^1 norm
$\|\cdot\|_2$	Euclidean norm
$\|\cdot\|_F$	Frobenius norm
$\|\cdot\|$	Absolute value
$\|\cdot\|$	Matrix determinant
Id	Identity matrix (or function)

Titre: Agrégation et méta-modélisation de procédés de méthanisation : théorie et algorithmes

Mots clés: Digestion anaérobie, Méta-apprentissage, PAC-Bayes, Bornes de généralisation, Statistiques Bayesienne, Quantification de l'incertitude

Résumé: Les technologies de digestion anaérobie (DA) jouent un rôle clé dans la transition vers une énergie propre en convertissant les déchets organiques en biogaz, remplaçant le gaz naturel fossile. Pour optimiser les systèmes de DA, une compréhension précise des mécanismes biologiques et physico-chimiques est nécessaire. L'utilisation de modèles biochimiques constitue une étape dans cette direction, mais ces derniers sont sujets au sur-apprentissage et sont souvent limités au processus sur lequel ils sont calibrés, ce qui limite leur pouvoir prédictif.

L'approche PAC-Bayes contrôle intrinsèquement le risque de sur-apprentissage grâce à son utilisation de garanties de généralisation. En tant qu'approche Bayésienne, elle offre de plus la possibilité d'intégrer des connaissances expertes utiles dans le processus d'apprentissage et fournit naturellement une quantification de l'incertitude, bien que cette dernière n'ait plus de garanties théoriques. Cependant, la nature corrélée des données et le coût de calcul élevé des simulations rendent l'application de PAC-Bayes à la DA difficile et nécessitent des adaptations.

Dans son état actuel, les connaissances expertes sont insuffisantes pour construire des modèles utilisables pour de nouveaux digesteurs, pour lesquels aucune donnée n'est disponible. Les stratégies de méta-apprentissage offrent une approche prometteuse pour améliorer ces connaissances expertes, en exploitant les informations provenant de plusieurs unités de DA. En apprenant les mécanismes communs entre différentes unités, le méta-apprentissage pourrait aboutir à des modèles plus robustes, notamment pour les sites pour lesquels peu ou pas de données sont disponibles. Cependant, les algorithmes de méta-apprentissage PAC-Bayes existants sont intensifs sur le plan du temps de calcul et mal adaptés aux longs temps de simulation des processus de DA.

Cette thèse se concentre sur le développement de méthodes efficaces pour appliquer la théorie PAC-Bayes et les approches de méta-apprentissage à la modélisation de la DA. Elle apporte des nouveaux éléments théoriques sur PAC-Bayes, présente de nouvelles implémentations efficaces des stratégies PAC-Bayes et de méta-apprentissage, et fournit des évaluations expérimentales des modèles stochastiques obtenus.

Title: Aggregation and Meta-modelling for anaerobic digestion processes: theory and algorithms

Keywords: Anaerobic digestion, meta-learning, PAC-Bayes, Generalisation bounds, Bayesian statistics, Uncertainty quantification

Abstract: Anaerobic Digestion (AD) technologies play a key role in the transition to cleaner energy by converting organic waste into biogas which can replace fossil natural gas. To optimize AD systems, a thorough understanding of the biological and physicochemical mechanisms is required. Biochemical computational models are a step in this direction but are prone to overfitting and are often specific to the process they are calibrated for, limiting their predictive power.
The PAC-Bayes framework inherently limits the risk of overfitting through its reliance on trainable generalisation guarantees. As an extended Bayesian approach, it moreover provides the opportunity to incorporate useful expert knowledge in the learning process, and inherently provides uncertainty quantification, albeit with no theoretical guarantees. However, the correlated nature of the data and the high computational cost of the simulation make the application of PAC-Bayes to AD challenging and require adaptation.
In its current state, expert knowledge is insufficient to construct meaningful models for new AD units for which no data is available. Meta-learning strategies offer a promising way to enhance this expert knowledge by leveraging information across multiple AD units. By learning common features across different units, meta-learning could result in more robust models, especially for units when few, or no data is available. Previous PAC-Bayes meta-learning algorithms are however quite computationally intensive, and ill-fitted to the high simulation time of AD processes.
This thesis focuses on developing efficient methods for applying PAC-Bayes theory and meta-learning approaches to AD modelling. It contributes theoretical insights into PAC-Bayes, presents efficient implementations of both PAC-Bayes and meta-learning strategies, and provides experimental assessments of the resulting models' predictive power and uncertainty quantification.

Université de Lille
L2EP / IRCICA
50 Avenue Halley, 59650 Villeneuve-d'Ascq, France