Reliable Uncertainty Quantification in Statistical Learning

DAVID WIDMANN
Dissertation presented at Uppsala University to be publicly examined in Häggsalen (10132), Ångströmlaboratoriet, Uppsala, Wednesday, 14 June 2023 at 09:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Dino Sejdinovic (University of Adelaide, School of Computer and Mathematical Sciences).

Abstract


Mathematical models are powerful yet simplified abstractions used to study, explain, and predict the behavior of systems of interest. This thesis is concerned with their latter application as predictive models. Predictions of such models are often inherently uncertain, as exemplified in weather forecasting and experienced with epidemiological models during the COVID-19 pandemic. Missing information, such as incomplete atmospheric data, and the very nature of models as approximations ("all models are wrong") imply that predictions are at most approximately correct.

Probabilistic models alleviate this issue by reporting not a single point prediction ("rain"/"no rain") but a probability distribution of all possible outcomes ("80% probability of rain"), representing the uncertainty of a prediction, with the intention to be able to mark predictions as more or less trustworthy. However, simply reporting a probabilistic prediction does not guarantee that the uncertainty estimates are reliable. Calibrated models ensure that the uncertainty expressed by the predictions is consistent with the prediction task and hence the predictions are neither under- nor overconfident. Calibration is important in particular in safety-critical applications such as medical diagnostics and autonomous driving where it is crucial to be able to distinguish between uncertain and trustworthy predictions. Mathematical models do not necessarily possess this property, and in particular complex machine learning models are susceptible to reporting overconfident predictions.

The main contribution of this thesis are new statistical methods for analyzing the calibration of a model, consisting of calibration measures, their estimators, and statistical hypothesis tests based on them. These methods are presented in the five scientific papers in the second part of the thesis. In the first part the reader is introduced to probabilistic predictive models, the analysis of calibration, and positive definite kernels that form the basis of the proposed calibration measures. The contributed tools for calibration analysis cover in principle any predictive model and are applied specifically to classification models, with an arbitrary number of classes, models for regression problems, and models arising from Bayesian inference. This generality is motivated by the need for more detailed calibration analysis of increasingly complex models nowadays. To simplify the use of the statistical methods, a collection of software packages for calibration analysis written in the Julia programming language is made publicly available and supplemented with interfaces to the Python and R programming languages.

Keywords: Reliability, Calibration, Uncertainty, Probabilistic Model, Prediction, Julia

David Widmann, Department of Information Technology, Division of Systems and Control, Box 337, Uppsala University, SE-75105 Uppsala, Sweden. Department of Information Technology, Artificial Intelligence, Box 337, Uppsala University, SE-75105 Uppsala, Sweden.

© David Widmann 2023

ISSN 1651-6214
URN urn:nbn:se:uu:diva-500736 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-500736)
To Rebekka, my wonderful wife.
Att planera sin framtid är inte lätt
fast lagret av data är stort.
Men det ordnar sig säkert på något sätt
för det har det alltid gjort.

Prognos
ALF HENRIKSON

Sammanfattning

George E. P. Box (1976), en brittisk statistiker, blev känt för sin mening att alla vetenskapliga modeller är fel (“all models are wrong”). Han påpekar att det är ytterst viktigt att komma ihåg att modeller är bara approximationer och förenklingar och inte några exakta kopior av processer och system som studeras. Detta märktes ganska tydligt under COVID-19 pandemin. Många matematiska modeller användes för att förutse hur infektionsläget skulle utvecklas och hur läget skulle påverkas av åtgärder. Det fanns stora skillnader mellan olika modeller – eftersom de baserades på olika antaganden och byggdes för olika behov. Några var mer komplexa (t.ex. den omdiskuterade modellen från Imperial College (Flaxman m. fl. 2020; Soltesz m. fl. 2020)) medan enklare modeller beskrev antalet mottagliga, infekterade och återhämtade personer i ett helt samhälle med bara två parametrar. Skillnader och felaktiga prognoser väckte stora debatter om hur pålitliga matematiska modeller faktiskt är (Gerlee, Thorén och Lundh 2020).

Väderprognoser är ett annat välkänd exempel av förutsägelser som baserar på matematiska modeller. Också i det fallet finns skillnader mellan prognoser av olika vädertjänster, och prognoserna stämmer såklart inte heller alltid som syns i Figur 2.2. Långsiktiga prognoser brukar vara fel ofta än prognoser för samma eller nästa dag (SMHI 2021a). Det finns några anledningar för varför prognoserna inte alltid stämmer (SMHI 2021b). Alla modeller, även komplexa, är bara approximationer av väderdynamiken. Dessutom kan man inte täta tillståndet i atmosfären överallt och därför är även utgångspunkten av alla dynamiska modeller osäker. Långsiktiga prognoser brukar vara mindre pålitliga eftersom fel i modellen kan ackumuleras och växa exponentiellt. Även om osäkerheten kan reduceras med mer komplexa modeller och kortsiktigare prognoser, är prognoser är alltid något osäkra.

Vad kan vi göra åt det här problemet? För att tydligt markera osäkrare väderprognoser, började Cooke (1906b) redan för mer än hundra år sen lägga till en vikt som indikerar osäkerheten i sina prognoser. Då vet man direkt vilka prognoser som är mer och vilka som är mindre pålitliga. Samma princip kan också användas för prognoser av andra matematiska modeller, inte bara för väderprognoser. Oftast anges osäkerheten med en sannolikhet nuförtiden.

Tyvärr garanterar matematiska modeller inte automatiskt att sannolikheten som rapporteras av modellen verkligen stämmer överens med osäkerheten som modellen borde uttrycka. Speciellt komplexa moderna modeller som den populära
språkmodellen ChatGPT är kända för att vara för övertygade om sina prognoser, d.v.s. sannolikheterna är för stora (OpenAI 2023). Det skapar en hel del problem eftersom det ibland är ganska svårt att upptäcka att en modell är för övertygad, även om prognosen är helt fel.


Min avhandling handlar om en aspekt av den här frågan: modellkalibrering. En modell är kalibrerad om rapporterade sannolikheter stämmer överens med verkligheten och den varken har för högt eller för lågt förtroende för sina egna

---

1Jag jämförde ChatGPT:s svar med andra recept som jag hittade i websökning och den verkade korrekt. I alla fall utan suspakta ingredienser.
prognoser. Därför kan man lita på prognoser av en kalibrerad modell och hantera säkra prognoser på ett annat sätt än opålitliga prognoser.

I min forskning fokuserar jag mest på hur man kan bestämma om en modell är kalibrerad eller inte. Eftersom många modeller nuförtiden är ganska komplexa, behövs också mer komplexa metoder för att analysera dem. Det finns ett stort antal metoder för att analysera enkla modeller med bara två möjliga utfall (t.ex. ”regn” och ”inget regn”) som man kan bygga på. Men de är inte tillräckliga om man vill analysera kalibrering av en modell med fler än två utfall i sin helhet och det inte räcker att några delaspekter av modellen är kalibrerade.

Analyser baseras på olika mått av kalibrering, s.k. kalibreringsfel. En kalibrerad modell har ett kalibreringsfel som är noll medan kalibreringsfel för icke-kalibrerade modeller är större än noll. Modeller kan jämföras med avseende på deras kalibreringsfel, där en modell med större fel anses vara mindre kalibrerad. Problemet är dock att olika mått av kalibrering kan resultera i olika rangordningar av modeller. En del av vår forskning handlar om alternativa kalibreringsmått som är byggda för komplexa modeller med flera och även oändligt många möjliga utfall.


Analyser av modellkalibrering hjälper inte direkt om en modell är inte kalibrerad. I så fall behövs metoder för att förbättra modellkalibrering, eller för att ställa upp en ny, bättre kalibrerad modell, som inte diskuterats i avhandlingen. Även i
I cannot tell you how thankful I am for our little infinity.

_The Fault in Our Stars_  
**JOHN GREEN**

# Acknowledgments

First and foremost, I would like to thank my supervisor Fredrik Lindsten for his support and guidance during my doctoral studies. I still remember how we met for the first time in spring 2017, and how unsure I was about whether I should apply for the PhD position you had announced since I had no prior experience in this area and no good feeling for what to expect. I do not know if you are aware of the fact that initially I had applied for another project at the Centre for Interdisciplinary Mathematics (CIM) and we were matched only thanks to Erik Ekström who called me and asked me to consider applying for your project instead. In retrospect, I made the right choice in coming back to Uppsala and starting this adventure together with you. I am grateful that you introduced me to the world of machine learning and statistical inference and I have learned a lot from you. Unfortunately, we had to get accustomed to virtual meetings with all their quirks and technical challenges (usually on my side) already prior to the pandemic. Nevertheless and most importantly, it has been a pleasure working with you and receiving your advice, regarding our research but also more generally regarding personal development and possible career paths.

I would also like to thank my co-supervisors Thomas Schön, Dave Zachariah, and Erik Sjöblom. Thomas, thank you for taking care of the division, forcing me to write in Swedish on Slack, and teaching me the importance of a clean and polished list of references. It is a bit unfortunate that you were replaced as co-supervisor already during the first half of my PhD studies and we only managed to work together on one project. Dave, thank you for agreeing to step in as co-supervisor when Fredrik moved to Linköping, for taking the time to attend many of our meetings, and sharing your advice and knowledge with me. I enjoyed in particular your frequentist perspective. Erik, I am sorry that we did not manage to collaborate on the initially proposed project more closely.

I would also like to thank my co-authors, Juozas, Calle, Jakob, Fredrik, Thomas, Dave, Pierre, and Arthur. The papers included in this thesis would not exist without you, and it would not have been possible to publish them without your sleep-depriving efforts in the days and final hours before the conference submission deadlines. Thank you, Juozas and Calle, for the long and stimulating discussions at the white boards. Thank you, Fredrik, Håkan, and Rebekka, for commenting on the draft of my thesis and helping me to improve it considerably.
I had a great time at the division and I feel sad that it will come to an end now. I would like to thank all my colleagues at SysCon who made me feel at home during my PhD studies. Fredrik G, Hákan, Juozas, Osama, and Paul, I enjoyed sharing an office with you. Thank you, Anna W and Calle A for becoming more than just colleagues. Andreas, Anna W, Calle A, Fredrik O, Koen, Juozas, and Niklas, thank you for the running sessions and motivating me to participate in Ekebyloppet. In addition to the colleagues mentioned above, thank you, Anna F, Antonio, Bengt, Bernhard, Calle J, Daniel, Diana, Dominik, Hans, Helena, Jack, Johan A, Johan W, Jonas, Ludvig, Maria, Pelle, Riccardo, Rubén, Ruoqui, Sofia, Tatiana, Torbjörn, and Viktor, for the interesting, fun, and sometimes controversial discussions during fika breaks, lunch, and after-works.

Thank you, Alex, Anna W, Calle A, Fredrik O, Jelena, Jim, Kristina, Lukas, and Simon, for our Ultimate practices and tournaments—and even SM!—we played with East Aros. Saturday at around noon was not an optimal time, but nevertheless our Ultimate sessions were one of my weekly highlights.

One drawback of moving to Sweden for my PhD studies was that I had to leave behind my friends in Germany. Benedikt, Fabian, Flo, Philipp, Titzi, Tobi, Wasti, and Zeno, you all have shaped my personality quite significantly, probably more than you and I can imagine. I am glad that we still managed to meet from time to time, and I am already looking forward to the next occasion when we can meet and chat, eat and drink, play schafkopf or compete in FIFA.

I would also like to thank my collaborators in the Julia community. It has been fun—and time-consuming—to try to improve the Julia package ecosystem together with you. I have learned a lot from you about software design, collaboration, and Julia internals. Hong, Tor, Cameron, Mohamed, and Philipp, it has been great fun to hack with you on the Turing ecosystem and learn hands-on about probabilistic programming languages!

Thank you, Mama and Papa, for raising me and making me the person I am, for your endless love, and all the care packages you have sent to Sweden. Lukas and Simon, we do not see each other very often right now but every time we meet, online or in real life, I am reminded of how much fun it is to be with you and how much I like you. Thank you, Tante Rosa, for your continuous love—I am still excited every time you come to visit us and sad when you leave!

Finally, I would like to thank my wife Rebekka and my daughter Solveig. Rebekka, you are the most wonderful person I have ever met. Thank you for being with me, regardless of whether life shows its bright or dark sides, and thank you for giving birth to Solveig. Solveig, thank you for showing me a completely new kind of unconditional love, recalibrating my priorities, and brightening up every day. And, last but not least, thank you for the cover image of this thesis!

David
Funding

My research was financially supported: by the Centre for Interdisciplinary Mathematics (CIM) at Uppsala University, Sweden; by the Swedish Research Council via the projects Learning of Large-Scale Probabilistic Dynamical Models (contract number: 2016-04278), Counterfactual Prediction Methods for Heterogeneous Populations (contract number: 2018-05040), and Handling Uncertainty in Machine Learning Systems (contract number: 2020-04122); by the Swedish Foundation for Strategic Research (SSF) via the projects ASSEMBLE (contract number: RIT15-0012) and Probabilistic Modeling and Inference for Machine Learning (contract number: ICA16-0015); by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation; and by the Excellence Center at Linköping–Lund in Information Technology (ELLIIT).
List of papers

This thesis is based on the following papers, which are referred to in the text by their roman numerals.


Reprints were made with permission from the publishers.
## Contents

1 Introduction  
1.1 Outline of the Thesis .................................................. 3  
1.2 Contributions ............................................................... 3  
1.3 Papers Included in the Thesis ......................................... 4  
1.4 Other Work ................................................................. 6  
1.5 Notation ................................................................. 7  

2 Models  
2.1 Prediction Tasks ........................................................ 9  
2.2 Probabilistic Predictive Models ...................................... 12  
2.3 Examples ................................................................. 15  
2.4 Bayesian and Frequentist Paradigms .............................. 28  

3 Calibration  
3.1 Reliable Uncertainties ................................................. 29  
3.2 Calibrated Models ....................................................... 32  
3.3 Scoring Rules ............................................................ 34  
3.4 Calibration Functions .................................................. 38  
3.5 Reliability Diagrams ...................................................... 42  
3.6 Calibration Errors ....................................................... 52  
3.7 Calibration Tests ....................................................... 55  

4 Kernel  
4.1 Positive Definite Kernel ............................................... 65  
4.2 Reproducing Kernel Hilbert Space .................................. 67  
4.3 Kernel Score ............................................................. 78  
4.4 Maximum-Mean Discrepancy ......................................... 80  

5 Concluding Remarks  
5.1 Conclusion ............................................................... 87  
5.2 Outlook ................................................................. 87  

A Notation List ............................................................. 89
B Measure Theory Basics

B.1 Probability Spaces ........................................ 93
B.2 Random Variables .......................................... 94
B.3 Probability Densities ...................................... 96
B.4 Conditional Expectations and Probabilities .......... 96
B.5 Conditional Distributions ................................. 98

References ..................................................... 101
CHAPTER 1

Introduction

This thesis marks the end of my PhD studies and summarizes my research during the last five and a half years.

When I started my studies in November 2017, I was completely new to the area of *machine learning*, and *deep learning* in particular.¹ My main impression was that it is related to statistics, “big data”, and GPU computing, and I was a bit worried whether the increasing interest in and expectations on machine learning already showed signs of a hype. Given these initial conditions, maybe it does not come as a complete surprise that my research has gradually transitioned from “Uncertainty-Aware Deep Learning for Image Classification in Pathology”, the original PhD project, to ”Reliable Uncertainty Quantification in Statistical Learning”, the title of the thesis you hold in hand. My studies have focused less and less on deep learning and specific applications of machine learning. Instead, I have turned towards the more fundamental question:

How can one ensure that a predictive model also reports some information about the uncertainty of its predictions, let it be weather forecasts, tumor grades, or COVID-19 infections, that is *interpretable* and *in line with reality*?

This question is at the core of my thesis and hence will be discussed in the upcoming chapters. However, already here I would like to stress a few points about this question.

First, the question aims at “predictive models”. In particular, it is not limited to deep learning or other classes of machine learning models. It does not matter whether a model is proposed and fine-tuned based on expert knowledge or large amounts of data, or both. “Predictive model” refers to any mathematical model that

---

¹Ironically, nevertheless I happened to become the first PhD student within the ”Machine Learning” PhD subject at Uppsala University.
outputs predictions, such as a model of weather forecasts, a model of tumor grades, a model of COVID-19 infections, or a language model that replies to a question.

Second, the question aims at the “uncertainty” of these predictions. Many predictions are inherently uncertain: As Box (1976) famously noted, “all models are wrong” and merely approximations of the modeled system, and thus typically predictions will be at most approximately correct. And this property is not unique to predictions of mathematical models, as de Finetti (2017) argued:

In almost all circumstances, and at all times, we all find ourselves in a state of uncertainty. Uncertainty in every sense. Uncertainty about actual situations, past and present […]. Uncertainty in foresight […]. Uncertainty in the face of decisions […].

Third, the question aims at uncertainties that are “interpretable” and “in line with reality”. Hopefully, the following chapters will clarify how these, in my view desirable, properties are defined precisely. On an intuitive level, the goal are uncertainty estimates that are reported in an intuitively understandable way and whose validity can be checked empirically.

Already the former is challenging as uncertainties and probabilities can be quite unintuitive, as evidenced by the “Monty Hall problem” (Selvin 1975): In a game show you are asked to select one out of three boxes, one containing the keys to a new car and the other two being empty, and you win whatever is in the box; After picking a box the game host, who knows which boxes are empty, opens one of the other two boxes that is empty and offers you to change your choice; Should you stick with your initial guess or change to the other unopened box? Even an expert in probability theory such as Paul Erdős did not believe that it is advantageous to switch your guess until he was shown a Monte Carlo simulation of the game (Vazsonyi 1999).

The desire for empirical validity is motivated by the problem of misleading model uncertainties: Some models, and in particular some complex modern models (Guo et al. 2017), tend to report too small uncertainties that mislead users into trusting their over-confident predictions too much. Over- and under-confident predictions are problematic whenever you would like to tell apart very confident from less confident predictions, e.g., to be able to handle only the former in a fully-automated way but require human interventions for more uncertain predictions. Unreliable uncertainties are a major concern in safety-critical applications such as autonomous driving but with the recent success and widespread use of language models such as GPT-4 (OpenAI 2023) and its predecessors these issues also show up in mainstream applications. The authors of GPT-4 remark in their report (OpenAI 2023, Section 5):

Despite its capabilities, GPT-4 has similar limitations as earlier GPT models. Most importantly, it still is not fully reliable (it “hallucinates”
1.1 Outline of the Thesis

This thesis is divided into two parts. The first part, including this chapter, contains an introduction to the topic of this thesis in Chapters 1 to 4 and finishes with concluding remarks in Chapter 5. The background material in these first chapters should make it easier for readers to study the five scientific papers that constitute the second part of the thesis. These articles cover the main part of my work on calibration and represent the public output of the research I conducted during my doctoral studies.

In the remaining sections of this first chapter, the contributions of this thesis and the included papers are summarized briefly. In Chapter 2, the class of mathematical models studied in this thesis are introduced and exemplified. In Chapter 3, the term "calibration", or "reliability", is clarified in the context of this thesis, and reliability diagrams and calibration tests are discussed as two common approaches for analyzing this statistical property. In Chapter 4, scalar- and operator-valued kernels and two kernel discrepancies, the maximum mean discrepancy and the kernel Stein discrepancy, are introduced. The first part ends with summarizing comments and an outlook with potential future work in Chapter 5.

1.2 Contributions

The main contributions of this thesis are:

- A critical review of and suggestions for calibration analysis of multi-class classification models, including calibration lenses and hypothesis tests (Paper I).

- A novel approach for calibration errors and tests of multi-class classification models based on matrix-valued kernels (Paper II).
• An alternative derivation of calibration errors and tests with scalar-valued kernels, enabling calibration analysis of probabilistic models beyond classification (Paper III).

• Novel calibration errors and tests of density models based on the score function and specially designed kernels, making it possible to analyze calibration of unnormalized models in particular (Paper IV).

• The implementation and demonstration of software packages for calibration analysis with the Julia, R, and Python programming languages (Paper V).

1.3 Papers Included in the Thesis

The five papers included in this thesis are briefly summarized below, together with a short explanation of my contributions.

Paper I: Evaluating Model Calibration in Classification


Summary In this paper we discuss how to formulate and evaluate calibration in multi-class classification in a rigorous way. Additionally, we point out weaknesses in current approaches for evaluating calibration and suggest hypothesis tests based on consistency resampling for robust model comparisons.

Contribution Juozas Vaicenavicius, Carl Anderson, and I came up with the underlying idea for this paper, which was then refined and extended in discussions with Fredrik Lindsten and Thomas Schön. The mathematical notation and the formulation of calibration lenses originated from discussions between Juozas and me. The theorems are results of discussions between Juozas, Carl, and me. Carl and I implemented the empirical studies, and the illustrations, apart from Figure 3, are my work. The paper was written mainly by Juozas, Carl, and me, but all authors contributed.

Paper II: Calibration Tests in Multi-Class Classification: A Unifying Framework

Summary In this paper we propose a framework of calibration errors for multi-class classification that covers commonly used forms of the expected calibration error. It allows to derive a so-called kernel calibration error which is based on matrix-valued kernels and yields consistent and unbiased estimators and well-founded hypothesis tests.

Contribution I came up with the underlying idea for this paper, which was then refined and extended in discussions with Fredrik and Dave. The theoretical derivations and the implementations are my work. The paper was written mainly by me, with help from Fredrik and Dave.

Paper III: Calibration Tests beyond Classification


Summary In this paper the calibration error framework is reformulated and extended to general probabilistic predictive models by constructing an extended product space of probability distributions and labels. This construction enables the formulation of the kernel calibration error with scalar-valued kernels. Moreover, it links the study of calibration more closely to existing research on integral probability metrics, maximum mean discrepancy, and optimal transport.

Contribution I came up with the underlying idea for this paper, which was then refined and extended in discussions with Fredrik and Dave Zachariah. The theoretical derivations and the implementations are my work. The paper was written mainly by me, with help from Fredrik and Dave Zachariah.

Paper IV: Fast and Scalable Score-Based Kernel Calibration Tests


Summary In this paper kernel calibration tests for density models are proposed that are only based on the score-function of the predicted distributions. This design enables calibration testing of models with unnormalized probability distributions such as energy-based models or Bayesian models. In particular, in contrast to the kernel calibration error it is not required to evaluate or approximate generally intractable expectations. Besides the proposed calibration discrepancy, another key component is the design of score-function based kernels of probability distributions.
Contribution The paper originated from my visit of Arthur’s research group at University College London and the discussions with Pierre during my stay. Pierre’s work on simulation-based inference motivated the calibration analysis of unnormalized Bayesian models, and he pointed us to existing work on kernel conditional goodness-of-fit testing. I derived the relation between expected coverage and calibration and the generic formulation of the kernel calibration conditional Stein discrepancy, and noticed that it is a special instance of the previously proposed kernel calibration error. Pierre designed the score-function-based kernels for density models and derived their properties. Pierre and I implemented the proposed methods with JAX, building on Pierre’s existing code base for kernel methods and simulation-based inference. The Gaussian experiments were designed by me and Pierre. The paper was written by me and Pierre, with help and feedback from Fredrik and Arthur.


Summary In this paper, we explain the design of CalibrationAnalysis.jl, a software suite for analyzing calibration of probabilistic models. The software packages are written in the Julia programming language but can also be accessed from Python and R with the provided interface packages. The capabilities of the software are illustrated with synthetic examples of calibrated and uncalibrated models. Moreover, a possible calibration analysis workflow is illustrated with three multi-class classification models for the Palmer penguins dataset.

Contribution I designed and implemented the software packages. The paper was written mainly by me. The initial draft and the demonstrations in the paper were refined based on discussions and feedback from Fredrik and Dave.

1.4 Other Work

In addition to the included papers, the following publication is also related to this thesis:

1.5. Notation

The following publications were written during my doctoral studies but are not relevant for my research on calibration:


1.5 Notation

The mathematical notation is consistent throughout the first part of this thesis. It is inspired by the books *Theory of Statistics* (Schervish 1995) and *Support Vector Machines* (Christmann and Steinwart 2008). The most important symbols are explained in Appendix A. The papers are written with different notations, partly since they target different readerships and partly since the notation has been changed over time. Therefore, it was not possible to use the same notation throughout the whole thesis and I decided to freely choose a notation for the introductory chapters without restricting myself to the notations in the papers.
In almost all circumstances, and at all times, we all find ourselves in a state of uncertainty.

---

Theory of Probability
BRUNO DE FINETTI

CHAPTER 2

Models

This thesis and the included papers are concerned with estimating and testing calibration of probabilistic predictive models. Hence, “calibration” and “probabilistic predictive models” are two core concepts in this work. This chapter is devoted to the explanation of the term “probabilistic predictive models”, and the term “calibration” is discussed in Chapter 3.

2.1 Prediction Tasks

Goethe (1906) said, “theory is in itself of no use, except in so far as it makes us believe in the connection of phenomena”. While applications have not been an integral part of my research, it has been motivated by practical problems. Therefore, two prediction tasks and their challenges are presented before diving into theoretical explanations.

Weather Forecasting

A core part of meteorology is predictive modeling, also often called forecasting. Hence, it is hardly surprising that this field has been very active and motivating for the research on the statistical analysis of predictive models (“forecasters”). A classic example of a prediction problem, illustrated in Figure 2.1, is:

Will it rain today?

Nowadays, short-term predictions tend to be quite accurate. For instance, SMHI, the Swedish meteorological and hydrological institute, reports that in the last twelve months on average around 93.2% of their predictions of precipitation in Uppsala county for the present day were correct.¹ Here, a prediction of “rain” (“no

¹They certainly helped me to ensure that my daughter is appropriately dressed for kindergarten.
rain”) is considered to be correct if within 3 hours more (less) than 0.3 mm rain is observed. However, long-range forecasts are less reliable: On average, around 90.0% of the predictions three days in advance were correct and around 84.8% of the predictions five days ahead. This difference is apparent also in Figure 2.2 where the accuracy of the predictions is plotted for the last twelve months.

Predicting the weather over larger time spans is still difficult, even though the quality of weather forecasts has been improving and nowadays “a five-day forecast is as good as a three-day forecast 10 years ago” (Roth 2022). One challenge is that the number of measurements is limited (Simmons and Singh 2015; SMHI 2021b): It is impossible to observe the atmosphere at every location at all altitudes at all times, and observations at a location are not necessarily representative for a larger area around it. Thus, the current and past state of the atmosphere are not fully known. Moreover, the mathematical models used for simulating the weather dynamics are abstractions of reality and hence not perfect (Simmons and Singh 2015). Typically, simulating these models is also computationally expensive and hence their application is limited by computational resources and time constraints (SMHI 2021b). The models are also susceptible to the “butterfly effect” (Lorenz 1972): Even small errors in the (not fully known!) initial conditions of the dynamical models can accumulate over time and lead to exponentially exploding errors in the predictions at future time points. Clearly, this phenomenon is problematic when trying to make predictions, as already noted by Poincaré (1913):
2.1. PREDICTION TASKS

Figure 2.3: Incidence rate of and deaths by malignant melanoma per 100000 persons in Sweden (source: Socialstyrelsen 2022).

But even when the natural laws should have no further secret for us, we could know the initial situation only approximately. If that permits us to foresee the subsequent situation with the same degree of approximation, this is all we require, we say the phenomenon has been predicted, that it is ruled by laws. But this is not always the case; it may happen that slight differences in the initial conditions produce very great differences in the final phenomena; a slight error in the former would make an enormous error in the latter. Prediction becomes impossible and we have the fortuitous phenomenon.

Tumor Detection

Inspired by the initially proposed project for my doctoral studies and my interest in medicine, let us turn to a different prediction problem: Tumor detection in medical images, or more precisely the detection of malignant melanoma in dermoscopy images.

Malignant melanoma is the most dangerous type of skin cancer (Codella et al. 2018; Socialstyrelsen and Cancerfonden 2018). In 2021, around 5000 new cases of malignant melanoma were diagnosed in Sweden, making it the third and fifth most common type of cancer in men and women, respectively (Socialstyrelsen 2022). The incidence rate of and deaths due to malignant melanoma in Sweden are shown in Figure 2.3.

Over the last few decades the prognosis for patients diagnosed with malignant melanoma has been improving, mainly since more tumors are detected at an earlier stage and hence can be treated by surgery (Socialstyrelsen and Cancerfonden 2018). Nowadays, in Sweden the 5-year survival rate is around 95% for women and 90% for men (Socialstyrelsen and Cancerfonden 2018).

However, studies have shown that inspection with the unaided eye leads to correct diagnoses in only around 60% of the cases, even in specialized centers (Kittler
et al. 2002). Dermoscopy, a non-invasive in-vivo microscopy technique that makes subsurface structures of the skin visible, can improve the diagnostic accuracy of trained experts (Kittler et al. 2002). A selection of dermoscopy images of malignant melanoma, seborrheic keratosis (benign skin tumor derived from keratinocytes), and nevi (benign skin tumor derived from melanocytes) are shown in Figure 2.4.

There is increasing interest in automated analysis of such dermoscopy images, to improve the diagnostic accuracy and reduce the workload for dermatologists, especially in areas with a shortage of those (Codella et al. 2018). To improve the automated diagnostic of melanoma, the International Skin Image Collaboration (ISIC) started to build a public archive of dermoscopy images with ground-truth annotations and conducted benchmark challenges (Codella et al. 2018). In the ISIC challenge 2017, one of the three tasks was a prediction problem, or more specifically multi-class classification problem, related to Figure 2.4:

Does the dermoscopy image show a melanoma, nevus or seborrheic keratosis?

Interestingly, the ISIC 2017 challenge asked participants to approach this task by considering two binary classification problems, melanoma vs. nevus and seborrheic keratosis (malignant vs. benign) and seborrheic keratosis vs. melanoma and nevus (keratinocytes vs. melanocytes) (ISIC 2017). In this thesis, however, typically multi-class classification problems are viewed as single entities and not reduced to binary problems.

### 2.2 Probabilistic Predictive Models

The two examples of prediction tasks in the previous section have one thing in common: It is unrealistic to assume that predictions are always correct, regardless of whether they are reported by a human forecaster or a machine. For instance, unfortunately it seems unavoidable that weather forecasts for longer time spans such as weeks or months are not as accurate as predictions for the current day. To make the uncertainty around forecasts apparent and to improve their trustworthiness, Cooke (1906a) started to report not only a prediction (e.g., “rain” or “no
rain”) but also his confidence in every forecast. He explained his motivation as follows (Cooke 1906b):

Now this is the point I wish to make clear. Those forecasts which were marked “doubtful” were the best I could frame under the circumstances. I could see no way of improving them at the time, and they would not have been expressed differently whether I weighted them or not. If I make no distinction between these and others, I degrade the whole. But if, on the other hand, I attach a figure which practically says “I’m sorry, but this is the best I can do for you to-day—do not attach too much importance to it”, I eliminate beforehand the adverse opinion which a great number of incorrect forecasts must produce, and I raise the bulk of the predictions to their true value. In particular, I create a series, marked with the maximum figure, which the public finds to be almost invariably correct, and thus raise the value of this particular series enormously.

Such a predictive model with confidence values is illustrated in Figure 2.5, highlighting the difference to predictive models of the form in Figure 2.1 with point predictions “rain” and “no rain”. Similarly, in the dermoscopy example one can complement the prediction of the lesion type with a confidence value.

These confidence values express how likely the predictions are in the view of the forecaster. A common approach, advocated for amongst others by Dawid (1984) and de Finetti (2017), is to state them in the form of probabilities, i.e., as non-negative real numbers that are normalized such that the confidence values of the classes sum up to one. To emphasize that these predicted probabilities represent “the degree of belief in the occurrence of an event attributed by a given person [or model] at a given instant and with a given set of information” (de Finetti 2017, p. 3), they are sometimes called subjective probabilities and distinguished from objective probabilities such as the probabilities of fair dice or “repeatable” events.

---

2The prediction task in the ISIC 2017 challenge even explicitly asked the participants to submit confidence values in the interval [0, 1] (ISIC 2017).
Chapter 2. Models

*Probabilistic predictive models* are predictive models where class probabilities are predicted, as in Figure 2.5, instead of single classes, as in Figure 2.1.

The problem of uncertain predictions, however, goes beyond classification problems. For instance, long-range predictions of the temperature are also attached with higher uncertainty than predictions for, say, tomorrow. However, predicting the temperature is not a classification but a regression task. In such cases, in general the confidence in the predictions are reported as *(subjective) probability distributions*. For instance, when predicting the temperature the confidence in the admissible temperature values can be expressed with a (truncated) normal distribution.\(^3\) Since sets of discrete probabilities form a probability distribution as well, more generally, *probabilistic predictive models* are predictive models with probability distributions as predictions.

This informal explanation of a probabilistic predictive model is made more precise now in a mathematical definition. The measure-theoretical concepts of probability spaces and random variables ensure that it covers all desired use cases in a theoretically solid way. Appendix B introduces the most important measure-theoretical definitions and notations. Note that the set of probability measures on a measurable space \(\mathcal{Z}\) is denoted by \(\mathcal{M}\)\(1(\mathcal{Z})\).

**Definition 2.1** (Probabilistic Predictive Model). Let \((\Omega,\mathcal{A},\mathbb{P})\) be a probability space, let \((\mathcal{X},\mathcal{A}_X)\) and \((\mathcal{Y},\mathcal{A}_Y)\) be measurable spaces, and let \((\mathcal{P},\mathcal{A}_\mathcal{P})\) with \(\mathcal{P} \subset \mathcal{M}\)\(1(\mathcal{Y})\) be a measurable space of probability measures on \(\mathcal{Y}\). Let \(X: (\Omega,\mathcal{A}) \to (\mathcal{X},\mathcal{A}_X)\) and \(Y: (\Omega,\mathcal{A}) \to (\mathcal{Y},\mathcal{A}_Y)\) be two random variables.

A measurable function \(m: (\mathcal{X},\mathcal{A}_X) \to (\mathcal{P},\mathcal{A}_\mathcal{P})\) is called a *probabilistic predictive model* with input\(^4\) \(X\) and target\(^5\) \(Y\). The random variable \(m(X)\) represents a *(subjective) probability distribution* of \(Y\) given input \(X\). It is denoted by \(\mathbb{P} := m(X)\) when its dependency on \(X\) is not relevant.

**Remark 2.1.** The mathematical formulation in Definition 2.1 makes it also possible to study predictions without a priori defined inputs within the same framework: Given random predictions \(\mathbb{P}: \Omega \to \mathcal{P}\), you can a posteriori define input space \((\mathcal{X},\mathcal{A}_X) = (\Omega,\mathcal{A})\), input \(X = \text{id}_\Omega\), and model \(m \equiv \mathbb{P}\) to obtain \(\mathbb{P} = m(X)\) as desired by Definition 2.1. In the thesis this special scenario is only considered in Box 3.1.

Throughout the thesis, \(\mathbb{P}\) denotes the probability measure of an underlying probability space \((\Omega,\mathcal{A},\mathbb{P})\), if not specified otherwise. Similarly, the definition of input \(X\), target \(Y\), and prediction \(P\), and their respective measurable spaces, is omitted if it is the same as in Definition 2.1 or clear from the context.

As discussed above, in binary classification you can identify predictions \(m(X)\) with the probability of one of the two targets, the reference class. For simplicity

\(^{\text{3}}\)Of course, alternatively you can work with a discretization (e.g., by rounding temperatures to the nearest integer value between \(-20^\circ\text{C}\) and \(+40^\circ\text{C}\)) and again report confidence as discrete probabilities.

\(^{\text{4}}\)Instead of “input”, it is also common to use the terms “feature” (in particular in machine learning), “independent variable”, or “explanatory variable” (Everitt 2002).

\(^{\text{5}}\)Other common terms are “output”, “dependent variable”, or “response variable” (Everitt 2002).
but without sacrificing generality, in this thesis it is assumed that a binary model has target space \( \{0, 1\} \) and that class 1 is the reference class.

**Definition 2.2** (Probabilistic Predictive Model for Binary Targets). A *binary probabilistic predictive model* is a probabilistic predictive model (see Definition 2.1) with target space \( \mathcal{Y} = \{0, 1\} \).

If \( \mathbf{m}(\cdot) \) is a binary probabilistic predictive model, it can be identified with the measurable function

\[
\mathbf{m}^1 : \mathcal{X} \to [0, 1], \quad \mathbf{m}^1(x) := \mathbf{m}(x)(\{1\}) = \mathbb{E}_{y \sim \mathbf{m}(x)} y.
\]

The random variable \( \mathbf{m}^1(X) \) represents the (subjective) probability of reference class 1 given input \( X \). It is denoted by \( \mathbf{P}^1 := \mathbf{m}^1(X) \) when its dependency on \( X \) is not relevant.

Similarly, if a binary probabilistic predictive model is specified in terms of \( \mathbf{m}^1(\cdot) \), it can be identified with the measurable function

\[
\mathbf{m} : \mathcal{X} \to \mathcal{M}_1(\mathcal{Y}), \quad \mathbf{m}(x) := \text{Ber}(\mathbf{m}^1(x)),
\]

where \( \text{Ber}(\mathbf{m}^1(x)) \) denotes a Bernoulli distribution with success probability \( \mathbf{m}^1(x) \).

### 2.3 Examples

So far, the definitions of a probabilistic predictive model are very abstract. In this section, you will see concrete examples of different—but not completely distinct—types of probabilistic predictive models.

**Parametric Models**

A large class of probabilistic predictive models belong to a *parametric family*

\[
\{\mathbf{m}(\cdot; \theta) : \theta \in \mathcal{T}\},
\]

of models where \( \mathcal{T} \subseteq \mathbb{R}^{n_{\theta}} \) is an \( n_{\theta} \)-dimensional parameter space. Typically, parameters \( \theta \) are optimized\(^6\) based on observed data.

**Classification** An example of a parametric model for classification is the *logistic regression* model (see, e.g., K. P. Murphy 2022, Chapter 10) which covers both binary and multi-class classification. Its inputs are real-valued with input space \( \mathcal{X} \subseteq \mathbb{R}^{n_x} \) (\( n_x \in \mathbb{N} \)). For binary classification, its predictions are defined as\(^7\)

\[
\mathbf{m}^1(x; \theta) := \frac{1}{1 + \exp(-\theta^\top x)} \tag{2.1}
\]

\(^6\)In machine learning, it is common to say that the parameters are “learned” or “trained”.

\(^7\)Commonly, the formulation of the binary logistic regression model includes an intercept \( \theta_0 \) but here it is assumed that \( x \) is prepended with a 1 in case an intercept is used.
with parameter space $\mathcal{F} \subset \mathbb{R}^{n_x}$. The behavior of the model for real-valued inputs is illustrated in Figure 2.6. The multinomial logistic regression model is a generalization of the model to multi-class classification with target space $Y = \{1, \ldots, n_y\}$ ($n_y \in \mathbb{N}_{\geq 2}$) and parameter space $\mathcal{F} \subset \mathbb{R}^{n_y \times n_x}$. It is defined by the model

$$m(x; \theta) := \text{Categorical}(\text{softmax}(\theta x)) \quad (2.2)$$

of categorical distributions, where $\text{softmax} : \mathbb{R}^{n_y} \rightarrow \mathbb{R}^{n_y}$ is the softmax function defined by

$$(\text{softmax}(z))_i := \frac{\exp(-z_i)}{\sum_{j=1}^{n_y} \exp(-z_j)}$$

for $i = 1, \ldots, n_y$. The formulation in Equation (2.2) is very popular within the machine learning community, but it yields a non-identifiable model: The softmax function is invariant under additions of a constant value to the logits $z$, i.e., $\text{softmax}(z) = \text{softmax}(z + c1_{n_y})$ for all $z \in \mathbb{R}^{n_y}$ and $c \in \mathbb{R}$. This problem can be alleviated by choosing a reference class, as in binary logistic regression, and setting its logit to 1. This leads, e.g., to the alternative multinomial logistic regression model

$$m(x; \theta) := \text{Categorical} \left( \text{softmax} \left( \begin{bmatrix} \theta x \\ 1 \end{bmatrix} \right) \right)$$

with parameter space $\mathcal{F} \subset \mathbb{R}^{(n_y-1) \times n_x}$.

**Regression** A classic example of a parametric model for regression is the (multiple) linear regression model (see, e.g., K. P. Murphy 2022, Chapter 1.2.2.1)

$$m(x; \theta) = \mathcal{N}(\theta^\top x, \sigma^2)$$

with input space $\mathcal{X} \subset \mathbb{R}^{n_x}$ ($n_x \in \mathbb{N}$), target space $\mathcal{Y} \subset \mathbb{R}$, parameter space $\mathcal{F} \subset \mathbb{R}^{n_x}$, and fixed variance $\sigma^2$. The variance can also be included in the model, possibly as
a parameter or as a (parametric) function of input $x$. A sample of predictions from
a linear regression model with real-valued input and parameter $\theta = 1/2$ is shown
in Figure 2.7.

**Feature Engineering** The flexibility of parametric models can be increased by
*feature engineering* (see, e.g., K. P. Murphy 2022, p. 11). The core idea is to transform
inputs $x$ to features $\phi(x)$ in a so-called *feature space*, and call the parametric model
with the extracted features $\phi(x)$ instead of $x$. A sketch of this approach is visualized
in Figure 2.8. For instance, in binary logistic regression commonly an intercept
term is included which corresponds to features

$$
\phi(x) = \begin{bmatrix} 1 \\ x \end{bmatrix}
$$

in the formulation in Equation (2.1). Another example is the polynomial regression
model which corresponds to a linear regression model of the monomial features

$$
\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^{n_\phi} \end{bmatrix}
$$

extracted from a real-valued input $x \in \mathbb{R}$. The feature map $\phi$ may be parametric as
well. For instance, instead of manually designing features such as in polynomial
regression you can use a neural network, i.e., a composition of parameterized
non-linear functions, as feature extractor and optimize its parameters.
Chapter 2. Models

Figure 2.8: Illustration of a feature map $\phi$, mapping inputs in the input space $\mathcal{X}$ to the feature space $\phi(\mathcal{X})$. The model $m$ operates on the features instead of the inputs directly.

Explicit Models

All examples of probabilistic predictive models presented so far share one property: The predicted distributions are specified via their parameters, their probability density function is available in closed form and can be evaluated easily, and samples can be drawn from the distributions efficiently.

In contrast, energy-based models (EBMs) (see, e.g., LeCun et al. 2007; K. P. Murphy 2023, Chapter 24) specify distributions not through their parameters but via an explicit unnormalized density function which allows them to cover a wide range of distributions. However, in general, their (normalized!) probability density function is intractable and has to be approximated, and samples can not be drawn efficiently. EBMs have a long history in physics, statistics, and machine learning (LeCun et al. 2007). They have found widespread use in different areas of machine learning such as image generation (Ngiam et al. 2011) and density estimation (Song et al. 2020; Wenliang et al. 2019). The following explanations, however, are focused on their use as discriminative models (Grathwohl et al. 2019; Kelly, Zemel, and Grathwohl 2021) with input $X$ and target $Y$ which matches the predictive setting in this thesis.

As hinted at by their name, EBMs model the real-valued energy function

$$E(x, y; \theta) \geq 0$$

which is parameterized by $\theta \in \mathcal{F}$. Note that here the energy function is a function of both input $x$ and target $y$ since the goal is to build a density model of the joint distribution of $X$ and $Y$, following the approach by Grathwohl et al. (2019) and Kelly, Zemel, and Grathwohl (2021). Let $\mu$ and $\nu$ be two base measures on the input space $\mathcal{X}$ and the target space $\mathcal{Y}$, respectively. Common choices are the Lebesgue measure on Euclidean vector spaces and the counting measure on countable sets. Then for each input $x$ and target $y$, such an EBM models the joint density of $X$ and $Y$ with respect to the product measure $\mu \otimes \nu$ on $\mathcal{X} \times \mathcal{Y}$ as the Gibbs distribution

$$\frac{d\mathbb{P}_{X,Y}}{d(\mu \otimes \nu)}(x, y) \approx \frac{\exp(-E(x,y; \theta))}{Z(\theta)}.$$
2.3. Examples

Usually, the normalizing constant $Z(\theta)$, known as partition function and given by

$$Z(\theta) = \int_{x \times y} \exp \left( -E(x, y; \theta) \right) d(\mu \otimes \nu)(x, y),$$

is intractable. The energy function implies a model for the density of $Y$ given $X$ as well:

$$\frac{dP_{Y|X}}{d\nu}(y | x) \approx \frac{\exp \left( -E(x, y; \theta) \right)}{\int_y \exp \left( -E(x, \xi; \theta) \right) d(\xi)}. $$

If $Y$ is a finite discrete set $\{y_1, \ldots, y_{n_y}\}$ ($n_y \in \mathbb{N}$) with the counting measure as base measure $\nu$, the integral is tractable. The predictions are categorical distributions $m(x; \theta) = \text{Categorical} \left( \text{softmax} \left( \left[ -E(x, y_i) \right]_{i=1}^{n_y} \right) \right)$, i.e., the predictive model $m(\cdot)$ is available explicitly for classification. Otherwise, typically the density of the predicted distributions is only known up to proportionality:

$$\frac{dm(x; \theta)}{d\nu}(y) \propto \exp \left( -E(x, y; \theta) \right).$$

Due to the intractable partition function, usually in such cases approximate methods are required when evaluating the likelihood and generating samples from the predicted distributions. Fortunately, for differentiable models the Stein score function

$$\nabla_y \log \frac{dm(x; \theta)}{d\nu}(y) = -\nabla_y E(x, y; \theta)$$

is tractable and can be evaluated solely based on the model of the energy function. Thus, the intractable predictions of the EBM are still amenable to, e.g., sampling with gradient-based Markov chain Monte Carlo methods such as Langevin MCMC (Besag 1994; Roberts and Tweedie 1996) and Hamiltonian Monte Carlo (Duane et al. 1987; Neal 1996). The approach for calibration analysis in paper IV is based on score functions and hence can be applied to such models.

**Example** In Figure 2.9, an example of a banana-shaped energy function $E(\cdot, \cdot)$ on $\mathbb{R} \times \mathbb{R}$ is shown:

$$E(x, y) = \frac{100(y - x^2)^2 + (1 - x)^2}{20}.$$

This energy function is known as the Rosenbrock function (Rosenbrock 1960), a classic test function for optimization algorithms. In particular, the variant with a scaling by $1/20$ here is also used for testing MCMC algorithms (Goodman and Weare 2010; Pagani, Wiegand, and Nadarajah 2022). Another special property of this energy function is that its partition function is tractable. It can be deduced that the energy function corresponds to the following model of input $X$ and target $Y$:

$$X \sim \mathcal{N}(1, 10),$$

$$Y \mid X = x \sim \mathcal{N}(x^2, 0.1).$$
Figure 2.9: Example of an energy function $E(\cdot, \cdot)$ (left), and the corresponding density of $Y$ given $X = 0$ together with a histogram of 10000 samples from the energy function drawn with the Metropolis-adjusted Langevin algorithm (right).

With the Metropolis-adjusted Langevin algorithm (MALA) (Besag 1994) 10000 samples are drawn from the energy function at $x = 0$, to approximate the distribution of the prediction for $X = 0$. A histogram of the samples is displayed in the right plot of Figure 2.9, together with the exact conditional density.

Bayesian Models

Another class of models with often unnormalized predictive distributions are Bayesian models. They are “fully probabilistic” models since all unknowns are viewed as random variables (Lindley 1975):

The distinguishing feature of Bayesian statistics is that all unknown quantities are random variables: not just the data, but other variables, like parameters, are, before they are observed, random. The act of observation changes the status of the quantity from a random variable to a number.

A Bayesian model specifies a joint distribution of all these random variables, typically by defining a prior distribution of the parameters and a family of conditional distributions for the data given the parameters. The densities of the latter, considered as a function of the parameters for some fixed data, are called the likelihood function.

At the core of Bayesian statistics, and hence Bayesian models, is Bayes’ theorem which establishes the relation between the densities of the prior distribution, the likelihood, and the posterior distribution, the conditional distribution of the parameters given some observed data.

**Theorem 2.1** (Bayes’ theorem (cf. Schervish 1995, Theorem 1.31)). Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and let data space $(\mathcal{Z}, \mathcal{A}_Z)$ and parameter space $(\mathcal{F}, \mathcal{A}_\mathcal{F})$ be
2.3. Examples

Borel spaces. Let $Z : \Omega \to \mathcal{Z}$ and $\Theta : \Omega \to \mathcal{F}$ be measurable. Suppose that $Z$ has a parametric family

$$\mathcal{P}_0 = \{P_\theta : \mathcal{A}_Z \to [0,1] \mid \theta \in \mathcal{F} \land \forall A : P_\theta(A) = \mathbb{P}_{Z|\Theta}(A \mid \theta)\}$$

of conditional distributions for $Z$ given $\Theta$. Suppose that there exists a measure $\nu$ on $\mathcal{Z}$, such that $P_\theta \ll \nu$ for all $\theta \in \mathcal{F}$, and let $f_{Z|\Theta}(z \mid \theta)$ be the conditional density of $Z$ given $\Theta = \theta$ (with respect to $\nu$). Let $\mathbb{P}_{\Theta}$ be the distribution of $\Theta$, called the prior distribution. Let $\mathbb{P}_{\Theta|Z}(\cdot \mid z)$ denote the conditional distribution of $\Theta$ given $Z = z$, called the posterior distribution of $\Theta$. Then

1. $\mathbb{P}_{\Theta|Z} \ll \mathbb{P}_{\Theta}$, almost surely with respect to $\mathbb{P}_Z$, and

2. the Radon-Nikodym derivative, referred to as the posterior density, is

$$\frac{d\mathbb{P}_{\Theta|Z}}{d\mathbb{P}_{\Theta}}(\theta \mid z) = \frac{f_{Z|\Theta}(z \mid \theta)}{\int_{\mathcal{F}} f_{Z|\Theta}(z \mid \theta) \, d\mathbb{P}_{\Theta}(\theta)}$$

for those $z$ for which the denominator is neither 0 nor infinite. For the other $z$ values the posterior can be defined arbitrarily.

The main takeaways from this theorem (which can be formulated even more generally (Schervish 1995, p. 16)) are that,

- under general assumptions on the parametric family and the spaces of parameters and observations, the posterior distribution is absolutely continuous with respect to the prior distribution (for almost all observations), and

- for almost all observations, the density of the posterior distribution with respect to the prior distribution is proportional to the likelihood.

In the following sections, it is assumed that the assumptions of Bayes’ theorem are satisfied. Note that the marginal distribution of $\mathbb{P}_Z$ is absolutely continuous with respect to measure $\nu$ on $\mathcal{Z}$ (Schervish 1995, p. 13), and its density is given by

$$f_Z(z) = \int_{\mathcal{F}} f_{Z|\Theta}(z \mid \theta) \, d\mathbb{P}_{\Theta}(\theta).$$

---

As discussed in Appendix B.5, Borel spaces are spaces that are sufficiently “similar” to $\mathbb{R}$. In particular, Euclidean spaces and countable discrete spaces with the Borel $\sigma$-algebra, i.e., the smallest $\sigma$-algebra that contains all open sets, are Borel spaces. In Bayes’ theorem, the most relevant property of Borel spaces is that they guarantee the existence of the desired regular conditional distributions exist (see Theorem B.5).
**Coin-Flip** To build some intuition for the Bayesian approach, consider a coin-flip experiment: You flip a coin $n \geq 1$ times and study whether it is a fair coin and, if not, how much it is biased towards heads or tails. Denote the probability of heads with a random variable $\Theta$ (parameters) taking values in $[0, 1]$. Then the probability of tails is $1 - \Theta$.

The outcome of $n$ coin-tosses can be modeled as a random variable $Z$ (data) taking values in $\{0, 1, \ldots, n\}$, where the event $Z = z$ corresponds to $z$ “heads” and $n - z$ “tails” amongst the first $n$ coin flips. The conditional distribution of $Z$ given probability $\Theta = \theta$ of heads is a binomial distribution of $n$ trials with success probability $\theta$: $Z \mid \Theta = \theta \sim \text{Binom}(n, \theta)$. This yields a class of parametric models $\mathcal{M}_{\Theta} = \{\text{Binom}(n, \theta) : \theta \in [0, 1]\}$ with conditional densities

$$f_{Z \mid \Theta}(z \mid \theta) = \binom{n}{z} \theta^z (1 - \theta)^{n-z}$$

with respect to the counting measure. In principle, the prior distribution on the parameter space $[0, 1]$ can be chosen arbitrarily but is often desired to use either “subjective” priors, incorporating prior knowledge (Tend my coins to be fair? Is this type of coin known to be biased?), or “objective” (weakly) uninformative reference priors. A particularly convenient choice is a beta distribution $\text{Beta}(\alpha, \beta)$ with parameters $\alpha, \beta > 0$ since it allows to derive a closed-form expression for the posterior distribution. The density of the prior distribution $P_{\Theta} = \text{Beta}(\alpha, \beta)$ with respect to the Lebesgue measure $\lambda$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is given by

$$f_{\Theta}(\theta) = \frac{dP_{\Theta}(\theta)}{d\lambda}(\theta) = \frac{\theta^{\alpha-1}(1 - \theta)^{\beta-1}}{B(\alpha, \beta)} \mathbb{1}[0 \leq \theta \leq 1]$$  

(2.3)

for $\theta \in \mathbb{R}$, where $B(\cdot, \cdot)$ is the Beta function and $\mathbb{1}[\cdot]$ is the Iverson bracket.

Bayes’ theorem yields that, given data $z \in \{0, \ldots, n\}$, the density of the posterior distribution $P_{\Theta \mid Z}(\cdot \mid z)$ with respect to the Lebesgue measure is given by

$$f_{\Theta \mid Z}(\theta \mid z) = \frac{dP_{\Theta \mid Z}(\theta \mid z)}{d\lambda}(\theta \mid z) = \frac{dP_{\Theta \mid Z}(\theta \mid z)}{dP_{\Theta}}(\theta) f_{\Theta}(\theta)$$

$$\propto \theta^{\alpha+z-1}(1 - \theta)^{\beta+n-z-1} \mathbb{1}[0 \leq \theta \leq 1]$$

for $\theta \in \mathbb{R}$. If you compare this result with Equation (2.3), you see that in this example both the prior and the posterior distributions are Beta distributions:

$$P_{\Theta \mid Z}(\cdot \mid z) = \text{Beta}(\alpha + z, \beta + n - z).$$  

(2.4)

---

9Here, the simplifying assumption is made that the only possible outcomes are heads and tails. However, generally the probability that a coin lands on edge is non-zero and is estimated to be approximately 1 in 6000 tosses for an American nickel (Murray and Teare 1993).

10The Beta function $B(\cdot, \cdot)$ is defined as $B(x, y) = \int_0^1 t^{x-1}(1 - t)^{y-1} dt$ for $x, y \in \mathbb{C}$ with $\mathfrak{R}(x), \mathfrak{R}(y) > 0$ (Abramowitz and Stegun 1972, Equation 6.2.1).

11The Iverson bracket $\mathbb{1}[\cdot]$ is defined as $\mathbb{1}[P] = 1$ if property $P$ is true, and $\mathbb{1}[P] = 0$ otherwise. It was initially introduced by Iverson (1962) in a less general way using parentheses. His formulation was extended by Knuth (1992), who changed the notation to square brackets and encouraged its use to “clarify the use of characteristic functions and Kronecker delta in sums and integrals”. In the thesis, the Iverson bracket is written with blackboard bold square brackets to make it stand out and less likely to be confused with regular brackets.
Thus, deriving the posterior distribution only amounts to updating the parameters of the prior distribution. Prior distributions, for which the posterior distribution with respect to a given likelihood belongs to the same family of distributions, are called conjugate prior distributions (for this likelihood). Such a special constellation allows efficient updates of the posterior distribution when new data is obtained. Figure 2.10 illustrates sequential updates of the distributions (“beliefs”) of $\Theta$ before flipping a coin, after one coin flip, and after two coin flips, starting with a uniform prior distribution $\text{Unif}(0, 1) = \text{Beta}(1, 1)$.

**Deriving A Predictive Model** The coin-flip experiment is an example of a generic Bayesian model with data $Z$ and parameters $\Theta$. How does this give rise to a probabilistic model of the form in Definition 2.1 with input $X$ and target $Y$? For a Bayesian model of $X, Y$, and parameters $\Theta$ (and possibly other random variables), where target space $Y$ is a Borel space\(^\text{12}\), Conditioning only on $X = x$ leads to the prior predictive model

$$m(x) = \mathbb{P}_{Y|X}(\cdot \mid x)$$

of the conditional distribution of $Y$ given only $X = x$. As in the coin-flip example, one core feature of Bayesian models is that the “beliefs” about the model parameters can be updated with data by applying Bayes’ theorem. Thus, if you have access to data, usually you want to incorporate it both in your Bayesian and your predictive model. This approach yields posterior predictive models of the form

$$m(x) = \mathbb{P}_{Y|X,Z}(\cdot \mid x, z)$$

where $Z$ is a random variable that represents data (e.g., samples of $(X, Y)$) and $z$ is the observed data.

\(^{12}\)Note once more that this ensures that a regular conditional distribution of $Y$ given $X$ (or some other random variables) exists.
Figure 2.11: Predictive distributions of the coin-flip experiment with \( n = 2 \) tosses and \( \alpha = \beta = 1 \).

**Coin-Flip (cntd.)** The coin-flip experiment can be recast as a prediction task with the number of heads amongst \( n \) coin tosses as target \( Y \) (target space \( Y = \{0, 1, \ldots, n\} \)) and a constant input, say, \( X \equiv 0 \) (input space \( \mathcal{X} = \{0\} \)). If the prior distribution is chosen as a Beta distribution, then the prior predictive distribution can be computed in closed form. For all \( x \in \mathcal{X} \) and \( y \in Y \), its density with respect to the counting measure is

\[
f_{Y \mid X}(y \mid x) = \int_{\mathbb{R}} f_{Y \mid X, \theta}(y \mid x, \theta) f_\theta(\theta) \, d\lambda(\theta)
\]

\[
= \binom{n}{y} \frac{B(\alpha + y, \beta + n - y)}{B(\alpha, \beta)}.
\]

For a uniform prior \( \text{Unif}(0, 1) = \text{Beta}(1, 1) \), i.e., if \( \alpha = \beta = 1 \), this expression simplifies to

\[
f_{Y \mid X}(y \mid x) = \frac{1}{n + 1}.
\]

Thus, with a uniform prior the prior predictive distribution is a discrete uniform distribution on \( Y = \{0, 1, \ldots, n\} \). Define observed data \( Z \) as number of heads amongst the previous \( m \) coin tosses (\( Z = \{0, 1, \ldots, m\} \)). Then for all \( x \in \mathcal{X} \), \( y \in Y \), and \( z \in Z \), the posterior predictive density with respect to the counting measure is

\[
f_{Y \mid X, Z}(y \mid x, z) = \int_{\mathbb{R}} f_{Y \mid X, \theta}(y \mid x, \theta) f_{\theta \mid Z}(\theta \mid z) \, d\lambda(\theta)
\]

\[
= \binom{n}{y} \frac{B(\alpha + z + y, \beta + m + n - z - y)}{B(\alpha + z, \beta + m - z)}.
\]

As expected, the posterior predictive distribution coincides with the prior predictive distribution when no data is observed (\( m = z = 0 \)). Figure 2.11 shows how the predictive distribution is affected by observing tails once or twice for \( n = 2 \): In line with the posterior distributions in Figure 2.10, a reduced number of heads is expected in the next two tosses.

**Discriminative Bayesian Models** Discriminative models with input \( X \) and target \( Y \), such as Bayesian linear regression models discussed below, focus by
design on the conditional distribution of \( Y \) given \( X \). They are stated as a parametric family of conditional distributions for target \( Y \) given input \( X \) and parameters \( \Theta \):

\[
\mathcal{P}_0 = \{ P_\theta : \mathcal{A}_Y \times \mathcal{X} \to [0, 1] \mid \theta \in \mathcal{T} \land \forall A, x : \mathbb{P}_Y(A \mid x) = \mathbb{P}_{Y \mid X, \theta}(A \mid x, \theta) \}.
\]

Moreover, it is assumed that input \( X \) and parameters \( \Theta \) are independent. In such a Bayesian model, the prior predictive distributions can be computed as

\[
\mathbb{P}_{Y \mid X}(\cdot \mid x) = \int_{\mathcal{T}} P_\theta(\cdot \mid x) \, d\mathbb{P}_\theta(\theta).
\]

Assume there is a measure \( \nu_x \) on \( (\mathcal{Y}, \mathcal{A}_Y) \) such that \( P_\theta(\cdot \mid x) \ll \nu_x \) for all \( \theta \in \mathcal{T} \), and denote the density of \( P_\theta(\cdot \mid x) \) with respect to \( \nu_x \) by

\[
f_{Y \mid X, \theta}(y \mid x, \theta) = \frac{dP_\theta}{d\nu_x}(y \mid x).
\]

Then the prior predictive distribution \( \mathbb{P}_{Y \mid X}(\cdot \mid x) \) is absolutely continuous with respect to \( \nu_x \) (see, e.g., Schervish 1995, p. 13) and has density

\[
f_{Y \mid X}(y \mid x) = \int_{\mathcal{T}} f_{Y \mid X, \theta}(y \mid x, \theta) \, d\mathbb{P}_\theta(\theta).
\]

Assume that \( Y \mid X \) and \( Z \) are conditionally independent given \( \Theta = \theta \). Then the posterior predictive distribution can be computed as

\[
\mathbb{P}_{Y \mid X, Z}(\cdot \mid x, z) = \int_{\mathcal{T}} P_\theta(\cdot \mid x) \, d\mathbb{P}_\theta(\theta \mid z),
\]

and its density with respect to \( \nu_x \) is

\[
f_{Y \mid X, Z}(y \mid x, z) = \int_{\mathcal{T}} f_{Y \mid X, \theta}(y \mid x, \theta) \, d\mathbb{P}_\theta(\theta \mid z).
\]

Unfortunately, typically these integrals are intractable and the prior and posterior predictive density have to be approximated, e.g., by a Monte Carlo estimate.

**Bayesian Linear Regression** A tractable example of such a discriminative Bayesian model is the *Bayesian linear regression* model (see, e.g., K. P. Murphy 2022, Chapter 11.7). Analogously to the non-Bayesian linear regression model, in its simplest form it consists of input \( X \) with input space \((\mathbb{R}^{n_x}, \mathcal{B}(\mathbb{R}^{n_x}))\), target \( Y \) with target space \((\mathbb{R}, \mathcal{B}(\mathbb{R}))\), parameters \( \Theta \) with parameter space \((\mathbb{R}^{n_x}, \mathcal{B}(\mathbb{R}^{n_x}))\), and the class

\[
\mathcal{P}_0 = \{ P_\theta : \mathcal{B}(\mathbb{R}) \times \mathcal{X} \to [0, 1] \mid \theta \in \mathbb{R}^{n_x} \land \forall x \in \mathcal{X} : P_\theta(\cdot \mid x) = \mathcal{N}(\theta^T x, \sigma^2) \}
\]

of parametric models of the conditional distributions of \( Y \) given \( X \) and \( \Theta \). For simplicity, variance \( \sigma^2 \) is fixed and a Gaussian prior \( \mathbb{P}_\Theta = \mathcal{N}(\mu, \Sigma) \) with mean
\(\mu \in \mathbb{R}^{n_x}\) and covariance matrix \(\Sigma \in \mathbb{R}^{n_x \times n_x}\) is chosen, as this leads to closed-form expressions for the predictive distributions. For all \(x \in \mathbb{R}^{n_x}\), the prior predictive distribution is

\[
P_{Y|X}(\cdot | x) = \int_{\mathbb{R}^{n_x}} \mathcal{N}(\cdot; \theta^\top x, \sigma^2) \mathcal{N}(\theta; \mu, \Sigma) \, d\lambda^n_x(\mathbb{R}^{n_x})
= \mathcal{N}(\mu^\top x, \sigma^2 + \mu^\top \Sigma \mu).
\]

Let \(Z = \{(X^{(i)}, Y^{(i)})\}_{i=1}^n\) be observations that are conditionally independent given \(\Theta = \theta\) with distributions \(P_{Y|X^{(i)}, \Theta}(\cdot | x, \theta) = \mathcal{N}(\cdot | x)\). Assume that the distribution of \((X^{(1)}, \ldots, X^{(n)})\) is independent of \(\Theta\). Then for all \(x \in \mathbb{R}^{n_x}\) and \(z\), the posterior predictive distribution is (see, e.g., K. P. Murphy 2022, p. 11.7.4)

\[
P_{Y|X,Z}(\cdot | x, z) = \mathcal{N}(m^\top x, \sigma^2 + m^\top Sm)
\]

where

\[
S = (\Sigma^{-1} + \sigma^{-2} X^\top X)^{-1}, \quad m = \Sigma(S^{-1} \mu + \sigma^{-2} X^\top y),
\]

\[
X = \begin{bmatrix} X^{(1)} \top \\ \vdots \\ X^{(n)} \top \end{bmatrix}, \quad y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}.
\]

**Predicting The Posterior** The discussion of Bayesian models is arguably a bit convoluted. This is mainly caused by the need for finding, or even designing, an input-target relation within the Bayesian model to obtain a model of conditional distributions. While the Bayesian linear regression model intuitively suggests a predictive model based on the prior or posterior predictive distribution, in the coin-flip example an artificial input had to be introduced. Another, arguably more natural, approach directly applicable to any Bayesian model with data \(Z\) and parameters \(\Theta\) is to define a predictive model that predicts the posterior distribution \(P_{\Theta|Z}(\cdot | z)\) given data \(Z = z\). That is, the data serves as input, \(X = Z\), and the parameters take the role of the target, \(Y = \Theta\). This setup matches the common scenario in Bayesian inference where the posterior distribution is the main object of interest.

**Coin-Flip (cntd.)** In the context of the coin-flip experiment, this corresponds to using the number of heads in \(n\) coin tosses as input \(X\), taking values in \(\{0, 1, \ldots, n\}\), and predicting the posterior distribution of the probability of heads, which represents target \(Y\) taking values in \([0, 1]\). Thus, the result in Equation (2.4) implies that the predictive model is given by

\[
m(x) = \text{Beta}(\alpha + x, \beta + n - x).
\]
2.3. Examples

**Approximate Posterior** In the coin-flip experiment and in Bayesian linear regression, the posterior distribution is available in closed form. However, in contrast to these examples, in many applications it is not possible or too restrictive to use conjugate prior distributions. Thus, unfortunately, often the posterior distribution is intractable and predictive models of the posterior distribution can not be evaluated exactly. Instead, typically you have to resort to approximations of the posterior. A large area of research is devoted to approximating probability distributions such as posterior distributions with discrete distributions. These approximations can be formed, e.g., by (weighted) samples obtained with Markov chain Monte Carlo methods (see, e.g., Neal 1993; MacKay 2003, Chapter 29). In the extreme case, the posterior distribution can even be approximated with a Dirac measure at its mode, the maximum a posteriori estimate.\(^{13}\) Alternatively, variational methods approximate the posterior distribution with a simpler variational distribution, e.g., a Gaussian distribution, whose parameters are optimized by minimizing the distance between the posterior and its approximation (see, e.g., MacKay 2003, Chapter 33). Typically, the Kullback–Leibler divergence is chosen as distance measure.

A common disadvantage of these approximate methods is that a new approximation of the posterior distribution has to be obtained when the predictive model is evaluated for a new input, i.e., when new data is available. Surrogate models of the posterior distribution such as conditional neural density estimators can alleviate this issue: A parameterized model \(q(\theta \mid z)\) of the posterior density \(f_\theta(z \mid \theta)\) is optimized with joint samples of data \(Z\) and parameters \(\Theta\), such that for any data \(z\) and parameter \(\theta\) the surrogate \(q(\theta \mid z)\) provides an approximation of the posterior density \(f_\theta(z \mid \theta)\). Such approaches are particularly popular in simulation-based inference (see, e.g., Cranmer, Brehmer, and Louppe 2020), also known as likelihood-free inference\(^{14}\), since the computational cost is amortized: “After an upfront simulation and training phase, new data can be evaluated very efficiently” (Cranmer, Brehmer, and Louppe 2020).

The main goal—and challenge—in simulation-based inference is to perform, frequentist or Bayesian, statistical inference in settings where the likelihood function is intractable. Instead, these statistical models are defined implicitly via a simulator, i.e., computer program, that takes a set of parameters and returns simulated data. Such models appear in different scientific domains such as particle physics, protein folding, population genetics, and astrophysics (Cranmer, Brehmer, and Louppe 2020).

---

\(^{13}\)MacKay (2003, p. 306) lists multiple “reason[s] for disliking the maximum a posteriori” method: The posterior density may be extremely large on very small volumes, and hence spikes might not be representative for the distribution, and the estimate is not invariant under changes of the basis.

\(^{14}\)According to Cranmer, Brehmer, and Louppe (2020), one motivation for the term “simulation-based inference” is that “likelihood-free inference” seems to be “a bit of a misnomer as typically one attempts to estimate the intractable likelihood”.

27
2.4 Bayesian and Frequentist Paradigms

Let me take the opportunity at the end of this chapter for a small detour and some personal comments, before moving on to the next chapter. There is a long history of debates between Bayesian statisticians, i.e., statisticians following the Bayesian paradigm and performing inference based on the posterior distribution, and classical frequentist statisticians, i.e., statisticians who do not follow this paradigm but perform inference based on, e.g., hypothesis tests and confidence intervals, following the “Fisherian theory and [...] the Neyman–Pearson–Wald (NPW) school of decision theory” (Efron 1986). Sometimes the discussions about which approach to statistics should be preferred culminated in quite harsh statements (Lindley 1975):

The only good statistics is Bayesian statistics. Bayesian statistics is not just another technique to be added to our repertoire alongside, for example, multivariate analysis; it is the only method that can produce sound inferences and decisions in multivariate, or any other branch of, statistics. It is not just another chapter to add to that elementary text you are writing; it is that text. It follows that the unique direction for mathematical statistics must be along the Bayesian road.

While some of these debates and arguments are interesting to hear, in my opinion the—partly exaggerated—dichotomy is not particularly helpful. I am not sure if I would call myself a statistician but if asked which camp I belong to, most certainly I would favor the side of the “pragmatists, who do not have an overarching philosophy and pick and choose what seems to work for the problem at hand” (Little 2006). Life seems easier if your toolkit contains both a hammer and a screwdriver, and you are able to choose the best tool for the problem at hand. Little (2006), however, argues that this pragmatist approach causes an “inferential schizophrenia” and that it underlines “our credibility as statisticians” that such fundamental questions are not settled, citing Efron (2005):

The physicists I talked with were really bothered by our 250-year-old Bayesian-frequentist argument. Basically, there’s only one way of doing physics, but there seems to be at least two ways to do statistics, and they don’t always give the same answers.

In my opinion, the ideas of Efron (2005), Little (2006), and Rubin (1984) and others to combine the strengths of Bayesian and frequentist ideas in a so-called calibrated Bayesian approach seem the most attractive way to resolve, or at least address, the Bayesian-frequentist debate. The core idea of the calibrated Bayes is that “inferences under a particular model should be Bayesian, but model assessment can and should involve frequentist ideas” (Little 2006). The focus of the next chapter will be one of this seemingly frequentist assessments: Model calibration.
In theory randomness is an intrinsic property, in practice, randomness is incomplete information.

---

**The Black Swan**

*NASSIM NICHOLAS TALEB*

CHAPTER 3

Calibration

In Chapter 2, probabilistic predictive models were introduced as a class of mathematical models whose defining property is that predictions are expressed as probability distributions. Hence, such models provide uncertainty estimates in the form of probability distributions. Per se these uncertainty estimates are subjective and come without any guarantees. For downstream applications and when using these models for decision-making, however, completely arbitrary uncertainty estimates are not trustworthy and not helpful. Thus, the uncertainty estimates should satisfy certain statistical properties that ensure that they are not only subjective but actually an appropriate model of the uncertainties involved in the prediction task. One such property is *calibration*, also called *reliability*. This chapter is devoted to the explanation of the term “calibration” in the context of probabilistic predictive models. Moreover, *reliability diagrams*, *calibration errors*, and *calibration tests* are introduced, three closely related approaches for analyzing the calibration of a model.

### 3.1 Reliable Uncertainties

In both examples in Chapter 2, weather forecasting and classification of dermoscopy images, probabilistic predictive models are supposed to improve the trustworthiness and usefulness of the predictions compared with models that predict only single classes. However, switching from a non-probabilistic to a probabilistic model does not generally imply such an improvement. For instance, typically a probabilistic model that yields the same predictions as the original non-probabilistic model with 100% confidence is not desired—such probabilistic predictions express that there is no uncertainty involved in the prediction task which usually is not a realistic assumption.

The definition of calibration is motivated by the desire to ground the predicted distributions, and hence the uncertainty represented by them, in real life experi-
CHAPTER 3. CALIBRATION

![Diagram showing the process from measurements to prediction]

**Empirical Observations**

<table>
<thead>
<tr>
<th></th>
<th>rain</th>
<th>no rain</th>
</tr>
</thead>
<tbody>
<tr>
<td>outcome</td>
<td>(y^{(1)}, y^{(3)}, \ldots)</td>
<td>(y^{(2)}, \ldots)</td>
</tr>
<tr>
<td>frequency</td>
<td>80%</td>
<td>20%</td>
</tr>
</tbody>
</table>

Figure 3.1: Empirical interpretation of a calibrated prediction “rain: 80%”: The empirical frequency of the observed outcomes matches the predicted confidence value.

...ences and observations. Predictions with, say, 60% confidence should be correct 60% of the time. Or put differently, the probability that a prediction with 80% confidence is correct should be 80%. If a prediction satisfies this property, it is called calibrated. This concept of calibrated predictions is illustrated in Figure 3.1: The prediction “rain: 80%” is calibrated if it rains on 80% of the days for which “rain: 80%” is predicted—which can be confirmed hypothetically by observing an infinite number of independent and identically distributed pairs of measurements \(x^{(i)}\) (of temperature, air pressure, etc.) and outcomes \(y^{(i)}\) (“rain” or “no rain”).

Similarly, the concept of calibration can be applied to probabilistic predictions for multi-class classification and even beyond classification where probability distributions are predicted instead of single probabilities. In these cases, a prediction is called calibrated if the conditional distribution of the target given this prediction is equal to the prediction. In binary classification this definition of calibration coincides with the one given above based on probabilities. This more general interpretation is visualized in Figure 3.2 for the classification of dermoscopy images: The prediction “melanoma: 90%, nevus: 5%, seborrheic keratosis: 5%” is calibrated if 90% of the images for which “melanoma: 90%, nevus: 5%, seborrheic keratosis: 5%” is predicted show a melanoma, 5% a nevus, and the remaining 5% a seborrheic keratosis.

If almost all predictions of a probabilistic model are calibrated, the model is called calibrated. Hence, if a probabilistic model is calibrated, the uncertainty expressed in its predictions has a clear interpretation: The reported probability distribution is equal to the conditional distribution of targets given this prediction.
## 3.1. Reliable Uncertainties

![Image](image_url)

**Empirical Observations**

<table>
<thead>
<tr>
<th>outcome</th>
<th>melanoma, $y^{(1)}$, $y^{(3)}$</th>
<th>nevus, $y^{(2)}$</th>
<th>seborrheic keratosis, $y^{(2)}$, $y^{(3)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
<td>90%</td>
<td>5%</td>
<td>5%</td>
</tr>
</tbody>
</table>

Figure 3.2: Empirical interpretation of a calibrated prediction “melanoma: 90%, nevus: 5%, seborrheic keratosis: 5%”: The empirical distribution of the observed outcomes matches the predicted distribution.

It is important to stress that this interpretation is not based on the input to the model but solely on the reported prediction. Thus even in cases such as weather forecasting or classification of dermoscopy images, where often the inputs are not available to the person receiving the forecast or the patient receiving the result of the examination, the uncertainty stated by the prediction has an empirically valid interpretation. Dawid (1982) argued that in general it is more appropriate to analyze the calibration of probabilistic predictions than to compare them with—possibly non-existing—conditional distributions of the target given the input.

The calibration criterion has some similarity with the frequency definition of probability, but does not require a background of repeated trials under constant conditions. In particular, it is rarely appropriate to interpret a subjective probability forecast as an estimate of some underlying “objective” probability; it is usually better considered as an estimate of (the indicator of) the forecast event itself. Thus we do not have to concern ourselves with the “true” probability of rain on a given day. Roberts (1968) has attempted to interpret such a concept by supposing that one could select a subset of all days that could be regarded, at the time of forecast, as identical in all relevant respects, and consider the limiting relative frequency of rain on such days as the “true” probability for any one of them. However, it is doubtful whether such a selection is practically meaningful, or whether differ-
ent forecasters would agree on it. The calibration approach avoids these difficulties.

While calibration seems to be a frequentist concept, it is by no means limited to non-Bayesian models. On the contrary, “calibrated Bayes” (Little 2006), i.e., Bayesian modeling with calibrated predictions, has been suggested by Bayesian statisticians such as Dawid (1982), Efron (2005), Gneiting, Balabdaoui, and Raftery (2007), Little (2006), and Rubin (1984) as a path towards merging Bayesian approaches with frequentist ideas, as mentioned in the previous chapter. For instance, Rubin (1984, p. 1160) argues that Bayesian predictions require justification that Bayes’ theorem does not provide:

In some formal sense, a Bayesian statement, such as a 95% posterior interval for an unknown, needs no justification since given the explicitly stated models, the statement follows from the laws of probability theory. This justification, however, is not very satisfying to the applied statistician for two related reasons. First, why should the models that are being conditioned upon be accepted? [...] Second, what does the stated Bayesian probability, e.g. 95%, mean objectively or empirically? The question of tying the 95% to real world events is addressed here via the concept of frequency calibration. A Bayesian is calibrated if his probability statements have their asserted coverage in repeated experience.

## 3.2 Calibrated Models

The general notion of calibrated probabilistic predictions, as discussed above for multi-class classification, is formalized in the following definition (cf., e.g., Bröcker 2009; DeGroot and Fienberg 1981).

**Definition 3.1.** Let $m(\cdot)$ be a probabilistic predictive model with $X$-valued input $X$, $Y$-valued target $Y$, and predictions $P = m(X)$ of probability distributions in $\mathcal{P} \subset \mathcal{M}_1(Y)$ (see Definition 2.1). Assume there exists a regular conditional distribution $P_{Y|P}$ of $Y$ given $P$.\(^1\) Model $m(\cdot)$ is called calibrated at a prediction $P \in \mathcal{P}$ if

$$P_{Y|P}(\cdot | P) = P.$$ 

The model is called calibrated if it is calibrated at $P$-almost all predictions $P \in \mathcal{P}$, i.e., if

$$P_{Y|P} = P \quad \text{almost surely.}$$

\(^1\)As discussed in Appendix B.5, a regular conditional distribution exists if the target space $Y$ is a Borel space, i.e., if it is sufficiently “similar” to $\mathbb{R}$. Fortunately, discrete spaces $Y$ (classification) and $Y = \mathbb{R}^d$ (regression) with their Borel $\sigma$-algebra are similar enough, and hence the concept of calibration can be studied in these common cases.
3.2. CALIBRATED MODELS

Note that the definition of calibration does not explicitly depend on the inputs $X$—they are only used to define random variable $P$ but, as discussed in Remark 2.1, both inputs and model can be defined a posteriori if $P$ is given. This illustrates once more the argument by Dawid (1982): The concept of calibration is not based on comparisons of the model with a version of the conditional distribution of $Y$ given $X$.

Definition 3.1 covers general probabilistic predictive models, including models for binary classification, multi-class classification, and regression problems (for instance, models for predicting the temperature). Historically, however, calibration has been pre-dominantly analyzed for binary classification models (Dawid 1982; A. H. Murphy and Winkler 1977, 1987). It is also common to analyze calibration of multi-class models by analyzing some binary aspects of the predictions such as the most-confident predictions (confidence calibration) (see, e.g., Guo et al. 2017; Kull, Perello Nieto, et al. 2019; Kumar, Liang, and Ma 2019) or predictions of each class separately (classwise or marginal calibration) (see, e.g., DeGroot and Fienberg 1981; Kull, Perello Nieto, et al. 2019; Kumar, Liang, and Ma 2019; Zadrozny and Elkan 2002). These forms of calibration are weaker than the notion in Definition 3.1; they are required but not sufficient for the latter (see, e.g., DeGroot and Fienberg 1981). A common characteristic of all these weaker notions of calibration is that they are based on predicted probabilities instead of predicted probability distributions. Thus, the special case of Definition 3.1 for binary targets stated explicitly in Corollary 3.1 covers these weaker notions, possibly after reducing the model to one with binary targets.

**Corollary 3.1.** Let $m(\cdot)$ be a binary probabilistic predictive model with predicted probabilities $P^1 \in \mathcal{P}^1 \subset [0,1]$ (see Definition 2.2). Let $p \in \mathcal{P}^1$. Then model $m(\cdot)$ is calibrated at prediction $\text{Ber}(p)$, also called calibrated at probability $p$, if and only if

$$
\mathbb{E}(Y \mid P^1 = p) = \mathbb{P}_{Y \mid P^1(\{1\} \mid p) = p}.
$$

The model is calibrated if and only if

$$
\mathbb{E}(Y \mid P^1) = P^1 \text{ almost surely.}
$$

**Calibration Is Not All You Need**

Calibration is a desirable property of a predictive model as it gives meaning to the reported uncertainty estimates by aligning them with the actual empirical distribution of targets. However, calibrated models are not necessarily informative: The constant model $m \equiv \mathbb{P}_Y$ that always predicts the marginal distribution of $Y$ satisfies

$$
\mathbb{P}_{Y \mid P} = \mathbb{P}_Y = P
$$
and hence is calibrated.\textsuperscript{2} Fortunately, this is not the only calibrated model as, e.g., the model $m(x) = \mathbb{P}_{Y\mid X}(\cdot \mid x)$ is calibrated as well.\textsuperscript{3}

These examples already illustrate that some calibrated models are more useful than others. DeGroot and Fienberg (1983) proposed a partial ordering of calibrated models to be able to rank them. Loosely speaking, according to the proposed ordering a calibrated model is more informative than another calibrated model if the latter can be expressed as a possibly noisy transformation of the first one.

### 3.3 Scoring Rules

More generally, probabilistic predictive models can be ranked with scoring rules such as the Brier score (Brier 1950) or the logarithmic score (Good 1952). As discussed by Gneiting and Raftery (2007), the main idea is to assess predictions by comparing them with observations of the target. Predictive models are rewarded with a high score if their predictions are consistent with the observations, and low scores otherwise. I shortly introduce the main definitions from the paper by Gneiting and Raftery (2007). Rewards may be infinite, e.g., when using the logarithmic score, and hence the following definitions are formulated to cover scoring rules whose co-domain is the extended real line $\mathbb{R} := \mathbb{R} \cup \{-\infty, \infty\}$.

**Definition 3.2** (Scoring Rule and Expected Score (cf. Gneiting and Raftery 2007, p. 360)). Let $(\Omega, \mathcal{A})$ be a measurable space, and fix a subclass $\mathcal{P} \subset \mathcal{M}_1(\Omega)$ of probability measures on $(\Omega, \mathcal{A})$. Let $s : \mathcal{P} \times \Omega \to \mathbb{R}$.

Then $s$ is called a *scoring rule* if for all $P \in \mathcal{P}$, $s(P, \cdot)$ is measurable with respect to $\mathcal{A}$ and quasi-integrable for all $Q \in \mathcal{P}$.\textsuperscript{4} The score $s(P, \omega)$ can be viewed as the "reward" of a prediction $P$ if outcome $\omega$ is realized.

Assume that $s$ is a scoring rule. Then the *expected score* of a prediction $P \in \mathcal{P}$ under the distribution $Q \in \mathcal{P}$ of outcomes with respect to $s$ is

$$S(P, Q) := \int_{\Omega} s(P, \omega) \, dQ(\omega).$$

Scoring rule $s$ is called *proper* with respect to $\mathcal{P}$ if

$$\forall P, Q \in \mathcal{P} : S(Q, Q) \geq S(P, Q),$$

\textsuperscript{2}In meteorology, this constant model is called the *climatology* of $Y$ (see, e.g., Bröcker 2009; Gneiting, Balabdaoui, and Raftery 2007; A. H. Murphy and Winkler 1987).

\textsuperscript{3}More generally, every regular conditional distribution of $Y$ defines calibrated predictions (see Bröcker 2009, and Box 3.1).

\textsuperscript{4}A function $f$ on a measure space $(\Omega, \mathcal{A}, Q)$ is $(Q)$-quasi-integrable, or equivalently it is said that the integral of $f$ exists, if it is $\mathcal{A}$-measurable and at least one of the functions $f^+ := f^\uparrow \mathbb{1}[f > 0]$ (positive part) or $f^- := -f^\uparrow \mathbb{1}[f < 0]$ (negative part) has a real integral (Bauer 2001, p. 64). As in the standard definition of the Lebesgue integral, the integral of a $(Q)$-quasi-integrable function $f$ is defined as $\int f \, dQ = \int f^+ \, dQ - \int f^- \, dQ$ which, however, is not necessarily a real number but only satisfies $\int f \, dQ \in \mathbb{R}$.
and it is called \textit{strictly proper} if additionally \( S(P, Q) = S(Q, Q) \) implies \( P = Q \).

Under a few additional assumptions on the scoring rule, the expected score gives rise to a divergence function.

\textbf{Definition 3.3} (Score Divergence and Entropy (cf. Gneiting and Raftery 2007, p. 361)). Let \( s : \mathcal{P} \times \Omega \to \mathbb{R} \) be a scoring rule and \( S : \mathcal{P} \times \mathcal{P} \to \mathbb{R} \) be the associated expected score function. Scoring rule \( s \) is called \textit{regular} with respect to \( \mathcal{P} \) if

1. for all \( Q \in \mathcal{P} \), \( S(Q, Q) \in \mathbb{R} \), and
2. for all \( P, Q \in \mathcal{P} \) with \( P \neq Q \), \( S(P, Q) \in \mathbb{R} \cup \{-\infty\} \).

If scoring rule \( s \) is proper with respect to \( \mathcal{P} \), we call

\[ H : \mathcal{P} \to \mathbb{R}, \quad H(P) := \sup_{Q \in \mathcal{P}} S(Q, P) = s(P, P), \]

the associated \textit{entropy function}. If \( s \) is proper and regular, then

\[ d : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_{\geq 0}, \quad d(P, Q) := S(Q, Q) - S(P, Q), \]

is called the associated \textit{divergence function}.

Examples of strictly proper scoring rules for a finite discrete sample space \( \Omega = \{\omega_1, \ldots, \omega_m\} \) and a class of categorical distributions

\[ \mathcal{P} = \{\text{Categorical}(p) : p \in \mathbb{R}_{\geq 0}^m, \|p\|_1 = 1\} \]

are

- the \textit{Brier (or quadratic) score} (Brier 1950; Gneiting and Raftery 2007; Selten 1998)

\[ s^{\text{BS}}(P, \omega) := -\sum_{i=1}^m \left( P(\{\omega_i\}) - \mathbb{E}[\omega = \omega_i] \right)^2, \]

- the \textit{logarithmic score} (Good 1952)

\[ s^{\text{LS}}(P, \omega) := \log P(\{\omega\}), \]

- and the \textit{pseudo-spherical score} for \( \alpha > 1 \) (Gneiting and Raftery 2007)

\[ s^{\text{PS}}(P, \omega) := \frac{P(\{\omega\})^{\alpha-1}}{\left( \sum_{i=1}^m P(\{\omega_i\})^{\alpha-1} \right)^{(\alpha-1)/\alpha}}. \]
The divergence of the Brier score is the squared Euclidean distance of the probability vectors,
\[ d_{BS}(P, Q) = \sum_{i=1}^{m} \left( P(\{\omega_i\}) - Q(\{\omega_i\}) \right)^2, \]
and the divergence associated with the logarithmic score is the Kullback-Leibler divergence
\[ d_{LS}(P, Q) = \sum_{i=1}^{m} Q(\{\omega_i\}) \log \left( \frac{Q(\{\omega_i\})}{P(\{\omega_i\})} \right). \]

By identifying predictions with their density with respect to a fixed base measure, the logarithmic score, the Brier score, and the pseudo-spherical score can also be generalized to predictions of continuous distributions (Gneiting and Raftery 2007). An alternative generalization of the Brier score, the kernel score (Dawid 2007), is presented in Section 4.3. The energy score and the continuous ranked probability score can be viewed as special instances of the kernel score (Gneiting and Raftery 2007). The Hyvärinen scoring rule (Hyvärinen 2005) is another alternative for predictions of continuous distributions with differentiable Stein score function. Since its computation does not involve evaluations of the density but only of its Stein score functions, the Hyvärinen scoring rule is particularly suitable for predictive models such as EBMs and Bayesian models seen in Section 2.3 for which only unnormalized densities of the predictions are available.

How are scoring rules related to calibration? It has been known for a long time that the Brier score can be decomposed into three components, called uncertainty of \( Y \), resolution term, and reliability term (see, e.g., A. H. Murphy 1973; A. H. Murphy and Winkler 1977). As shown by Bröcker (2009), for problems with finite discrete target spaces a similar decomposition exists for any proper scoring rule. This result also generalizes to other target spaces (see, e.g., Widmann 2020).

**Lemma 3.1 (Decomposition of Expected Scores).** Let \( m(\cdot) \) be a probabilistic predictive model with \( X \)-valued input \( X \), \( Y \)-valued target \( Y \), and predictions \( P = m(X) \) of probability distributions in \( \mathcal{P} \subset \mathcal{M}_1(Y) \) (see Definition 2.1). Assume there exists a regular conditional distribution of \( Y \) given \( P \).

Let \( \mathcal{A} \subset \mathcal{M}_1(Y) \) be such that \( \mathcal{P} \subset \mathcal{A} \), \( \mathbb{P}_Y \in \mathcal{A} \), and \( \mathbb{P}_{Y|P} \in \mathcal{A} \) almost surely. Let \( s : \mathcal{A} \times Y \rightarrow \mathbb{R} \) be a proper and regular scoring rule that is measurable with respect to the second argument, and let \( d \) and \( H \) denote the associated score divergence and entropy, respectively.

---

5There exist alternative decompositions of the expected score, with the most well-known being a decomposition into a "reliability" component, also called "calibration" or "validity", and a "resolution" component, also called "sharpness" or "refinement", where the latter combines the uncertainty and resolution terms in the decomposition with three components (DeGroot and Fienberg 1983; Kull and Flach 2015; A. H. Murphy 1973; Sanders 1963).

6Note that this formulation follows the convention by Gneiting and Raftery (2007) of maximizing scores whereas Bröcker (2009) used the convention of minimizing scores.
Then the expected score \( \mathbb{E} s(P, Y) \) can be decomposed into the three terms

\[
\begin{align*}
\text{RES} & := \mathbb{E} d(\mathbb{P}_Y, \mathbb{P}_Y|P) \in \mathbb{R}_{\geq 0} \quad \text{(resolution),} \\
\text{REL} & := \mathbb{E} d(P, \mathbb{P}_Y|P) \in \mathbb{R}_{\geq 0} \quad \text{(reliability), and} \\
\text{UNC} & := H(\mathbb{P}_Y) \in \mathbb{R} \quad \text{(uncertainty of } Y),
\end{align*}
\]

as

\[
\mathbb{E} s(P, Y) = \text{RES} - \text{REL} + \text{UNC}.
\]

\textbf{Proof.} As above, let \( S(P, Q) \) denote the expected score of prediction \( P \in \mathcal{P} \) under the distribution \( Q \) with respect to the scoring rule \( s(\cdot, \omega) \). By assumption \( s \) is regular, and hence \( S(P, \mathbb{P}_Y|P) \in \mathbb{R} \cup \{-\infty\} \) almost surely. Thus, by the law of total expectation

\[
\mathbb{E} s(P, Y) = \mathbb{E} \mathbb{E}(s(P, Y) | P) = \mathbb{E} S(P, \mathbb{P}_Y|P) \in \mathbb{R} \cup \{-\infty\}.
\]

Moreover,

\[
\mathbb{E} S(\mathbb{P}_Y|P, \mathbb{P}_Y|P) \in \mathbb{R}
\]

and, using the law of total expectation once more,

\[
\mathbb{E} S(\mathbb{P}_Y, \mathbb{P}_Y|P) = \mathbb{E} \mathbb{E}(s(\mathbb{P}_Y, Y) | P) = \mathbb{E} s(\mathbb{P}_Y, Y) = S(\mathbb{P}_Y, \mathbb{P}_Y) \in \mathbb{R}.
\]

Thus, Equations (3.3) and (3.4) can be added and subtracted from Equation (3.2). Rearranging and using the definition of the score divergence and entropy yields

\[
\begin{align*}
\mathbb{E} s(P, Y) &= \mathbb{E} S(P, \mathbb{P}_Y|P) + \mathbb{E} S(\mathbb{P}_Y|P, \mathbb{P}_Y|P) - \mathbb{E} S(\mathbb{P}_Y|P, \mathbb{P}_Y|P) \\
&= \mathbb{E} d(\mathbb{P}_Y|P, \mathbb{P}_Y|P) + H(\mathbb{P}_Y).
\end{align*}
\]

The same calculations also reveal the range of the three terms.

An immediate observation of the decomposition in Equation (3.1) is that constant models, including the completely uninformative calibrated model \( \mathbf{m} \equiv \mathbb{P}_Y \), yield a minimal resolution term of zero, all calibrated models yield a minimal reliability term of zero, and the uncertainty term is not affected by the choice of model. Moreover, the decomposition highlights a few additional points regarding comparisons of two probabilistic predictive models with expected scores (Bröcker 2009):

1. If both models are calibrated, or more generally their reliability term is equal, then the model with larger expected score has better resolution, i.e., is more informative. This corresponds to the partial ordering proposed by DeGroot and Fienberg (1983).

2. If one model has better resolution/calibration than the other, then in general it is unclear whether its expected score is larger since the improvement in resolution/calibration could be outweighed by a decrease in calibration/resolution.
3. In general a larger expected score does neither imply better resolution nor better calibration since one component can be traded off for the other.

### 3.4 Calibration Functions

The main question discussed in the remaining sections of this chapter is: How can you assess if a model is calibrated?

Let $m(\cdot)$ be a predictive model with $\mathcal{X}$-valued input $X$, $\mathcal{Y}$-valued target $Y$, and predictions $P = m(X)$ in $\mathcal{P} \subset \mathcal{M}_1(Y)$, and suppose that a regular conditional distribution $\mathbb{P}_{Y|P}$ of $Y$ given $P$ exists. The calibration of $m(\cdot)$ can be studied by inspecting the calibration function\(^7\) (Bröcker 2008) $\rho : \mathcal{P} \to \mathcal{M}_1(Y)$, given for all predictions $P \in \mathcal{P}$ by

$$\rho(P) := \mathbb{P}_{Y|P}(\cdot \mid P).$$

The calibration function is of interest for at least three reasons. First, Definition 3.1 implies that model $m(\cdot)$ is calibrated if and only if the calibration function is equal to the identity function $\text{id}_P$ almost surely. Second, it provides information not only about the calibration of the model but also about the calibration of specific (regions of) predictions of interest and hence is a helpful tool for analyzing calibration. Third, it allows you to re-calibrate possibly uncalibrated predictions: It is well-known (see, e.g., Bröcker 2009) and verified easily (see Box 3.1) that the model $\rho \circ m$ formed by the composition of a model $m(\cdot)$ with its calibration function $\rho$ is always calibrated, regardless of whether model $m(\cdot)$ is calibrated or not.

<table>
<thead>
<tr>
<th>Box 3.1: Conditional Distributions Are Calibrated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analogously to the proof by Bröcker (2009, Appendix A) for finite discrete target spaces, you can show that all calibrated models correspond to regular conditional distributions.</td>
</tr>
</tbody>
</table>

**Theorem 3.1.** Let $m(\cdot)$ be a probabilistic predictive model with $\mathcal{X}$-valued input $X$, $\mathcal{Y}$-valued target $Y$, and predictions $P = m(X)$ of probability distributions in $\mathcal{P} \subset \mathcal{M}_1(Y)$ (see Definition 2.1). Suppose that $\mathcal{Y}$ is a Borel space.

Then model $m(\cdot)$ is calibrated if and only if there exists a measurable function $h$ on $\mathcal{X}$ such that

$$P = \mathbb{P}_{Y|h(X)} \quad \text{almost surely.}$$

**Proof.** First, note that regular conditional distributions of $Y$ exist since $\mathcal{Y}$

---

\(^7\)It seems there is no consensus on the name of this function in the literature even though its estimation is a core interest in the study of calibration. Instead of the name “calibration function” (used, e.g., by Bröcker 2008; Siegert 2014; Wilks 2019) some authors use the term “reliability curve” (see, e.g., Jolliffe and Stephenson 2011). The notation in this thesis is inspired by the use of $\rho(\cdot)$ for this function in DeGroot and Fienberg 1981, 1983.
3.4. Calibration Functions

is a Borel space (see Appendix B.5). (Alternatively, you could assume the existence of the required regular conditional distributions explicitly.)

If model \( \mathbf{m}(\cdot) \) is calibrated, then by Definition 3.1 \( \mathbb{P}_Y|\mathbf{P} = \mathbf{P} \). Thus, Equation (3.6) is satisfied for the measurable function \( h = \mathbf{m} \).

Suppose now that predictions \( \mathbf{P} \) can be expressed in the form of Equation (3.6) for some measurable function \( h \) on \( \mathcal{X} \). Note that then random variable \( \mathbf{P} \) is measurable with respect to the sub-\( \sigma \)-algebra \( \sigma(h(X)) \subset \sigma(X) \subset \mathcal{A} \) generated by \( h(X) \), and by the law of total probability (Schervish 1995, Theorem B.70)

\[
\mathbb{P}_Y|\mathbf{P} = \mathbb{E}(\mathbb{E}(\mathbb{E}(Y \mathbb{1}_\cdot | h(X)) | \mathbf{P}) = \mathbb{E}(\mathbb{E}(\mathbb{E}(Y \mathbb{1}_\cdot | h(X)) | \mathbf{P})
\]

where equalities are to be understood in the almost sure sense. Plugging in the definition of \( \mathbf{P} \) yields

\[
\mathbb{P}_Y|\mathbf{P} = \mathbb{E}(\mathbf{P} | \mathbf{P}) = \mathbf{P} \quad \text{almost surely,}
\]

and hence the model \( \mathbf{m}(\cdot) \) is calibrated. \( \square \)

Theorem 3.1 is formulated for models with an input \( X \). As explained in Remark 2.1, predictive models without input \( X \) can be viewed as predictive models in the sense of Definition 2.1 by constructing an input \( X = \text{id}_\Omega \) with input space \( \mathcal{X} = \Omega \). Thus, such models are also covered by Theorem 3.1. Moreover, since in this generic case \( h(X) = h \) denotes an arbitrary random variable, you see that any regular conditional distribution of \( Y \) with respect to some random variable yields calibrated predictions (cf. Bröcker 2009, Appendix A).

Choosing a constant function, say, \( h \equiv 0 \), shows that the constant model \( \mathbf{m} \equiv \mathbb{P}_Y \) is calibrated. Another interesting choice is \( h = \text{id}_\mathcal{X} \) which implies that the model \( \mathbf{m}(x) = \mathbb{P}_Y|X(\cdot | x) \) is calibrated. Finally, setting \( h = \tilde{\mathbf{m}} \) for a probabilistic model \( \tilde{\mathbf{m}}(\cdot) \) with calibration function \( \tilde{\rho} \) yields that the model \( \tilde{\rho} \circ \tilde{\mathbf{m}} \) is calibrated.

For binary models with predicted probabilities \( \mathbf{P}^1 \) in \( \mathcal{P}^1 \subset [0,1] \), instead of the calibration function \( \rho \) you can equivalently study the simpler calibration function \( \rho^1 : \mathcal{P}^1 \to [0,1] \) given for all \( p \in \mathcal{P}^1 \) by

\[
\rho^1(p) := \rho(\text{Ber}(p))(\{1\}) = \mathbb{E}_{y \sim \text{Ber}(p)} y = \mathbb{E}(Y | \mathbf{P}^1 = p).
\]

Thus, the calibration function of binary models can be inspected visually by plotting the graph of \( \rho^1 \) in \([0,1]^2\): The binary model is calibrated if and only if \( \mathbb{P}_P \)-almost all points of the graph lie on the principal diagonal \( \{(p, p) : p \in [0,1]\} \). In the following sections, I refer to \( \rho^1(p) \) as the \textit{conditional target frequency} (of class 1) given a predicted probability \( p \).


**CHAPTER 3. CALIBRATION**

**Estimation**

In practice, usually the conditional target distribution $\mathbb{P}_{Y \mid X}$ is not known and hence the calibration function is not available. As Bröcker (2008) rightfully notes, “if it were, there would no longer be any excuse for not issuing fully calibrated forecasts”.

Hence, typically you have to estimate the calibration function from empirical data, commonly provided as samples $(x^i, y^i)$ of inputs and targets or as samples $(P^i, y^i)$ of probabilistic predictions and targets. As usual, from now on I assume that the samples are independent and identically distributed according to $\mathbb{P}_{X,Y}$ or $\mathbb{P}_{P,Y}$, respectively. Typically, in practice samples of inputs and targets are given. However, for the analysis of calibration samples of predictions are sufficient since the calibration condition in Definition 3.1 does not explicitly depend on input $X$ but only implicitly via prediction $\mathbf{P} = \mathbf{m}(X)$. Thus, often it is assumed that samples of predictions and targets are available. This assumption is not restrictive since you can recover samples of predictions from samples of inputs by computing the prediction for each input.

Estimating the calibration function from empirical data is straightforward if only a finite set of probabilities are predicted and for each prediction multiple samples of targets are observed: For each prediction the conditional target frequency, and hence the value of the calibration function, can be estimated by the empirical frequency of the observed targets (A. H. Murphy and Winkler 1977).

Unfortunately, often the predictions are not confined to a finite set and only a set of distinct predictions is available, typically each with a single or very few target observations. In these cases, it is popular to estimate the calibration function $\rho_1$ of a binary model with an ordinary histogram regression estimate (Atger 2004; Bröcker 2008; Bröcker and Smith 2007; Dimitriadis, Gneiting, and Jordan 2021; A. H. Murphy and Winkler 1977), referred to as “binning and counting” in the calibration literature (see, e.g. Bröcker 2008; Dimitriadis, Gneiting, and Jordan 2021). In this approach the calibration function is approximated with a step-function on the interval $[0,1]$. Predictions are partitioned into different bins, often subintervals of equal size, and on each bin the calibration function is estimated with the empirical frequency of target class 1. More formally, assume that $n$ observations

$$D := \{(P^{1,i}, y^i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{P^{1,Y}}$$

of predicted probabilities and targets are given. Then the calibration function $\rho_1$ is approximated at probability $p \in [0,1]$ by

$$\hat{\rho}_1(p) = \frac{\sum_{i=1}^n y^i \mathbb{I} [P^{1,i} \in B(p)]}{\sum_{i=1}^n \mathbb{I} [P^{1,i} \in B(p)]},$$

where $B(p)$ denotes the bin of prediction $p$. Thus, the estimate of the calibration function is constant on each bin.

An alternative approach (Bröcker 2008) is discussed in Chapter 4.
3.4. Calibration Functions

Examples
In the following sections different methods for estimating the calibration function and analyzing calibration are discussed. Throughout, I use three binary models for illustration purposes.

In all examples, the input space is set to \( \mathcal{X} = [0, 1] \) and the conditional distribution of target \( Y \) given input \( X = x \) is \( P_{Y \mid X}(\cdot \mid x) = \text{Ber}(x) \). Two choices for the distribution \( P_X \) of input \( X \) are considered: The uniform distribution Unif(0, 1) and the beta distribution Beta(5, 1), whose mass is skewed towards 1. Finally, the three models A, B, and C are defined as

A. \( m^1(x) := x \),

B. \( m^1(x) := \sqrt{x} \), and

C. \( m^1(x) := \begin{cases} 2x^2 & \text{if } 0 \leq x \leq 0.5, \\ 1 - 2(1 - x)^2 & \text{if } 0.5 < x \leq 1. \end{cases} \)

Since all models are bijections, in all three cases the calibration function satisfies for all \( p \in [0, 1] \)

\[ \rho^1(p) = \mathbb{E}(Y \mid P^1 = p) = \mathbb{E} \left( Y \mid X = (m^1)^{-1}(p) \right) = (m^1)^{-1}(p). \]

Thus, the corresponding calibration functions, shown in Figure 3.3, are given for \( p \in [0, 1] \) by\(^8\)

A. \( \rho^1(p) = p \),

B. \( \rho^1(p) = p^2 \), and

C. \( \rho^1(p) = \begin{cases} \sqrt{p/2} & \text{if } 0 \leq p \leq 0.5, \\ 1 - \sqrt{(1 - p)/2} & \text{if } 0.5 < p \leq 1. \end{cases} \)

Model A is a calibrated model, whereas models B and C are uncalibrated. All probabilities predicted by model B, apart from predictions of 0% and 100%, are larger than the actual conditional probability of class 1. Thus, the predictions of model B are biased towards class 1. On the other hand, predictions of model C are biased towards the class with larger confidence. If the predicted probability of class 0/1 is in (0.5, 1), it is larger than the actual conditional probability of class 0/1. Hence, model C is generally too overconfident.

For both scenarios, uniformly distributed inputs and inputs distributed with Beta(5, 1), a set of \( n = 500 \) independent input-target pairs \( (x^i, y^i) \) is drawn. For all models, for each input \( x^i \) the resulting predicted probability \( P^{1,i} \) is computed.

---

\(^8\)Even though all functions are monotonous and “in practice, nonmonotonic calibration functions are rarely encountered” (Bröcker 2008), in principle calibration functions can be nonmonotonic “since there is no general reason why calibration functions have to be monotonous” (Bröcker 2008).
From the empirical data the calibration function is estimated using the binning and counting approach in Equation (3.7). Three different choices of bins are compared: A single bin [0, 1], five right-open bins [0, 0.2), ..., [0.8, 1], and ten right-open bins [0, 0.1), ..., [0.9, 1]. The resulting estimates of the calibration function are shown in Figure 3.4.

The visualization highlights some challenges of the histogram regression estimation. First, the calibration function cannot be estimated on empty bins, i.e., for probabilities \( p \) in which the corresponding bin \( B(p) \) does not contain any predictions \( p^{1,i} \), since in this case the right-hand side of Equation (3.7) is not defined (see, e.g., Bröcker 2008). Second, the estimates are highly dependent on the number of bins (see, e.g., Dimitriadis, Gneiting, and Jordan 2021). Third, the quality of the estimates obtained with bins of equal size decreases if the predicted probabilities are non-uniformly distributed. The first issue can be alleviated by setting the histogram regression estimate to zero on empty bins, i.e., by extending the definition in Equation (3.7) with the special case \( \hat{ρ}^1(p) = 0 \) if \( B(p) = \emptyset \) (see, e.g., Nobel 1996). The second and third issue can be alleviated by choosing bins, not necessarily of the same size, in a principled way using existing methods for selecting bins of histogram (regression) estimators (see, e.g., Scott 2015, and the references therein).

### 3.5 Reliability Diagrams

Reliability diagrams (Bröcker and Smith 2007; A. H. Murphy and Winkler 1977, 1987; Sanders 1963) also visualize estimates of the calibration function \( ρ^1 \) for binary models. However, in contrast to the approach in the previous chapter, in a reliability diagram not only the conditional target frequencies but also the predicted probabilities are discretized.

Again, the unit interval is partitioned into non-overlapping bins, often of equal size. Let \( B \) be a non-empty bin. As above in Equation (3.7), the conditional target
3.5. RELIABILITY DIAGRAMS

Figure 3.4: Estimates of the calibration functions of models A, B, and C based on 500 randomly sampled input-target pairs. The inputs are sampled from the uniform distribution (left column) and the Beta distribution Beta(5, 1) (right column). The predicted probabilities are partitioned into 1 (top row), 5 (middle row), and 10 (bottom row) right-open bins of equal size. The marks indicate the values at the discontinuities of the histogram regression estimates.
Chapter 3. Calibration

frequency \( \hat{\rho}_B^1 \) in bin \( B \) is estimated by the empirical frequency of the targets for the predictions falling into this bin,

\[
\hat{\rho}_B^1 = \frac{\sum_{i=1}^{n} y_i \mathbb{1}[P_{1,i} \in B]}{\sum_{i=1}^{n} \mathbb{1}[P_{1,i} \in B]}.
\]

That is, for all \( p \in B \), \( \hat{\rho}_B^1(p) = \hat{\rho}_B^1 \).

Additionally, in a reliability diagram the bin \( B \) is represented by a single prediction: The arithmetic mean of \( B \), \( \int_B p \, dp \), or, arguably better, the average of the predictions in \( B \) (Bröcker and Smith 2007),

\[
\hat{\rho}_B^1 = \frac{\sum_{i=1}^{n} P_{1,i} \mathbb{1}[P_{1,i} \in B]}{\sum_{i=1}^{n} \mathbb{1}[P_{1,i} \in B]}.
\]

Plotting \( \hat{\rho}_B^1 \) versus \( \hat{\rho}_B^1 \), as recommended by Bröcker and Smith (2007), ensures that for a calibrated model the representative prediction and the empirical target frequency converge to the same value as the number of samples goes to infinity.

Figure 3.5 shows reliability diagrams for models A, B, and C estimated from the \( n = 500 \) randomly sampled input-target pairs used for estimating the calibration function in the previous section. As already visible in the estimates of the calibration function in Figure 3.4, the quality of the estimates depends on the number of bins and the distribution of the inputs, and hence of the predicted probabilities. With a single bin, the deviations of the calibration function of model B from the principal diagonal are averaged out. With reduced number of samples per bin, by increasing the number of bins or modifying the distribution of predictions, the bias of the estimators decreases whereas the variance increases (Bröcker 2008). The stochastic nature of the samples and the resulting randomness of the estimators implies that even for the calibrated model A the points in the reliability diagram do not lie on the diagonal. If you would not know that model A is calibrated, would you believe that its deviations in the reliability diagram indicate that it is uncalibrated? If not, would you that the deviations of model B which seem comparable in the case of uniformly distributed inputs are caused by a miscalibrated model?

The partitioning of the predictions and the randomness of the samples both make the interpretation of the reliability diagrams more challenging.

Infinite Sample Size

Consider first the implications of partitioning the predictions. The effect of the discretization is visualized in the “exact” reliability diagrams in the infinite sample size limit in Figure 3.6: For each bin \( B \), the exact conditional target frequency \( \mathbb{E}(Y \mid P_1 \in B) \)—the infinite data limit of \( \hat{\rho}^B \)—is plotted against the mean predicted probability \( \mathbb{E}(P_1 \mid P_1 \in B) \)—the infinite data limit of \( \hat{\rho}^B \).

As expected, if an infinite number of samples is available the estimate of the calibration function improves with increasing number of bins. Generally, the

44
Figure 3.5: Reliability diagrams for models A, B, and C based on 500 randomly sampled input-target pairs. The inputs are sampled from the uniform distribution (left column) and the Beta distribution Beta(5, 1) (right column). The predicted probabilities are partitioned into 1 (top row), 5 (middle row), and 10 (bottom row) right-open bins of equal size.
estimates of the calibration function are closer to the actual calibration function than the estimates in Figure 3.5 based on 500 samples. For uniformly distributed inputs, apparently both model A and model B yield estimates on the diagonal in the infinite limit. In this case the deviations of the calibration function of model B for predicted probabilities smaller and greater than 50% cancel out exactly. This examples can be generalized to less trivial settings with multiple bins in which the calibration function deviates positively and negatively from the diagonal but both contributions are canceled out. Thus, due to the partitioning in general it is possible even in the infinite sample size limit that all points of an uncalibrated reliability diagram come to lie on the principal diagonal.

**Finite Sample Size**

Unfortunately, however, the infinite sample size limit is unrealistic. In contrast, typically each bin only contains a finite, and possibly sometimes rather small, number of predictions. In the infinite sample size limit, the estimates depend only on the choice of the binning scheme, and hence in particular they are not random for a fixed binning scheme. However, if the number of samples is finite, the randomness of the samples propagates to the estimates $\hat{\rho}^1_B$ and $\hat{\rho}^2_B$ of the conditional target frequencies and mean predicted probabilities, respectively. Thus, the estimation of the calibration function conditional target frequencies becomes more challenging.

For simplicity, assume for now that the predictions are fixed. Moreover, as usual, assume that the targets are conditionally independent given the predicted probabilities. Then in each bin the estimator of the target frequency is distributed according to a scaled Poisson-binomial distribution (see Box 3.2 for a brief introduction to the Poisson-binomial distribution): If a bin $B$ contains $m$ predicted probabilities with distinct indices $b_1, \ldots, b_m \in \{1, \ldots, n\}$, then the scaled estimate of the conditional target frequency is distributed according to

$$m\hat{\rho}^1_B \mid P^{1,1}, \ldots, P^{1,n} \sim \text{PoiBinom}\left(\rho^1(P^{1,b_1}), \ldots, \rho^1(P^{1,b_m})\right),$$

given the predictions $P^{1,1}, \ldots, P^{1,n}$. If the model is calibrated, the expression simplifies to

$$m\hat{\rho}^1_B \mid P^{1,1}, \ldots, P^{1,n} \sim \text{PoiBinom}\left(P^{1,b_1}, \ldots, P^{1,b_m}\right).$$

### Box 3.2: The Poisson-Binomial Distribution

The Poisson-binomial distribution is the distribution of a sum of independent but not necessarily identically distributed Bernoulli random variables. It is parameterized by the number $n \geq 0$ of Bernoulli random variables and their success probabilities $p_1, \ldots, p_n \in [0, 1]$, and has discrete support $[0,1,\ldots,n]$. The distribution appears in the analysis of calibration (DeChant and Moradkhani 2015) but also different other fields such as genetics (Melton et al. 2015) and ecology (Calabrese et al. 2014).
3.5. **Reliability Diagrams**

![Reliability Diagrams](image)

Figure 3.6: Reliability diagrams for models A, B, and C in the infinite sample limit. The input is distributed according to the uniform distribution (left column) and the Beta distribution Beta(5, 1) (right column). The predicted probabilities are partitioned into 1 (top row), 5 (middle row), and 10 (bottom row) right-open bins of equal size.
A random variable \( Z \sim \text{PoiBinom}(p_1, \ldots, p_n) \) has mean \( \mathbb{E}Z = \sum_{i=1}^{n} p_i \) and variance \( \text{var} \, Z = \sum_{i=1}^{n} p_i(1 - p_i) \). The evaluation of the probability mass function and the cumulative distribution function are computationally more expensive as they involve summing over sets of subsets of the Bernoulli trials. For instance, the probability mass function is given by

\[
P(Z = k) = \sum_{S \subseteq \{1, \ldots, n\} : |S| = k} \prod_{i \in S} p_i \prod_{i \in S^c} (1 - p_i), \quad k \in \{0, 1, \ldots, n\}.
\]

Common approaches for efficiently computing the probability mass function and the cumulative distribution function are based on the discrete Fourier transform (Biscarri, Zhao, and Brunner 2018; Fernandez and S. Williams 2010; Hong 2013) and recursive algorithms (Barrett and Gray 2014; Biscarri, Zhao, and Brunner 2018; Thomas and Taub 1982). Alternatively, the Poisson-binomial distribution can be approximated with a Poisson distribution (Le Cam 1960), binomial distribution (Choi and Xia 2002), or normal distribution (Berry 1941; Volkova 1996).

The binomial distribution \( \text{Binom}(n, p) \) can be viewed as a special instance of the Poisson-binomial distribution with identical success probabilities \( p_1 = \cdots = p_n = p \). As shown in Figure 3.7, the additional degrees of freedom of the Poisson-binomial distribution make it possible to tune mean and variance separately, whereas the binomial distribution (shown in green) is fully determined by the number of Bernoulli trials and its expected value.

![Figure 3.7: Probability mass function of four Poisson-binomial distributions with \( n = 10 \) Bernoulli trials and mean \( \mathbb{E}Z = 4 \).](image)

The results are illustrated in Figure 3.8 for the calibrated model \( A \) with the fixed sets of \( n = 500 \) inputs, and hence fixed sets of predicted probabilities of model \( A \), used in the previous sections. For each bin \( B \), the mean\(^9\) and the 95\% highest-density region (Hyndman 1996), i.e., the smallest region with 95\% coverage, of the

\(^9\)Note that the mean is equal to the average of the predictions, \( \hat{P}_B \).
estimator $\hat{\rho}_B^1$ are plotted. The visualization indicates that with increasing number of bins the bias of the estimator decreases but the variance increases. Similarly, the distribution of the predicted probabilities affects the distribution, and in particular the variance, of the estimator.

In Figure 3.8 it was exploited that model A is calibrated and hence Equation (3.8) can be applied. The same analysis for models B and C requires to know the calibration function which is not known in practice (otherwise the estimation would not be needed). However, even in such cases Equation (3.8) can be valuable: It allows you to reason about and visualize the distribution of $\hat{\rho}_B^1$ if the predictions are fixed and the model would be calibrated. Thus, it can be used to study whether an estimated deviation from the diagonal is (un)likely if the model is calibrated.

Interestingly, except for the approach proposed by DeChant and Moradkhani (2015), most calibration assessments are based not on Equation (3.8) but instead on its binomial approximation

$$m\hat{\rho}_B^1 | P^{1,1}, \ldots, P^{1,n} \approx \text{Binom}(m, \hat{\rho}_B^1).$$

For instance, Bröcker and Smith (2007) suggested evaluating the deviation from the principal diagonal in a reliability diagram relative to this approximation and plotting the resulting quantile level on probability paper. The binomial approximation corresponds to the assumption that the probability of reference class 1 is the same for all targets in a bin $B$ and approximated by the representative predicted probability $\hat{\rho}_B^1$. However, in particular if the number $m$ of data points in a bin is small, it can differ noticeable from the Poisson-binomial distribution. While the mean of both distributions is equal to $m\hat{\rho}_B^1$ the variance of the binomial approximation is always greater than or equal to the Poisson-binomial distribution, as can be shown by Jensen’s inequality (see Box 3.3).\(^\text{10}\)

### Box 3.3: Variance of the Binomial Approximation

**Lemma 3.2.** Let $n \in \mathbb{N}$ and $p_1, p_2, \ldots, p_n \in [0, 1]$. Define $Z \sim \text{PoiBinom}(p_1, \ldots, p_n)$ and $\hat{Z} \sim \text{Binom}(n, \sum_{i=1}^{n} p_i/n)$. Then

$$\forall Z \leq \forall \hat{Z}$$

with equality if and only if $p_1 = p_2 = \cdots = p_2$.

**Proof.** The non-linear function $f: \mathbb{R} \to \mathbb{R}$, $f(x) := x(1 - x)$, is concave on $\mathbb{R}$. Thus, Jensen’s inequality implies that

$$\forall Z = \sum_{i=1}^{n} f(p_i) \leq n f\left(\frac{1}{n} \sum_{i=1}^{n} p_i\right) = \forall \hat{Z}$$

\(^\text{10}\)As discussed by DeChant and Moradkhani (2015), who gave an alternative proof, this implies that the power of statistical tests based on the Poisson-binomial distribution is greater than the power of tests based on the binomial approximation.
Figure 3.8: Means and 95% highest-density regions of the estimators $\hat{p}_B^1$ of the conditional target frequencies for model A with fixed sets of $n = 500$ randomly sampled inputs. The inputs are distributed according to the uniform distribution (left column) and the Beta distribution Beta(5, 1) (right column). The predicted probabilities are partitioned into 1 (top row), 5 (middle row), and 10 (bottom row) right-open bins of equal size.
with equality if and only if \( p_1 = p_2 = \cdots = p_n \).

Probably it is less well-known that the same relation holds for all higher order moments (Hoeffding 1956, Theorem 3) (at least I learned about these results only recently while writing my thesis). There exists also a similar statement based on probabilities (Hoeffding 1956, Theorem 5): If \( p_1, \ldots, p_n, Z, \text{ and } \tilde{Z} \) are defined as in Lemma 3.2 and \( a \) and \( b \) are two integers with \( 0 \leq a \leq \sum_{i=1}^{n} p_i \leq b \leq n \), then

\[
P(a \leq Z \leq b) \geq P(a \leq \tilde{Z} \leq b)
\]

with equality if and only if \( p_1 = p_2 = \cdots = p_n \), unless \( a = 0 \) and \( b = n \).

### Extensions

Figure 3.8 shows that the variance of the estimator of the conditional frequency is increased if the predicted probabilities are non-uniformly distributed. This effect can be reduced by using bins of (approximately) equal mass, i.e., with (approximately) equal number of predictions, instead of bins of equal size (Bröcker and Smith 2007). Unfortunately, even with bins of equal mass, reliability diagrams still vary depending on the numbers of bins. Recently, Dimitriadis, Gneiting, and Jordan (2021) proposed a method for creating stable and statistically consistent reliability diagrams to address this issue. Although commonly the calibration function is estimated using histogram regression, less popular alternatives such as, e.g., kernel density estimation (Bröcker 2008) exist as well and can help to overcome such binning-related issues (see also the discussion in Chapter 4).

Per se, a reliability diagram does not contain any information about the distribution of the predicted probabilities themselves. Therefore, sometimes the number of samples in each bin is indicated, e.g., by the size of the scatter points (Hagedorn, Doblas-Reyes, and Palmer 2005) or by annotating them (A. H. Murphy and Winkler 1977; Sanders 1963), or the distribution of predictions is plotted as well (see, e.g., Dimitriadis, Gneiting, and Jordan 2021; Guo et al. 2017). The latter is also known as calibration diagram (Bröcker and Smith 2007), but this term does not seem to be used consistently in the literature. Dimitriadis, Gneiting, and Jordan (2021) suggested adding estimates of a decomposition of the Brier score to the reliability diagram as well.

Additionally, Bröcker and Smith (2007) proposed to add so-called consistency bars to reliability diagrams. These bars highlight the range of estimates of the conditional target frequency that are expected from a calibrated model. They are approximated by consistency resampling, a two-step parametric bootstrap method that repeatedly performs the following two sampling steps: Predictions are bootstrapped by sampling with replacement from the given dataset, and then in a second step for each prediction a consistent target is sampled from the prediction. Alternat-
tively, instead of classic reliability diagrams, Bröcker and Smith (2007) suggested plotting the “distance in probability from the 50% quantile” of the estimator \( \hat{\rho}^B \) of the conditional target frequency on probability paper\(^{11} \), assuming the model was calibrated. When creating such plots on reliability paper, Bröcker and Smith (2007) calculated the distance in probability based on a binomial approximation of the second step of the consistency resampling procedure to improve computational efficiency.

### 3.6 Calibration Errors

Reliability diagrams are visual aids for inspecting the empirical (mis-)calibration of binary probabilistic predictive models. In contrast, calibration errors quantify miscalibration and reduce the deviation of the calibration function from the identity function to a real-valued statistic. The most common calibration errors within the statistics and machine learning community are defined as \( L^p \)-distances \( (p \in [1, \infty]) \) between the left- and right-hand side of the calibration condition in Definition 3.1, or rather Corollary 3.1, the special case for binary models (see, e.g., Blattenberger and Lad 1985; Filho et al. 2023; Guo et al. 2017; Kull, Perello Nieto, et al. 2019; Kumar, Liang, and Ma 2019; A. H. Murphy 1973; Naeini, Cooper, and Hauskrecht 2015; Nixon et al. 2019; Seidenfeld 1985).

**Definition 3.4** \( (L^p\text{-Calibration Errors for Binary Models}) \). Let \( m(\cdot) \) be a binary probabilistic predictive model with predicted probabilities \( P^1 \) (see Definition 2.2). Let \( p \in [1, \infty] \). The \( L^p \)-calibration error of model \( m(\cdot) \) is defined as

\[
\| \rho^1(P^1) - P^1 \|_{L^p} = \begin{cases} 
\left( \mathbb{E} |\rho^1(P^1) - P^1|^p \right)^{1/p} & \text{if } 1 \leq p < \infty, \\
\text{ess sup} |\rho^1(P^1) - P^1| & \text{if } p = \infty.
\end{cases}
\]

If \( p = 1 \), the calibration error is called expected calibration error (ECE). If \( p = \infty \), the calibration error is called maximum calibration error (MCE).

The \( L^p \)-calibration error is zero if and only if a model is calibrated according to Corollary 3.1. This can also be observed in Table 3.1 which lists the ECE, \( L^2 \)-calibration errors, and MCE of the three models studied above. Note that the results are also consistent with the ordering expected from Hölder’s inequality: For a fixed model and distribution of predictions, the \( L^p \)-calibration error is a monotonically increasing function of \( p \).

\(^{11}\)Probability paper is a “graph paper scaled so that the cumulative distribution function for a specified probability distribution will be a straight line” (Clapham and Nicholson 2009). As used by Bröcker and Smith (2007), calibrated predictions form a straight line and a reliability diagram on probability paper shows the “distance in probability of the observed relative frequencies from that expected for a reliable forecast system”.

52
3.6. Calibration Errors

Table 3.1: ECE, $L^2$-calibration error, and MCE for models A, B, and C. The input is distributed according to the uniform distribution Unif(0, 1) and the beta distribution Beta(5, 1).

<table>
<thead>
<tr>
<th>Model</th>
<th>$L^1$ (ECE)</th>
<th>$L^2$</th>
<th>$L^\infty$ (MCE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unif</td>
<td>Beta</td>
<td>Unif</td>
</tr>
<tr>
<td>A</td>
<td>0.000 00</td>
<td>0.000 00</td>
<td>0.000 00</td>
</tr>
<tr>
<td>B</td>
<td>0.166 67</td>
<td>0.076 34</td>
<td>0.045 45</td>
</tr>
<tr>
<td>C</td>
<td>0.083 33</td>
<td>0.078 35</td>
<td>0.059 15</td>
</tr>
</tbody>
</table>

Score Divergences

As discussed above, scoring rules such as the Brier score or the logarithmic score do not solely measure calibration but can trade off calibration for resolution (Bröcker 2009). However, the reliability term in the decomposition in Equation (3.1) measures only calibration and is not affected by changes in resolution. Actually, the $L^2$-calibration error is closely related to the reliability term in the standard decomposition of the Brier score:

$$REL^{BS} = \mathbb{E} d^{BS}(\mathbf{P}, \mathbb{P}_Y|\mathbf{P}) = 2 \mathbb{E} \left( \rho^1(\mathbf{P}^1) - \mathbf{P}^1 \right)^2 = 2 \left\| \rho^1(\mathbf{P}^1) - \mathbf{P}^1 \right\|_L^2.$$

The reliability term in the decomposition motivates also the definition of more general ECEs. Since it measures only calibration, it can be used to define an ECE-like calibration error for general probabilistic predictive models of the form in Definition 3.1 as

$$\mathbb{E} d(\mathbf{P}, \mathbb{P}_Y|\mathbf{P})$$

where $d$ is a score divergence (or some other statistical discrepancy). For discrete target spaces $Y = \{y_1, \ldots, y_m\}$, this construction yields the calibration error

$$\mathbb{E} d^{BS}(\mathbf{P}, \mathbb{P}_Y|\mathbf{P}) = \sum_{i=1}^m \left\| \mathbb{P}_Y|\mathbf{P}(y_i) \mid \mathbf{P} - \mathbf{P}(\{y_i\}) \right\|_L^2$$

(3.9)

based on the Brier score.

Estimation

Since $L^1$-calibration errors are based on the calibration function, typically their estimation procedure is quite similar to the estimation of the calibration function discussed above. Calibration errors are also estimated from, supposedly identically and independently distributed, samples of predictions and observed targets, typically with the same “binning and counting” approach (Bröcker 2008). The predictions are binned and the expectations in Definition 3.4 (or the supremum
Table 3.2: Estimates of the ECE, $L^2$-calibration error, and MCE for models A, B, and C based on 500 randomly sampled input-target pairs. The inputs are sampled from the uniform distribution Unif(0, 1) and the beta distribution Beta(5, 1). The predicted probabilities are partitioned into 10 right-open bins of equal size.

<table>
<thead>
<tr>
<th>Example</th>
<th>$L^1$ (ECE)</th>
<th>$L^2$</th>
<th>$L^\infty$ (MCE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.04714</td>
<td>0.01166</td>
<td>0.05407 0.02398</td>
</tr>
<tr>
<td>B</td>
<td>0.15553</td>
<td>0.07441</td>
<td>0.18150 0.09617</td>
</tr>
<tr>
<td>C</td>
<td>0.09416</td>
<td>0.08193</td>
<td>0.10395 0.08830</td>
</tr>
</tbody>
</table>

for $p = \infty$) are estimated as the empirical average (or maximum) of the deviations $|\hat{\beta}_B - \beta_B|$ in each bin $B$, weighted by the number of samples per bin. Of course, also the calibration error estimates are highly dependent on the choice, and in particular number, of bins. To improve estimation when the distribution of predicted probabilities is non-uniform, Nixon et al. (2019) suggested to use bins of equal mass.

For the squared $L^2$-error, i.e., the reliability term of the Brier score, bias-reduced estimators exist (Bröcker 2012; Ferro and Fricker 2012). They come, however, at the cost of increased variance (Siegent 2014). Alternatives to the classic decomposition of the Brier score yield different reliability terms which can also serve as a calibration statistic (Dimitriadis, Gneiting, and Jordan 2021; Kull and Flach 2015; Siegert 2017).

**Examples**

Table 3.2 lists the estimates of the ECE, $L^2$-calibration error, and MCE for models A, B, and C obtained from the $n = 500$ randomly sampled input-target pairs that also the calibration function and reliability diagrams were estimated from in the previous sections. Even when neglecting the challenges involved in estimating the calibration errors, an immediate problem is that the estimates are non-zero also for the calibrated model A. Generally, a non-zero estimate does not necessarily imply that a model is uncalibrated.

**Extensions and Alternatives**

A natural generalization of the $L^p$-calibration errors to multi-class classification would be

$$\|\hat{\mathbf{P}}(\mathbf{P}) - \mathbf{P}\|_{L^p} = \|\hat{\mathbf{P}}_{Y \mid \mathbf{Y} \mid \mathbf{P}(\mathbf{P})} - \mathbf{P}\|_{L^p}$$  \hspace{1cm} (3.10)

where $\mathbf{P}$ denotes the probability vector of a categorical distribution $P$ in the probability simplex. For $p = 2$, this error would be the square root of the calibration error based on the Brier score in Equation (3.9).
Multiple authors suggested extensions of the calibration errors to multi-class classification that are different from this generalization and less strict. Kumar, Liang, and Ma (2019) proposed a “top-label calibration error” \( (p = 2) \) that is equivalent to an \( L^2 \)-calibration error of a reduced binary model with only the most-confident predictions. Thus, as the “confidence-ECE” \( (p = 1) \) and “confidence-MCE” \( (p = \infty) \) proposed by Kull, Perello Nieto, et al. (2019) and Filho et al. (2023) it measures whether a model is “confidence-calibrated” (Kull, Perello Nieto, et al. 2019).

Additionally, Filho et al. (2023), Kull, Perello Nieto, et al. (2019), and Kumar, Liang, and Ma (2019) proposed calibration errors that measure whether a model is “classwise-calibrated” (Kull, Perello Nieto, et al. 2019), a weaker notion of multi-class calibration that goes back to at least DeGroot and Fienberg (1981) (“marginally well-calibrated”) and Zadrozny and Elkan (2002). These classwise calibration errors are called “classwise-ECE” \( (p = 1) \) (Kull, Perello Nieto, et al. 2019), “marginal calibration error” \( (p = 2) \) (Kumar, Liang, and Ma 2019), and “classwise-MCE” (Filho et al. 2023). These calibration errors are almost identical to the generalization in Equation (3.10). The crucial difference is that in Equation (3.10) the conditional probabilities are computed with respect to \( P \) whereas in the classwise-ECE, marginal calibration error, and classwise-MCE the conditioning is performed with respect to \( P(\hat{y}_i) \), i.e., the predicted probability of the \( i \)th class.

Instead of computing the average of the deviation for each class, one can also study the contribution of the \( j \)th class to the classwise-ECE separately with the so-called “class-j-ECE” (Kull, Perello Nieto, et al. 2019). This can be viewed as the ECE, in the sense of Definition 3.4, of a reduced binary model of class \( j \) versus the other classes.

For confidence-calibration, Kumar, Sarawagi, and Jain (2018) proposed the maximum mean calibration error as an alternative to the \( L^p \)-calibration errors based on kernels (see Chapter 4). Its estimators are differentiable and hence can be included as a penalty term in the gradient-based optimization procedures. It can be viewed as a special case of the kernel calibration errors proposed in Paper II, as discussed in the paper.

### 3.7 Calibration Tests

Calibration errors such as the \( L^p \)-calibration errors are real-valued statistics that reveal in a compact way whether a model is calibrated or not. Since a model is calibrated if and only if its \( L^p \)-error is zero, a quick look at Table 3.1 is sufficient to know that model A represents a calibrated model whereas all other examples correspond to uncalibrated models. Additionally, since the statistic is real-valued its estimators can, e.g., be added—at least hypothetically—to the objective function when optimizing a model. Clearly, the same can not be done with plots of the calibration function.

However, there is one big caveat: Table 3.1 is computed by numerical integration
(for \( p = 1 \) and \( p = 2 \)) and optimization (for \( p = \infty \)) of the absolute deviation \(|\hat{\rho}^1(\cdot) - \cdot|\) of predicted probabilities and conditional target frequencies. Thus, the table shows the exact values of the calibration errors (modulo round-off errors and numerical inaccuracies)! These calculations can only be performed accurately since the calibration function is known. In practice, however, the \( IP \)-calibration errors have to be estimated from samples of predictions and targets. Unfortunately, typically the estimates of the calibration errors, such as observed in the previous section and listed in Table 3.2, are non-zero even for calibrated models. Additionally, generally it is unclear when a non-zero calibration error indicates that a model is actually uncalibrated: The calibration error estimates for model B are also rather small but does this indicate that the model is uncalibrated?

The same problems surfaced also in the discussion of reliability diagrams in Section 3.5. As visualized in Figure 3.8, the stochastic nature of the empirical data implies that the estimates, and hence the reliability diagrams, can look very different for two different sets of samples. As Bröcker and Smith 2007, Figure 3 convincingly shows, due to this randomness also merely comparing two reliability diagrams by how close their points are to the diagonal can be misleading: The empirical deviation from the diagonal displayed by a calibrated model can be larger than the deviation of an uncalibrated model.

Thus, a principled analysis of calibration and solid interpretation of calibration errors should take into account that the estimator is randomly distributed even when a model is calibrated. Statistical hypothesis tests of calibration, so-called calibration tests, are designed to solve this problem.

**Definition 3.5** (Calibration Tests). Let \( m(\cdot) \) be a probabilistic predictive model (see Definition 2.1). A calibration test is a statistical hypothesis test that tests the null hypothesis

\[
H_0 : \text{“model } m(\cdot) \text{ is calibrated”}
\]

that model \( m(\cdot) \) is calibrated with samples of predictions and observations of the target.

Calibration tests of this general form do not make any parametric assumption about the underlying distributions of inputs, predictions, and targets, and hence belong to the class of nonparametric, or more precise distribution-free, hypothesis tests (Hollander, Wolfe, and Chicken 2013, p. 2). Typically, the calibration null hypothesis \( H_0 \) is tested against the natural alternative hypothesis

\[
H_1 : \text{“model } m(\cdot) \text{ is not calibrated”}
\] (3.11)

but other choices are possible as well by, e.g., restricting the space of considered calibration functions.

Equivalently, the calibration hypothesis can also be formulated based on a real-valued statistic. For instance, since a binary model is calibrated if and only if
its expected calibration error is zero the calibration hypothesis for a binary model can be formulated as

\[ H_0 : \text{ECE} = 0 \]  \hspace{1cm} (3.12)

by means of the expected calibration error ECE of the model.

As all hypothesis tests, calibration tests can be viewed as decision rules (see, e.g., Casella and Berger 2008, Chapter 8; Schervish 1995, Chapter 4): They specify for which samples the null hypothesis \( H_0 \) is rejected, and for which it is not. Commonly, the rejection region, also called critical region, is based on a function \( T \) of the samples, the test statistic. For instance, a calibration test based on 500 samples

\[ \{(p^{1,i}, y^i)\}_{i=1}^{500} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{p^1, Y} \]

could use an estimator of a calibration error as test statistic, say,

\[ T_{L^2}(D) = \text{"L}^2\text{-calibration error estimated from } D \text{ with 10 bins of equal size"}, \]  \hspace{1cm} (3.13)

and define the critical region as

\[ \{D = \{(p^{1,i}, y^i)\}_{i=1}^{500} \mid T_{L^2}(D) > 0.6\}. \]

Such a calibration test would reject the calibration hypothesis \( H_0 \) whenever the \( L^2 \)-calibration error estimate is greater than 0.6. Per se, it is unclear though if this test is a reasonable decision rule or if you should rather use a different test.

Calibration tests based on the \( IP \)-calibration error have been proposed in the literature (see, e.g., Lee et al. 2022, and references therein), but the history of calibration tests reaches back to at least the publication by Cox (1958) in which calibration tests for logistic-regression models are discussed. A whole line of research is devoted to the study of calibration tests for logistic-regression models, including the Hosmer-Lemeshow test (Hosmer and Lemeshow 1980) with a chi-squared test statistic of the difference between expected and observed outcomes. The consistency-resampling method by Bröcker and Smith (2007) gives naturally rise to a calibration test as well. DeChant and Moradkhani (2015) proposed calibration tests based on the Poisson-binomial distribution instead of the more common binomial approximation.

**Test Errors**

The quality and validity of a calibration test depends on how often its decision is wrong. There are four possible outcomes of a calibration test, listed in Table 3.3:

1. Null hypothesis is **not rejected** and model is **calibrated** (top left).
2. Null hypothesis is **rejected** and model is **uncalibrated** (bottom right).
3. Null hypothesis is **not rejected** but model is **uncalibrated** (top right).
Table 3.3: Possible outcomes of a calibration test.

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Do not reject</th>
<th>Reject</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calibrated Model</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Uncalibrated Model</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

4. Null hypothesis is rejected but model is calibrated (bottom left).

The last two scenarios, marked in red in Table 3.3, are the ones you have to be concerned about. In both cases the test makes an error, the test result is false. In the third scenario (top right entry in Table 3.3), the test result is false because the test does flag the model as uncalibrated (positive) even though it is calibrated. In the fourth scenario (bottom left entry in Table 3.3), the test result is false because the test does not flag the model as uncalibrated (negative) even though it is calibrated. Thus, in the third scenario the test result is false positive whereas in the fourth scenario it is false negative. In a very non-descriptive way, commonly the error of a false positive test result is called type I error and the error of a false negative result is called type II error (Casella and Berger 2008, p. 382).

Actually, both the probability of a type I error and the probability of a type II error can be expressed in terms of the probability that the test statistic is contained in the critical region: If the model is calibrated, then the probability that test statistic is in the critical region is exactly the probability of a type I error, and if the model is not calibrated, then it is the probability of not making a type II error. Therefore, the probabilities of type I and type II errors can be studied by means of the power function defined by

$$\beta(P) := P(T \in R)$$

for every probability measure $P$ under the null or alternative hypothesis (Schervish 1995, p. 215; Casella and Berger 2008, p. 338).12

Ideally, a calibration test is always correct, i.e., the probabilities of a type I and type II error are both zero. Equivalently, its power function should be zero under the null hypothesis and one under the alternative hypothesis. Unfortunately, typically this can not be achieved and usually even minimizing both errors simultaneously is not possible (Casella and Berger 2008, p. 385). Commonly, you select a test for which the probability of a type I error is upper bounded and hence the probability of false positive test results is controlled. Thus, in calibration testing you control the probability of rejecting the calibration hypothesis for a model that is calibrated.

---

12If the considered probability measures form a parametric family $\{P_\theta\}$, the power function can be specified as a function $\beta(\theta) = P_\theta(T \in R)$ of the parameters.
3.7. Calibration Tests

Clearly, only bounding the probability of a type I error is not sufficient though: A calibration test that never rejects the calibration hypothesis has a zero probability of a type I error but is rather useless since it does not detect any uncalibrated model, implying that the probability of a type II error is 1. Thus, amongst the calibration tests with controlled type I error, you try to select the test with the smallest probability of a type II error.

Some authors such as Schervish (1995, p. 215) and Casella and Berger (2008, p. 358) distinguish between size $\alpha$ tests and level $\alpha$ tests to highlight that in both categories of tests the probability of a type I error is upper bounded by $\alpha$ but that the bound of a level $\alpha$ test can possibly be sharpened, in contrast to the tight bound of size $\alpha$ tests. Formally, a test with power function $\beta(\cdot)$ and the set $\mathcal{P}_0$ of probability measures under the null hypothesis is called a size $\alpha$ test if $\sup_{P_0 \in \mathcal{P}_0} \beta(P_0) = \alpha$, and it is called a level $\alpha$ test if $\sup_{P_0 \in \mathcal{P}_0} \beta(P_0) \leq \alpha$. Thus, level $\alpha$ tests are the union of all tests of size $\alpha' \leq \alpha$.

Another class of tests that ensures that the probability of a type I error is controlled are asymptotic level $\alpha$ tests. These tests are based on the asymptotic distribution of the test statistic in the large sample limit. Therefore, asymptotic level $\alpha$ tests only ensure that the probability of a type I error is upper bounded by $\alpha$ in the limit (Vaart 1998). As noted by Vaart (1998, p. 3), while asymptotic tests are useful in practice, in particular if the distribution of the test statistic under the null hypothesis is analytically intractable, you have to keep in mind that they are only approximations:

Clearly, a theorem that can be interpreted as saying that a statistical procedure works fine for $n \to \infty$ is of no use if the number of available observations is $n = 5$.

$p$-values

As discussed so far, calibration tests report their decision—reject or do not reject the calibration hypothesis—for the pre-specified significance level $\alpha$. However, it is unknown how strongly the decision is supported by the empirical data (Schervish 1995, p. 279):

A common criticism of hypothesis-testing methodology is that the decision to “reject” or “accept” a hypothesis is not informative enough. One should also provide a measure of the strength of evidence in favor of (or against) the hypothesis. The posterior probability of the hypothesis is an obvious candidate to provide the strength of evidence in favor of the hypothesis, but the posterior is not available in a classical analysis. In fact, there is no theory for strength of evidence or degree of support in the classical theory. Instead, some alternatives to testing hypotheses are available.
One alternative that Schervish (1995, p. 279) refers to is the p-value. The p-value is a test statistic of the empirical data $D$ that attains values in $[0, 1]$ and is small if the data is unlikely under the null hypothesis. The p-value $p(\cdot)$ of a test is valid if for every level $\alpha \in [0, 1]$ and every probability measure $P_0 \in \mathcal{P}_0$ under the null hypothesis $P_0(p(D) \leq \alpha) \leq \alpha$ (Casella and Berger 2008, Definition 8.3.26).

Thus, a valid p-value provides more information than the decision for a single pre-specified significance level $\alpha$: For every $\alpha \in [0, 1]$, the level $\alpha$ test “reject $H_0$ if and only if $p(D) \leq \alpha$” can be constructed from the reported p-value. Moreover, “a p-value reports the results of a test on a more continuous scale” (Casella and Berger 2008, p. 397) than the decision “reject $H_0$” or “do not reject $H_0$”.

How can you compute a valid p-value for a calibration test? Assume that inference is based on the empirical data

$$D = \{(P^i, y^i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{P,Y}$$

of predictions and targets, and denote a given set of observations, i.e., a specific realization of $D$, by

$$D^{\text{obs}} = \{(P^{i, \text{obs}}, y^{i, \text{obs}})\}_{i=1}^n.$$ If a calibration test employs an estimator $T$ of a calibration error as test statistic, such that large values indicate miscalibrated models, then a valid p-value (Casella and Berger 2008, Theorem 8.3.27) is

$$p(D^{\text{obs}}) = \sup_{P_0 \in \mathcal{P}_0} P_0(T(D) \geq T(D^{\text{obs}}))$$

where the supremum is taken over the set $\mathcal{P}_0$ of all possible probability measures under the calibration hypothesis. However, taking the supremum over all possible probability measures under the calibration hypothesis, and hence all possible distributions of $D$, is difficult if no additional assumptions on the set of admissible probability measures are introduced.

Under the same assumptions on the test statistic $T$, a more feasible alternative is the p-value

$$p(D^{\text{obs}}) = P_0(T(D) \geq T(D^{\text{obs}}) \mid P^1 = P^{1, \text{obs}}, \ldots, P^n = P^{n, \text{obs}})$$

obtained by conditioning on the predictions with respect to an arbitrary (!) probability measure $P_0$ under the null hypothesis. Note that under the calibration hypothesis, the conditional distribution of $D$ given $P^1, \ldots, P^n$ is independent of the probability measure $P_0 \in \mathcal{P}_0$ since the calibration hypothesis implies that $y^i \sim P^i (i = 1, \ldots, n)$. This p-value is valid as well, as discussed by Casella and Berger (2008, p. 399). It connects back to the discussion in Section 3.5, and the accompanying Figure 3.8, regarding the probable deviations in the reliability diagram under the calibration hypothesis for fixed predictions. One main difference is that the analysis and visualization in Section 3.5 is performed for each bin separately, without summarizing.
the deviations in a single real-valued statistic such as the test statistic $T$ in the formulation of the $p$-value in Equation (3.14).

Typically, the distribution in Equation (3.14) is not available in closed form. It can be approximated, e.g., with samples of the test statistic obtained from datasets of predictions and targets generated with a parametric bootstrap (see, e.g., Efron and Hastie 2016): A dataset $\{(P^{i,\text{obs}}, \tilde{y}^j)\}_{i=1}^n$ of predictions and targets under the calibration hypothesis is bootstrapped by sampling a new target $\tilde{y}^j \sim P^{i,\text{obs}}$ ($i = 1, \ldots, n$) for every observed prediction. Consistency resampling (Brücker and Smith 2007) is a variant of the parametric bootstrap in which as a first step, before sampling new targets, a new set of predictions is bootstrapped by sampling with replacement from the observed predictions.

Examples

After this short introduction to calibration tests, the chapter concludes with an empirical example. For the three models studied already in the previous sections valid $p$-values are estimated from two datasets of input-target pairs. Recall that each dataset contains 500 samples and the inputs are sampled a uniform distribution in the first dataset and a beta distribution Beta(5, 1) in the second. The $p$-values are computed according to Equation (3.14) based on the test statistic $T_{L_2}$ in Equation (3.13). For each model and dataset, the conditional distribution of the test statistic under the calibration hypothesis is approximated with the empirical cumulative distribution function (CDF) of $10^5$ samples of the test statistic, each generated with a parametric bootstrap. The empirical cumulative distribution functions of these estimates are shown in Figure 3.9, together with the estimates from the two datasets. The $p$-value approximations can be deduced visually from these plots and their numerical values are listed in Table 3.4. The plot and the table show that for both distributions of inputs considered the $p$-values of model A are large: With high probability, at least as large values would be observed if the model would be calibrated. Hence, for common significance levels such as $\alpha = 5\%$ or $\alpha = 1\%$, the calibration tests corresponding to these $p$-values do not reject the calibration hypothesis of model A. However, the estimated $p$-values of models B and C are almost zero since the calibration error estimates of the observed data are larger than most bootstrap estimates: If these models would be calibrated, then with close to 100% probability the calibration error estimate would be smaller than the actually observed estimate. Thus, based on these estimates, for almost every level $\alpha \in [0, 1]$ the corresponding level $\alpha$ calibration test rejects the calibration hypothesis.

Finally, note that an important point is deliberately omitted from the discussion in this section: The choice of the test statistic is crucial to maximize the test power and to obtain non-trivial $p$-values. A chi-squared test statistic or estimators of the Brier score or of $L^P$-calibration errors suggest themselves, but a priori it is unclear which estimator and which hyperparameters, such as the binning scheme, should be used. To alleviate the dependency on the binning scheme, adaptive calibration
Figure 3.9: Approximate distribution functions of the test statistic $T_{t_2}$ for models A, B, and C under the calibration hypothesis and the considered distributions of inputs. The inputs are distributed according to a uniform distribution (left column) and a Beta distribution Beta(5, 1) (right column). The calibration error estimates match the results in Table 3.2.
3.7. Calibration Tests

Table 3.4: Approximate p-values for models A, B, and C based on the test statistic $T_{L2}$ in Equation (3.13) and estimated from 500 randomly sampled input-target pairs. The inputs are sampled from the uniform distribution Unif(0, 1) and the beta distribution Beta(5, 1).

<table>
<thead>
<tr>
<th>Model</th>
<th>Unif(0, 1)</th>
<th>Beta(5, 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.53688</td>
<td>0.99703</td>
</tr>
<tr>
<td>B</td>
<td>0.00001</td>
<td>0.00013</td>
</tr>
<tr>
<td>C</td>
<td>0.00086</td>
<td>0.01194</td>
</tr>
</tbody>
</table>

tests (see, e.g., DeChant and Moradkhani 2015; Lee et al. 2022) combine tests with different number of bins, while maintaining the desired significance level.
CHAPTER 4

Kernel

As illustrated in the previous chapter, the calibration function defined in Equation (3.5) is a central object in the analysis of calibration of probabilistic predictions, in particular for models with binary targets. Reliability diagrams but also calibration statistics such as the $L_2$-calibration error and calibration tests with these statistics are all based on an estimator of the calibration function. Commonly, for binary models the histogram-regression estimator in Equation (3.7) is used (“binning and counting” (Bröcker 2008)). A drawback of this estimator is that slight changes in the number of bins can lead to largely different estimates if many predictions are close to the edges of the bins (Dimitriadis, Gneiting, and Jordan 2021). This problem can be avoided by constructing a continuous approximation of the calibration function, without binning the predictions.\(^1\) One such approach is based on kernel density estimation of the cumulative distribution function of the joint distribution of predictions and targets (Bröcker 2008). The resulting estimate of the binary calibration function $\hat{\rho}(p)$ at a probability $p \in [0, 1]$ is a weighted average that takes into account how close (“similar”) the existing predictions are to $p$, known as the Nadaraya-Watson-estimator (Nadaraya 1964; Scott 2015; Watson 1964; Zwanzig and Mahjani 2019):

$$
\hat{\rho}(p) = \frac{\sum_{i=1}^{n} y_{i} k(p, P_{1,i})}{\sum_{i=1}^{n} k(p, P_{1,i})},
$$

where function $k$, called “kernel”, measures the similarity between probabilities $P_{1,i}$ and $p$. The dependency on the binning scheme in the histogram-regression estimator is exchanged for a dependency on the kernel. The kernel estimator inherits some properties of the kernel: The estimator is continuous if the kernel is continuous in

\(^1\)A continuous approximation can be constructed by, e.g., linearly interpolating between the calibration function estimates of each bin with nodes at the average prediction in each bin (Bröcker 2008). However, this approach still leads to unstable estimates if many predictions are close to the boundary of the bins.
CHAPTER 4. KERNEL

the first argument $p$, and it is differentiable if the kernel is differentiable in the first argument $p$.

The binning estimator is a special case of the kernel estimator and can be recovered with the kernel

$$k(p, P^{1,i}) = \mathbb{I}[P^{1,i} \in B(p)].$$

Alternatively, by using a uniform kernel

$$k(p, P^{1,i}) = \frac{\mathbb{I}[|p - P^{1,i}| < h/2]}{h}$$

bins of width $h > 0$ are placed around each prediction $P^{1,i}$ (Silverman (1986) called the corresponding density estimator “the naive estimator”). This kernel requires to select the bin width $h$ instead of a binning scheme $B(\cdot)$. The resulting kernel estimator avoids the instability problems of the binning estimator but is still discontinuous.

Bröcker (2008) placed “Gaussian bumps”\(^2\)

$$k(p, P^{1,i}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(p - P^{1,i})^2}{2\sigma^2}\right)$$

around each prediction, resulting in an infinitely differentiable estimator of the calibration function. The standard deviation $\sigma > 0$ is a bandwidth parameter whose role Bröcker (2008) explained as follows:

The bandwidth controls the smoothness of the estimate, with a large bandwidth giving rather smooth estimates, but diminishing the influence of individual samples, while a small bandwidth has the opposite effect. It should be noted that the bandwidth plays a role similar to that of the bin diameter in the binning and counting approach (somewhat inverse to the number of bins).

Bröcker (2008) selected the bandwidth based on leave-one-out cross-validation (see, e.g., Scott 2015, p. 247): The bandwidth was chosen as the minimizer of the leave-one-out estimate

$$\frac{1}{n} \sum_{i=1}^{n} \left(y^i - \hat{\rho}_{k,-i}^{1}(P^{1,i}) \right)^2$$

of the average squared error $\mathbb{E}(Y - \hat{\rho}_{k}^{1}(P^{1}))$, where $\hat{\rho}_{k,-i}^{1}$ is the Nadaraya-Watson estimator of all but the $i$th sample $(P^{1,i}, y^i)$. As noted by Bröcker (2008), the average squared error is actually the Brier score, divided by a factor of 2.

As intended, in all these examples the kernel plays the role of a similarity measure: The closer probabilities $p$ and $P^{1,i}$, the larger $k(p, P^{1,i})$. But what exactly defines a “kernel”? And can kernels also be used to not only compare real-valued predictions but more general objects such as probability distributions? The following sections are devoted to answering these questions.

\(^2\)Bandwidth parameter $\delta$ in the formulation by Bröcker (2008, Eq. (20)) corresponds to $\sigma/\sqrt{2}$. 

66
4.1 Positive Definite Kernel

Motivation

How would you check if two real-valued vectors \( x^*, x^o \in \mathbb{R}^m \) \((m \geq 1)\) are similar? Of course, you can compute their Euclidean distance \( \|x^* - x^o\|_2 \): If the distance is close to or even equal to 0, then it seems justified to call the vectors similar. Instead of the Euclidean distance you could also use other metrics such as \( p \)-norm distances with \( p \neq 2 \).\(^3\) However, a similarity measure does not necessarily have to be a metric: For instance, the squared Euclidean distance does not satisfy the triangle inequality and hence is not a metric (see Box 4.1).

Box 4.1: Metric

Let \( S \) be a non-empty set. A function \( d : S \times S \rightarrow [0, \infty) \) is called a metric if it satisfies the following properties for all \( x, y, z \in S \) (see, e.g., Christmann and Steinwart 2008, Definition A.2.2):

1. \( d(x, y) = 0 \) if and only if \( x = y \) (coincidence),
2. \( d(x, y) = d(y, x) \) (symmetry), and
3. \( d(x, z) \leq d(x, y) + d(y, z) \) (triangle inequality).

If \( d(\cdot, \cdot) \) is a metric on \( S \), the pair \( (S, d) \) is called a metric space.

Another possible similarity measure is the inner product

\[
\langle x^*, x^o \rangle_2 = \sum_{i=1}^m x_i^* x_i^o.
\]

It is not a metric but quantifies similarity based on the direction of \( x^* \) and \( x^o \): As illustrated in Figure 4.1 for \( m = 2 \) dimensions, the inner product satisfies

\[
\langle x^*, x^o \rangle_2 = \cos(\angle(x^*, x^o)) \|x^*\|_2 \|x^o\|_2,
\]

where \( \angle(x^*, x^o) \) is the angle enclosed by \( x^* \) and \( x^o \). This implies that

\[
-\|x^*\|_2 \|x^o\|_2 \leq \langle x^*, x^o \rangle_2 \leq \|x^*\|_2 \|x^o\|_2,
\]

which also follows more generally from the Cauchy-Schwarz inequality, with the following special cases:

- If \( x^* \) and \( x^o \) are collinear and have the same direction, i.e., \( x^o = cx^* \) for some \( c > 0 \), then \( \langle x^*, x^o \rangle_2 = \|x^*\|_2 \|x^o\|_2 \) and hence the inner product is maximized amongst all vectors with the same norm as \( x^* \) and \( x^o \).

\(^3\)For \( p \geq 1 \), the \( p \)-norm distance of \( x^*, x^o \in \mathbb{R}^m \) is \( \|x^* - x^o\|_p := \left( \sum_{i=1}^m |x_i^* - x_i^o|^p \right)^{1/p} \).
Figure 4.1: Sketch of the relation between the inner product of two non-zero vectors $x^*, x^0 \in \mathbb{R}^2$ and the enclosed angle $\alpha = \angle(x^*, x^0)$. For simplicity, $x^*$ and $x^0$ are chosen as unit vectors, the general statement follows by appropriate scaling.

- If $x^*$ and $x^0$ are collinear and have the opposite direction, i.e., $x^0 = cx^*$ for some $c < 0$, then $\langle x^*, x^0 \rangle_2 = -\|x^*\|_2\|x^0\|_2$ and hence the inner product is minimized amongst all vectors with the same norm as $x^*$ and $x^0$.

- If $x^*$ and $x^0$ are orthogonal, then $\langle x^*, x^0 \rangle_2 = 0$.

Inner products can be defined not only on $\mathbb{R}^m$ but on more general vector spaces, including infinite-dimensional spaces such as the space of square-summable sequences and the space of square-integrable functions (see Box 4.2). These similarity measures can be generalized even further by composing the inner product with a feature map $\phi$: Inputs $x^*$ and $x^0$ are mapped to points $\phi(x^*)$ and $\phi(x^0)$ in a pre-Hilbert space $H$, and the similarity between $x^*$ and $x^0$ is computed as $\langle \phi(x^*), \phi(x^0) \rangle_H$.

Feature maps were already encountered in Chapter 2 and Figure 2.8. While there no assumptions were made about the feature space in the context of feature engineering, here it has to be a pre-Hilbert space, as visualized in Figure 4.2.

**Box 4.2: Hilbert Space**

Let $H$ be a $\mathbb{K}$-vector space, i.e., either an $\mathbb{R}$-vector space or a $\mathbb{C}$-vector space. A function $\langle \cdot, \cdot \rangle : H \times H \to \mathbb{K}$ is called an inner product if it satisfies the following properties for all $x, y, z \in H$ and numbers $\alpha, \beta \in \mathbb{K}$ (see, e.g., Christmann and Steinwart 2008, Definition A.5.8):

1. $\langle x, x \rangle > 0$ if $x \neq 0$ and $\langle x, x \rangle = 0$ if $x = 0$ (positive definiteness),
Figure 4.2: Illustration of the feature map $\phi$. The feature space is a pre-Hilbert space and hence has more structure than the generic feature space shown in Figure 2.8.

2. $\langle x, y \rangle = \overline{\langle y, x \rangle}$ (conjugate symmetry),

3. $\langle ax + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$ (linearity).

If $\langle \cdot , \cdot \rangle$ is an inner product on $H$, the pair $(H, \langle \cdot , \cdot \rangle)$ is called a pre-Hilbert space.

An inner product $\langle \cdot , \cdot \rangle$ induces the norm $\|x\| = \sqrt{\langle x, x \rangle}$ and thereby also the metric $d(x, y) = \|x - y\| = \sqrt{\langle x - y, x - y \rangle}$. A pre-Hilbert space that is complete with respect to the induced metric is called a Hilbert space.

**Basic Theory**

The following definition formalizes the notion of “kernels” and covers even complex-valued kernels since symbol $\mathbb{K}$ represents either $\mathbb{R}$ or $\mathbb{C}$.

**Definition 4.1** (Kernel (Christmann and Steinwart 2008, Definition 4.1)). Let $S$ be a non-empty set. A function $k : S \times S \to \mathbb{K}$ is called a kernel on $S$ if there exists a $\mathbb{K}$-Hilbert space $H$ and a map $\phi : S \to H$ such that for all $x, y \in S$

$$k(x, y) = \langle \phi(x), \phi(y) \rangle.$$  

Function $\phi$ is called a feature map and $H$ a feature space of $k$.

In general, the feature map and feature space are not unique (see, e.g., Minh, Niyogi, and Yao 2006; Christmann and Steinwart 2008, p. 113): For instance, the
function $k : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}, k(x, y) \equiv xy$, can be written as

$$k(x, y) = \langle x, y \rangle_2$$

for $x, y \in \mathbb{R}$, and therefore it is a kernel on $\mathbb{R}$ with feature space $\mathbb{R}$ and the identity function $\text{id}_\mathbb{R}$ as feature map. However, for all $x, y \in \mathbb{R}$, the function can also be expressed as

$$k(x, y) = \left\langle \frac{x}{\sqrt{2}}, \frac{y}{\sqrt{2}} \right\rangle_2,$$

and thus it can also be defined with feature space $\mathbb{R}^2$ and feature map $x \mapsto x/\sqrt{21_2}$.

As both the intuitive motivation and the formal Definition 4.1 illustrate, a kernel maps its arguments with an—often non-linear—feature map to a feature space in which an inner product can serve as similarity measure. Many applications within statistics and machine learning such as support vector machines (Christmann and Steinwart 2008), Gaussian processes (Rasmussen and C. K. I. Williams 2006), and density estimation (Silverman 1986), however, do not use the representation in the feature space explicitly but are only based on the evaluations of the inner product in the feature space. Often the kernel can be evaluated efficiently even though the feature space is infinite-dimensional. Thus, it can be hugely beneficial to work only with the kernel and never evaluate the feature map explicitly, if possible. This approach is known as the kernel trick in machine learning (K. P. Murphy 2022, Section 17.3.4; Christmann and Steinwart 2008, p. 111).

Kernels can be characterized not only as an inner product on a feature space. For an alternative characterization in Theorem 4.1, first a new definition is required (unfortunately, the literature does not fully agree on it, see Footnote 5).

**Definition 4.2** (Positive Definite Function (Christmann and Steinwart 2008, Definition 4.15; Berlinet and Thomas-Agnan 2004, Definition 2)). Let $S$ be a non-empty set. A function $k : S \times S \rightarrow \mathbb{K}$ is called positive definite\footnote{Some authors also use the names “positive definite kernel” (Berg 1984, Definition 3.1.1), “positive type function” (Berlinet and Thomas-Agnan 2004, Definition 1), and “positive semi-definite” function (Christmann and Steinwart 2008, p. 117; Berg 1984, p. 66). Another source of disagreement is that some authors (see, e.g., Christmann and Steinwart 2008, Definition 4.15) call real-valued functions positive definite if Equation (4.1) holds for real coefficients $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ whereas others demand that Equation (4.1) holds for complex coefficients $\alpha_1, \ldots, \alpha_n \in \mathbb{C}$ as well (see, e.g., Berlinet and Thomas-Agnan 2004, Lemma 4; Berg 1984, Section 3.1.6).} if for all $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{C}$, and $x_1, \ldots, x_n \in S$,

$$\sum_{i,j=1}^n \alpha_i \overline{\alpha_j} k(x_i, x_j) \in \mathbb{R}_{\geq 0}. \quad (4.1)$$

The definition implies the following connection between positive definite functions and positive semi-definite matrices: A function $k : S \times S \rightarrow \mathbb{K}$ on a non-empty
set $S$ is a positive definite function if and only if for all $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in S$, the matrix\(^6\)

$$(k(x_i, x_j))_{i,j} \in \mathbb{C}^{n \times n}$$

is positive semi-definite (as a complex matrix) (Berg 1984, Remark 3.1.2). For real-valued functions, positive definiteness can be defined also without involving complex numbers:

**Corollary 4.1** (Real-Valued Positive Definite Function (Berlinet and Thomas-Agnan 2004, Lemma 4; Berg 1984, Section 3.1.6)). *Let $S$ be a non-empty set. A function $k : S \times S \to \mathbb{R}$ is positive definite\(^7\) if and only if $k$ is symmetric (i.e., for all $x, y \in S$, $k(x, y) = k(y, x)$) and if for all $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, and $x_1, \ldots, x_n \in S$,

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0.$$  

Obviously every kernel is positive definite (see, e.g., Berlinet and Thomas-Agnan 2004, Lemma 1): For all $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{C}$, and $x_1, \ldots, x_n \in S$,

$$\sum_{i,j=1}^{n} \alpha_i \overline{\alpha_j} k(x_i, x_j) = \sum_{i,j=1}^{n} \alpha_i \overline{\alpha_j} (\phi(x_j), \phi(x_i))_H$$  

$$= \sum_{i,j=1}^{n} \langle \overline{\alpha_j} \phi(x_j), \overline{\alpha_i} \phi(x_i) \rangle_H = \left\| \sum_{i=1}^{n} \overline{\alpha_i} \phi(x_i) \right\|_H^2 \geq 0.$$  

Interestingly the reverse implication holds as well:

**Theorem 4.1** (Positive Definite Kernel (cf. Christmann and Steinwart 2008, Theorem 4.16; Berg 1984, Section 3.1.6)). *Let $S$ be a non-empty set. A function $k : S \times S \to \mathbb{K}$ is a kernel if and only if $k$ is positive definite.*

Roughly, to show that a positive definite function $k$ is a kernel you construct a pre-Hilbert space

$$H_{\text{pre}} := \left\{ \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \middle| n \in \mathbb{N}, \alpha_1, \ldots, \alpha_n \in \mathbb{K}, x_1, \ldots, x_n \in S \right\}$$  \hspace{1cm} (4.2)

with inner product

$$\langle f, g \rangle := \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \overline{\beta_j} k(y_j, x_i)$$  \hspace{1cm} (4.3)

for $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \in H_{\text{pre}}$ and $g = \sum_{j=1}^{m} \beta_j k(\cdot, y_j) \in H_{\text{pre}}$. You have to show that the definition in Equation (4.3) is independent of the representation of $f$ and

---

\(^6\)If function $k$ is a kernel, this matrix is known as the *kernel matrix*. Definition 4.1 implies that a kernel matrix is a matrix of pairwise inner products, which explains why it is also commonly called the *Gram matrix* (see, e.g. Christmann and Steinwart 2008, p. 117).

\(^7\)Christmann and Steinwart (2008, Definition 4.15) do not require a real-valued positive definite function to be symmetric.
g and that it satisfies the properties of an inner product. Finally, you take the completion \( H \) of \( \mathcal{H}_{\text{pre}} \) as feature space and define the feature map as a composition of the mapping \( x \mapsto k(\cdot, x) \) from \( S \) to \( \mathcal{H}_{\text{pre}} \) and the isometric embedding of \( \mathcal{H}_{\text{pre}} \) in \( H \).

Theorem 4.1 provides an important alternative characterization of kernels. It implies that you do not necessarily have to find a feature space and feature map to show that a function is a kernel. Instead, you can prove that the function is positive definite which sometimes can be a simpler task. For instance, using this approach it is easy to show that the pointwise limit of kernels is a kernel as well (Christmann and Steinwart 2008, Corollary 4.17).

**Examples**

Kernels can be combined in multiple ways to construct new kernels (see, e.g., Duvenaud 2014; Rasmussen and C. K. I. Williams 2006, Section 4.2.4; Christmann and Steinwart 2008, Chapter 4.1). If you are familiar with properties of positive semi-definite matrices, the link between kernels and positive semi-definite matrices in Theorem 4.1 makes it easy to see that the following approaches yield kernels.

Let \( k_1 : S \times S \to \mathbb{K} \), \( k_2 : S \times S \to \mathbb{K} \), and \( k_3 : T \times T \to \mathbb{K} \) be kernels on non-empty sets \( S \) and \( T \). Let \( f : S \to T \) and \( \alpha \in \mathbb{R}_{\geq 0} \). Then also the following functions are positive definite kernels (Christmann and Steinwart 2008, Lemma 4.5, Lemma 4.6):

- \( \alpha k_1 \) (scaling) and \( k_1 + k_2 \) (sum).
- \( k_1 k_2 \) (product)\(^8\) and \( k_1 \otimes k_3 \) (tensor product)\(^9\).
- \( k_3(f(\cdot), f(\cdot)) \) (composition/embedding).

So what are concrete examples of kernels? You saw already a few kernels at the beginning of this chapter. Here is a small selection of other kernels on \( S = \mathbb{R}^d \) \((d \in \mathbb{N})\), visualized in Figure 4.3 for the case \( d = 1 \):

**Constant Kernel** Arguably the simplest kernel is the constant kernel

\[
k_c^{\text{CONST}}(x, y) := c
\]

for some constant \( c \geq 0 \). It can be constructed with the feature space \( H = \mathbb{R} \) and the feature map \( \phi \equiv \sqrt{c} \).

\(^8\) Positive definiteness follows from the Schur product theorem (Schur 1911, Theorem VII); interestingly, the implication for positive definite kernels is even discussed in Schur’s publication (Theorem VII*).

\(^9\) The tensor product kernel \( k_1 \otimes k_3 : (S \times T) \times (S \times T) \to \mathbb{K} \) is defined by \( (k_1 \otimes k_3)((s, t), (s', t')) := k_1(s, s')k_3(t, t') \) for all \( s, s' \in S \) and \( t, t' \in T \).
4.1. Positive Definite Kernel

**White Noise Kernel** Another rather simple kernel is the *white noise kernel*

\[ k^\text{EYE}(x, y) := \|x - y\|. \]

It can be constructed with the feature space \( H = L^2(\mathbb{R}^d) \) of equivalence classes of square-integrable real-valued functions on \( \mathbb{R}^d \) and the feature map \( \phi : \mathbb{R} \rightarrow L^2(\mathbb{R}^d) \), \( \phi(s) := \delta_s \). Clearly, the formulation of the positive definite kernel as an inner product on the infinite-dimensional Hilbert space \( L^2(\mathbb{R}^d) \) is surprisingly complex given the simplicity of the explicit kernel function.

**Linear Kernel** If the set \( S \) is already a Hilbert space with inner product \( \langle \cdot, \cdot \rangle_S \), such in these examples with \( S = \mathbb{R}^d \), then you can choose \( S \) as the feature space and the identity function \( \text{id}_S \) as feature map. This yields the *linear kernel* (Christmann and Steinwart 2008, p. 115)

\[ k^\text{LIN}(x, y) := \langle x, y \rangle_2. \]

**Gaussian Kernel** The *Gaussian kernel*, also called exponentiated quadratic kernel and squared-exponential kernel (K. P. Murphy 2022, Section 17.1), “is probably the most widely-used kernel within the kernel machines field” (Rasmussen and C. K. I. Williams 2006, Section 4.2). It is defined by

\[ k^\text{EQ}_\ell(x, y) := \exp\left(-\frac{\|x - y\|^2}{2\ell^2}\right) \]

with bandwidth \( \ell > 0 \). The kernel can be represented (Steinwart, Hush, and Scovel 2006, Lemma 2) as an inner product on the feature space \( H = L^2(\mathbb{R}^d) \) of square-integrable real-valued functions on \( \mathbb{R}^d \), composed with the feature map

\[ \phi : \mathbb{R}^d \rightarrow L^2(\mathbb{R}^d), \quad \phi(x) := \left(\frac{2}{\pi \ell^2}\right)^{d/4} \exp\left(-\frac{\|x - \cdot\|^2}{\ell^2}\right). \]

As discussed at the beginning of this chapter, the value of the bandwidth parameter \( \ell \) is an important choice. Decreasing the bandwidth reduces \( k(x, y) \) for \( x \neq y \), and hence their implied similarity, with the limit

\[ \lim_{\ell \rightarrow 0} k^\text{EQ}_\ell(x, y) = \|x - y\| = k^\text{EYE}(x, y). \]

On the other hand, by increasing the bandwidth the kernel function \( k(x, y) \) increases as well, with the limit approaching

\[ \lim_{\ell \rightarrow \infty} k^\text{EQ}_\ell(x, y) = 1 = k_1^\text{CONST}(x, y). \]
**Matérn Kernels**  Another family of kernels I would like to mention\(^{10}\) are *Matérn kernels*, named after the Swedish mathematician Bertil Matérn (Stein 1999, p. 31). These kernels are of the form (Matérn 1960, p. 18; Handcock and Wallis 1994; Stein 1999, p. 50; Rasmussen and C. K. I. Williams 2006, pp. 84–85)

\[
k_{\nu,\ell}^{\text{MA}}(x, y) := \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \frac{\sqrt{2\nu} \| x - y \|_2}{\ell} \right)^\nu K_\nu \left( \frac{\sqrt{2\nu} \| x - y \|_2}{\ell} \right)
\]

(4.4)

with bandwidth \( \ell > 0 \) and degree parameter \( \nu \geq 0 \).\(^{11}\) Function \( K_\nu(\cdot) \) denotes the modified Bessel function of the second kind of order \( \nu \) (Abramowitz and Stegun 1972, Section 9.6). The degree parameter \( \nu \) is directly linked to the smoothness of the kernel as \( t \mapsto (t/\ell)^\nu K_\nu(t/\ell) \) (\( t \geq 0 \)) is \( 2m \) times differentiable if and only if \( \nu > m \) (Stein 1999, p. 32). Stein (1999) (“Use the Matérn model”) argues that this property makes the Matérn kernel a more suitable model for the covariance function of random fields than the Gaussian kernel:

> The smoothness of a random field plays a critical role in interpolation problems. Furthermore, there is often no basis for knowing a priori the degree of smoothness of some physical process modeled as a random field. Thus, it is prudent to use classes of models that allow for the degree of smoothness to be estimated from the data rather than restricted a priori. The Matérn model does allow for great flexibility in the smoothness of the random field while still keeping the number of parameters manageable.

For half-integer degrees \( \nu = m + 1/2 \ (m \in \mathbb{N}_0) \), the Matérn kernel can be written as\(^{12}\)

\[
k_{m+1/2,\ell}^{\text{MA}}(x, y) = \exp \left( -\frac{\sqrt{2m+1} \| x - y \|_2}{\ell} \right) \cdot \frac{m!}{(2m)!} \sum_{l=0}^{m} \frac{(m + l)!}{l!(m - l)!} \left( \frac{2\sqrt{2m+1} \| x - y \|_2}{\ell} \right)^{m-l},
\]

---

\(^{10}\)Not only but also since this PhD thesis is written at a Swedish university.

\(^{11}\)Matérn (1960, p. 18) used the form \( (x, y) \mapsto 2^{1-\nu}(\| x - y \|_2/\ell)^\nu K_\nu(\| x - y \|_2/\ell) / \Gamma(\nu) \) but as discussed by Stein (1999, p. 49) this leads to undesirable asymptotics as \( \nu \to \infty \). Thus, commonly the form in Equation (4.4) is preferred (see, e.g., Handcock and Wallis 1994; Stein 1999, p. 50; Rasmussen and C. K. I. Williams 2006, pp. 84–85) which avoids this problem and converges to the Gaussian kernel as \( \nu \to \infty \).

\(^{12}\)The result is well-known (see, e.g., Rasmussen and C. K. I. Williams 2006, Equation (4.16)). It can be derived in a straightforward way using the identities \( K_{m+1/2}(z) = \sqrt{\pi} / (2z) \sum_{n=0}^{m} [(m + l)! / (l!(m - l)!)] (2z)^{-1} \) (Abramowitz and Stegun 1972, Equation 10.2.15) and \( \Gamma(m + 1/2) = (2m)! / (m! 2^{2m}) \sqrt{\pi} \) (Abramowitz and Stegun 1972, Equations 6.1.8 and 6.1.12) for \( m \in \mathbb{N}_0 \).
4.1. Positive Definite Kernel

![Kernel Functions](image)

Figure 4.3: Different kernel functions on $\mathbb{R}$, with second argument fixed to 1.

i.e., as the product of an exponential and a polynomial of $\|x - y\|_2$. A notable special case of the Matérn kernel is the exponential kernel (Rasmussen and C. K. I. Williams 2006, p. 85) obtained for $\nu = 1/2$:

$$k_{1/2, \ell}^{MA}(x, y) = \exp\left( -\frac{\|x - y\|_2}{\ell} \right).$$

The Matérn kernel is also related to the Gaussian kernel which is recovered as the degree approaches infinity and the Matérn kernel becomes infinitely differentiable (Rasmussen and C. K. I. Williams 2006, p. 85):

$$\lim_{\nu \to \infty} k_{\nu, \ell}^{MA}(x, y) = k_{\ell}^{EQ}(x, y).$$

Operator-Valued Kernels

All kernels considered so far are real- or complex-valued functions. In applications such as learning of vector-valued functions (Micchelli and Pontil 2005), however, it is desirable to consider kernels with more general co-domains. When learning vector-valued functions, in principle you can learn each component independently and reduce the problem of learning one vector-valued function to the problem of learning multiple scalar-valued functions. If the components are not independent though, e.g., if they represent pixels in an image, this procedure is suboptimal as it does not take into account the relation between the components (Micchelli and Pontil 2005). The correlation structure between components can be captured by matrix-valued kernels $K(\cdot, \cdot)$ where value $K(\cdot, \cdot)_{i,j}$ in the $i$th row and $j$th column represents the relation between the $i$th and the $j$th component.

More generally, you can consider operator-valued kernels, i.e., kernels whose co-domain is a space of linear operators. They can be defined analogously to scalar-valued kernels in Definition 4.1.\textsuperscript{13}

\textsuperscript{13}As for scalar-valued kernels, operator-valued kernels can be introduced and defined in multiple equivalent ways. In this thesis, operator-valued kernels are defined analogously to scalar-valued kernels in Definition 4.1, and therefore the definition by Carmeli, De Vito, and Toigo (2006, Definition 2.2) is broken up into two equivalent definitions in Definitions 4.3 and 4.4.
**Chapter 4. Kernel**

**Definition 4.3** (Operator-Valued Kernel). Let $S$ be a non-empty set and $\mathcal{K}$ be a $\mathbb{K}$-Hilbert space. Denote the Banach space of bounded linear operators on $\mathcal{K}$ by $\mathcal{L}(\mathcal{K})$.

A function $K: S \times S \to \mathcal{L}(\mathcal{K})$ is called an $\mathcal{L}(\mathcal{K})$-valued kernel on $S$ if there exists a $\mathbb{K}$-Hilbert space $H$ and a map $\phi: S \to \mathcal{L}(\mathcal{K}, H)$ such that for all $x, y \in S$,

$$K(x, y) = \phi(x)^*\phi(y),$$

where $\mathcal{L}(\mathcal{K}, H)$ is the Banach space of bounded linear operators from $\mathcal{K}$ to $H$ and $\phi(x)^*$ is the Hermitian adjoint\(^{14}\) of $\phi(x)$.

By choosing $\mathcal{K} = \mathbb{K}$, you recover Definition 4.1 of scalar-valued kernels as a special case. You can also adapt Definition 4.2 to operator-valued functions. In this context, some authors use the term “positive type” (Berg 1984; Carmeli, De Vito, and Toigo 2006) and some prefer the term “non-negative” (Kadri et al. 2016; Lian 2007; Saha and Palaniappan 2020), whereas typically the term “positive definite” refers to a subset of the former (Lian 2007; Saha and Palaniappan 2020).

**Definition 4.4** (Operator-Valued Non-Negative Function (Carmeli, De Vito, and Toigo 2006, Definition 2.2)). Let $S$ be a non-empty set and $\mathcal{K}$ be a $\mathbb{K}$-Hilbert space.

A function $K: S \times S \to \mathcal{L}(\mathcal{K})$ is called non-negative if for all $n \in \mathbb{N}$, $\alpha_1, \ldots, \alpha_n \in \mathbb{C}$, and $x_1, \ldots, x_n \in S$,

$$\forall y \in \mathcal{K}: \sum_{i,j=1}^{n} \alpha_i\overline{\alpha_j}(K(x_i, x_j)y, y)_{\mathcal{K}} \in \mathbb{R}_{\geq 0}.$$

Again, the scalar analog of this definition in Definition 4.2 can be obtained by setting $\mathcal{K} = \mathbb{K}$. Similar to Corollary 4.1, operator-valued non-negative functions for real vector spaces ($\mathbb{K} = \mathbb{R}$) can be defined in terms of real coefficients but as in the scalar case additionally you have to assume that the functions are symmetric, i.e., for all $x, y \in S$, $K(x, y) = K(y, x)$ (Carmeli, De Vito, and Toigo 2006, p. 382). Probably it does not come at a surprise that Definitions 4.3 and 4.4 are equivalent. Hence, “kernel”, “non-negative function”, and “non-negative kernel” may be used interchangeably.

**Theorem 4.2** (Operator-Valued Non-Negative Kernel). Let $S$ be a non-empty set and $\mathcal{K}$ be a $\mathbb{K}$-Hilbert space. A function $K: S \times S \to \mathcal{L}(\mathcal{K})$ is an $\mathcal{L}(\mathcal{K})$-valued kernel if and only if $K$ is non-negative.

\(^{14}\)The Hermitian adjoint $\phi(x)^*$ of $\phi(x)$ is the uniquely defined operator in $\mathcal{L}(H, \mathcal{K})$ that satisfies for all $u, v \in \mathcal{K}$, $\langle \phi(x)u, v \rangle_H = \langle u, \phi(x)^*v \rangle_{\mathcal{K}}$. 

76
4.1. Positive Definite Kernel

**Proof.** Again, it is clear that every kernel is non-negative: For all \( n \in \mathbb{N}, \alpha_1, \ldots, \alpha_n \in \mathbb{C}, x_1, \ldots, x_n \in S, \) and \( y \in \mathcal{K}, \)

\[
\sum_{i,j=1}^{n} \alpha_i \overline{\alpha_j} \langle K(x_i, x_j) y, y \rangle_{\mathcal{K}} = \sum_{i,j=1}^{n} \alpha_i \overline{\alpha_j} \langle \phi(x_j) y, \phi(x_j) y \rangle_H
\]

\[
= \sum_{i,j=1}^{n} \langle \overline{\alpha_j} \phi(x_j) y, \overline{\alpha_i} \phi(x_i) y \rangle_H = \left\| \sum_{i=1}^{n} \overline{\alpha_i} \phi(x_i) y \right\|_{H}^{2} \geq 0.
\]

The reverse implication follows from Carmeli, De Vito, and Toigo (2006, p. 380 and Proposition 2.3).

The following statement provides an interesting connection between operator-valued kernels and scalar-valued kernels.

**Proposition 4.1** (Tilde Correspondence (Carmeli, De Vito, and Toigo 2006, Proposition 2.7)) pedrickTheoryReproducingKernels1957. *Let \( S \) be a non-empty set and \( \mathcal{K} \) be a \( \mathbb{K} \)-Hilbert space. Let \( K : S \times S \to \mathcal{L}(\mathcal{K}) \) be an \( \mathcal{L}(\mathcal{K}) \)-valued kernel on \( S. \) Define the function \( k : (S \times \mathcal{K}) \times (S \times \mathcal{K}) \to \mathbb{K} \) by

\[
k((s, \kappa), (s', \kappa')) := \langle K(s, s') \kappa', \kappa \rangle_{\mathcal{K}}
\]

for all \( s, s' \in S \) and all \( \kappa, \kappa' \in \mathcal{K}. \) Then \( k \) is a \( \mathbb{K} \)-valued kernel.

Does this imply that it is sufficient to study scalar-valued kernels and that operator-valued kernels are superfluous? As noted by Carmeli, De Vito, and Toigo (2006, p. 384) and Pedrick (1957), this link between operator-valued kernels and scalar-valued kernels is weaker than it might seem initially. For instance, in general proofs cannot be simplified by referring to the corresponding scalar-valued kernel, and the ground spaces \( S \) and \( S \times \mathcal{K} \) might not share the same properties (Carmeli, De Vito, and Toigo 2006, p. 384).

Examples of matrix- and operator-valued kernels are given, e.g., by Álvarez, Rosasco, and Lawrence (2012), Kadri et al. (2016), and Micchelli and Pontil (2005). Particularly simple operator-valued kernels are kernels of the form

\[
K(x, y) = k(x, y) \text{id}_{\mathcal{K}} \in \mathcal{L}(\mathcal{K})
\]

where \( k(\cdot, \cdot) \) is a scalar-valued kernel. These are instances of a larger class of *separable kernels* (Álvarez, Rosasco, and Lawrence 2012; Kadri et al. 2016; Micchelli and Pontil 2005) of the form

\[
K(x, y) = k(x, y)A \in \mathcal{L}(\mathcal{K})
\]

where \( k(\cdot, \cdot) \) is a scalar-valued kernel and \( A \) is a bounded linear operator in \( \mathcal{L}(\mathcal{K}). \)

For instance, if \( \mathcal{K} = \mathbb{R}^{m} \) \( (m \in \mathbb{N}), \) then \( A \) may be an \( m \times m \)-dimensional symmetric positive semi-definite matrix. In such separable kernels “the contribution of input
[in set $S$] and output [in space $\mathcal{K}$] is decoupled” (Álvarez, Rosasco, and Lawrence 2012). If $A$ is chosen as the identity operator, then each component in $\mathcal{K}$ is assumed to be independent, an assumption that might be too strict sometimes and can be alleviated by choosing an operator $A$ that is able to model relations between components in $\mathcal{K}$ (Kadri et al. 2016).

## 4.2 Reproducing Kernel Hilbert Space

In the previous sections, it was argued that kernels are similarity measures that can be viewed and constructed as an inner product in a feature space which therefore has to be a Hilbert space. The Euclidean vector spaces $\mathbb{R}^d$ $(d \in \mathbb{N})$ are well-known finite-dimensional Hilbert spaces and feature spaces of, e.g., the constant kernel $k_c^{\text{CONST}}$ and the linear kernel $k^\text{LIN}$ defined in the previous section. Examples of infinite-dimensional Hilbert spaces are $l^2$-spaces of countably infinite square-summable sequences and $L^2$-spaces of equivalence classes of square-integrable functions, which could serve as feature spaces for the white noise kernel $k^\text{EYE}$ and the Gaussian kernel $k^\text{EQ}$.

Another particularly interesting class of Hilbert spaces of real- or complex-valued functions are reproducing kernel Hilbert spaces.

**Definition 4.5** (Reproducing Kernel Hilbert Space (Christmann and Steinwart 2008, Definition 4.18; Berlinet and Thomas-Agnan 2004, Definition 1)). Let $(H, \langle \cdot, \cdot \rangle)$ be a Hilbert space of $\mathbb{K}$-valued functions on a non-empty set $S$. A function $k : S \times S \to \mathbb{K}$ is called a reproducing kernel of $H$ if for all $x \in S$ and $f \in H$:

1. $k(\cdot, x) \in H$, and
2. $f(x) = \langle f, k(\cdot, x) \rangle$ (reproducing property).

The space $H$ is called a reproducing kernel Hilbert space (RKHS) over $S$ if it possesses a reproducing kernel $k : S \times S \to \mathbb{K}$.$^{15}$

Note that the name reproducing “kernel” is justified since every reproducing kernel is a kernel in the sense of Definition 4.1:

**Lemma 4.1** (Christmann and Steinwart (2008, Lemma 4.19)). Let $H$ be an RKHS over a non-empty set $S$ with reproducing kernel $k : S \times S \to \mathbb{K}$.

Then $H$ is a feature space of $k$ with feature map $\phi : S \to H$ defined by

$$\phi(x) := k(\cdot, x)$$

for all $x \in S$. Function $\phi$ is called the canonical feature map.

---

$^{15}$In contrast to Berlinet and Thomas-Agnan (2004, Definition 1), Christmann and Steinwart (2008, Definition 4.18) define an RKHS as a Hilbert space of $\mathbb{K}$-valued functions in which all evaluation functionals are continuous. Theorem 4.4 shows that these definitions are equivalent.
4.2. Reproducing Kernel Hilbert Space

Proof. By the reproducing property, for all $x, y \in S$,

$$k(x, y) = \langle k(\cdot, y)k(\cdot, x) \rangle = \langle k(\cdot, x), k(\cdot, y) \rangle$$

(note that by definition $k(\cdot, x) \in H$ and $k(\cdot, y) \in H$).

From the defining properties of a reproducing kernel, it is clear as well that an RKHS $H$ has a unique reproducing kernel: If $k_1 : S \times S \to \mathbb{k}$ and $k_2 : S \times S \to \mathbb{k}$ are reproducing kernels of an RKHS $(H, \langle \cdot, \cdot \rangle)$ over $S$, then by the reproducing property and the conjugate symmetry of $\langle \cdot, \cdot \rangle$, for all $x, y \in S$

$$k_1(x, y) = \langle k_1(\cdot, y)k(\cdot, x) \rangle = \langle k_2(\cdot, x), k_1(\cdot, y) \rangle = k_2(\cdot, x)(y) = \langle k_2(\cdot, x), k_2(\cdot, y) \rangle = \langle k_2(\cdot, y), k_2(\cdot, x) \rangle = k_2(x, y).$$

Thus, every RKHS has a unique reproducing kernel and this reproducing kernel is a kernel in the sense of Definition 4.1. The famous Moore-Aronszajn theorem (Aronszajn 1943; Moore and Barnard 1935) shows that the reverse implication is true as well: Every kernel is the reproducing kernel of a uniquely defined RKHS.

Theorem 4.3 (Moore-Aronszajn theorem (Aronszajn 1943; Moore and Barnard 1935)). Let $k : S \times S \to \mathbb{k}$ be a kernel on a non-empty set $S$. Then there exists a unique RKHS $H$ of $\mathbb{k}$-valued functions on $S$ with $k$ as reproducing kernel.

The subspace $H_{\text{pre}} \subset H$ spanned by the functions $\{k(\cdot, x) | x \in S\}$ as defined in Equation (4.2) is dense in $H$, and $H$ is the completion of $H_{\text{pre}}$ with the inner product in Equation (4.3).

The important conclusion is: There is a one-to-one correspondence between positive definite kernels and reproducing kernels. Additionally, “the RKHS $H$ of a given kernel $k [...][is] the ‘smallest’ feature space of $k$ in the sense that there is a canonical metric surjection $V$ from any other feature space $H_0$ of $k$ onto $H’” (Christmann and Steinwart 2008, Theorem 4.21).

An RKHS is a Hilbert space of functions but the same is true, e.g., for $L^2$-spaces. However, one main difference between an RKHS and other Hilbert spaces of $\mathbb{k}$-valued functions is:

Theorem 4.4 (Evaluation Functionals Are Continuous (Berlinet and Thomas-Agnan 2004, Theorem 1)). A Hilbert space $H$ of $\mathbb{k}$-valued functions on a non-empty set $S$ is an RKHS if and only if for all $x \in S$ the evaluation functional $e_x : H \to \mathbb{k}$, $e_x(f) := f(x)$, is continuous.

A consequence of this characteristic property is that in an RKHS convergence in the norm also implies pointwise convergence. That is, if two functions in an RKHS are close to each other (with respect to the norm of the RKHS), also all their values are close to each other.

Figure 4.4 shows a classic example that highlights that this property is not satisfied in the Hilbert space $L^2([0, 1])$ of square-integrable functions on the unit
interval. The sequence of monomials \( (f_n)_{n \in \mathbb{N}} \), \( f_n(x) := x^n \), converges to the zero function \( f \equiv 0 \) in \( L^2([0, 1]) \):

\[
\lim_{n \to \infty} \| f_n - f \|_{L^2([0, 1])} = \lim_{n \to \infty} \left( \int_{[0,1]} |x|^{2n} \, d\lambda(x) \right)^{1/2} = \lim_{n \to \infty} (2n + 1)^{-1/2} = 0.
\]

However, the value \( f_n(1) \) of the functions \( f_n \) at \( x = 1 \) is constant and does not converge to \( f(1) = 0 \):

\[
\lim_{n \to \infty} |f_n(1) - f(1)| = \lim_{n \to \infty} 1 = 1.
\]

Multiple properties of a kernel carry over to its RKHS, sometimes under additional assumptions on the domain of the kernel. This includes, e.g., boundedness, measurability, continuity, integrability, and differentiability (Christmann and Steinwart 2008, Chapter 4.3).

### 4.3 Kernel Score

At this point, you might be a bit lost in the theoretical details of kernels and associated RKHSs. It is helpful to recall what the initial goal was: To measure the similarity of mathematical objects such as predicted probabilities in order to be able to account for how close or distant they are when, e.g., estimating the calibration function. The kernel approach defines similarity based on an inner product in a feature space. Fortunately, typically the feature space is not needed explicitly, and it is sufficient to evaluate the kernel function (known as the “kernel trick”).

In this section, the focus shifts to evaluations and comparisons of general probability distributions, not limited to predicted probabilities. The importance of the one-to-one correspondence between kernels and RKHSs will become apparent and hopefully thereby the motivation for the theory on RKHSs will become clearer.

In Chapter 3, scoring rules such as the Brier score (Brier 1950) and the logarithmic score (Good 1952) were presented as a common tool for comparing probabilistic predictions and estimating how well they are aligned with observations of the
targets. Classically, the Brier score is defined for discrete target spaces but, as mentioned in Section 3.3, it can be generalized to other spaces. One such generalization is the kernel score.

**Definition 4.6** (Kernel Score (cf. Steinwart and Ziegel 2021, Definition 1.1)). Let \((\Omega, \mathcal{A})\) be a measurable space and let \(\mathcal{M}_1(\Omega)\) denote the class of all probability measures on \((\Omega, \mathcal{A})\). Let \(k : \Omega \times \Omega \to \mathbb{R}\) be a measurable kernel on \(\Omega\), and define\(^{16}\)

\[
\mathcal{M}_1^k(\Omega) := \left\{ P \in \mathcal{M}_1(\Omega) \left| \int_\Omega \sqrt{k(\omega, \omega)} \, dP(\omega) < \infty \right\} \subset \mathcal{M}_1(\Omega).
\]

The scoring rule \(s_{k}^{\text{KS}} : \mathcal{M}_1^k(\Omega) \times \Omega \to \mathbb{R}\) defined\(^{17}\) by

\[
s_{k}^{\text{KS}}(P, \omega) := \int_\Omega \int_\Omega k(z, \tilde{z}) \, d(P - \delta_\omega)(z) d(P - \delta_\omega)(\tilde{z})
\]

is called the *kernel score* \(s_{k}^{\text{KS}}\) associated with kernel \(k\) on \(\Omega\).

The Brier score \(s_{\text{BS}}\) on the discrete space \(\Omega = \{\omega_1, \ldots, \omega_m\}\) (see Section 3.3) can be recovered as a special instance of the kernel score with the white noise kernel (cf. Gneiting and Raftery 2007, Section 5.1). The *energy score* and the *continuous ranked probability score*, two scoring rules for predictions of continuous distributions, can also be viewed as special instances of the kernel score (Gneiting and Raftery 2007).

**Kernel Mean Embedding**

Recall that (strictly) proper scoring rules ensure that the expected score \(S(P, Q)\) is maximized (only) if \(P\) and \(Q\) are equal. Thus, a strictly proper scoring rule ensures that the expected score \(S(\cdot, Q)\) with fixed distribution \(Q\) attains a unique maximum at \(P\). The uniqueness property is desirable, e.g., when optimizing \(P\) or when comparing two distributions \(P\) and \(Q\) and checking whether they are equal. Thus, usually strictly proper scoring rules are desired, and hence an immediate question is, for which kernels the kernel score \(s_{k}^{\text{KS}}\) is a proper scoring rule with respect to \(\mathcal{M}_1^k(\Omega)\).

To be able to answer this question without having to assume that the kernel is continuous (Gneiting and Raftery 2007, Theorem 4), an object is needed that is widely useful even beyond this specific question in, e.g., two-sample testing (Gretton, Borgwardt, et al. 2012) or tests of conditional independence (Fukumizu, Gretton, Sun, et al. 2007):

\(^{16}\)Note that \(\mathcal{M}_1^k(\Omega) = \mathcal{M}_1(\Omega)\) if kernel \(k\) is bounded.

\(^{17}\)The definition of the kernel score is based on Steinwart and Ziegel (2021, Definition 1.1), modulo a scaling with factor 2. The scaling simplifies the link between the kernel score and the maximum-mean discrepancy, and matches the less general definitions of the kernel score by Dawid (2007) and Gneiting and Raftery (2007) which only considered continuous kernels.
**Definition 4.7** (Kernel Mean Embedding (cf. Steinwart and Ziegel 2021, Theorem 1.1)). Let $k : \Omega \times \Omega \to \mathbb{R}$ be a measurable kernel on $(\Omega, \mathcal{A})$ with reproducing kernel Hilbert space $H$. Function $\Phi : \mathcal{M}_k^1 \to H$ defined by

$$
\Phi(P) := \int_{\Omega} k(\cdot, \omega) \, dP(\omega) = \mathbb{E}_{Z \sim P} k(\cdot, z)
$$

is called the **kernel mean embedding**.

Thus, the kernel mean embedding provides a way to map probability measures in $\mathcal{M}_k^1$ to elements in the RKHS of kernel $k$. Hence, if the kernel mean embedding is injective, then each probability measure can be identified by its kernel mean embedding and hence operations with these measures can be transferred to operations with their mean embeddings in the RKHS. Intuitively, this approach is similar to working with characteristic functions of distributions—but mean embeddings are not based on probability density functions but on expectations of the feature map with respect to the probability measures.

In some cases the kernel mean embedding can be computed in closed form and hence the function in the RKHS that the distribution is identified with can be stated explicitly. For instance, with a Gaussian kernel $k_{\ell^2}^{\text{Gauss}}$ the kernel mean embedding of a normal distribution with mean $\mu$ and standard deviation $\sigma > 0$ is given by

$$
\Phi(\mathcal{N}(\mu, \sigma^2)) = \frac{\ell}{\sqrt{\ell^2 + \sigma^2}} \exp\left(-\frac{(\mu - \cdot)^2}{2(\ell^2 + \sigma^2)}\right),
$$

and of a Laplace distribution with mean $\mu$ and scale parameter $\beta > 0$ by

$$
\Phi(\text{Laplace}(\mu, \beta)) = \sqrt{\frac{\pi \ell}{8 \beta}} \left( \exp\left(\frac{\ell^2}{2\beta^2} + \frac{\cdot - \mu}{\beta}\right) \text{erfc}\left(\frac{\ell}{\sqrt{2\beta}} + \frac{\cdot - \mu}{\sqrt{2\ell}}\right) 
+ \exp\left(\frac{\ell^2}{2\beta^2} + \frac{\cdot - \mu}{\beta}\right) \text{erfc}\left(\frac{\ell}{\sqrt{2\beta}} + \frac{\mu - \cdot}{\sqrt{2\ell}}\right) \right)
$$

where $\text{erfc}(\cdot)$ is the complimentary error function, defined for $x \in \mathbb{R}$ by

$$
\text{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} \, dt.
$$

Figure 4.5 shows the kernel mean embeddings of a Normal and a Laplace distribution with zero mean and unit variance with respect to a Gaussian kernel with bandwidth $\ell = 0.5$. It confirms visually that it is possible to tell apart both distributions in the RKHS when using a Gaussian kernel, even though their first two moments are equal.

Alternatively, if the kernel mean embedding is not available in closed form, its definition as the expectation of the feature map implies that it can be approximated by samples: If you have access to $n$ i.i.d. samples $z_1, \ldots, z_n$ from a distribution $P$, you can approximate the kernel mean embedding $\Phi(P)$ by the Monte Carlo estimate

$$
\hat{\Phi}(P) = \frac{1}{n} \sum_{i=1}^n k(\cdot, z_i).
$$
Figure 4.5: The kernel mean embedding of a normal distribution $N(0, 1)$ and a Laplace distribution Laplace(0, 1/\sqrt{2}) for a Gaussian kernel $k_\ell^Q$ with bandwidth $\ell = 0.5$.

**Kernel Score Divergence**

The kernel mean embedding is closely related to the kernel score, as the following theorem shows.

**Theorem 4.5** (Proper Kernel Score (cf. Steinwart and Ziegel 2021, Theorem 1.1)). Let $k : \Omega \times \Omega \to \mathbb{R}$ be a measurable kernel on $(\Omega, \mathcal{A})$ with reproducing kernel Hilbert space $H$. Then the score divergence $d_k^{KS}$ of the kernel score $s_k^{KS}$ satisfies for all $P, Q \in M_1^k(\Omega)$

$$d_k^{KS}(P, Q) = \|\Phi(P) - \Phi(Q)\|_H^2.$$  

In particular, $s_k^{KS}$ is a proper scoring rule with respect to $M_1^k(\Omega)$, and it is strictly proper if and only if the kernel mean embedding $\Phi$ is injective.

Bounded measurable kernels with injective mean embedding are known as *characteristic kernels* (see, e.g., Fukumizu, Gretton, Schölkopf, et al. 2008; B. Sriperumbudur, Fukumizu, and Lanckriet 2010). Many common kernels, and in particular the Gaussian and Laplacian kernel on $\mathbb{R}^m (m \in \mathbb{N})$, are characteristic (Fukumizu, Gretton, Sun, et al. 2007, Theorem 2). This explains also why the mean embeddings of the normal distribution and the Laplace distribution in Figure 4.5 are different.

### 4.4 Maximum-Mean Discrepancy

In the machine learning literature, the kernel divergence $d_k^{KS}$ is known as the squared maximum mean-discrepancy (MMD) (Gretton, Borgwardt, et al. 2012). Based on Theorem 4.5, it can be rewritten for $P, Q \in M_1^k(\Omega)$ as

$$\text{MMD}_k(P, Q) = (d_k^{KS}(P, Q))^{1/2} = \sup_{f \in H : \|f\|_H \leq 1} \left| \mathbb{E}_{y \sim P} f(y) - \mathbb{E}_{y \sim Q} f(y) \right|. \quad (4.6)$$
Integral Probability Metrics

Statistical distances of the generic form

\[ d(P, Q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{y \sim P} f(y) - \mathbb{E}_{y \sim Q} f(y) \right| \]

where \( \mathcal{F} \) is a space of integrable real-valued functions, are called integral probability metrics (Müller 1997). They can be contrasted with \( \phi \)-divergences of the form

\[ D(P \parallel Q) = \begin{cases} \mathbb{E}_Q \left( \phi(dP/dQ) \right) & \text{if } P \ll Q, \\ \infty & \text{otherwise,} \end{cases} \]

where \( \phi : [0, \infty) \to (-\infty, \infty] \) is a convex function (B. K. Sriperumbudur et al. 2010). Generally, \( \phi \)-divergences—such as the Kullback-Leibler divergence—are not symmetric and require for finite evaluation that \( P \) is absolutely continuous with respect to \( Q \), indicated by \( P \ll Q \). On the other hand, integral probability metrics do not assume absolute continuity.

Besides the maximum-mean discrepancy, other well-known examples of integral probability metrics are the Kantorovich metric (if \( \mathcal{F} \) is taken as the set of Lipschitz continuous function with Lipschitz constant less than or equal to 1) and the total variation distance (\( \mathcal{F} \) is the set of bounded functions with norm less than or equal to 1) (B. K. Sriperumbudur et al. 2010).

Estimation

Assume you have access to two sets of samples, \( \{x^i\}_{i=1}^m \overset{\text{i.i.d.}}{\sim} P \) and \( \{y^i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} Q \), from two not necessarily distinct probability measures \( P \) and \( Q \). How can you estimate from these samples the distance of \( P \) and \( Q \)?

The squared MMD, or equivalently the kernel divergence, is a particularly appealing distance measure since by Theorem 4.5 it can be expressed as (see, e.g., Gretton, Borgwardt, et al. 2012, Lemma 6)

\[
\text{MMD}_k^2(P, Q) = \langle \Phi(P) - \Phi(Q), \Phi(P) - \Phi(Q) \rangle_H \\
= \langle \Phi(P), \Phi(P) \rangle_H - 2\langle \Phi(P), \Phi(Q) \rangle_H + \langle \Phi(Q), \Phi(Q) \rangle_H \\
= \mathbb{E}_{x \sim P, x' \sim P} k(x, x') - 2\mathbb{E}_{x \sim P, y \sim Q} k(x, y) + \mathbb{E}_{y \sim Q, y' \sim Q} k(y, y').
\]

The first and the last term can be estimated with a U-statistic (Vaart 1998, Chapter 12) and the remaining term with a sample average, leading to the unbiased consistent estimator (Gretton, Borgwardt, et al. 2012, Lemma 6)

\[
\frac{2}{m(m-1)} \sum_{1 \leq i < j \leq m} k(x^i, x^j) + \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} k(y^i, y^j) - \frac{2}{mn} \sum_{i=1}^m \sum_{j=1}^n k(x^i, y^j).
\]

Rewriting the squared MMD as

\[
\text{MMD}_k^2(P, Q) = \mathbb{E}_{x \sim P, x' \sim P, y \sim Q, y' \sim Q} h(x, y, (x', y'))
\]
with
\[ h((x, y), (x', y')) = k(x, x') - k(x, y') - k(x', y) + k(y, y') \]
shows that in the case \( m = n \) an alternative unbiased estimator of the squared MMD is the U-statistic
\[ \frac{2}{m(m-1)} \sum_{1 \leq i < j \leq m} h((x_i, y_i), (x_j, y_j)). \]
However, the variance of this estimator is increased due to omitting the terms of the form \( k(x_i, y_i) \) (Gretton, Borgwardt, et al. 2012).

Generally, other unbiased estimators with reduced sample complexity, such as the linear estimator studied by Gretton, Borgwardt, et al. (2012), improve computational efficiency at the cost of increased variance.

As discussed in the context of calibration tests, in some scenarios only estimating the distance between two distributions from their samples is not sufficient. Instead, sometimes principled approaches are desired that take into account the distribution of the estimators and provide better guidance, since as in the calibration setting typically the estimates are non-zero even if the two distributions are equal. Multiple two-sample tests with estimators of the maximum mean discrepancy as test statistic have been proposed, of different sample complexity and test power (Gretton, Borgwardt, et al. 2012). Apart from the estimator, also the choice of the kernel, and in particular of its bandwidth, is crucial for maximizing the power of these tests (Gretton, Sejdinovic, et al. 2012). The median heuristic has been used as a simple guideline (Garreau, Jitkrittum, and Kanagawa 2018; Gretton, Borgwardt, et al. 2012) but other approaches such as optimization over the kernel parameters (Gretton, Sejdinovic, et al. 2012) and aggregated tests (Schrab, Kim, et al. 2022) can yield tests with improved power.

**Witness Function**

This chapter concludes with an example taken from Gretton, Borgwardt, et al. (2012, pp. 729–730) that tries to provide some additional intuition for the maximum mean discrepancy. As seen in Equation (4.6), the maximum mean discrepancy can be formulated as an integral probability metric, and hence it can be viewed as the maximum of an optimization problem in the unit ball of the RKHS. But what is the function in the RKHS for which the maximum is attained?

The function \( f^* \) in the RKHS with \( \|f^*\| \leq 1 \) that maximizes the expression
\[ \mathbb{E}_{z \sim P} f^*(z) - \mathbb{E}_{z \sim Q} f^*(z) = \langle f^*, \Phi(P) - \Phi(Q) \rangle \]
in Equation (4.6) is called the *witness function* (Gretton, Borgwardt, et al. 2012, p. 729). From the Cauchy-Schwarz inequality, it follows immediately that \( f^* \) satisfies
\[ f^* \propto \Phi(P) - \Phi(Q), \]
Figure 4.6: Unnormalized versions of the exact witness function (green, solid line) and its estimate (green, dotted line) maximizing the kernel score of a normal distribution $\mathcal{N}(0, 1)$ (orange, solid line) and a Laplace distribution $\text{Laplace}(0, 1/\sqrt{2})$ (orange, dashed line) for a Gaussian kernel $k_\ell^{\text{GQ}}$ with bandwidth $\ell = 0.5$.

i.e., it is proportional to the difference of the kernel mean embeddings.

Figure 4.5 shows the kernel mean embeddings of a normal distribution and a Laplace distribution with respect to a Gaussian kernel with bandwidth $\ell = 0.5$. We return to this example and now study how the maximum mean discrepancy distinguishes between the two distributions, or rather between both kernel mean embeddings. In Figure 4.6, the witness function (or, to be precise, a function that is proportional to the witness function) is shown for this example: An empirical estimate based on $n = 10^4$ samples from each distribution (cf. Gretton, Borgwardt, et al. 2012, Figure 1) is plotted alongside the exact witness function. As observed by Gretton, Borgwardt, et al. (2012, p. 730), the witness function is “smooth, negative where the Laplace density exceeds the Gaussian density (at the center and tails), and positive where the Gaussian density is larger”.

86
Concluding Remarks

This final chapter of the thesis summarizes the previous chapters and points out possible directions for future research. In case you have not read the included papers yet, I suggest reading them first before turning back to this chapter.

5.1 Conclusion

The main focus of my doctoral research has been the estimation and hypothesis testing of calibration of probabilistic predictive models. It has been driven by my interest in medical applications such as the analysis of pathological images, for which trustworthy and interpretable uncertainty estimates are crucial. Even though I have not worked on such applications myself, I think the theoretical contributions in this thesis are useful for such scenarios: Mathematical and statistical models tend to become more and more complex, and hence methods for assessing model calibration beyond binary classification, such as the ones proposed in papers I-IV, seem increasingly important. Initially, in papers I and II, the calibration evaluation was still focused on (multi-class) classification. The methods proposed in papers III and IV, however, can be applied to more general classes of models, including models for regression and unnormalized models such as energy-based models and Bayesian models. With the public software packages for calibration analysis described in paper V I tried to bridge the gap between my research and its application.

5.2 Outlook

One thing I have noticed during the last few years: There are always additional questions to ask and problems to solve, even if a project is finished or paper accepted. Similarly, even though this thesis marks the end of my doctoral studies, there are many open questions, problems, and ideas left that are unresolved and unexplored.
For instance, I would like if there existed better interpretable calibration tests—or other calibration assessments—that take into account the impact of miscalibration. While we would like a probabilistic predictive model to be calibrated (and refined), probably many models encountered in practice do not live up to this ideal. However, often “a bit of” miscalibration might be acceptable for downstream applications, similarly to how other deficiencies such as suboptimal accuracy are accepted. Thus, instead of testing whether a model is calibrated (and eventually reject the null hypothesis of calibration once the number of samples is large enough) it might be even more useful to test whether a model is not too miscalibrated. Intuitively, I think the admissible degree of miscalibration should be application-specific: If a model is deployed in a safety-critical application, probably calibration should be enforced more strictly than if it is applied in a less critical environment.

Due to time constraints I have also postponed some incremental but probably still worthwhile extensions of my research. For instance, the kernel calibration conditional Stein discrepancy proposed in paper IV is motivated by and formulated for calibration evaluation of unnormalized density models, such as energy-based models and Bayesian models. However, it would be interesting to generalize the approach to other families of models, such as models for discrete target spaces, and to optimize it for specific parametric families such as Gaussian distributions. Similarly, exploration, implementation, and application of optimized—possibly analytically tractable—kernels, beyond the kernels for Gaussian, Laplace, and mixture distributions presented in paper III, might be useful for the kernel calibration error more generally.

Probably, the kernel calibration tests could also be improved by performing aggregated calibration tests in the spirit of Schrab, Guedj, and Gretton (2022) and Schrab, Kim, et al. (2022) with different kernel bandwidths instead of selecting a bandwidth with the median heuristic or by optimization on a held-out dataset. Generally, maintaining the software packages and adding new methods would probably be useful for both practitioners and other researchers, even though such work is not appropriately rewarded in academia.

Hence, many interesting questions I have not been able to answer, or even think about, during my doctoral studies. Probably Feynman (1955) is right and to a large extent science is a continuously evolving process with steadily improving solutions, for days, years, and generations:

We are at the very beginning of time for the human race. It is not unreasonable that we grapple with problems. But there are tens of thousands of years in the future. Our responsibility is to do what we can, learn what we can, improve the solutions, and pass them on.

Certainly, I have learned a lot in the last years, but it is up to others to decide whether I have succeeded in contributing my small share to the existing knowledge base.
Notation List

The following table lists the mathematical notation used in the introductory chapters. It differs from the notation used in the included papers.

### Spaces

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathbb{C} )</td>
<td>Set of complex numbers</td>
</tr>
<tr>
<td>( H )</td>
<td>Hilbert space</td>
</tr>
<tr>
<td>( \mathbb{K} )</td>
<td>Either ( \mathbb{R} ) or ( \mathbb{C} )</td>
</tr>
<tr>
<td>( \mathcal{L}(U) )</td>
<td>Banach space of bounded linear operators on ( U, \mathcal{L}(U, U) )</td>
</tr>
<tr>
<td>( \mathcal{L}(U, V) )</td>
<td>Banach space of bounded linear operators from ( U ) to ( V )</td>
</tr>
<tr>
<td>( \mathbb{N} )</td>
<td>Set of natural numbers, ( {1, 2, \ldots} )</td>
</tr>
<tr>
<td>( \mathbb{N}_0 )</td>
<td>Set of non-negative integers, ( \mathbb{N} \cup {0} )</td>
</tr>
<tr>
<td>( \mathbb{R} )</td>
<td>Set of real numbers</td>
</tr>
<tr>
<td>( \overline{\mathbb{R}} )</td>
<td>Closure of ( \mathbb{R} ), ( \mathbb{R} \cup {-\infty, \infty} )</td>
</tr>
<tr>
<td>( \mathbb{R}_{\geq 0} )</td>
<td>Set of non-negative real numbers, ( [0, \infty) )</td>
</tr>
<tr>
<td>( \overline{\mathbb{R}_{\geq 0}} )</td>
<td>Closure of ( \mathbb{R}<em>{\geq 0} ), ( \mathbb{R}</em>{\geq 0} \cup {\infty} )</td>
</tr>
<tr>
<td>( \mathcal{X} )</td>
<td>Input space, co-domain of ( X )</td>
</tr>
<tr>
<td>( y )</td>
<td>Target space, co-domain of ( Y )</td>
</tr>
</tbody>
</table>
APPENDIX A. NOTATION LIST

Norms

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>( | \cdot |_2 )</td>
<td>Euclidean norm</td>
</tr>
<tr>
<td>( | \cdot |_{L^p} )</td>
<td>( L^p )-norm</td>
</tr>
</tbody>
</table>

Measures and Probability Distributions

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>((\Omega, \mathcal{A}))</td>
<td>Measurable space with ( \sigma )-algebra ( \mathcal{A} )</td>
</tr>
<tr>
<td>((\Omega, \mathcal{A}, \mathbb{P}))</td>
<td>Probability space</td>
</tr>
<tr>
<td>(\mathcal{B}(A))</td>
<td>Borel ( \sigma )-algebra of a topological space ( A )</td>
</tr>
<tr>
<td>(\mu, \nu)</td>
<td>Measures</td>
</tr>
<tr>
<td>(\lambda)</td>
<td>Lebesgue measure on ((\mathbb{R}, \mathcal{B}(\mathbb{R})))</td>
</tr>
<tr>
<td>(\delta_x)</td>
<td>Dirac measure centered on ( x )</td>
</tr>
<tr>
<td>(\mathbb{P}_Z)</td>
<td>Distribution (\mathbb{P}(Z^{-1}(\cdot))) of random variable ( Z )</td>
</tr>
<tr>
<td>(\mathbb{P}_{Z</td>
<td>A})</td>
</tr>
<tr>
<td>(\mathcal{M}_1(A))</td>
<td>Space of probability measures on a measurable space ( A )</td>
</tr>
<tr>
<td>(\mathcal{P})</td>
<td>Set of probability measures</td>
</tr>
<tr>
<td>(P, Q)</td>
<td>Probability measures</td>
</tr>
</tbody>
</table>

Random Variables and Operators

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\mathbb{E} Z, \mathbb{E}_P Z)</td>
<td>Expectation of random variable ( Z ) with respect to, possibly implicit, probability measure ( P )</td>
</tr>
<tr>
<td>(\mathbb{E} f, \mathbb{E}<em>P f, \mathbb{E}</em>{Z\sim P} f(z))</td>
<td>Expectation of function ( f ) with respect to, possibly implicit, probability measure ( P )</td>
</tr>
<tr>
<td>(\mathbb{E}(Z \mid A))</td>
<td>Conditional expectation of random variable ( Z ) given random variable ( A )</td>
</tr>
<tr>
<td>(m(\cdot))</td>
<td>Probabilistic predictive model, measurable function ( m : \mathcal{X} \to \mathcal{P} )</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m^1(\cdot)$</td>
<td>Probabilistic prediction for target 1 of a probabilistic predictive model with binary target space ${0, 1}$, measurable function $m^1 : \mathcal{X} \to [0, 1], m^1(x) = \mathbb{E}(m(x))$</td>
</tr>
<tr>
<td>$\forall Z$</td>
<td>Variance of random variable $Z$</td>
</tr>
<tr>
<td>$X$</td>
<td>Input, $\mathcal{X}$-valued random variable</td>
</tr>
<tr>
<td>$Y$</td>
<td>Target, $Y$-valued random variable</td>
</tr>
<tr>
<td>$Z$</td>
<td>Random variable</td>
</tr>
</tbody>
</table>

**Distributions**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Ber}(p)$</td>
<td>Bernoulli distribution with success probability $p$</td>
</tr>
<tr>
<td>$\text{Binom}(n, p)$</td>
<td>Binomial distribution with $n$ trials and success probability $p$</td>
</tr>
<tr>
<td>$\text{Categorical}(p)$</td>
<td>Categorical distribution with probability vector $p$</td>
</tr>
<tr>
<td>$\text{PoiBinom}(p_1, \ldots, p_n)$</td>
<td>Poisson-binomial distribution with success probabilities $p_1, \ldots, p_n$</td>
</tr>
</tbody>
</table>

**Scoring Rules**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s(\cdot, \cdot)$</td>
<td>Scoring rule</td>
</tr>
<tr>
<td>$s^{\text{BS}}(\cdot, \cdot)$</td>
<td>Brier score</td>
</tr>
<tr>
<td>$s^{\text{KS}}(\cdot, \cdot)$</td>
<td>Kernel score</td>
</tr>
<tr>
<td>$s^{\text{LS}}(\cdot, \cdot)$</td>
<td>Logarithmic score</td>
</tr>
<tr>
<td>$s^{\text{PS}}(\cdot, \cdot)$</td>
<td>Pseudo-spherical score</td>
</tr>
<tr>
<td>$S(\cdot, \cdot)$</td>
<td>Expected score of a scoring rule</td>
</tr>
</tbody>
</table>
## Appendix A. Notation List

### Miscellaneous

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\langle \cdot, \cdot \rangle$</td>
<td>Inner product</td>
</tr>
<tr>
<td>$\langle \cdot, \cdot \rangle_H$</td>
<td>Inner product in pre-Hilbert space $H$</td>
</tr>
<tr>
<td>$A^*$</td>
<td>Hermitian adjoint of operator $A$</td>
</tr>
<tr>
<td>$f \otimes g$</td>
<td>Tensor product $(f \otimes g)(x, y) = f(x)g(y)$ of functions $f$ and $g$</td>
</tr>
<tr>
<td>$1_d$</td>
<td>Vector in $\mathbb{R}^d$ with each element being equal to 1</td>
</tr>
<tr>
<td>$\text{id}_A$</td>
<td>Identity function on $A$</td>
</tr>
<tr>
<td>$[P]$</td>
<td>Iverson bracket, $[P] = 1$ if $P$ is true, $[P] = 0$ otherwise</td>
</tr>
<tr>
<td>$\Re(x)$</td>
<td>Real part of $x$</td>
</tr>
</tbody>
</table>
A faulty judgement has caused mathematicians to equate elegance and conciseness at the cost of intelligibility.

Mathematics and the Physical World
MORRIS KLINE

APPENDIX B

Measure Theory Basics

This chapter explains some basic measure-theoretical concepts of probability theory that are used in this thesis. Measure theory provides a mathematical framework for assigning measures to mathematical objects such as sets, intervals, manifolds etc. in a consistent way. It is used in probability theory since probabilities are a special non-negative measure with the property that the total measure of the space of outcomes is one.

B.1 Probability Spaces

First, one has to define the sets that can be assigned a probability. Their collection has to satisfy some properties since otherwise the probabilities might be inconsistent and lack desired properties such as additivity. Formally, the set $\mathcal{A}$ of measurable subsets of a space $\Omega$ has to be a $\sigma$-algebra:

**Definition B.1** ($\sigma$-algebra (cf. Bauer 2001, Definition 1.1)). Let $\mathcal{A}$ be a collection of subsets of a set $\Omega$. The set $\mathcal{A}$ is called a $\sigma$-algebra (in $\Omega$) if

- $\mathcal{A}$ contains $\Omega$: $\Omega \in \mathcal{A}$.
- $\mathcal{A}$ is closed under complementation: For all $A \in \mathcal{A}$, $\Omega \setminus A \in \mathcal{A}$.
- $\mathcal{A}$ is closed under countable unions: For all $(A_n)_{n \in \mathbb{N}} \subset \mathcal{A}$, $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{A}$.

One instance of a $\sigma$-algebra in a set $\Omega$ is its power set, i.e., the set of all subsets of $\Omega$. It is a popular choice in particular if $\Omega$ is a finite discrete space. If $\Omega$ is a topological space, i.e., if there exists a collection of open (and closed) subsets of $\Omega$, then a common choice for $\mathcal{A}$ is the so-called Borel $\sigma$-algebra:

**Definition B.2** (Borel $\sigma$-algebra (cf. Bauer 2001, Definition 25.1)). Let $\Omega$ be a topological space. Then the smallest $\sigma$-algebra that contains all open sets is called the Borel $\sigma$-algebra in $\Omega$ and denoted by $\mathcal{B}(\Omega)$. 

93
A space $\Omega$ together with a $\sigma$-algebra $\mathcal{A}$ of measurable sets forms a measurable space $(\Omega, \mathcal{A})$ (Bauer 2001, p. 34). Next, one has to define the measure of the measurable sets.

**Definition B.3** (Measure (cf. Bauer 2001, Definitions 3.1 and 3.3; Schervish 1995, Definition A.18)). Let $(\Omega, \mathcal{A})$ be a measurable space. A function $\mu : \mathcal{A} \to \mathbb{R}_{\geq 0} = [0, \infty]$ is called a measure (on $\mathcal{A}$) if

- $\mu(\emptyset) = 0$, and
- for every sequence $(A_n)$ of pairwise disjoint sets from $\mathcal{A}$, then $\mu(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$ (\(\sigma\)-additivity).

The function value $\mu(A)$ of $\mu$ at an $A \in \mathcal{A}$ is called the ($\mu$-)measure of $A$, and the triple $(\Omega, \mathcal{A}, \mu)$ is called a measure space. If $\mu(\Omega) < \infty$ (and consequently $\mu(A) < \infty$ for all $A \in \mathcal{A}$), then measure $\mu$ is called finite.

Probability measures are special measures that assign every measurable set a probability. As usual, probabilities are expected to be real numbers between zero and one, and thus a probability measure should assign each set at most a measure of one. Moreover, the probability that any event occurs should be 100%, i.e., the probability measure of the whole space should be one.

**Definition B.4** (Probability Space (cf. Bauer 2001, pp. 31, 34; Schervish 1995, Definition A.18)). Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a measure space. If $\mathbb{P}(\Omega) = 1$, then $\mathbb{P}$ is called a probability (measure) and $(\Omega, \mathcal{A}, \mathbb{P})$ is called a probability space.

In a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, $\Omega$ is set of outcomes (“sample space”), $\sigma$-algebra $\mathcal{A}$ defines the set of “events”, and probability measure $\mathbb{P}$ assigns every “event” a probability.

### B.2 Random Variables

Often, you do not work with the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ directly but rather some, e.g., discrete or real-valued, random variables. In the measure-theoretical framework a random variable is a function on $\Omega$. In this sense, the realization of a random variable corresponds to a function evaluation at a specific outcome $\omega \in \Omega$. The function on $\Omega$ must not be completely arbitrary though: The probabilities of different realizations of a random variable have to be consistent with the probability measure $\mathbb{P}$ on $(\Omega, \mathcal{A})$. As discussed below, this consistency can be achieved by defining the probabilities of the random variable based on the probability measure on $\mathcal{A}$. Thus, the function that defines the random variable has to be measurable: It has to be possible to trace back every measurable set of realizations of the random variable to a measurable set in $\mathcal{A}$. This property is formalized in the following two definitions.
**Definition B.5** (Measurable Function (Bauer 2001, Definition 7.1)). Let $(\Omega, \mathcal{A})$ and $(\Omega', \mathcal{A}')$ be measurable spaces, and $T : \Omega \to \Omega'$ be a mapping of $\Omega$ into $\Omega'$. Then $T$ is called $\mathcal{A}$-$\mathcal{A}'$-measurable if for all $A' \in \mathcal{A}'$

$$T^{-1}(A') \in \mathcal{A}.$$ 

The $\mathcal{A}$-$\mathcal{A}'$-measurability of $T$ is expressed symbolically by $T : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$.

**Definition B.6** (Random Variable (cf. Bauer 1996, Definition 3.1)). Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(\Omega', \mathcal{A}')$ a measurable space. Then every $\mathcal{A}$-$\mathcal{A}'$-measurable mapping $Z : \Omega \to \Omega'$ is called a random variable (with values in $\Omega'$) or a $(\Omega', \mathcal{A}')$-random variable.

Following the categorization by Bauer (1996, p. 12), random variables whose co-domain $(\Omega', \mathcal{A}')$ is $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ or $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ or $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ are called real(-valued) or numerical or $\mathbb{R}^d$-valued random variables. Any measurable function allows lifting a measure on its domain to a measure on its co-domain:

**Theorem B.1** (Pushforward measure (Bauer 2001, Theorem 7.5, Definition 7.6)). Let $T : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ be a measurable function. Then for every measure $\mu$ on $\mathcal{A}$, $\mu' := \mu \circ T^{-1}$ defines a measure on $\mathcal{A}'$. Measure $\mu'$ is called the image, or pushforward, of $\mu$ under the mapping $T$.

The probability distribution of $Z$ is defined as the image of $\mathbb{P}$ under function $Z$ and denoted by $\mathbb{P}_Z$:

**Definition B.7** (Distribution (cf. Bauer 1996, Definition 3.2)). Let $Z$ be a $(\Omega', \mathcal{A}')$-random variable on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Then the image measure $Z(\mathbb{P})$ is called the distribution (or the probability law) of $Z$ (with respect to the probability measure $\mathbb{P}$). We set

$$\mathbb{P}_Z := Z(\mathbb{P}).$$

An important statistical property of a random variable is its expectation.

**Definition B.8** (Expectation (Bauer 1996, Definition 3.3)). Let $Z$ be a numerical random variable on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. If either $Z \geq 0$ or $Z$ is $\mathbb{P}$-integrable, we call

$$\mathbb{E}(Z) := \mathbb{E}_\mathbb{P} Z = \mathbb{E}_{Z \circ \mathbb{P}_Z} Z = \int_{\Omega} Z(\omega) \, d\mathbb{P}(\omega)$$

the expected value of $Z$. 

95
APPENDIX B. MEASURE THEORY BASICS

B.3 Probability Densities

Another important concept are probability densities. In the measure-theoretical framework they arise from comparing two measures on the same probability space.

This requires the concept of absolute continuity: Let $\mu$ and $\nu$ be two measures on the same measurable space $(\Omega, \mathcal{A})$. Then $\nu$ is called absolutely continuous with respect to $\mu$, written $\nu \ll \mu$, if for all $A \in \mathcal{A}$, $\mu(A) = 0$ implies $\nu(A) = 0$, i.e., if every null set of $\mu$ is also a null set of $\nu$.

Assume now that $\nu \ll \mu$ and that $\mu$ is $\sigma$-finite, i.e., there exists a countable number of measurable sets $A_1, A_2, \ldots$ with finite measure $\mu(A_i) < \infty$ such that $\Omega = \bigcup_{i=1}^{\infty} A_i$. Then there exists a measurable function $f : \Omega \to [0, \infty]$, called the Radon-Nikodym derivative of $\nu$ with respect to $\mu$ and written as $d\nu/d\mu$, such that for every measurable set $A \in \mathcal{A}$

$$\nu(A) = \int_A f(x) \, d\mu(x).$$

The Radon-Nikodym derivative is unique $\mu$-almost everywhere, and finite $\mu$-almost everywhere if $\nu$ is $\sigma$-finite.

Of particular interest here is it if $\nu$ is a probability measure. Then the Radon-Nikodym derivative of $\nu$ with respect to $\mu$ is called the probability density function of $\nu$ with respect to the base measure $\mu$. Since every probability measure is $\sigma$-finite, the probability density function is finite $\mu$-almost everywhere. Common choices for the base measure $\mu$ are the Lebesgue measure $\lambda$ and the counting measure #.

B.4 Conditional Expectations and Probabilities

The measure-theoretical definition of conditional expectations and probabilities is based on conditioning of a real-valued random variable with respect to a sub-$\sigma$-algebra. The following theorem ensures the existence of a conditional expectation for every sub-$\sigma$-algebra.

**Theorem B.2** (Existence of Conditional Expectation (cf. Bauer 1996, Definition 15.2, Theorem 15.1)). Let $Z$ be a numerical random variable on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ which is either non-negative or integrable. Then to every sub-$\sigma$-algebra $\mathcal{C} \subset \mathcal{A}$ corresponds a real-valued random variable $Z_0$ on $\Omega$, non-negative or integrable as the case may be, which is $\mathcal{A}$-measurable, unique to within almost sure equality, and satisfies for all $C \in \mathcal{C}$

$$\int_C Z_0(\omega) \, d\mathbb{P}(\omega) = \int_C Z \, d\mathbb{P}(\omega).$$

If $Z$ is both integrable and almost surely non-negative, then so is $Z_0$.  

96
Definition B.9 (Conditional Expectation (Bauer 1996, Definition 15.2)). Any random variable \( Z_0 \) derived from \( Z \) and \( \mathcal{C} \) as described in Theorem B.2 is called a conditional expectation of \( Z \) given \( \mathcal{C} \), in symbols

\[
E(Z \mid \mathcal{C}) := Z_0.
\]

One usually speaks of the conditional expectation \( E(Z \mid \mathcal{C}) \) since it is almost surely uniquely determined. Different, almost surely equal but not identical, random variables that are conditional expectations of \( Z \) given \( \mathcal{C} \) are called versions of the conditional expectation.

Conditioning with respect to random variables, which is more common in basic probability theory, is a special case of conditioning with respect to a \( \sigma \)-algebra.

Definition B.10 (Conditional Expectation Given Random Variables (Bauer 1996, p. 113)). Let \( Z \) be a numerical random variable on the probability space \((\Omega, \mathcal{A}, \mathbb{P})\) which is either non-negative or integrable. If \((Z_i)_{i \in I}\) is a family of random variables \( Z_i : (\Omega, \mathcal{A}) \to (\Omega_i, \mathcal{A}_i) \) with values in measurable spaces \((\Omega_i, \mathcal{A}_i)\), and \( \mathcal{C} := \sigma(Y_i; i \in I) \) is the sub-\( \sigma \)-algebra generated by these random variables, then instead of \( E(Z \mid \mathcal{C}) \) we also write

\[
e(Z \mid Z_i, i \in I)
\]

and speak of the conditional expectation of \( Z \) given the random variables \( Z_i (i \in I) \).

In case \( I = \{1, \ldots, n\} (n \in \mathbb{N}) \), we write for short

\[
e(Z \mid Z_1, \ldots, Z_n).
\]

The conditional probability of an event can be defined as the conditional expectation of its indicator variable.

Definition B.11 (Conditional Probability (Bauer 1996, Definition 15.7)). Let \((\Omega, \mathcal{A}, \mathbb{P})\) be a probability space, \( \mathcal{C} \subset \mathcal{A} \) be a sub-\( \sigma \)-algebra, and \( A \in \mathcal{A} \) an event. Then

\[
\mathbb{P}(A \mid \mathcal{C}) := E(1_A \mid \mathcal{C})
\]

is called the conditional probability of \( A \) given \( \mathcal{C} \).

Similar to the definition of conditional expectations, \( \mathbb{P}(A \mid Z_i, i \in I) \) denotes the conditional probability \( \mathbb{P}(A \mid \sigma(Z_i, i \in I)) \) and is called the conditional probability of \( A \) given random variables \( Z_i (i \in I) \). It is very important to note that, possibly surprising, \( A \mapsto \mathbb{P}(A \mid \mathcal{C})(\omega) \) is not necessarily a probability measure on \( \mathcal{A} \) for almost every \( \omega \in \Omega \) (Bauer 1996, p. 121).

Recall that conditional expectations and conditional probabilities given random variables, such as \( E(Z \mid Z_1, \ldots, Z_n) \) and \( \mathbb{P}(Z \mid Z_1, \ldots, Z_n) \) considered above, are random variables themselves. Outside of measure theory, however, it is more common to work with conditional expectations and conditional probabilities given
realizations of random variables, and thus these expectations and probabilities are not random variables but fixed numbers. Due to the factorization lemma (Bauer 2001, Lemma 11.7; Schervish 1995, Theorem A.42), it is possible to rigorously define conditional expectations given realizations of random variables in the measure-theoretical framework.

**Theorem B.3** (Bauer (cf. 1996, Theorem 15.9)). Let $Z$ denote an integrable real random variable on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Moreover, let $W: (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ be a $(\Omega', \mathcal{A}')$-random variable on $\Omega$ and $\mathbb{E}(Z \mid W)$ a real (integrable) version of the conditional expectation of $Z$ given $W$. The factorization lemma supplies a measurable real function $g$ on $(\Omega', \mathcal{A}')$, which may depend on the particular version of the conditional expectation, such that

$$\mathbb{E}(Z \mid W) = g \circ W.$$

(B.1)

Function $g$ is characterized as follows:

Every $\mathcal{A}'$-measurable real function $g$ which satisfies Equation (B.1) for a real version $\mathbb{E}(Z \mid W)$ is $\mathbb{P}_W$-integrable and satisfies for all $A' \in \mathcal{A}'$

$$\int_{A'} g(w) \, d\mathbb{P}_W(w) = \int_{W^{-1}(A')} Z(\omega) \, d\mathbb{P}(\omega).$$

(B.2)

Via these equations it is $\mathbb{P}_W$-almost surely uniquely determined. If, conversely, $g$ is a real $(\mathcal{A}'$-measurable) $\mathbb{P}_W$-integrable function on $\Omega'$ which satisfies Equation (B.2), then $g \circ W$ is a version of the conditional expectation of $Z$ given $W$.

**Definition B.12** (Bauer (cf. 1996, Definition 15.10)). For an integrable real random variable $Z$ let $g$ be an $\mathcal{A}'$-measurable, $\mathbb{P}_W$-integrable real function satisfying Equation (B.2). Then for every $w \in \Omega'$, $g(w)$ is called the conditional expectation of $Z$ given $W = w$, in symbols

$$\mathbb{E}(Z \mid W = w) := g(w).$$

These results carry over naturally to conditional probabilities of an event $A$ by considering the conditional expectation of the indicator variable of $A$.

### B.5 Conditional Distributions

As noted above, in general it is not guaranteed that versions of conditional probabilities $\mathbb{P}(A \mid \mathcal{C})$ ($A \in \mathcal{A}$) given a sub-$\sigma$-algebra $\mathcal{C} \subset \mathcal{A}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ are compatible such that $A \mapsto \mathbb{P}(A \mid \mathcal{C})(\omega)$ is a probability measure for almost every $\omega \in \Omega$. Thus, to satisfy this property, a special collection of versions of conditional probabilities, a so-called regular conditional probability, has to be considered.

---

1As noted by Bauer (1996, p. 122), this assumption can be relaxed to $Z$ being a non-negative numerical random variable.
B.5. Conditional Distributions

Definition B.13 (Markov Kernel (cf. Bauer 1996, Definition 36.1)). Let measurable spaces \((\Omega, \mathcal{A})\) and \((\Omega', \mathcal{A}')\) be given. A function \(K: \Omega \times \mathcal{A} \to \mathbb{R}_{\geq 0}\) is a Markov kernel from \((\Omega, \mathcal{A})\) to \((\Omega', \mathcal{A}')\)—or, briefly, from \(\Omega\) to \(\Omega'\)—if it has the following two properties:

1. For all \(A' \in \mathcal{A}'\), the function \(\omega \mapsto K(\omega, A')\) is \(\mathcal{A}\)-measurable.

2. For all \(\omega \in \Omega\), the function \(A' \mapsto K(\omega, A')\) is a probability measure on \(\mathcal{A}'\).

Definition B.14 (Regular Conditional Probability (Bauer 1996, p. 387; Schervish 1995, Definition B.29)). Let \((\Omega, \mathcal{A}, \mathbb{P})\) be a probability space, and let \(\mathcal{E} \subseteq \mathcal{A}\) be a sub-\(\sigma\)-algebra. Then a regular conditional probability for \(\mathcal{E}\) is a Markov kernel \(K: (\Omega, \mathcal{E}) \to (\Omega, \mathcal{A})\) with the property that for each \(A \in \mathcal{A}\) the function \(\omega \mapsto K(\omega, A)\) is a version of the conditional probability \(\mathbb{P}(A | \mathcal{E})\).

As above, you can consider conditional probabilities given random variables as a special case.

Definition B.15 (Regular Conditional Distribution (Bauer 1996, Definition 44.1)). Let \(Z: (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')\) be a random variable on a probability space \((\Omega, \mathcal{A}, \mathbb{P})\) with values in a measurable space \((\Omega', \mathcal{A}')\) and let \(\mathcal{E}\) be a sub-\(\sigma\)-algebra of \(\mathcal{A}\). Every Markov kernel \(K\) from \((\Omega, \mathcal{E})\) to \((\Omega', \mathcal{A}')\) such that for all \(A' \in \mathcal{A}'\), \(\omega \mapsto K(\omega, A')\) is a version of the conditional probability \(\mathbb{P}(Z^{-1}(A') | \mathcal{E})\), is called a (regular) conditional distribution of \(Z\) given \(\mathcal{E}\). If \(\mathcal{E}\) is generated by a random variable \(W: (\Omega, \mathcal{A}) \to (\Omega'', \mathcal{A}'')\), i.e., if \(\mathcal{E} = Z^{-1}(A'')\), then \(\mathbb{P}_{Z|W}\) is often written instead of \(\mathbb{P}_{Z|\mathcal{E}}\).

Two immediate questions are: Are regular conditional distributions unique and under which conditions do they exist?

Theorem B.4 (Uniqueness of Regular Conditional Distributions (Bauer 1996)). If \(K\) and \(K^*\) are two conditional distributions of a random variable \(Z: (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')\) given \(\mathcal{E} \subseteq \mathcal{A}\), then for every \(A' \in \mathcal{A}'\), for \(\mathbb{P}\)-almost all \(\omega \in \Omega\)

\[ K(\omega, A') = K^*(\omega, A'). \]

If the \(\sigma\)-algebra \(\mathcal{A}'\) has a countable generator, then there is a \(\mathbb{P}\)-null set \(N \in \mathcal{A}\) such that for every \(A' \in \mathcal{A}'\) and every \(\omega \in \Omega \setminus N\)

\[ K(\omega, A') = K^*(\omega, A'). \quad (B.3) \]

The following theorem provides sufficient conditions for the existence of a conditional distribution.

Definition B.16 (Schervish (1995, Definition B.31)). Let \((Z, \mathcal{A}_Z)\) be a measurable space. If there exists a bimeasurable\(^2\) function \(f: Z \to R\) where \(R\) is a Borel subset of \(\mathbb{R}\), then \((Z, \mathcal{A}_Z)\) is called a Borel space.

---

\(^2\) A function \(f\) is called bimeasurable if it is measurable, bijective, and \(f^{-1}\) is measurable.
Appendix B. Measure Theory Basics

**Theorem B.5** (Schervish (1995, Theorem B.32)). Let \((\Omega, \mathcal{A}, \mathbb{P})\) be a probability space and let \(\mathcal{C}\) be a sub-\(\sigma\)-algebra of \(\mathcal{A}\). Let \((Z, \mathcal{A}_Z)\) be a Borel space and let \(Z : \Omega \to Z\) be a random variable. Then there exists a regular conditional distribution of \(Z\) given \(\mathcal{C}\).

Note that there are no conditions on \(\mathcal{C}\) in the theorem, for the existence of a regular conditional distribution of \(Z\) given \(\mathcal{C}\) it is sufficient the co-domain of \(Z\) is a Borel space. Fortunately, many common spaces are Borel spaces:

**Lemma B.1** (Schervish (1995, Lemma B.40)). If \((Z, \mathcal{A}_Z)\) is a Polish space with the Borel \(\sigma\)-algebra and metric \(d\), then it is a Borel space.

In this case, also the stronger form of uniqueness in Equation (B.3) holds.

**Theorem B.6** (Bauer (1996, Theorem 44.3)). Let \(Z : (\Omega, \mathcal{A}) \to (E, \mathcal{B}(E))\) be a random variable on the probability space \((\Omega, \mathcal{A}, \mathbb{P})\) with values in a Polish space \(E\). Then a conditional distribution \(\mathbb{P}_{Z|\mathcal{C}}\) exists for every \(\sigma\)-algebra \(\mathcal{C} \subset \mathcal{A}\). It is uniquely determined in the sense of Equation (B.3).
References


REFERENCES


102
REFERENCES


References


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 2275

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)