When is subjective objective enough? Frequentist guarantees for Bayesian methods

Stefan Franssen

COLOPHON

Stefan Franssen
When is subjective objective enough?
Frequentist guarantees for Bayesian methods, Delft, 2023

Printing: TODO: Drukker
Cover: Dalle-2 and postprocessing

An electronic version of this dissertation is available at:

`http://repository.tudelft.nl/`

When is subjective objective enough? Frequentist guarantees for Bayesian Methods

Dissertation

For the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus Prof.dr.ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates,
to be defended publicly on
Tuesday 6 June 2023 at 12:30 o'clock

by

Stefan Esther Mariëlle Patrick FRANSSEN

Master of Science in Mathematical Sciences
Universiteit Utrecht, Nederland
born in Maastricht, Nederland

This dissertation has been approved by the promotors.

Composition of the doctoral committee

Rector Magnificus	chairperson
Prof.dr. A.W. van der Vaart	Technische Universiteit Delft, promotor
Dr. B.T. Szabó	Universitá Bocconi, promotor

Independent members:

Prof.dr. I. Pruenster	Bocconi U., Italy
Prof.dr. A.J. Schmidt-Hieber	U. Twente, NL
Prof.dr.ir. F.H. van der Meulen	VU. Amsterdam, NL
Prof.dr.ir. G. Jongbloed	Technische Universiteit Delft
Prof.dr. A.J. Cabo	Technische Universiteit Delft, Reserve member

Het onderzoek in dit proefschrift werd (mede) gefinancierd door NWO.

Contents

Chapter 1

Introduction

1.1 A gentle introduction to statistics

To understand the thesis, we need to introduce a few basic notions from statistics. We start with a non-rigorous introduction aimed at a general audience in this section. We first introduce the goals of statistics. Next, we introduce the concept of estimators and how to measure their quality. Then we introduce Bayesian methods as we study Bayesian methods. Finally, we introduce the general ideas of studying Bayesian methods from a frequentist point of view.

1.1.1 A first introduction to statistics

We constantly make decisions. What shall we wear today? What job should we do? What study should we take? In order to make good decisions, we need to be able to estimate outcomes. Statistics is the science of making accurate estimates. In order to make the best decisions, we need to have the best statistical tools available. Ideally, we want to make estimates that are as accurate as possible and quantify the uncertainty of these estimates.

To clarify our language, let us introduce a few concepts. The thing we want to study is the estimand. Examples of estimands are the weather tomorrow, the job market in a year or the effect of eating sugar on blood pressure. To figure out what the estimand should be, we use estimators. These take the data and produce an outcome. The estimate is the outcome of the estimator given the data.

1.1.2 Estimators

To give the best possible estimates we need to have the best estimators out there. So we need to think about quality of estimators. Is every estimator good? To talk about how good an estimator is, we first need to talk about what it means to be a good estimator. There are various philosophies on what it means to be a good estimator. We will study estimators from a frequentist point of view. That is, we assume that there is a true process that we are observing. Moreover, we could have gotten a different distribution of observations. This allows us to give meaning to the quality of estimators.

One way of measuring the quality of an estimator is the expected loss. In order to introduce the expected loss, we need to give a few names to things. We are trying to find the process f_0, our estimand. We do not assume we know the process f_0 in reality. Instead, we use it as a hypothetical object to define other things. To introduce the expected loss, we need to define some things.

We start with a loss. The loss is a way of measuring how wrong we were, given our estimate and the truth. This loss can be an arbitrary function. Let's give this function a name, L. For example, say we are predicting how warm it will be tomorrow. Suppose we predict the temperature to be 20 degrees C. The day after, we observed the temperature, 21 degrees. We would be 1 degree off. However, if we instead predicted 40 C, we would have been 20 degrees off. The 40C answer is usually more wrong than the 20C answer. However, there might be much more damage if we underestimate the temperature than if we overestimate. Hence in some cases, the 40C answer might have led to a lower loss.

To define the expected loss, we need to give a few more things a name. We start with hypothetical data coming from the process f_0. We will evaluate the estimator in the data. We call the outcome $\hat{f}$. Using this, we can compute the loss $L(\hat{f}, f_0)$.

$\hat{f}$ is a random variable because it depends on the data. Now we can ask for the expected value of the loss. The expected value of a random variable is the long-term average of independent instances of this random variable. The formula for the expected loss is given by

$$\mathbb{E}_{f_0}[L(f_0, \hat{f})].$$

This definition gives us one way of measuring the quality of our estimator.

Imagine we are shooting arrows at a target. Each place where we hit our target is an estimate. In reality, we only get one estimate. However, imagine that we can conduct our experiment all over again. Due to randomness, we could get slightly different data and hence different conclusions. This randomness leads to a spread of our outcomes. The expected value of our estimator can still be wrong. Because of this, we can have a statistical bias. So in the ideal case, we want to minimise the bias and spread. In Figure 1.1 we can see an illustration. The ideal situation is a low bias and a low variance. In certain situations, we can gain accuracy by trading a small amount of bias for a larger reduction in variance. In these situations, we have a so-called bias-variance trade-off.

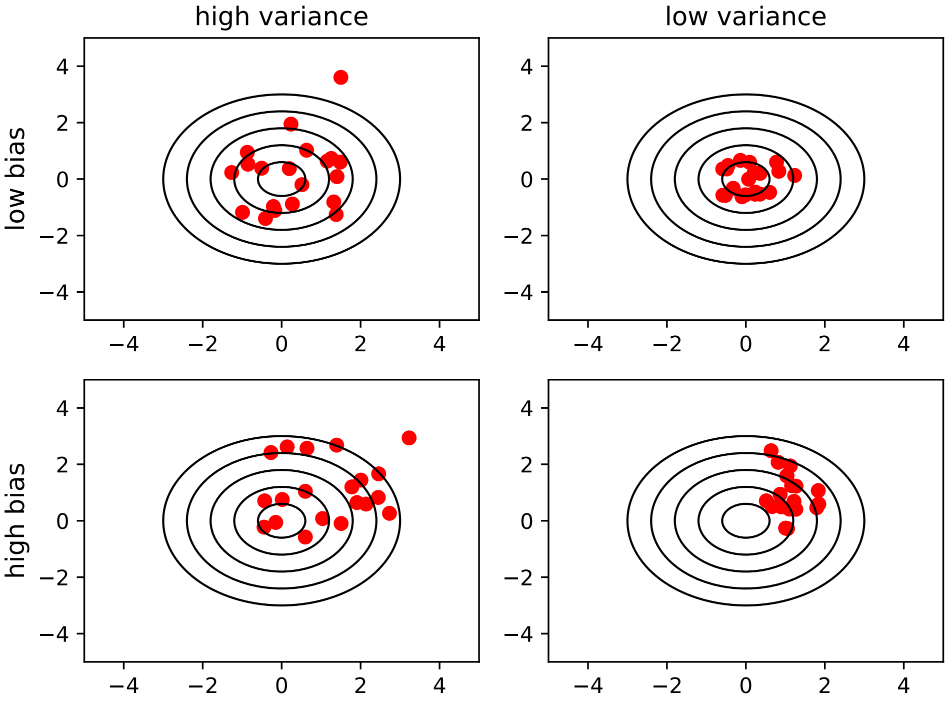


Figure 1.1: Illustration of bias, variance and accuracy

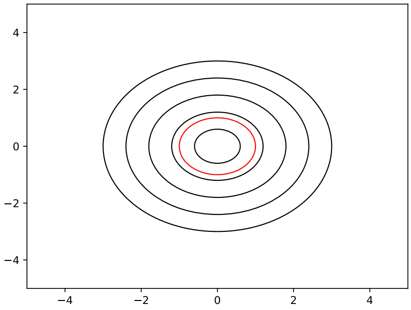


Figure 1.2: Consistency

The main obstacle in statistics is that the truth is usually unknown. Moreover, we do not repeat the experiment. If we could repeat the experiment, we could collect all our data into one bigger dataset and get more accurate predictions. Hence we only get to see one of these points. And without convenient targets to let us know where the truth is. Because of this, we cannot experimentally verify the quality of our estimators. Therefore we need to use alternative methods to verify the quality. Two such alternatives are theory and simulation studies. In this dissertation, we focus on theory and use simulation studies to illustrate our theoretical findings.

To study the quality of the estimator we need to introduce various concepts. These concepts are consistency, convergence rates and uncertainty quantification. The idea behind consistency is as follows: We want our estimators to eventually come close to the true generating process. In Figure 1.2, you can see a red circle, call its radius ϵ. Roughly, we say that an estimator is consistent if for a given red circle of a specific radius, and any error tolerance $\delta > 0$, we can find a minimum data size n such that if our data has size at least n the probability that our estimator produces an answer back that is outside our red circle is at most δ. Recall that we call the true process f_0 and the outcome of our estimator $\hat{f}$. $\hat{f}$ would be in the circle if the distance between the centre f_0 and $\hat{f}$ would be less than ϵ. Thus, in formula this would give, where d is the distance metric:

$$\mathbb{P}(d(\hat{f}, f_0) > \epsilon) < \delta$$

Intuitively, no matter how close we demand our estimator to be, we will get there if we have enough data. To get the formal definition see Section 1.2.2 Definition 1.2.1.

Next is the notion of a contraction rate. Again, imagine the red circle. Now imagine this circle shrinking at a certain speed. Let's call the radius ϵ_n. Roughly, we say that an estimator has a contraction rate ϵ_n if for every $\delta > 0$ we can find an $M > 0$ such that the probability that the distance between the truth and the outcome of our estimator is more than $M\epsilon_n$, is less than δ. In formulas, this would give

$$\mathbb{P}(d(\hat{f}, f_0) > \epsilon_n) < \delta.$$

Intuitively this means that our estimator comes closer to the truth at a speed of ϵ_n. For a formal definition, see again Section 1.2.2 Definition 1.2.2.

Next up is the notion of uncertainty quantification. Because we only get to see one dataset, we only see one realisation of the estimator. This means that we do not know how accurate the estimator is. Consider in Figure 1.1 the difference between the top right and the bottom left picture. One is much more accurate than the other. If we want to know how accurate we are, we have to develop techniques and give theoretical guarantees for them. One idea to quantify uncertainty is the notion of a confidence set. A confidence set of level $1-\alpha$ is a random set C_α, such that the probability that the truth f_0 is contained in this set, $f_0 \in C_\alpha$, is at least $1-\alpha$. To be as informative as possible, we want this set to be as small as possible.

How can we make such confidence sets? We can do the same trick as we did for the contraction rates. We made a circle of a certain radius around the truth, such that the probability that the estimator would produce an outcome in this circle is at least our prescribed level α. Such an outcome is in the circle when the distance between the truth f_0 and our estimator $\hat{f}$ is less than $M\epsilon_n$. Because the distance $d(f_0, \hat{f})$ is equal to $d(\hat{f}, f_0)$, we can also make our random set C to be a ball of radius $M\epsilon_n$. This construction requires knowing the contraction rate and the constant M. In certain situations, this is possible. However, sometimes this is impossible.

Another idea to construct confidence sets is to use a technique called bootstrap. In the bootstrap, you make new artificial datasets which resemble the original dataset. Then by studying the spread by the variance introduced by bootstrapping, we can get an idea of the spread of our estimator. Bootstrapping, however, does not always explain the bias. Therefore we should be careful employing bootstrap in cases where we use a bias-variance trade-off.

A third idea to construct confidence sets is to study a Bayesian alternative of uncertainty quantification.

1.1.3 A first introduction to Bayes

Bayesian statistics is a way of doing statistics. In Bayesian statistics, you start with your prior beliefs. These prior beliefs you encode in a probability distribution Π. This probability distribution is called the prior or prior probability distribution. Then you model the process as follows. Reality draws the generating process f randomly following the prior Π, $f \sim \Pi$. Then, condition on f, we draw our dataset from f: $X|\, f \sim f$.

Because our observation X depends on the generating process f, we can learn about f by looking at the data. This dependence means that, given our data, our beliefs

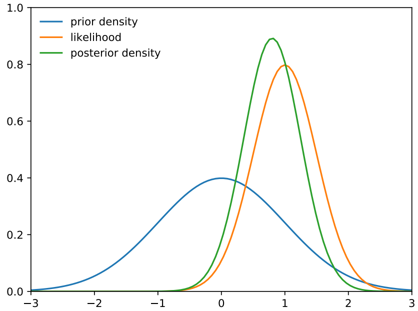


Figure 1.3: Prior, likelihood and posterior

on f can change. To update our beliefs on f, we compute the conditional probability distribution of f given the data X. In formulas, this is

$$\Pi(\cdot |\, X).$$

This probability distribution represents the beliefs that you should have as a Bayesian.

In general, there is no direct way of finding the posterior. However, often we can use Bayes' formula to compute the posterior. To apply Bayes' formula, we need to be able to know something called the likelihood. The likelihood depends on the data and our hypothetical process f, not f_0. Roughly speaking, the likelihood is how likely we would see this data if the generating process was f. If we denote the likelihood by $\mathrm{Lik}(X, f)$ Bayes formula becomes

$$\Pi(A |\, X) = \frac{\int_A \mathrm{Lik}(X, f)\Pi(f)}{\int \mathrm{Lik}(X, f)\Pi(f)}.$$

The intuition is that our new belief in some hypothesis f is how likely the data was if f was the generating process times how likely we think f was a priori.

In Figure 1.3 there is a graphical illustration of how it looks to update your beliefs. In blue we can see the density of the prior distribution. The likelihood of the data is drawn in orange. If we multiply these and then normalise the result we the posterior density in green. As you can see, this is shifted to look more and more like the likelihood.

This posterior distribution can be a complicated object, so you give summary statistics instead, for example, the posterior mean and variance. You might lose information, but they give a quick way of summarising the posterior.

Bayesians also want to quantify uncertainty. They do this by constructing credible sets. A credible set of level $1-\alpha$ is a set C such that the posterior assigns probability at least $1 - \alpha$ to the event $f \in C$:

$$\Pi(f \in C|\, X) > 1 - \alpha.$$

Because your prior beliefs are your personal subjective beliefs, the Bayesian methodology is inherently subjective. However, we can study these methods from a Frequentist point of view.

1.1.4 A first introduction to Frequentist Bayes

We still want to give theoretical guarantees to Bayesian methods, even though they are inherently subjective. It turns out we can do this. There is just one complication. The Bayesian posterior distribution is not a point estimate like we represented our estimators earlier in the chapter. It is a probability distribution.

Thus we have to modify our definitions for consistency and convergence rates. To illustrate consistency, we can go back to Figure 1.2. Let us pick a circle of radius ϵ. We can ask: what is the probability $\Pi(d(f_0, f) < \epsilon|\, X)$ that f is in this circle according to the posterior distribution? Because our data is random, the posterior distribution will also be. There are several mathematically equivalent ways of giving this definition, but the cleanest concept is via the expected value. Let us start with a error tolerance $\delta > 0$. For every such error tolerance δ, we specify a minimum data size N. This minimum data size N has to guarantee that, if our data size is larger than N, we have that

$$\mathbb{E}_{f_0}[\Pi(d(f_0, f) > \epsilon|\, X)] < \delta.$$

Intuitively this means that the posterior will start assigning a higher and higher probability to each ball around the truth. To see the formal set-up see Section 1.4.1 Definition 1.4.1.

We can do the same for convergence rates. The Bayesian counterpart of convergence rates is called contraction rates. We say that a posterior contracts at rate ϵ_n if for every $\delta > 0$ there exists a datasize n and $M > 0$ such that

$$\mathbb{E}_{f_0}[\Pi(d(f_0, f) > M\epsilon_n|\, X)] < \delta.$$

Intuitively this means that the posterior will concentrate most of its mass within a circle of radius ϵ_n. For a more formal set-up, see Section 1.4.1 Definition 1.4.4.

The more complicated question is the one surrounding uncertainty quantification. Bayesian credible sets give a subjective form of uncertainty quantification. With enough data, the subjectiveness might become small enough to be essentially objective. This objectiveness would lead to asymptotically valid confidence sets. Thus we could use these credible sets for uncertainty quantification. Therefore we must study when this subjectiveness disappears.

In the beginning, it is not so clear why uncertainty quantification can fail. Suppose we know the posterior contracts at the best rate possible. We then know that the spread of the posterior distribution cannot be too big. Most of the mass has to fit inside the circle of a radius proportional to the contraction rate. However, the spread might be too small. One can imagine that the posterior concentrates inside the circle. Moreover, all its mass concentrates on a small region within this circle. The bulk of the mass converges fast to the truth. But the truth will not be within the bulk of the mass. Hence uncertainty quantification can fail even when we have contraction rate guarantees. Thus we need different techniques for studying uncertainty quantification.

One collection of techniques is the so-called Bernstein-von Mises theorem. They make precise what happens with the posterior in large data sizes. The Bernstein-von Mises theorem allows us to study the posterior distribution in great detail. Then we use our knowledge about the posterior distribution to study the validity of our uncertainty quantification.

Precisely because the Bernstein-von Mises theorem is a detailed analysis of the posterior, it can be hard to prove. Moreover, even if we can prove them, we cannot always conclude that credible sets give valid uncertainty quantification. For more details on the Bernstein-von Mises theorem, see Section 1.4.3.

Another approach to studying the behaviour of the posterior distribution is extending Schwartz' theorem. This theorem is simpler than the Bernstein-von Mises theorem, but this makes it less precise. This lack of precision makes it hard to study the validity of uncertainty quantification. There is one such extension which gives some knowledge about the posterior distribution. This extension requires strict assumptions on the form of the model and the prior. With these assumptions, we can study the frequentist coverage of credible sets. For more details, see [67].

1.2 Frequentist statistics

There are different philosophies of probability. Because statistics depends on probability theory, these differences lead to different philosophies of statistics. The main philosophy of probability theory is frequentism. In Frequentism, you can, in theory, repeat your experiments arbitrarily often. The probability of an event is the same as the long-run frequency of this event.

1.2.1 Limit theorems

For frequentism to be an interpretation of probability theory, the claims of frequentism should be theorems of probability theory. Limit theorems provide that. They verify the correspondence between frequencies and probabilities.

The strong law of large numbers states that the sample average of independent identically distributed random variables will converge to the mean, provided it exists. So with probability 1

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}[X_1].$$

The strong law of large numbers states how fast this convergence will take place. This convergence can be arbitrarily slow. Under some additional conditions, namely the existence of the variance, the central limit theorem makes precise how fast it converges. It tells us more than just how fast it converges. Suppose that we could pick many samples. Because they are random, we can study the distribution of this sample around the expected value. This spread would be approximately a Gaussian distribution:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_1]\right) \rightsquigarrow \mathcal{N}(0, \mathrm{Var}(X_1)).$$

The power of the central limit theorem is that it gives a precise asymptotic distribution of the sample mean under very weak assumptions. The properties of estimators can be studied using this convergence. One example of an estimator that can be studied using the central limit theorem is the maximum likelihood estimator. Another important example is the bootstrap method. For more details on this, see [77, chapters 5 to 9] for the maximum likelihood estimator, and [78] for the bootstrap.

1.2.2 Frequentist properties of estimators

Suppose we want to estimate some object θ_0 based on observations $X_1, \dots, X_n$. We will denote an estimator with $\hat{\theta}$. Then $\hat{\theta}$ can depend on our data. With mild abuse of notation, we suppress this dependence. We will write:

$$\hat{\theta} = \hat{\theta}_k = \hat{\theta}(X_1, \dots, X_k).$$

Note that $\hat{\theta}_k$ only depends on the first k observations.

We will call an estimator unbiased for θ if its expected value is θ:

$$\mathbb{E}[\hat{\theta}] = \theta_0.$$

Ideally, we want to be unbiased whatever the value of θ_0 is. Suppose the distribution depends on θ. We would want that our estimate $\hat{\theta}$ has expected value θ_0:

$$\mathbb{E}_\theta[\hat{\theta}] = \theta.$$

Our estimator $\hat{\theta}$ is called unbiased in this case.

Suppose we could repeat this experiment often. Because of the unbiasedness, the average of the estimates would be close to θ_0. This unbiasedness does not mean that a given estimate is close to θ_0.

The bias of an estimator $\hat{\theta}$ for estimating θ_0 can be given as

$$B_{\hat{\theta}} = \mathbb{E}_\theta[\hat{\theta}] - \theta$$

An unbiased estimator $\hat{\theta}$ has $B_{\hat{\theta}} = 0$. Having a bias means that the long-term average of repeating your experiment will not be θ, but if the bias is small, its long-term average will still be close. In *nonparametric statistics*, it is often valuable to trade a small amount of bias for a lot of reduction in the spread to make the final estimator more accurate.

Consistency

Unbiasedness is not enough to determine the quality of an estimator. As a next step, we will consider *consistency*. This consistency is not a property of a single estimator. It is a property of a sequence of estimators. We call a sequence of estimators $\hat{\theta}_k$ consistent if simultaneously with enough data

- The bias can be made arbitrarily small
- The spread can be made arbitrarily small

Consistency means that the predictions will become more and more accurate the more data you have. If we generalise a bit by allowing general metrics, we get the following definition for consistency:

Definition 1.2.1. *We say an estimator is consistent for some distance metric d if*

$$\lim_{n\to\infty} \mathbb{E}_{\theta_0}[d(\hat{\theta}_n, \theta_0)] = 0.$$

Rates of convergence

Consistency is the first requirement for being a good estimator. Consistency means that, eventually, we will be accurate in finding the value we want to estimate. However, it does not tell us how fast we will be accurate. This speed will be the next step in the quality of estimators: the rate of convergence. We can measure the speed by measuring the distance between the estimate and the estimand. Roughly, we want that this distance is on average small.

Definition 1.2.2. *We say a rate ϵ_n is a rate of convergence if for every sequence $M_n \to \infty$*

$$\lim_{n\to\infty} \mathbb{P}_{\theta_0}\left(d(\hat{\theta}_n, \theta_0) > M_n\epsilon_n\right) = 0.$$

This extra sequence M_n is a technical nuisance. We need this extra sequence M_n for intuition and the definition to match. Note that we said a rate, since if ϵ_n is a rate of convergence, then also any sequence ϵ'_n such that $\epsilon'_n \geq \epsilon_n$.

Now one can wonder, how fast can I learn? Is there a limit? The limiting rate depends on how complex the statistical problem at hand is. The fastest rate you can achieve for a specific problem is called the minimax rate. For simple models, the *parametric* models (see [80]) the minimax rate is usually $\frac{1}{\sqrt{n}}$.

To understand the name minimax, one has to think about when the best rate possible makes sense. If all you have to estimate one value θ_0, which is fixed but unknown, the estimator $\hat{\theta} = \theta_0$ is always at distance 0, and therefore converges at rate 0 (the fastest possible). However, this estimator does not depend on the data and requires knowing θ_0. Hence we would want to exclude this somehow. Moreover, this estimator is only good if we are estimating θ_0. If reality had some other value θ_0', we would never learn. This indicates that our concepts are not matching what we want. What we actually want is to estimate θ in a class of options Θ. For every estimator $\hat{\theta}$ find the worst rate for estimating any θ in Θ, and then find the estimator that minimises this. To make this explicit, let E denote the set of all estimators. Then the minimax rate is given by

$$r_n = \inf_{\hat{\theta} \in E} \sup_{\theta \in \Theta} \mathbb{E}_\theta d(\hat{\theta}_n, \theta)$$

Example 1.2.3. *One important example where rates appear are in nonparametric regression. In nonparametric regression, the classes we consider are smoothness classes. One such example consist of the classes of β-Sobolev functions. Let $\Theta_\beta = \mathcal{H}^\beta = \{f \colon f \text{ is } \beta \text{ Sobolev}\}$ be the class of all β-Sobolev functions on $[0,1]^d$. This roughly means that f has β derivatives. Then the minimax rate of estimation in the L^2-distance is given by*

$$\inf_{\hat{f} \in E} \sup_{f_0 \in \Theta_\beta} \mathbb{E}_{f_0} \|\hat{f}_n - f_0\|_2 \asymp n^{\frac{-\beta}{2\beta+d}}.$$

Adaptation
For completeness, we will say a few words about adaptation. The goal in adaptation is to achieve the minimax rate of estimation as if we knew the correct class. Adapting to the correct complexity class is one example for which we can do adaptation. So even though we do not know how smooth the true function is, we want to find a procedure that can learn as if it knew how smooth it was. It turns out we can do this. There exists estimators $\hat{f}$ such that

$$\forall \beta \sup_{f_0 \in \mathcal{H}_\beta} \mathbb{E}_{f_0}[\|\hat{f} - f_0\|_2] \asymp n^{-\frac{\beta}{2\beta+d}}.$$

Given that adaptive estimation is possible, one might hope that adaptive uncertainty quantification is possible. However, it turns out that this, in general, is not possible (see [34, Chapter 8.3]). We will not consider adaptive estimation or uncertainty quantification in this thesis.

Efficient estimators
Often in statistics, it is the goal to find estimators which achieve the minimax rate. This is not the end of the story. The rate is a crude measure of accuracy. We

will restrict to parametric models for this discussion. We can extend the theory of efficiency to semiparametric models with relative ease. For more details, see [77, Chapter 7, 8 9, and 25].

Suppose we are in a parametric model, which is dominated and has densities p_θ. It turns out that the log-likelihood $\ell_\theta(x) = \log p_\theta(x)$ plays a central role in the analysis. We find that $P_\theta \dot{\ell}_\theta = 0$ and $P_\theta \ddot{\ell}_\theta = I_\theta$ some matrix. We call I_θ the Fisher information for θ.

By the almost-everywhere Convolution theorem [77, p. 8.9] and Andersons Lemma [77, Lemma 8.5], we know that, under some conditions,

$$\sqrt{n}\,(T_n - \theta) \rightsquigarrow N(0, I_\theta^{-1})$$

for almost every θ is the "best" asymptotic distribution that can be achieved. Estimators that achieve this will be called efficient estimators. Efficient estimators are thus, in some sense, the "asymptotic best" estimators. Hence they play an important role in the theory of asymptotic statistics.

1.2.3 Testing

In statistics, we do not only want to *estimate* objects of interest. Often we want to test hypotheses as well. We often use *confidence sets* to test hypotheses. A confidence set C_α of level $1 - \alpha \in (0, 1)$ is a random set such that $\mathbb{P}(\theta_0 \in C) \geq 1 - \alpha$.

Suppose we want to test a null hypothesis $H_0\colon \theta_0 \in \Theta_0$ against an alternative hypothesis $H_1\colon \theta_0 \in \Theta_1$. We first need to specify a significance level α. This significance level is the chance that we reject the null hypothesis while the null hypothesis is true. To be more precise, for every $\theta_0 \in \Theta_0$, we want the probability of rejecting the null hypothesis to be at most α. Moreover, if the alternative hypothesis is true we want to reject the null hypothesis with as high probability as possible. This probability is called the power. Suppose $\theta_0 \in \Theta_1$. The power is the chance of rejecting the null hypothesis under P_θ. We want the power to be as high as possible.

Asymptotic testing theory
To understand what we should aim for when designing and studying tests, we must know what the best is we can hope to achieve. From an asymptotic point of view, [77, Theorem 15.1, asymptotic representation theorem] gives us a tool for understanding the asymptotics. This representation theorem gives us the tools to study testing from an asymptotic point of view. From an asymptotic point of view it is sufficient to study the power in the limit experiment.

We often consider experiments that converge to a Gaussian limit experiment $X \sim \mathbb{G}$, where $\mathbb{G}$ is the standard Gaussian distribution. Since Gaussian experiments are explicit, it is easier to analyze them. In particular, we can understand the quality of closed convex confidence sets. Let C be a closed, convex set. Suppose we want to test the null hypothesis $H_0\colon \theta_0 = 0$ versus $H_1\colon \theta_0 \neq 0$. Then by [72, Theorem 30.4], it follows that the tests that reject the null if our data $X \notin C$ is *admissible*. We call a test of level α admissible if we cannot improve the power under a given alternative without losing power under another alternative hypothesis. Admissibility means that we cannot strictly improve our test ϕ. Therefore we can use admissibility of tests as an optimality criterion for designing our tests. While this is not completely satisfactory, admissibility gives a strong motivation for designing our confidence sets.

Motivating confidence and credible balls
We want to design our credible sets to be as powerful as possible. We have seen that balls are admissible for testing in the Gaussian experiment. That means that no test can beat the power of such limiting tests. Thus from an asymptotic point of view, it suffices to study confidence sets that have power converging to this limiting power.

Fix some level $1 - \alpha$. If we now look at balls, centred on some efficient estimator $\hat{\theta}_n$ and pick radius $\hat{\rho}_n$ as small as possible while maintaining level $1 - \alpha$. By the weak convergence and the efficiency of the estimators, these balls achieve the asymptotic optimal power. For more details of this argument, see Section 1.4.3.1. Hence, we want our credible sets to be balls.

1.3 Bayesian statistics

1.3.1 Bayes

We aim to give a short overview of the theoretical results of Bayesian statistics. While we aim to give a broad overview of the techniques, we have to skip extensions for brevity. We will also skip the proofs and try to aim for the big picture and the motivation of what the assumptions mean.

1.3.1.1 What is Bayesian inference

Bayesian inference is a methodology of inference. First, we define a model of reality which depends on some parameters. A Bayesian would start by specifying a prior distribution on these parameters. This prior distribution encodes all our beliefs and uncertainty about reality. To learn, a Bayesian collects observations and combines these with the prior. A Bayesian computes the conditional probability of the parameters given the data. The posterior distribution then encodes the beliefs you should have about reality.

1.3.1.2 Why Bayesian inference

Bayesian inference is based on combining data with your prior beliefs and updating by creating a new probability distribution that captures your new knowledge about the parameters. This inherently gives rise to some (subjective) form of uncertainty quantification and includes prior knowledge. Furthermore, Bayesian inference follows from various interpretations of probability and rationality. These are all advantages over frequentist techniques which lack these motivations. However, frequentist point estimators are usually faster and easier to compute. They are easier to represent since they do not require you to specify a probability distribution but a point estimator.

1.3.2 Examples

We introduce two examples: the model based on flipping coins and the regression model.

Example 1.3.1 (Coin flips)**.** *Suppose we observe $X_1, \dots, X_n$ independent flips of a coin whose probability of landing on heads is p, and we want to estimate p. We start with a prior, the uniform prior $p \sim U[0,1]$. The uniform prior is a beta distribution, namely Beta$(1,1)$. There are many other priors you could use. However, we chose this prior for its simplicity. Another prior you could consider is Jeffreys' prior. Jeffreys' prior for this model is the Beta$(\frac{1}{2}, \frac{1}{2})$ distribution. Jeffreys' prior has many properties that make it a good candidate for a choice of default prior, but this goes beyond the scope of this thesis.*

As our second example, we will consider the regression model. In regression, we observe independent data pairs of data from a distribution (X, Y) with $Y = f(X) + \epsilon$, where X is either in a fixed grid (fixed design regression) or random (random design), and ϵ is some noise independent of X. We want to estimate f given many pairs (X, Y). We will first see the example of parametric regression, but the problem we are more interested in is nonparametric regression.

Example 1.3.2 (Linear regression)**.** *In linear regression, we assume that the function is linear $f_\beta(X) = f_\beta(X) = \beta^T X$. Our goal is to estimate the unknown vector β. We then can put a prior on the weights β and use this to infer something on the function f_β.*

In general, this linearity assumption might not hold, and we might want to use broader classes of functions. One example is smoothness conditions, which state that the function we want to estimate has to be smooth enough. Under these assumptions, we can build various estimators. One example of models that one can use to build estimators is series models. Another method can be Deep Learning. Deep learning is applied, and Bayesian deep learning can learn these models. While the final purpose is to give empirical Bayesian estimators which yield valid uncertainty estimation, we will not do this in these notes, although the approach works along similar lines.

Example 1.3.3 (Nonparametric regression: series estimation)**.** *In our setup, we will use a basis expansion on a suitable basis. We need to be able to control the bias and variance properly. We can control the bias and variance by choosing the right number of basis functions and using the right type of basis functions. For various purposes, it helps if the basis functions are, in some sense, nearly orthogonal. Using that the basis functions are nearly orthogonal allows for translations of different metrics, in particular, allows one to do the entropy computations within the model itself using*

Euclidean distances instead of the L^2 distance in function space. This equivalence of metrics allows you to use the tools we will develop later.

If we assume the basis of basis functions B to be nearly orthogonal in the sense that for all θ

$$\|\theta\|_2^2 \lesssim \|\theta^T B\|_2^2 \lesssim \|\theta\|_2^2$$

Here $\lesssim$ means less than or equal to up to some constant factor independent of n on which our basis functions depend.

We can pick a finite number of basis functions and increase the dimension with the number of data points. Hence we put a finite-dimensional prior on the coefficients. Then the previous condition implies that Euclidean balls generated by the metric on parameters are of comparable size to the balls generated by the L^2 metric. Thus we can do all the entropy conditions with the Euclidean metric instead of the L^2 and only have to change ϵ to $\xi\epsilon$ for some fixed (but possibly unknown) ξ.

1.3.3 The posterior

In Bayesian inference, we compute a posterior distribution. The posterior distribution is the conditional probability of the parameters given the prior and the data. In general, these conditional probabilities are not uniquely defined. Hence we cannot uniquely define the posterior probabilities. However, the different versions of a posterior distribution have to agree almost everywhere. Moreover, we will often work with a prefered posterior distribution. One way of constructing such a prefered posterior distribution is Bayes Theorem.

Theorem 1.3.4. *Suppose that $X|\theta \sim P_\theta$, where P_θ is a dominated family of measures with measurable densities p_θ with respect to some σ-finite dominated measure, then one version of the posterior distribution in the model that $\theta \sim \Pi$ is:*

$$\Pi(\theta \in B|X) = \frac{\int_B p_\theta(X)\, d\Pi(\theta)}{\int p_\theta(X)\, d\Pi(\theta)}$$

In general, such dominating measures might not exist. Thus we cannot always apply Bayes' theorem but need to use other tools. One example where you need to use other tools to compute the posterior distribution is distribution estimation with the Dirichlet or Pitman-Yor process. In distribution estimation, there is no σ-finite dominating measure. Therefore we cannot apply Bayes' formula.

An important concept to note is posterior conjugacy. Suppose we pick a prior which belongs to a family of distributions $\mathcal{D}$. We say the prior distribution is conjugate if the posterior distribution is also in $\mathcal{D}$. This conjugacy often allows for explicit computations. These computations can make your life much easier. See, for example, the coin-flipping example. Here, you can compute everything without using the tools developed in the coming sections. However, in many examples, no conjugate priors are known. They might not even exist in any useful sense, so the tools are needed.

Example 1.3.5 (Coin flips continued)**.** *In the coin flip example, we used a Beta*$(1, 1)$ *prior. This prior is a conjugate prior for this model. Hence the posterior is again a Beta distribution. If you have* h *heads and* t *tails, the posterior distribution of* p *given* $X_1, \ldots, X_n$ *is given by Beta*$(1 + h, 1 + t)$*.*

1.4 Frequentist analysis of Bayesian methods

We can study Bayesian methods as any other method from a frequentist point of view. This gives rise to the theory of frequentist Bayes.

1.4.1 Consistency, rates and uncertainty quantification for Bayes

In the Bayesian methodology, there is no reference to the true generating process, just data and your beliefs. We can now wonder if a Bayesian can find the truth if it exists. We will start with Freedman's inconsistency theorem. This result shows that truth finding is not automatic. Hence you must be careful in selecting your priors to match the problem.

We begin with Freedman's inconsistency theorem [29, 33]. This theorem tells us that most priors have no chance of finding the truth. To illustrate this, consider the collection of pairings of priors Π and "truths" P_0. Call a prior compatible with truth P_0 if, with infinite data from P_0, a Bayesian with prior Π would conclude that the true distribution is P_0. Then the collection of compatible pairs is small. As an even stronger statement, they construct prior and truth combinations such that every non-empty open set of parameter values will get posterior probability arbitrarily close to 1 infinitely often. This behaviour means that, for these combinations, the posterior distribution will wander around without converging.

Based on this result, one might conclude that the Bayesian analysis might be hopeless. However, there are a lot of positive results as well. It shows that one should be more careful than just permitting any prior: you need a more refined approach. We make consistent priors for models. Consistency means that, with infinite data, we will find the true process. To make precise what consistency means for posterior distribution, we define it and then discuss its meaning.

Definition 1.4.1. *The posterior distribution* $\Pi_n(\cdot|X^{(n)})$ *is said to be consistent at* $\theta_0 \in \Theta$ *if* $\Pi_n(U^c|X^{(n)}) \to 0$ *in* $P_{\theta_0}^{(n)}$ *probability, as* $n \to \infty$*, for every open set* U *containing* θ_0*.*

Let's take a closer look at this definition. The purpose of this definition is to state that all the posterior mass will concentrate on the truth. The posterior probability of the true parameter θ_0 can be zero. For example, this happens when the posterior distribution admits a Lebesgue density. Since the posterior probability of θ_0 can be zero, we cannot ask that the posterior probability of θ_0 has to converge to 1. This limitation would rule out many models that people want to use. The next best thing is to ask if all parameter values θ that are in some sense near θ_0 will get a large posterior mass.

We are analyzing Bayesian inference from a frequentist point of view. Hence we can wonder if we can use the posterior distribution to create an estimator which is frequentistically valid and looks more like the usual point estimators typically seen in frequentist literature. The following result gives such a construction, which works without any extra assumptions other than consistency. We can replace the mass bound of 1/2 to any arbitrary value between 0 and 1, see [33, Theorem 6.7].

Theorem 1.4.2. *Suppose that the posterior distribution $\Pi_n(\cdot|X^{(n)})$ is consistent at θ_0 relative to the metric d on Θ. Then $\hat{\theta}_n$, defined as the centre of a (nearly) smallest ball that contains posterior mass at least 1/2 satisfies $d(\hat{\theta}_n, \theta_0) \to 0$ in $P_{\theta_0}^{(n)}$-probability.*

This result implies that if frequentist methods cannot solve a problem, Bayesian methods cannot solve it either. Therefore, Bayesian methods do not solve problems like the no-free-lunch theorem. A slightly more refined result also shows that the frequentist minimax theory will apply to Bayesian methods, which we will see in the next section.

Before we continue, let us study consistency with explicit tools in the coin flip example.

Example 1.4.3 (Consistency for coin flips by explicit computation)**.** *In Example 1.3.5, we saw the explicit posterior. The posterior allows us to do computations directly. Using the formulas for the mean and variance of a Beta distribution we can compute these. The posterior mean is $\frac{1+h}{2+n}$ and the posterior variance is $\frac{(1+h)(1+t)}{(n+2)^2(n+3)}$. If $0 < p < 1$ then by the strong law of large numbers, $\frac{1+h}{n} \to p$ and $\frac{1+t}{n} \to 1-p$. Therefore the posterior variance converges to zero, while the posterior mean converges to p. Thus we can apply Markov's inequality to conclude that the posterior is consistent.*

Now we have seen how to find the truth eventually. But we can wonder how quickly we can learn it as well. Intuitively, it would come down to rejecting all hypotheses that are far away from the truth. We want to quantify how fast we would zoom in on the truth. Let's start with the formal definition first.

Definition 1.4.4. *We say that the posterior contracts at a rate ϵ_n with respect to a pseudo-metric d if for all $M_n \to \infty$*

$$\Pi\left(d(\theta, \theta_0) \geq M_n \epsilon_n | X^{(n)}\right) \to 0,$$

in $P_0^{(n)}$ probability.

Note that we talk about "a" rate and not "the" rate. We do that since rates are not unique: any larger ϵ_n will again be a valid rate. We want to show ϵ_n to be as small as possible. This minimality means we are focussing on the target as fast as possible. One way of finding the optimal rate is when you are neither overfitting nor underfitting.[1]

One should reduce the bias coming from the models as much as possible, but not at the cost that the variance starts increasing more than the bias decreases. The precise balance depends on the underlying complexity of the model. In parametric models, one can often have zero bias while achieving an optimal rate, while in nonparametric statistics, a bias variance trade-off is often necessary. In reality, one does not know the correct complexity and does not know how to tune it before seeing the data. However, you can extend these results given here. These extensions can make adaptive rates of concentration possible. You can make adaptive estimators that "learn" the underlying complexity in the sense that you can use these estimates to get a (near) optimal contraction rate.

As with the consistency example, contraction rates of the posterior allow us to find a point estimator with matching convergence rates. This shows that the usual theory of lower bounds and minimax estimators of the frequentist literature also applies to Bayesian estimators. We have the following corresponding result of Theorem 1.4.2, which makes this notion precise. See also [33, Theorem 8.7].

Example 1.4.5 (Rates of contraction by explicit computations)**.** *To show that the posterior concentrates on balls of radius $\sqrt{n}$ around p_0 we can reason as follows. First, we know the posterior explicitly by Example 1.3.5. Denote $\hat{p}_n$ the posterior mean. To show a contraction rate $\frac{1}{\sqrt{n}}$ we need to show that a ball of radius $\frac{M_n}{\sqrt{n}}$ contains an arbitrarily high amount of mass. Consider a ball around $\hat{p}_n$ with radius $\|\hat{p}_n - p_0\| + \frac{M_n}{3\sqrt{n}}$. Then p_0 is inside this ball. Now we want to show the following two facts:*

- *The posterior assigns arbitrary high mass to this ball.*
- *This ball is contained in the ball around p_0 with radius M_n.*

To show the second, we use the properties of the Maximum likelihood estimator $\hat{p}_n$. This states that the convergence rate of $\hat{p}_n$ is $\frac{1}{\sqrt{n}}$. Hence it follows that, for every $M_n \to \infty$, $\|\hat{p}_n - p_0\| \leq \frac{M_n}{3\sqrt{n}}$. In particular for our choice of M_n. Hence we are contained in the ball with arbitrarily high probability. Next is to show that the posterior assigns arbitrary high mass to this ball. For this, we can use properties of the Beta distribution and the Markov inequality. This step gives a bound of

$$\Pi\left(\|p - \hat{p}_n\| \geq \|\hat{p}_n - p_0\| + \frac{M_n}{3\sqrt{n}} | X\right) \leq \frac{\frac{(1+h)(1+n-h)}{(2+n)^2(3+n)}}{\left(\frac{M_n}{3\sqrt{n}} + \|\hat{p}_n - p_0\|\right)^2}.$$

[1] You overfit when you have a too small bias at the cost of too much variance. You underfit when you have a too-small variance at the cost of too much bias.

Note that $\frac{(1+h)(1+n-h)}{(2+n)^2}$ converges almost surely to $p(1-p)$. We have also seen that with high probability $\|\hat{p}_n - p_0\| \leq \frac{M_n}{3\sqrt{n}}$. Hence with high probability, we can bound our upper bound into

$$\frac{3}{2M_n}.$$

This upper bound converges to zero as $M_n \to 0$. Hence our rate of contraction is $\sqrt{n}$.

1.4.2 Schwartz Theorem

Our goal was to motivate the Bayesian methodology from the frequentist point of view. One method of doing that is the Schwartz theorem. Compared to the Bernstein-von Mises theorems we will cover in Section 1.4.3 we tend to have weaker assumptions and conclusions. However, the conclusions are still very positive. Paraphrasing:

- If the truth can be recovered by any method, it can be recovered by any posterior given that the prior puts positive mass at every neighbourhood of the truth.

- If the truth can be recovered at minimax rates by any method, it can be recovered by any posterior given that the prior mass put in shrinking neighbourhoods does not decay too quickly.

1.4.2.1 Kullback-Leibler divergence and variation

In dominated models, the likelihood of the data plays an important role in statistics. The likelihood shows up in frequentist statistics since the maximum likelihood estimator is one of the core tools that people use. In Bayesian statistics, the likelihood shows up in Bayes' rule. Since the logarithm is increasing and maps products to sums, the average log-likelihood will converge to its expected value if the latter exists. If we want to apply central limit type arguments we also want to control the variance of the log-likelihood.

Since the likelihood is driving our decision making, we want to study which models have a similar (expected) likelihood. Since log-likelihood is better behaved, we introduce a distance based on the expected log-likelihood. This leads to the Kullback-Leibler divergence and variation as distances. The Kullback-Leibler variation distance is a bit stronger, so harder to control, while the Kullback-Leibler divergence is weaker, and harder to use. For our purposes, the Kullback-Leibler divergence will be enough.

Definition 1.4.6. *Assume that P_θ is a dominated family with respect to some σ-finite dominating measure.*

- *The Kullback-Leibler divergence is*

$$KL(p_0, p_\theta) = P_0 \log(\frac{p_0}{p_\theta}).$$

- *The Kullback-Leibler variation is*

$$V(p_0, p_\theta) = P_0 \left(\log(\frac{p_0}{p_\theta}) - P_0 \log(\frac{p_0}{p_\theta}) \right)^2 .$$

We now want to look at models for which there is positive mass in any Kullback-Leibler neighbourhood of the truth, which will become the Kullback-Leiber property.

Definition 1.4.7. *A density p_0 is said to possess the Kullback-Leibler property relative to a prior Π if $\Pi(p\colon KL(p_0, p) < \epsilon) > 0$ for all $\epsilon > 0$. This is denoted $p_0 \in KL(\Pi)$.*

This definition states that the prior is putting some mass in any neighbourhood of the parameter θ_0. Let us see how this looks in our two examples.

Example 1.4.8 (KL divergence in coinflips)**.** *The KL divergence between two Bernoulli random variables is given by $KL(p, q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$.*

Example 1.4.9 (KL divergence in fixed design regression)**.** *In the fixed design regression model, we get instead the squared empirical ℓ^2 distance,*

$$d_n(f_0, f)^2 = \frac{1}{n} \sum_{i=1}^{n} (f_0(X_i) - f(X_i))^2$$

as the KL divergence.

1.4.2.2 Entropy

Entropy gives us a way of measuring how complex a space is. Roughly speaking, it tries to capture how many different things you can see at a given resolution. There are various kinds of entropy, and they each have their uses. For our current purposes, we introduce the covering numbers. This measures how many metric balls you need to use to cover the entire space. We say a collection of balls covers a space if the space is a subset of the union of these balls.

Definition 1.4.10. *We say that a space Θ has covering number $N(\epsilon, \Theta, d)$ with respect to a metric d if you need at least $N(\epsilon, \Theta, d)$ d-balls of radius ϵ to cover the entire space.*

One useful result is the covering number of euclidean balls:

Lemma 1.4.11. *Let $\|x\|_p^p = \sum_i |x_i|^p$. Then for any M and $0 \leq \epsilon \leq M$,*

$$N(\epsilon, \{x \in \mathbb{R}^d \colon \|x\|_p \leq M\}, \| \cdot \|_p) \leq \left(\frac{3M}{\epsilon} \right)^d .$$

Let us use Lemma 1.4.11 to bound the entropy in our two examples.

Example 1.4.12 (Entropy for coinflips). *In the coin flip example, the parameters live in the interval* $[0,1]$*. By Lemma 1.4.11 we get entropy bound of*

$$N(\epsilon, \{x \in [0,1]\}, \| \cdot \|) = \frac{3}{\epsilon}.$$

Example 1.4.13 (Entropy in fixed design regression). *In the fixed design model with series estimators, we have* $K_n = n^{1/(2\beta+d)}$ *many variables. Under the near orthogonality assumption, it follows that the Euclidean distance within the model parameters satisfies, for some* $C \geq 1$ *independent of* n $\frac{1}{C}\|\theta\| \leq \|\theta^T \phi\| \leq C\|\theta\|$*. So the local entropy is given by*

$$N(\epsilon, \{x \in \mathbb{R}^{K_n}, \|x\|_2 \leq M\}) \leq \left(\frac{3MC}{\epsilon}\right)^{K_n}.$$

1.4.2.3 Testing

For the theory of Bayesian nonparametrics, it turns out that testing is a fundamental concept for understanding consistency and rates. A test is a measurable function from the space of observables to the interval $[0,1]$. We want to construct tests that are powerful enough to use for the rest of the theory. The current results are adapted from [33, Appendix D].

We begin with the easier statement that gives consistent testing.

Theorem 1.4.14. *Given a distance d that generates convex balls and satisfies* $d(p_0, p) \leq d_H(p_0, p)$ *for every p. Suppose that*

$$\log N(\epsilon, \Theta, d) \leq n\epsilon_n^2.$$

Then there exists tests ϕ_n *such that*

$$P_0^n \phi_n \leq e^{-n\epsilon^2}, \qquad \sup_{p: d(p_0,p) > 4\epsilon} P^n(1 - \phi_n) \leq e^{-2n\epsilon^2}.$$

This theorem gives some conditions for any metric which is dominated by the Hellinger metric. For stronger results, we need to do a bit more work, but we can go beyond the Hellinger metric by being a bit more precise in our other assumptions.

Theorem 1.4.15. *Suppose that for universal constants* $\xi, K > 0$ *and for any* $\epsilon > 0$ *and any density* p_1 *with* $d(p_0, p_1) > \epsilon$ *there exist a test* ϕ_n *with*

$$P_0^n \phi_n \leq e^{-Kn\epsilon^2}, \qquad \sup_{p: d(p,p_1) < \xi\epsilon} P^n(1 - \phi_n) \leq e^{-Kn\epsilon^2},$$

and that for a sequence ϵ_n

$$\sup_{\epsilon > \epsilon_n} \log N(\xi\epsilon, \{p \colon d(p, p_0) \leq 2\epsilon\}, d) \leq n\epsilon_n^2.$$

Then there exist tests such that, for every $j \in \mathbb{N}$,

$$P_0^n \phi_n \to 0, \qquad \sup_{p: j\epsilon_n \leq d(p,p_0) \leq (j+1)\epsilon_n} P^n(1-\phi_n) \leq e^{-Kn\epsilon_n^2 j^2}.$$

The first condition tells us that we can test locally on balls. The other part where it differs is that it builds the tests using the local entropy around the truth, and the global entropy can be higher. The strategy for making the global test is based on glueing local tests together. This result is precise enough to give the tests we need in the contraction rates section.

Example 1.4.16 (Testing in coin flips and nonparametric regression)**.** *We have seen the entropy bounds for our examples in Examples 1.4.12 and 1.4.13. They both satisfy the assumptions of Theorem 1.4.15 with rates $\epsilon_n = \frac{1}{\sqrt{n}}$ and $\epsilon_n = n^{-\beta/(2\beta+d)}$ respectively. Hence we get tests with these specified rates.*

1.4.2.4 Schwartz theorem

With the help of the Kullback-Leibler property and the existence of tests, we can formulate the result of Schwartz theorem [33, Theorem 6.16].

Theorem 1.4.17. *If $p_0 \in KL(\Pi)$ and for every neighbourhood U of p_0 there exists tests ϕ_n such that $P_0^n \phi_n \to 0$ and $\sup_{\theta \in U^c} P_\theta^n(1-\phi_n) \to 0$, then the posterior distribution $\Pi_n(\cdot|X^{(n)})$ in the model $X_1, \ldots, X_n | p \overset{i.i.d.}{\sim} p$ and $p \sim \Pi$ is strongly consistent at p_0.*

Schwartz's theorem states that all you need for consistency are two conditions. The first condition is that there should be some prior mass near the truth. The second one is a condition that roughly translates to being able to recover the truth in this model with some method. This condition means that if your prior is flat enough, as soon as one method can find the truth, Bayesian methods can also find it. It should not be too big of a surprise that the testing condition is necessary. It talks about the recovery of the model with any method, and we could, in principle, create tests using the consistency result and Theorem 1.4.2.

This result shows that if there is something to be learned, Bayesian methods will also be able to learn. At least if we design our priors to be flat enough. Now we can wonder if we can find easier tools to take care of the testing part. The following theorem is as stated in [33, Theorem 6.23]

Theorem 1.4.18. *Under the assumptions of Theorem 1.4.14, for any $p_0 \in KL(\Pi)$, the posterior distribution in the model $X_1, \ldots, X_n | \theta \sim P_\theta$ and $\theta \sim \Pi$ is consistent relative to d.*

Since Schwartz's theorem is quite important, we can take a peek under the hood at how it works. Since we are working in dominated models, we can use Bayes' formula to find an expression for the posterior probability of U^c:

$$\frac{\int_{U^c} \prod_{i=1}^n p_\theta(X_i) \, \mathrm{d}\, \Pi(\theta)}{\int \prod_{i=1}^n p_\theta(X_i) \, \mathrm{d}\, \Pi(\theta)}.$$

We want to show that this goes to zero, so if we can show that the denominator is large and the numerator is small, we are done. If we multiply and divide by $\prod_{i=1}^n p_0(X_i)$ we can reorganise this to read

$$\frac{\int_{U^c} \prod_{i=1}^n \frac{p_\theta}{p_0}(X_i) \, \mathrm{d}\, \Pi(\theta)}{\int \prod_{i=1}^n \frac{p_\theta}{p_0}(X_i) \, \mathrm{d}\, \Pi(\theta)}.$$

If we use the almost sure evidence lower bound [33, Theorem 6.16, Lemma 6.26] we find that for all $c > 0$ eventually almost surely

$$\int \prod_{i=1}^n \frac{p_\theta}{p_0}(X_i) \, \mathrm{d}\, \Pi(\theta) \geq e^{-cn}$$

Meanwhile, using the tests to make exponentially powerful tests, we can find a $C > c$ such that eventually almost surely:

$$\int_{U^c} \prod_{i=1}^n \frac{p_\theta}{p_0}(X_i) \, \mathrm{d}\, \Pi(\theta) \leq e^{-Cn}.$$

Combining this yields Schwartz's theorem. Now we can apply this to get consistency in the coin flip model.

Example 1.4.19 (Consistency for coin flips)**.** *In the coin flip example, the Kullback-Leibler balls are open sets. The uniform prior gives positive mass to all non-empty open sets. Thus every open set containing p gets a positive mass. This means that p is in the Kullback-Leibler support of the prior. By Example 1.4.16, there exist tests. Therefore we can apply Schwartz's theorem and get consistency.*

1.4.2.5 Refinements of Schwartz theorem

Schwartz theorem was our tool to show consistency. It requires that the true distribution satisfies the Kullback-Leibler condition and that consistent tests exist. We will refine both conditions. We will control how fast the mass of Kullback-Leibler neighhours decays. And we will use a refined argument based on the existence of tests. Furthermore, we can simplify this by constructing tests using entropy conditions.

Assumption 1.4.20. *For the constant K in Theorem 1.4.21, we assume that the metric balls satisfy this probability bound:*

$$\frac{\Pi_n\left(p\colon j\epsilon_n < d(p, p_0) \leq 2j\epsilon_n\right)}{\Pi_n(p\colon KL(p_0, p) \leq \epsilon_n)} \leq e^{Kn\epsilon_n^2 j^2/2}$$

This assumption compares the prior mass of the parameters which are close to the true parameter p_0 according to the metric, and the mass of the true parameters which are close according to the Kullback-Leibler divergence. It tries to measure how much the likelihood can distinguish between hypotheses that are close to each other under our metric. If this assumption is satisfied, and we can use some testing arguments, we can expect good rates. See also [33, Theorem 8.12].

Theorem 1.4.21. *Assume that Assumption 1.4.20 holds for constants $\bar{\epsilon}_n \leq \epsilon_n$ with $n\bar{\epsilon}_n^2 \geq 1$ and every sufficiently large j, and in addition there exist tests ϕ_n such that, for some constant $K > 0$ and all large enough j*

$$P_0^n \phi_n \to 0, \qquad \sup_{p: j\epsilon_n \leq d(p,p_0) \leq 2\epsilon_n} P^n(1-\phi_n) \leq e^{-Kn\epsilon_n^2 j^2},$$

then the posterior rate of contraction at p_0 is ϵ_n.

The proof of this theorem is a slightly refined proof of the Schwartz theorem. The Kullback-Leibler argument tells us that the prior mass should not decrease too fast in a neighbourhood of the true parameter, since if we make the prior think parameters which are near the truth are very unlikely, we shouldn't expect great rates. The testing assumption is almost necessary. Under an additional conditional on the exponential decay of the posterior tail you can prove that, if one has contraction rates, you can construct tests.

Now we can wonder how we can construct such tests. We can construct test using entropy numbers as in Schwartz's theorem. This construction requires the assumptions of the testing theorems. If we combine these assumption we get the following result (see also [33, Theorem 8.11])

Theorem 1.4.22. *Assume that the assumptions in Theorem 1.4.15 hold and that Assumption 1.4.20 holds with the K as in Theorem 1.4.15. Then the posterior contracts at rate ϵ_n.*

We can even go a bit further than these theorems. You can balance the prior probability and entropy conditions, but this is beyond the scope of this note.

Example 1.4.23 (Contraction rates for coin flips)**.** *In the coin-flip model, suppose that $0 < p_0 < 1$. By Example 1.4.8 we know how the KL balls look. There exists a $C_p > 0$ such that the ball of radius $\frac{C_p}{\sqrt{n}}$ around p_0 is included in the Kullback-Leibler ball around p_0. As we have seen in Example 1.4.16, there exists tests at $\frac{1}{\sqrt{n}}$ rate. This prior mass result combined with the entropy bounds from Example 1.4.12, allow us to apply Theorem 1.4.22. This uses the less refined bound for entropy, and this gives a contraction rate of $\sqrt{\frac{\log(n)}{n}}$. If we were to use Lemma 1.4.11 directly to the set $B(p_0, C\epsilon)$, we would get a $\sqrt{\frac{1}{n}}$ contraction rate instead.*

Example 1.4.24 (Contraction rates for fixed design regression)**.** *In the fixed design regression model, we can again verify the assumptions of Theorem 1.4.22. Though, the computations become more involved. See for example [67].*

1.4.3 Bernstein-von Mises theorems

There are several ways in which you can give theoretical guarantees for the posterior distribution. In this section we are going to study the so-called *Bernstein-von Mises* (BvM) theorem. We first explain the rough idea behind the BvM theorem and then go into the details. The posterior distribution is a probability distribution. We want to understand the asymptotic behaviour of this distribution. To understand this, we want to show that asymptotically the posterior distribution is, in some sense, close to a limit distribution. Preferably, we want to have that the limiting distribution is "easy to understand". This limiting distribution will have to depend on the data. Let us denote this distribution by $W_n(X^{(n)})$. We will denote our measure of closeness by a function d. So, in the end, we want to prove a statement that looks like

$$d\left(\Pi\left(\cdot|X^{(n)}\right), W_n(X^{(n)})\right) \overset{P_0}{\rightsquigarrow} 0$$

Typical choices of a distance measure will be the total variation distance and the bounded Lipschitz distance. The total variation distance can be seen as the L^1-distance between the two densities, while the bounded Lipschitz metric metrises the weak topology.

To understand the finer details of the posterior distribution, we might want to "rescale" it. We often have a posterior distribution that starts concentrating near a specific point, usually the true parameter θ_0. We want to study how it starts concentrating around that point. To do this, we need to be able to see finer details. We can rescale the posterior. This rescaling will allow us to see finer details. We can do the rescaling as follows. We pick a centring point $\hat{\theta}_n$ and a rescaling rate r_n. Now, look at the distribution of

$$r_n(\theta - \hat{\theta}_n),$$

where $\theta \sim \Pi(\cdot|X^{(n)})$. This rescaling makes the differences between θ and $\hat{\theta}_n$ bigger. And in turn, this allows us to see more details. We can use this to reformulate our goal slightly. We again have some target limiting distribution W_n. The limiting distribution depends on our data $X^{(n)}$. However, we rescale the posterior to allow for finer details. We want to prove statements of the form:

$$d\left(\Pi\left(r_n\left(\cdot - \hat{\theta}_n\right)|X^{(n)}\right), W_n(X^{(n)})\right) \overset{P_0}{\rightsquigarrow} 0$$

The centring point $\hat{\theta}_n$ and the asymptotic distribution $W_n(X^{(n)})$ will both be important in our analysis. Usually, we want to show that the centring point is a good estimator. However, this is not always the case. Sometimes the centring point has a bias. To make this bias clear, we can write $\hat{\theta}_n = \tilde{\theta}_n + B_n$ where $\tilde{\theta}_n$ would be some good estimator and B_n some bias term.

Often, we want that the asymptotic distribution does not depend on the data except for the true data-generating process P_0. This independence is often the case. In that case, we can drop the dependence and simplify our desired statement. This simplification would lead to a statement of the form

$$d\left(\Pi\left(r_n\left(\cdot - \hat{\theta}\right)|X^{(n)}\right), W\right) \overset{P_0}{\rightsquigarrow} 0.$$

If the distance is the bounded Lipschitz metric, the statement often gets rephrased. Since we require the distance to converge weakly to zero, it also converges in probability to zero. The bounded Lipschitz metric metrises the weak topology. Hence we have "weak convergence in probability". Then we say that

$$r_n\left(\theta - \tilde{\theta}_n - B_n\right) \mid X^{(n)} \rightsquigarrow W$$

in P_0-probability.

We need two conditions to use BvM theorems to motivate Bayesian methods from a frequentist point of view. First, the limiting distribution should have "good" properties from a frequentist point of view. Secondly, there needs to be a way to translate these properties from the limiting distribution to the posterior distribution.

1.4.3.1 Using BvM to give frequentist guarantees

Suppose we have a BvM theorem for the Bounded Lipschitz metric. That is, we know that:

$$r_n\left(\theta - \tilde{\theta}_n - B_n\right) \mid X^{(n)} \rightsquigarrow W$$

in P_0-probability. Here $\tilde{\theta}_n$ is a "good" estimator and B_n is a bias term. We assume that the asymptotic distribution is centred at zero. Our goal is to study consistency, contraction rates and coverage.

We will start with consistency. Lemma 1.4.25 states that if the posterior distribution satisfies a BvM in the bounded Lipschitz metric, the posterior distribution will be consistent.

Lemma 1.4.25. *Suppose that either $r_n \to \infty$ and W is continuous or $r_n = O(1)$ and W is a point mass at zero. Assume that $\hat{\theta}_n \overset{P_0}{\rightsquigarrow} \theta_0$. Suppose that*

$$r_n\left(\theta - \tilde{\theta}_n - B_n\right) \mid X^{(n)} \rightsquigarrow W$$

in P_0-probability. If, moreover,

$$\tilde{\theta} + B_n \rightsquigarrow \theta^*$$

the posterior will concentrate on θ^*.

We sketch a proof with a few minor details missing for case that $r_n \to \infty$. The other case uses that $W = \delta_0$ to give the conclusion almost immediately.

Proof. We want to show that the posterior concentrates on θ^*. That means that we have to show that for every open neighbourhood U of θ^* the posterior probability of U^c converges to zero. Because $\hat{\theta}_n + B_n$ converges weakly to θ^*, we know that eventually $\hat{\theta}_n + B_n \in U$. Pick $0 < \alpha < 1$. Let V_α be a neighbourhood of 0 such that W assigns probability at least α to V. Condition on the event that $\theta_n + B_n \rightsquigarrow \theta_0$. Then eventually $\hat{\theta}_n + B_n + \frac{1}{r_n} V_\alpha$ is a subset of U. Hence the posterior probability of U is at least

$$\Pi\left(\hat{\theta}_n + B_n + \frac{1}{r_n} V_\alpha |\, X^{(n)}\right).$$

Moreover, because of the BvM result, we know that the rescaled posterior converges weakly to W. By the portmanteau Lemma, it follows that the difference between

$$\Pi\left(\cdot \in \hat{\theta}_n + B_n + \frac{1}{r_n} V_\alpha |\, X^{(n)}\right)$$

and

$$W(V_\alpha)$$

converges to zero. Hence $\Pi\left(U |\, X^{(n)}\right) \geq \Pi\left(U \cap V_\alpha |\, X^{(n)}\right) \rightsquigarrow \alpha$. Hence

$$\Pi\left(U^c |\, X^{(n)}\right) \rightsquigarrow 1.$$

This is what we wanted to show. ■

This result can be used to show that a posterior distribution is consistent when the estimator it starts concentrating on converges to the true parameter θ_0:

Corollary 1.4.26. *Assume in addition that $\theta_n + B_n \rightsquigarrow \theta_0$. Then the posterior is consistent.*

This result follows directly because the posterior concentrates on $\theta^* = \theta_0$.

In a similar fashion, one can prove the following result for contraction rates. If the posterior distribution satisfies a BvM theorem at rate r_n, the posterior contracts at rate $\frac{1}{r_n}$.

Lemma 1.4.27. *Let $r_n \to \infty$. Assume that W is a continuous distribution. Assume that the model is well specified, i.e. there exists a θ_0 such that $P_{\theta_0} = P_0$. Suppose that $\mathbb{E}[d(\hat{\theta}_n, \theta_0)] \lesssim \frac{1}{R_n}$. Suppose that the BvM holds, i.e.*

$$r_n\left(\theta - \hat{\theta}_n\right) |\, X^{(n)} \rightsquigarrow W$$

in P_0-probability. Then the posterior contracts at rate $\max(\frac{1}{r_n}, \frac{1}{R_n})$.

In both previous results, the asymptotic distribution W was not very important. This changes when we want to study the validity of uncertainty quantification. To get the most information, we split the centring point $\hat{\theta}_n$ into two, $\tilde{\theta}_n$ and B_n. We will use another distribution Z. This distribution is the asymptotic distribution of $\tilde{\theta}_n$. Instead of focusing on θ_0, we will focus on a sequence of centring points θ_n^*. This sequence of centring points preferably will equal θ_0. However, in some of our results, we need to pick a different choice of θ_n^*. To give a frequentist interpretation, we allow for these choices of centring points in our results. That is, we assume that:

$$r_n \left(\tilde{\theta}_n - \theta_n^*\right) \rightsquigarrow Z.$$

Because $\tilde{\theta}_n = \hat{\theta}_n - B_n$ we also know that

$$r_n \left(\hat{\theta}_n - B_n - \theta_n^*\right) \rightsquigarrow Z.$$

Thus we can use B_n as a bias correction. We will study the coverage of θ_n^* by specific credible sets. If θ_n^* converges fast enough, that is, faster than rate r_n^{-1}, to θ_0 this will also give coverage of θ_0. However, the coverage might not be the same as the credibility level used to construct the credible set.

Lemma 1.4.28. *Let $r_n \to \infty$. Assume that W possesses a continuous distribution. Let Z be possesses continuous distribution such that $r_n \left(\tilde{\theta}_n - \theta_n^*\right) \rightsquigarrow Z$. Suppose that the posterior distribution satisfies the BvM theorem, i.e.*

$$r_n \left(\theta - \hat{\theta}_n\right) \mid X^{(n)} \rightsquigarrow W$$

in P_0-probability. Let $0 < \alpha < 1$. Let $C_{\alpha,n}$ be a ball centred at a point c_n and radius $\frac{\rho_n}{r_n}$ such that

- $r_n \left(c_n - \hat{\theta}_n\right) \rightsquigarrow 0$*;*
- *$C_{\alpha,n}$ is a credible ball of level $1 - \alpha$: $\Pi \left(C_{\alpha,n} \mid X^{(n)}\right) \geq 1 - \alpha$;*
- *For all $\rho < \rho_n$ the credibility level of the ball $B(c_n, \rho)$ centered at c_n with radius ρ is less than $1 - \alpha$:*

$$\forall \rho < \rho_n \Pi \left(B(c_n, \rho) \mid X^{(n)}\right) < 1 - \alpha.$$

Then $\rho_n \to \rho_\alpha$, the radius of a ball centred at zero such that $W(B(0, \rho_\alpha)) = 1 - \alpha$,

$$\mathbb{P}(\theta_n^* \in C_{\alpha,n} - B_n) \to Z\left(B(0, \rho_\alpha)\right)$$

If in addition $r_n(\theta_n^* - \theta_\infty^*) \rightsquigarrow 0$, then also

$$\mathbb{P}(\theta_\infty^* \in C_{\alpha,n} - B_n) \to Z\left(B(0, \rho_\alpha)\right).$$

We give a sketch of a proof. The proof can be made rigorous by taking care of the conditioning.

Proof. W and Z are continuous distributions. Thus the maps $\rho \mapsto W(B(0,\rho))$ and $\rho \mapsto Z(B(0,\rho))$ are continuous. We are first going to show that the radius of the credible ball converges to the right radius. To do that we use a sandwiching argument. For any level β we will study the credibility of balls $B_n(\beta)$ centred on c_n with radius ρ_β / r_n where we pick ρ_β such that

$$W(B_n(\beta)) = 1 - \beta.$$

By the BvM theorem, it follows that

$$\Pi\left(B_n(\beta) \middle| X^{(n)}\right) \overset{P_0^n}{\to} W(B_n(\beta)) = 1 - \beta.$$

Let ρ_n be the radius of our credible ball $C_{n,\alpha}$. If $\limsup \rho_n > \rho_\beta$, it follows that infinitely often $B_n \subset C_{\alpha,n}$. But then for all n large enough

$$\Pi\left(C_{\alpha,n} \middle| X^{(n)}\right) > 1 - \beta > 1 - \alpha.$$

This result contradicts our assumption on $C_{\alpha,n}$. Similarly, we can bound the liminf and conclude that $\rho_n \rightsquigarrow \rho_\alpha$.

Now the probability that θ_n^* is contained in the debiased credible set $C_n - B_n$ is

$$\begin{aligned}
\mathbb{P}\left(\theta_n^* \in C_n - B_n\right) &= \mathbb{P}\left(\theta_n^* \in B(c_n - B_n, \rho_n / r_n)\right. \\
&= \mathbb{P}\left(r_n(\theta_n^* - c_n + B_n) \in B(0, \rho_n)\right) \\
&= \mathbb{P}\left(r_n(\theta_n^* - c_n + B_n) \in B(0, \rho_\alpha)\right) + o(1) \\
&= \mathbb{P}\left(r_n(\hat{\theta}_n - \theta_n^* - B_n) \in B(0, \rho_\alpha)\right) + o(1) \\
&\to Z(B(0, \rho_\alpha)).
\end{aligned}$$

By sandwiching, using Slutsky's Lemma applied to $r_n(\hat{\theta}_n - c_n) \rightsquigarrow 0$ and using the asymptotic distribution of $\hat{\theta}_n$. If in addition $r_n(\theta_n^* - \theta_\infty^*) \rightsquigarrow 0$, then also, by another application of Slutsky's Lemma:

$$\mathbb{P}\left(\theta_\infty^* \in C_n - B_n\right) \to Z(B(0, \rho_\alpha)).$$

■

This statement tells us the exact coverage level of the credible balls. If $Z = W$, in particular, it follows that credible sets of level $1 - \alpha$ are asymptotic confidence sets of level $1-\alpha$. Thus this theorem allows us to conclude that if the BvM theorem holds and the asymptotic distribution of the posterior W matches the asymptotic distribution Z of the estimator $\tilde{\theta}_n$, the Bayesian uncertainty quantification will be asymptotically valid frequentist uncertainty quantification. Moreover, in many cases, the estimator $\hat{\theta}$ will be an efficient estimator, so not only does the Bayesian methodology give valid uncertainty quantification, it will give efficient estimation as well.

We needed to talk about specific credible sets: balls centred at a good estimator and with a specific radius. To illustrate that this is needed, we will construct a credible set which has a specified credibility level $1-\alpha$ but has 0 coverage. This example means that we cannot take any credible set. Extra consideration is needed.

Example 1.4.29. *To construct our credible set without coverage, we will first make a ball centred at the true θ_0, and with a maximal radius such that the posterior assigns less than mass α to it. So we pick ρ such that*

$$\Pi\left(B(\theta_0,\rho)|\, X^{(n)}\right) \leq \alpha$$

and for any $\rho' > \rho$

$$\Pi\left(B(\theta_0,\rho')|\, X^{(n)}\right) \geq \alpha$$

Our credible set will be the complement of this ball: $C = B(\theta_0,\rho)^c$. Hence

$$\Pi(C|\, X^{(n)}) \geq 1-\alpha$$

However, by construction, $\theta_0 \notin C$. Hence

$$\mathbb{P}(\theta_0 \notin C) = 1$$

Thus not every credible ball gives valid uncertainty quantification.

Examples of Bernstein-von Mises theorems

There have been various examples of Bernstein-von Mises theorems. We will present the BvM theorem for well-specified parametric models, and give references to other examples. This version of the theorem comes from [77, Theorem 10.1].

Theorem 1.4.30. *Let the experiment $\{P_\theta\colon \theta\in\Theta\}$ be differentiable in* quadratic mean *at θ_0 with nonsingular Fisher information matrix I_{θ_0}, and suppose that for every $\epsilon > 0$ there exists a sequence of tests ϕ_n such that*

$$P^n_{\theta_0}\phi_n \to 0, \qquad \sup_{\|\theta-\theta_0\|>\epsilon} P^n_\theta(1-\phi_n) \to 0.$$

Furthermore, let the prior measure be absolutely continuous in a neighbourhood of θ_0 with a continuous positive density at θ_0. Then the corresponding posterior distributions satisfy

$$\|\Pi(\sqrt{n}\,(\theta-\Delta_n)\in\cdot\,|\, X^{(n)}) - N(0, I^{-1}_{\theta_0})\| \overset{P^n_{\theta_0}}{\to} 0.$$

This shows that we have a convergence in the total variation distance between two objects:

- The rescaled posterior centred at an efficient estimator;
- The normal distribution with the inverse Fisher information matrix as the covariance matrix.

The asymptotic distribution of the efficient estimator is exactly the limiting distribution of the posterior. Hence we can use the Bernstein-von Mises theorem to give an asymptotic motivation from a frequentist perspective. We can use this to study the coin flip example.

Example 1.4.31 (Bernstein-von Mises for the coin flip example)**.** *The posterior mean is an efficient estimator. The Beta prior is smooth, the experiment P_θ is differentiable in quadratic mean and by Example 1.4.16, there exist tests. Hence we can apply Theorem 1.4.30. The BvM gives that the posterior is asymptotically normal, with the variance equal to the inverse Fisher information and centred on the posterior mean. Then we find that the posterior is consistent by Lemma 1.4.25, and it contracts at rate $\frac{1}{\sqrt{n}}$ by Lemma 1.4.27. Define C_α to be the smallest balls centred on the posterior mean with posterior mass $\Pi(C_\alpha | X) \geq 1-\alpha$. Then by Lemma 1.4.28 we find that C_α also gives asymptotically valid uncertainty quantification.*

The story becomes more complicated once we go beyond the correctly specified parametric models.

The well-specified parametric Bernstein-von Mises theorem is just one of the examples. Other models have also been studied. Several models have been studied by Bernstein-von Mises theorems. One such example is the misspecified parametric models [44]. Here we do not assume the true distribution is in the model. In this situation, the frequentist coverage of the credible set does not need to be equal to its level.

Another example is semiparametric models [9, 11, 66]. Here the Bernstein-von Mises theorem can be used to study the asymptotic distribution. Depending on the precise model we might or might not get the correct asymptotic distribution and coverage.

For the last highlighted example, in distribution estimation, we have a Bernstein-von Mises theorem for certain priors. For the Dirichlet prior [49]; For the Pitman-Yor processes when the true distribution is continuous [41]. One of my own contributions has been to extend the latter result to general true distributions, see Chapter 3 and [27].

1.5 Overview

We have seen how to use the Bernstein-von Mises theorem to prove that specific credible sets give valid uncertainty quantification in Section 1.4.3. The next step we need to take is to prove the Bernstein-von Mises theorem for the specific models we want to study. One prior that Bayesians use is the Pitman-Yor prior. We will introduce this prior in Chapter 2. Once we have introduced this prior, we prove Berstein-von Mises theorems for the prior itself and the hyperprior in Chapters 3 and 4 respectively. Deep learning is a popular tool used by the machine learning community. Mathematicians have developed tools to study deep learning. We will introduce these tools in Chapter 5. Finally, we propose a new Bayesian method for deep learning. We give theoretical guarantees and prove that specific credible sets provide valid uncertainty quantification in Chapter 6.

Chapter 2

Prerequisite theory for the Pitman-Yor papers

2.1 The Pitman-Yor processes

Suppose we want to estimate a distribution P_0 based on i.i.d. observations. There are several methods that people use. One such class of methods are species sampling processes. Proper species sampling processes can be defined as follows. We need to specify a random weight vector W and an atomless measure G, the centre measure. A species process prior is the distribution of

$$\sum_{i=1}^{\infty} W_i \delta_{\theta_i},$$

where $\theta_1, \theta_2, \ldots \overset{\text{iid}}{\sim} G$. To specify such a prior, we need to specify the distribution of W. There are many choices possible here.

One key example of the species sampling processes is the Dirichlet process. To define the Dirichlet process, we need to give the distribution of W. We can give this distribution in terms of the stick-breaking construction. Let $M > 0$. Then we define random variables W as follows

$$V_1, V_2, \ldots \overset{\text{iid}}{\sim} \mathrm{B}(1, M), \qquad W_i = V_i \prod_{j=1}^{i-1} (1 - V_j).$$

The Dirichlet process plays an important role in Bayesian nonparametrics. There are several reasons for this. Important reasons are the frequentist properties of this estimator. There is a BvM theorem which shows that the posterior has strong guarantees. Not only this, the prior is computationally very easy. These reasons makes it easy to motivate such a prior.

To make this prior more general, one can try to allow more general distributions on the W. One can try to keep the relative stick-breaking weights V independent but allow for different distributions. This generelisation would lead to a much larger class of priors. If one studies the posterior distribution, it turns out that the *size-biased permutations* of W are the important object to understand. If the weights W are invariant under size-biased permutations, many of the computations turn out to be easier. However, the only distributions with i.i.d. relative stick-breaking weights are exactly those in the Dirichlet process. If we allow for independent but no longer require i.i.d. relative stick-breaking weights, we can allow for a small bit more. By a result of Pitman [60], the only distributions invariant under size-biased permutations and with independent relative stick-breaking weights are given by

$$V_i \overset{\text{ind}}{\sim} \mathrm{B}(1-\sigma, M+j\sigma), W_i = V_i \prod_{j=1}^{i-1}(1-V_j),$$

for $\sigma \in [0,1)$ and $M > -\sigma$, or one of 3 families of finitely discrete distributions. In particular the Pitman-Yor process is the only one with infinitely many weights nonzero. The Pitman-Yor process uses the distribution defined in the previous display to define the random weights W. We will denote this distribution by $\mathrm{PY}(\sigma, M, G)$.

You can either choose the two parameters σ and M or you can try to learn them from the data. The σ is the type parameter and influences the power-law behaviour of the data. If $\sigma = 0$, the expected number of distinct observations K_n in a sample $X_1, \ldots, X_n | P \sim P$ and $P \sim \mathrm{PY}(\sigma, M, G)$ satisfies

$$\frac{K_n}{M \log(n)} \overset{\text{a.s.}}{\to} 1.$$

For more details, see [33, Proposition 4.8]. If $\sigma \in (0,1)$ we get a different behaviour:

$$\frac{K_n}{n^\sigma} \overset{\text{a.s.}}{\rightsquigarrow} Z_\sigma.$$

Here Z_σ is an almost sure finite random variable, called the σ-diversity. For more details, see [33, Theorem 14.50]. This result means that the Pitman-Yor prior gives rise to random probability measures with specific power-law behaviour. Because this power-law behaviour is expected in specific applications, people like to use the Pitman-Yor prior in these settings. Moreover, if you try to learn the σ from the data, you might hope to learn the power law of the true distribution. This comes up in applications.

To do inference, we need to be able to compute the posterior distribution. We can compute this due to the structure of the Pitman-Yor process. To write down the posterior distribution, we first need to introduce notation. Suppose we have K_n

distinct observations. Let $\tilde{X}_i$ be the i-th distinct observation. We denote the number of times we have observed $\tilde{X}_i$ in the sample by $N_{i,n}$. The posterior distribution of the Pitman-Yor process is the distribution of

$$R_n \sum_{j=1}^{K_n} \hat{W}_j \delta_{\tilde{X}_j} + (1 - R_n) Q_n,$$

where $R_n \sim \mathrm{B}(n - K_n\sigma, M + K_n\sigma)$, $(\hat{W}_1, \dots, \hat{W}_{K_n}) \sim \mathrm{Dir}(K_n; N_{1,n} - \sigma, \dots, N_{K_n,n} - \sigma)$ and $Q_n \sim \mathrm{PY}(\sigma, M + \sigma K_n, G)$. The fact that the explicit posterior distribution is known is important for the analysis of the posterior distribution. We use explicit properties of this to derive the BvM theorem.

2.2 BvM for distribution estimation

We will use a BvM theorem to motivate the usage of the PY process from a frequentist point of view. To do so, we first have to figure out what our precise goal will be. Recall that in distribution estimation, we wanted to estimate a distribution P_0, given a sample $X_1, \dots, X_n \overset{\text{iid}}{\sim} P_0$. One of the estimators is the empirical distribution $\mathbb{P}_n$. This estimator is given by

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

This estimator is the distribution you get by putting a point mass of $1/n$ on each observation. If f is a function such that $P_0|f| < \infty$, then by the strong law of large numbers, it follows that

$$\mathbb{P}_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \overset{P_0 \text{ a.s.}}{\to} P_0 f.$$

Suppose our goal is to compute some estimate of expectation of $f(X)$ under P_0. Then $\mathbb{P}_n(f)$ will converge almost surely to $P_0(f)$, the right answer. If in addition $P_0 f^2 < \infty$, then also

$$\sqrt{n}\,(\mathbb{P}_n(f) - P_0 f) \rightsquigarrow N(0, P_0(f - P_0 f)^2).$$

This means we converge at $\frac{1}{\sqrt{n}}$ rate. Moreover, we can use this to construct confidence intervals for our estimates. Often we are not interested in just estimating one functional, but many at the same time. Then we would want that this convergence holds uniformly over a whole class of functions $\mathcal{F}$. To illustrate that uniform convergence is not automatic, consider the class of indicator functions of finite sets. The set $S_n = \{X_1, \dots, X_n\}$ is finite. Hence the indicator function $\mathbb{1}_S$ of S is contained in our class. But since $X_i \in S_n$, it follows that $\mathbb{1}_S(X_i) = 1$. Thus $\mathbb{P}_n(\mathbb{1}_S) = 1$. However, if the true distribution is continuous, for example, a standard normal distribution, $P_0(\mathbb{1}_{S_n}) = P_0(S_n) = 0$. Hence the difference is surely 1. Thus we have no uniform convergence. This example implies that not all classes of functions satisfy the uniform convergence. We will give a name to two classes of functions that have some form of uniform convergence. A class of functions $\mathcal{F}$ is called Glivenko-Cantelli if $\sup_{f\in\mathcal{F}} \|\mathbb{P}_n(f) - P_0(f)\|$ converges to zero. A class of functions $\mathcal{F}$ is called Donsker if

$$\mathbb{G}_n = \sqrt{n}\,(\mathbb{P}_n - P_0) \rightsquigarrow \mathbb{G}, \qquad \text{in } \ell^\infty(\mathcal{F}).$$

Here $\mathbb{G}$ is a P_0 Brownian Bridge. This Brownian Bridge is the counterpart to the Gaussian distribution for classes of functions. For a discussion on all these concepts, see [78]. It turns out that the empirical process is asymptotically efficient for estimating these expectations. See for example [77, Example 25.24].

When we formulated the general BvM results, we wanted to centre on a good estimator. Since, in a certain sense, the efficient estimators are the "best", we want to centre on those. Thus for any BvM result in distribution estimation, we want to centre on the empirical process. The asymptotic distribution of the empirical process is the Brownian Bridge $\mathbb{G}_{P_0}$. Hence this would also be the ideal asymptotic distribution of the posterior from a frequentist point of view.

We already knew that the Pitman-Yor processes did satisfy a BvM result in case the true distribution P_0 was atomless. In that case, the BvM result stated that:

$$\begin{aligned}\sqrt{n}\Big(P - (1-\sigma)\mathbb{P}_n + \sigma G\Big)\Big|\, X_1, \dots, X_n \rightsquigarrow \sqrt{1-\sigma}\mathbb{G}_{P_0} \\ + \sqrt{\sigma(1-\sigma)}\mathbb{G}_G + \sqrt{(1-\sigma)\sigma}\Big(P_0 - G\Big)\, Z_1.\end{aligned}$$

Hence, in that case, we have a bias term $\sigma(G - \mathbb{P}_n)$. The asymptotic distribution of the posterior is a mixture of different Gaussians and not the ideal Brownian Bridge. However, $\sigma(G - \mathbb{P}_n)$ converges almost surely to $\sigma(G - P_0)$. If this was the general bias term, it would mean that the posterior would always converge to $(1-\sigma)P_0 + G$. However, we know that the posterior is consistent for discrete P_0, and hence the bias term, in general, must be more complicated.

As we have seen before, the type σ of the Pitman-Yor process defines the power-law behaviour. This behaviour makes the parameter σ of independent interest. So we aim to prove a BvM theorem for this as well. One might hope this gives a BvM centred at an efficient estimator for the power-law behaviour. However, this is not exactly the case. It centres on the Marginal maximum likelihood estimator as expected. However, the estimator does not in general estimate the power law efficiently. Because σ is a 1-dimensional parameter, we can expect to use the tools from the parametric misspecified BvM theorem [44]. Indeed, we will find a misspecified BvM result. For more details, see Chapter 4.

Chapter 3

The BvM for PY

This chapter is an adaptation of a paper published as: S. Franssen, A. van der Vaart, "Bernstein-von Mises theorem for the Pitman-Yor process of nonnegative type", [27].

3.1 Introduction

The Pitman-Yor process [59, 54] is a random probability distribution, which can be used as a prior distribution in a nonparametric Bayesian analysis. It is characterised by a *type* parameter σ, which in this paper we take to be positive. The Pitman-Yor process of type $\sigma = 0$ is the Dirichlet process [24], which is well understood, while negative types correspond to finitely discrete distributions and were considered in [17]. The Pitman-Yor process is also known as the two-parameter Poisson-Dirichlet Process, is an example of a Poisson-Kingman process [57], and a species sampling process of Gibbs type [18].

The easiest definition is through *stick-breaking* ([54, 40]), as follows. The family of nonnegative Pitman-Yor processes is given by three parameters: a number $\sigma \in [0, 1)$, a number $M > -\sigma$ and an atomless probability distribution G on some measurable space $(\mathcal{X}, \mathcal{A})$. We say that a random probability measure P on $(\mathcal{X}, \mathcal{A})$ is a Pitman-Yor process (of nonnegative type), denoted $P \sim \mathrm{PY}(\sigma, M, G)$, if P can be represented as

$$P = \sum_{i=1}^{\infty} W_i \delta_{\theta_i},$$

where $W_i = V_i \prod_{j=1}^{i-1}(1 - V_j)$ for $V_i \overset{\text{iid}}{\sim} \mathrm{B}(1 - \sigma, M + i\sigma)$, independent of $\theta_i \overset{\text{iid}}{\sim} G$, and B the beta distribution.

It is clear from this definition that the realisations of P are discrete probability measures, with countably many atoms at random locations, with random weights. If one first draws $P \sim \mathrm{PY}(\sigma, M, G)$, and next given P a random sample $X_1, \ldots, X_n$ from P, then ties among the latter observations are possible, or even likely. It is known ([57]) that the number K_n of different values among $X_1, \ldots, X_n$ is almost surely of the order n^σ if $\sigma > 0$, whereas it is logarithmic in n if $\sigma = 0$. This suggests that the Pitman-Yor process is a reasonable prior distribution for a dataset in which similar patterns are expected (or observed). In particular, when a large number of clusters is expected, a Pitman-Yor process of positive type could be preferred over the standard Dirichlet prior, which corresponds to $\sigma = 0$. Applications in genetics or topic modelling can be found in [83, 76, 35, 3]. The Pitman-Yor process has also been proposed as a prior for estimating the probability that a next observation is a new species [23], with applications in e.g. forensic statistics [12, 13]. The papers [8, 7] study hierarchical versions of Pitman-Yor processes, which are useful to discover structure in data beyond clustering.

In this paper we consider the properties of the Pitman-Yor posterior distribution to estimate the distribution of a random sample of observations. By definition this posterior distribution is the conditional distribution of P given $X_1, \ldots, X_n$ in the Bayesian hierarchical model $P \sim \mathrm{PY}(\sigma, M, G)$ and $X_1, \ldots, X_n |\, P \overset{\text{iid}}{\sim} P$. We assume that in reality the observations $X_1, \ldots, X_n$ are an i.i.d. (i.e. independent and identically distributed) sample from a distribution P_0 and investigate the use of the posterior distribution for inference on P_0. It was shown in [41, 18] that in this setting, as $n \to \infty$,

$$P |\, X_1, \ldots, X_n \rightsquigarrow \delta_{(1-\lambda)P_0^d + \lambda(1-\sigma)P_0^c + \sigma\lambda G}, \tag{3.1}$$

where $\rightsquigarrow$ denotes weak convergence of measures, δ_Q denotes the Dirac measure at the probability distribution Q, and $P_0 = (1-\lambda)P_0^d + \lambda P_0^c$ is the decomposition of P_0 in its discrete component $(1-\lambda)P_0^d$ and the remaining (atomless) part λP_0^c. In the case that P_0 is discrete, we have $\lambda = 0$ and the measure $(1-\lambda)P_0^d + \lambda(1-\sigma)P_0^c + \sigma\lambda G$ in the right side reduces to $P_0^d = P_0$, and hence (3.1) expresses that the posterior distribution collapses asymptotically to the Dirac measure at P_0. The posterior distribution is said to be consistent in this case. However, if P_0 is not discrete, then the posterior distribution recovers P_0 asymptotically only if $\sigma = 0$ (the case of the Dirichlet prior) or if $G = P_0^c$. The last case will typically fail and hence in the case that $\sigma > 0$ the posterior distribution will typically be consistent if and only if P_0 is discrete. This reveals the Pitman-Yor prior of positive type as a reasonable prior only for discrete distributions.

Besides for recovery, a posterior distribution is used to express remaining uncertainty, for instance in the form of a credible (or Bayesian confidence) set. To justify such a procedure from a non-Bayesian point of view, the posterior consistency must be refined to a distributional result of Bernstein-von Mises type. Such a result was obtained by [41] in the case that the true distribution P_0 is atomless, the case that the posterior distribution is inconsistent and the Pitman-Yor prior is better avoided. In the present paper we study the case of general distributions P_0, including the case of most interest that P_0 is discrete. It turns out that discreteness per se is not enough for valid inference, but it is also needed that the weights of the atoms in P_0 decrease fast enough. In the other case, ordinary Bayesian credible sets are not valid confidence sets. For the latter case our result suggests a bias correction.

Since the type parameter σ determines the number of distinct values in a sample from the prior, it might be interpreted as influencing the discreteness of the prior, smaller σ favouring fewer distinct values and hence a more discrete prior. In the asymptotic result the type parameter plays only a secondary role. At first thought counterintuitively, a larger σ, which gives a less discrete prior, increases the bias in the posterior distribution that arises when the atoms in P_0 decrease too slowly.

In practice one may prefer to estimate the type parameter from the data. The empirical Bayes method maximizes the marginal likelihood of $X_1, \ldots, X_n$ in the Bayesian setup over σ. We show that in the consistent case, substitution of this estimator in the posterior distribution for given type parameter does not change the asymptotics of the Pitman-Yor posterior. Alternatively, we may equip σ itself with a prior, resulting in a mixture of Pitman-Yor processes as a prior for P. We show that this too results in the same posterior behaviour. Thus estimating the type parameter does not solve the inconsistency problem.

We can conclude that the Pitman-Yor process is an appropriate prior for estimating a distribution only if the sizes of the atoms of this distribution decrease sufficiently rapidly. Our results show that the speed of decay depends on the aspect of interest, for instance different for the distribution function than for the mean.

Our results depend heavily on the characterisation of the posterior distribution given in [60] (see Section 3.4).

3.2 Main result

The nonparametric maximum likelihood estimator of the distribution P of a sample of observations $X_1, \ldots, X_n$ is the empirical measure $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$, the discrete uniform measure on the observations. Therefore in analogy with the case of classical parametric models (e.g. Theorem 10.1 and page 144 in [77]), in this setting a Bernstein-von Mises theorem would give the approximation of the posterior distribution of $\sqrt{n}(P - \mathbb{P}_n)$ given $X_1, \ldots, X_n$ by the normal distribution obtained as the limit of $\sqrt{n}(\mathbb{P}_n - P_0)$. To give a precise meaning to such a distributional statement, we may evaluate all the measures involved on a collection of sets, and interpret $\sqrt{n}(P - \mathbb{P}_n)$ and $\sqrt{n}(\mathbb{P}_n - P_0)$ as stochastic processes indexed by sets. For instance, in the case that the sample space is the real line, we could use the sets $(-\infty, t]$, for $t \in \mathbb{R}$, corresponding to the distribution functions of the measures P, $\mathbb{P}_n$ and P_0.

More generally, we may evaluate these measures on measurable functions $f\colon \mathcal{X} \to \mathbb{R}$, as

$$Pf = \int f\, dP, \qquad \mathbb{P}_n f = \int f\, d\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^n f(X_i), \qquad P_0 f = \int f\, dP_0.$$

Given a collection $\mathcal{F}$ of such functions, the Bernstein-von Mises can then address the distributions of the stochastic processes $\bigl\{\sqrt{n}(Pf - \mathbb{P}_n f)\colon f \in \mathcal{F}\bigr\}$ and $\bigl\{\sqrt{n}(\mathbb{P}_n f - P_0 f)\colon f \in \mathcal{F}\bigr\}$, the first one conditionally given the observations $X_1, \ldots, X_n$. For instance, in the case that $\mathcal{X} = \mathbb{R}$, we might choose the collection $\mathcal{F}$ to consist of all indicator functions $x \mapsto 1_{x \leq t}$, for t ranging over $\mathbb{R}$, but we can also add the identify function $f(x) = x$, yielding the means of the measures.

For a set $\mathcal{F}$ of finitely many functions, these processes are just vectors in Euclidean space and their distributions can be evaluated as usual. Furthermore, the limit law of $\bigl\{\sqrt{n}(\mathbb{P}_n f - P_0 f)\colon f \in \mathcal{F}\bigr\}$ is a multivariate normal distribution, in view of the multivariate central limit theorem (provided $P_0 f^2 < \infty$, for every $f \in \mathcal{F}$). It is convenient to write the latter as the distribution of a Gaussian process $\{\mathbb{G}_{P_0} f\colon f \in \mathcal{F}\}$, determined by its mean and covariance function

$$\mathbb{E}\mathbb{G}_{P_0} f = 0 \qquad \mathbb{E}\mathbb{G}_{P_0} f \mathbb{G}_{P_0} g = P_0(f - P_0 f)(g - P_0 g).$$

The process $\mathbb{G}_{P_0}$ is known as a P_0-*Brownian bridge* (see e.g. [62, 78, 77]).

An appropriate generalisation (and strengthening) of the central limit theorem to sets $\mathcal{F}$ of infinitely many functions is Donsker's theorem (e.g. [77], Chapter 19). The Bernstein-von Mises theorem can be strengthened in a similar fashion. For the case of indicator functions on the real line, Donsker's theorem was derived by [22], and the corresponding Bernstein-von Mises theorem for the Dirichlet process by [49, 48]. A precise formulation (in the general case, which is not more involved than the real case) is as follows.

A class of functions $\mathcal{F}$ is called P_0-*Donsker* if the sequence $\sqrt{n}\,(\mathbb{P}_n - P_0)$ converges in distribution to a tight, Borel measurable element in the metric space $\ell^\infty(\mathcal{F})$ of bounded functions $z\colon \mathcal{F} \to \mathbb{R}$, equipped with the uniform norm $\|z\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |z(f)|$. The limit is then a version of the Gaussian process $\mathbb{G}_{P_0}$. The Bernstein-von Mises theorem involves conditional convergence in distribution given the observations $X_1, \ldots, X_n$, which is best expressed using a metric. The bounded Lipschitz metric (see for example [78], Chapter 1.12) is convenient, and leads to defining conditional convergence in distribution of the sequence $\sqrt{n}(P - \mathbb{P}_n)$ in $\ell^\infty(\mathcal{F})$ given $X_1, \ldots, X_n$ to $\mathbb{G}_{P_0}$ as

$$\sup_{h \in \mathrm{BL}_1} \Bigl| \mathbb{E}\Bigl(h\bigl(\sqrt{n}(P - \mathbb{P}_n)\bigr) \,|\, X_1, \ldots, X_n\Bigr) - \mathbb{E}h(\mathbb{G}_{P_0})\Bigr| \to 0,$$

where the convergence refers to the i.i.d. sample $X_1, X_2, \ldots$ from P_0, and can be in (outer) probability or almost surely. The supremum is taken over the set BL_1 of all functions $h\colon \ell^\infty(\mathcal{F}) \to [0,1]$ such that $|h(z_1) - h(z_2)| \leq \|z_1 - z_2\|_{\mathcal{F}}$, for all $z_1, z_2 \in \ell^\infty(\mathcal{F})$. For simplicity of notation and easy interpretation, we write the preceding display as

$$\sqrt{n}(P - \mathbb{P}_n) \,|\, X_1, \ldots, X_n \rightsquigarrow \mathbb{G}_{P_0}.$$

Conditional convergence in distribution of other processes is defined and denoted similarly. For finite sets $\mathcal{F}$, the complicated definition using the bounded Lipschitz metric reduces to ordinary weak convergence of random vectors. Also, a finite set $\mathcal{F}$ is P_0-Donsker if and only if $P_0 f^2 < \infty$, for every $f \in \mathcal{F}$. There are many examples of infinite Donsker classes (see e.g. [78]), with the set of indicators of cells $(-\infty, t]$ as the classical example.

We are ready to formulate the main result of the paper. Let $\tilde{X}_1, \tilde{X}_2, \ldots$ be the distinct values in $X_1, X_2, \ldots$ in the order of appearance, let K_n be the number of distinct elements among $X_1, \ldots, X_n$, and set

$$\tilde{\mathbb{P}}_n = \frac{1}{K_n} \sum_{i=1}^{K_n} \delta_{\tilde{X}_i}. \tag{3.2}$$

All limit results refer to a sample $X_1, X_2, \ldots, X_n$ drawn from a measure P_0. This can always be written as $P_0 = (1-\lambda)P_0^d + \lambda P_0^c$, where P_0^d is a discrete and P_0^c an atomless distribution and $\lambda \in [0,1]$ is the weight of the discrete part in P_0. The decomposition is unique unless $\lambda = 0$ or $\lambda = 1$, when P_0^c or P_0^d is arbitrary.

Theorem 3.2.1. *Let $P_0 = (1-\lambda)P_0^d + \lambda P_0^c$ where P_0^d is a discrete and P_0^c an atomless probability distribution. The posterior distribution of P in the model $P \sim \mathrm{PY}(\sigma, M, G)$ and $X_1, \ldots, X_n | P \overset{iid}{\sim} P$ satisfies for every finite collection $\mathcal{F}$ of functions with $(P_0 + G)f^2 < \infty$, for every $f \in \mathcal{F}$, almost surely under P_0^∞,*

$$\begin{aligned}\sqrt{n}\Big(P - \mathbb{P}_n - \frac{\sigma K_n}{n}(G - \tilde{\mathbb{P}}_n)\Big)\Big|\, X_1, \ldots, X_n \\ \rightsquigarrow \sqrt{1-\lambda}\,\mathbb{G}_{P_0^d} + \sqrt{(1-\sigma)\lambda}\,\mathbb{G}_{P_0^c} + \sqrt{\sigma(1-\sigma)\lambda}\,\mathbb{G}_G \\ + \sqrt{(1-\sigma\lambda)\sigma\lambda}\Big(\frac{(1-\lambda)P_0^d + (1-\sigma)\lambda P_0^c}{1-\sigma\lambda} - G\Big) Z_1 \\ + \frac{\sqrt{(1-\sigma)\lambda(1-\lambda)}}{\sqrt{1-\sigma\lambda}}(P_0^c - P_0^d)\, Z_2.\end{aligned}$$

Here $\mathbb{G}_{P_0^d}$, $\mathbb{G}_{P_0^c}$ and $\mathbb{G}_G$ are independent Brownian bridge processes, independent of the independent standard normal variables Z_1 and Z_2. More generally this is true, with convergence in $\ell^\infty(\mathcal{F})$ in probability, for every P_0-Donsker class of functions $\mathcal{F}$ for which the $\mathrm{PY}(\sigma, \sigma, G)$ process satisfies the central limit theorem in $\ell^\infty(\mathcal{F})$. If in addition $P_0^* \|f - P_0 f\|_{\mathcal{F}}^2 < \infty$, then the convergence is also P_0^∞-almost surely.

The proof of the theorem is deferred to Section 3.4.1. The condition that the $\mathrm{PY}(\sigma, \sigma, G)$ process satisfies the central limit theorem in $\ell^\infty(\mathcal{F})$, is satisfied, for instance, for all classes $\mathcal{F}$ that are suitably measurable with finite uniform entropy integral and for all classes $\mathcal{F}$ with finite G-bracketing integral. This follows from Theorems 2.11.9 or 2.11.1 in [78].

The limit process in the theorem is Gaussian, but it is the P_0-Brownian bridge $\mathbb{G}_{P_0}$ only if $\lambda = 0$, i.e. if $P_0 = P_0^d$ is discrete. In addition, the behaviour of the Pitman-Yor posterior deviates from the "desired" behaviour by the presence on the left side of the term

$$\sqrt{n}\,B_n(f){:}=\frac{\sigma K_n}{\sqrt{n}}(G-\tilde{\mathbb{P}}_n). \tag{3.3}$$

Given the observations $X_1,\ldots,X_n$ this term is deterministic, and we can only expect it to disappear if $K_n/\sqrt{n}$ tends to zero. While $K_n/n\to 0$ almost surely for any discrete distribution P_0, the more stringent convergence to zero of $K_n/\sqrt{n}$ is valid only if the sizes of the atoms of P_0 decrease fast enough. This relationship was made precise in [42] (also see the corollary below) in terms of the function

$$\alpha_0(u)=\#\{x{:}\,1/P_0\{x\}\leq u\}. \tag{3.4}$$

If α_0 is regularly varying at $u=\infty$ (in the sense of Karamata, see e.g. [6] or the appendix to [19]) with exponent $\gamma_0\in(0,1)$, then $K_n/\alpha_0(n)\to\Gamma(1-\gamma_0)$, almost surely, and $\alpha_0(n)$ is n^{γ_0} up to a slowly varying factor. In this case, for $K_n/\sqrt{n}$ to tend to zero, it is necessary that the exponent be smaller than 1/2 and sufficient that it is strictly smaller than 1/2. For instance, if the ordered atoms $P_0\{x_j\}$ of P_0 decrease proportionally to $1/j^\alpha$, then $K_n/\sqrt{n}\to 0$ in probability if and only if $\alpha>2$.

For bounded functions f, the convergence $K_n/\sqrt{n}\to 0$ is also enough to drive the additional term (3.3) to zero, as the terms $(G-\tilde{\mathbb{P}}_n)f$ will remain bounded in that case. For unbounded functions f, a still more stringent condition on P_0 is needed to make the term $(K_n/\sqrt{n})\tilde{\mathbb{P}}_nf$ go away. For instance, for the posterior mean of a distribution on $\mathbb{N}$ with atoms $P_0\{j\}$ of the order $1/j^\alpha$, the next corollary implies that $\alpha>4$ is needed.

We conclude that for a large class of discrete distributions P_0, but not all, the Bernstein-von Mises theorem takes its standard form, and this also depends on which aspect of the posterior distribution we are interested in.

Corollary 3.2.2. *Under the conditions of Theorem 3.2.1, if P_0 is a discrete probability distribution, then $\sqrt{n}\big(P-\mathbb{P}_n-(\sigma K_n/n)(G-\tilde{\mathbb{P}}_n)\big)|X_1,\ldots,X_n\rightsquigarrow\mathbb{G}_{P_0}$ in $\ell^\infty(\mathcal{F})$, in probability or almost surely.*

- *(i) If the class of functions $\mathcal{F}$ is uniformly bounded and the atoms $\{x_j\}$ of P_0 satisfy $P_0\{x_j\}\leq Cj^{-\alpha}$, for some constants C and $\alpha>2$, then also $\sqrt{n}(P-\mathbb{P}_n)|X_1,\ldots,X_n\rightsquigarrow\mathbb{G}_{P_0}$, in probability. If the class of functions $\mathcal{F}$ is uniformly bounded and the function $u\mapsto\alpha_0(u)=\#\{x{:}\,1/P_0\{x\}\leq u\}$ is regularly varying at $u=\infty$ of exponent strictly smaller than 1/2, then this is also true almost surely.*
- *(ii) If the atoms $\{x_j\}$ of P_0 and the function f satisfy $P_0\{x_j\}\leq Cj^{-\alpha}$ and $f(x_j)\asymp j^p$, for some $p>0$, then $\sqrt{n}(P-\mathbb{P}_n)f|X_1,\ldots,X_n\rightsquigarrow\mathbb{G}_{P_0}f$, in probability if $\alpha>2p+2$.*

Proof. The first assertion merely specializes the limit in Theorem 3.2.1 to the case of a discrete distribution, by setting $\lambda=0$. Assertions (i) and (ii) follow from this if the term (3.3) tends to zero, in probability or almost surely.

For bounded functions f, as assumed in (i), the term (3.3) tends to zero provided $K_n/\sqrt{n}$ tends to zero. The almost sure convergence is immediate from [42], Theorems 9 and 1‘, which show that $K_n/\alpha_0(n) \to \Gamma(1-\gamma_0)$, almost surely, for γ_0 the exponent of regular variation. For the convergence in probability, we note that $K_n = \sum_{j=1}^{\infty} 1_{j\in\{X_1,\ldots,X_n\}}$, whence $\mathbb{E}K_n = \sum_{j=1}^{\infty}\bigl(1-(1-P_0\{x_j\})^n\bigr)$. By the inequality $(1-p)^n \geq 1-np$, for $p \geq 0$, we find that $\mathbb{E}K_n \leq \sum_{j=1}^{\infty}(nCj^{-\alpha}\wedge 1)$, which can be seen to be $o(\sqrt{n})$ if $\alpha > 2$.

The assertion in (ii) follows provided $K_n/\sqrt{n} \to 0$ and $(K_n/\sqrt{n})\tilde{\mathbb{P}}_n f \to 0$, in probability. Reasoning as before, we find

$$\mathbb{E}\Bigl(\frac{K_n}{\sqrt{n}}\tilde{\mathbb{P}}_n f\Bigr) = \frac{1}{\sqrt{n}}\sum_{j=1}^{\infty} f(x_j)\bigl(1-(1-P_0\{x_j\})^n\bigr).$$

Under the given assumptions on f and the atoms, this is bounded above by

$$\frac{1}{\sqrt{n}}\int_{C^{1/\alpha}}^{\infty} u^p\Bigl(1-\Bigl(1-\frac{C}{u^\alpha}\Bigr)^n\Bigr)\,du = \frac{n^{(p+1)/\alpha}}{\sqrt{n}}\int_{C/n}^{\infty} v^{(p+1)/\alpha-1}\Bigl(1-\Bigl(1-\frac{C}{nv}\Bigr)^n\Bigr)\,dv.$$

The integrand is bounded above by $Cv^{(p+1)/\alpha-1}$ and hence the integral converges near 0. By again the inequality $(1-p)^n \geq 1-np$, for $p \geq 0$, the integrand is also bounded above by $v^{(p+1)/\alpha-2}$ and hence the integral converges near infinity if $(p+1)/\alpha < 1$. The middle part of the integral always gives a non-vanishing contribution and hence the full expression can tend to zero only if the leading factor tends to zero. This is true under the more stringent condition that $(p+1)/\alpha < 1/2$. ∎

For $\lambda = 1$ and $P_0 = P_0^c$, Theorem 3.2.1 was obtained by [41]. In this case all observations are distinct and the left side of the theorem reduces to $\sqrt{n}\bigl(P-(1-\sigma)\mathbb{P}_n-\sigma G\bigr)$, since $K_n = n$. As noted in the introduction, the posterior distribution is not even consistent, i.e. the asymptotic limit is "wrong" even without the $\sqrt{n}$ multiplier.

The Bernstein-von Mises theorem is important for the validity of credible sets. A credible interval for Pf, for a given function f, could be formed as the interval between two quantiles of the marginal posterior distribution of Pf given $X_1,\ldots,X_n$. For instance, for f equal to the indicator of a given set $A \in \mathcal{A}$, this gives a credible interval for a probability $P(A)$, and for $f(x) = x$, we obtain a credible interval for the mean. Simultaneous credible sets, for instance a credible band for a distribution function can be obtained similarly.

By the inconsistency of the posterior distribution in the case that the true distribution possesses a continuous component ($\lambda > 0$), there is no hope that in this case such an interval for Pf will cover a true value $P_0 f$ with the desired probability. However, also in the case of a discrete distribution P_0, the coverage may not tend to the nominal value, due to the presence of the bias term (3.3). We need at least that $K_n/\sqrt{n}$ tends to zero, and more for unbounded functions f.

Because the bias $B_n(f) = (\sigma K_n/n)(Gf - \tilde{\mathbb{P}}_n f)$ is observed (and σ and the center measure G are fixed by our prior choices), it is possible to correct a credible interval by shifting it by minus this amount. Thus for $Q_{n,\alpha}(f)$ the α-quantile of the posterior distribution of Pf given $X_1, \ldots, X_n$, we consider both the credible intervals $\big[Q_{n,\alpha}(f), Q_{n,\beta}(f)\big]$ and corrected intervals $\big[Q_{n,\alpha}(f) - B_n(f), Q_{n,\beta}(f) - B_n(f)\big]$, for given $\alpha < \beta$.

Corollary 3.2.3. *Under the conditions of Theorem 3.2.1, if P_0 is a discrete probability distribution, then $\mathrm{P}_{P_0}\big(Q_{n,\alpha}(f) - B_n(f) \leq P_0 f \leq Q_{n,\beta}(f) - B_n(f)\big) \to \beta - \alpha$, for every f with $(P_0 + G)f^2 < \infty$. If $\sqrt{n}B_n(f) \to 0$, in probability, then also $\mathrm{P}_{P_0}\big(Q_{n,\alpha}(f) \leq P_0 f \leq Q_{n,\beta}(f)\big) \to \beta - \alpha$, for every such f. For bounded functions f, the latter is true if the atoms of P_0 satisfy $P_0\{x_j\} \leq Cj^{-\alpha}$, for some constants C and $\alpha > 2$. For $f(x) = x$, this is true for $\alpha > 4$.*

Proof. The α-quantile $Q_{n,\alpha}(f)$ of the posterior distribution of Pf is equal to $n^{-1/2}\bar{Q}_{n,\alpha}(f) + \mathbb{P}_n f + B_n(f)$, for $\bar{Q}_{n,\alpha}(f)$ the α-quantile of the posterior distribution of $\sqrt{n}\big(Pf - \mathbb{P}_n f - B_n(f)\big)$. By Theorem 3.2.1, the latter posterior distribution tends to a normal distribution with mean zero and variance $\tau^2(f) = \operatorname{var} \mathbb{G}_{P_0} f$. It follows that

$$Q_{n,\alpha}(f) = \mathbb{P}_n f + B_n(f) + \frac{\tau(f)}{\sqrt{n}}\xi_\alpha + o_P\Big(\frac{1}{\sqrt{n}}\Big),$$

where ξ_α is the α-quantile of the standard normal distribution. Thus the event $P_0 f \geq Q_{n,\alpha}(f) - B_n(f)$ can be rewritten as $P_0 f \geq \mathbb{P}_n f + \tau(f)/\sqrt{n}\,\xi_\alpha + o_P(n^{-1/2}$. The probability of the latter event tends to tends to $1 - \alpha$, by the central limit theorem applied to $\sqrt{n}(\mathbb{P}_n f - P_0 f)$.

If $\sqrt{n}B_n(f)$ tends to zero in probability, then in the preceding display $B_n(f)$ can be incorporated into the $o_P(n^{-1/2})$ remainder term, and the remaining argument works for the uncorrected interval as well.

The final assertions follow from Corollary 3.2.2. ■

Example 3.2.4. *The following explicit counterexample illustrates that the coverage can fail. Let G be the normal distribution with both mean and variance 1, let $P_0 = \sum_{j=1}^\infty p_j\delta_j$, for $p_j = 6/(\pi j)^2$, and consider the function $f = \mathbb{1}_{(1,\infty)} - \mathbb{1}_{(-\infty,1]}$. Since $Gf = 0$, we get $(\sigma K_n/\sqrt{n})(G - \tilde{\mathbb{P}}_n)f = (\sigma/\sqrt{n})\sum_{i=1}^{K_n} f(\tilde{X}_i)$. Eventually the atom $\{1\}$ will be among the observations. Since $f(1) = -1$ and $f(j) = 1$ for all atoms $j \geq 2$, $(\sigma K_n/\sqrt{n})(G - \tilde{\mathbb{P}}_n)f = (\sigma/\sqrt{n})(-1 + (K_n - 1)) \to \sigma\sqrt{6/\pi}$, almost surely, by [42], Theorem 8, Theorem 1' and Example 4. The coverage of the uncorrected interval $\big[Q_{n,\alpha}(f), Q_{n,\beta}(f)\big]$ will tend to $\Phi(-\xi_\alpha - \sigma\sqrt{6/\pi}) - \Phi(-\xi_\beta - \sigma\sqrt{6/\pi})$.*

The joint convergence in collections of functions f allows to study simultaneous credible sets and credible bands, besides univariate intervals. For instance, in the case that the sample space is the real line, we can take $\mathcal{F}$ equal to the set of all indicators of cells $(-\infty, t]$, and obtain a credible band for the distribution function $F_0(t) = P_0(-\infty, t]$, as follows. Let $F(t) = P(-\infty, t]$ be the distribution function of the posterior process, and for $m_n(t)$ and $s_n(t)$ two functions dependent on $X_1, \ldots, X_n$, let $\xi_{n,\alpha}$ be the α-quantile of the posterior distribution of $\sup_{t\in\mathbb{R}}\bigl|(F(t) - m_n(t))/s_n(t)\bigr|$. Consider the credible band of functions

$$C_n(\alpha) := \bigl\{F : m_n(t) - \xi_{n,1-\alpha}s_n(t) \le F(t) \le m_n(t) + \xi_{n,1-\alpha}s_n(t), \forall t\bigr\}.$$

Possible choices for the functions m_n and s_n are the pointwise posterior mean $m_n(t) = \mathbb{E}\bigl(F(t)|\, X_1, \ldots, X_n\bigr)$ and the pointwise posterior standard deviation $s_n(t) = \mathrm{sd}\bigl(F(t)|\, X_1, \ldots, X_n\bigr)$. The quantiles $\xi_{n,\alpha}$ will typically be computed approximately from an MCMC sample from the posterior distribution, or approximated using tables for the limiting Brownian bridge process.

Corollary 3.2.5. *If P_0 is a discrete probability distribution with atoms such that $P_0\{x_j\} \le Cj^{-2-\varepsilon}$, for some constants C and $\varepsilon > 0$, then $\mathrm{P}_{P_0}\bigl(F_0 \in C_n(\alpha)\bigr) \to 1 - 2\alpha$.*

Proof. Because the class of indicator functions is Donsker, both the classical empirical process $\{\sqrt{n}(\mathbb{F}_n - F_0)(t) : t \in \mathbb{R}\}$ and the posterior empirical process $\{\sqrt{n}(F - \mathbb{F}_n)(t) : t \in \mathbb{R}\}|\, X_1, \ldots, X_n$ tend to the process $\mathbb{G}_U \circ F_0$, for $\mathbb{G}_U$ a standard (classical) Brownian bridge process, by Theorem 3.2.1. The result follows from this along the same lines as the proof of Corollary 3.2.3. ■

The bias term (3.3) vanishes as $\sigma \downarrow 0$, which is in agreement with the fact that in this case the Pitman-Yor prior approaches the Dirichlet prior, which is well known to give asymptotically correct inference for any distribution P_0. The bias term increases with σ, which is counterintuitive, as the bias appears only for heavy-tailed P_0 (having many large atoms), while large σ gives more different atoms in the prior.

One might hope that a data-dependent choice of σ could solve this bias problem. The empirical Bayes method is to estimate σ by the maximum likelihood estimator based on observing $X_1, \ldots, X_n$ in the Bayesian model, i.e. the maximiser of the marginal likelihood, and plug this into the posterior distribution of P for known σ. The hierarchical Bayes method is to put a prior on σ, and given σ, put the Pitman-Yor prior on P. Disappointingly, these methods do not change the limit behaviour of the posterior distribution of P. This is explained by the fact that these methods yield a reasonable estimator of a value of σ connected to the discreteness of the true distribution P_0, and we already noted the counterintuitive fact that a better match of discreteness does not solve the bias problem, but even makes it worse.

A sample $X_1, \ldots, X_n$ from a realisation of the Pitman-Yor process induces a (random) partition of the set $\{1, 2, \ldots, n\}$ through the equivalence relation $i \equiv j$ if and only if $X_i = X_j$. An alternative way to generate the sample is to generate first the partition and next attach to each set in the partition a value generated independently from the center measure G (see e.g. [33], Lemma 14.11 for a precise statement), duplicating this as many times as there are indices in the set, in order to form the observations $X_1, \ldots, X_n$. Because the parameter σ enters only in creating the partition, the partition is a sufficient statistic for σ. Because of exchangeability, the vector $(N_{n,1}, \ldots, N_{n,K_n})$ of cardinalities of the partitioning sets is already sufficient for σ and hence the empirical Bayes estimator and posterior distribution of σ based on observations $(X_1, \ldots, X_n)$ or on observations $(K_n, N_{n,1}, \ldots, N_{n,K_n})$ are the same.

The likelihood function for σ is therefore equal to the probability of a particular partition, called the exchangeable partition probability function (EPPF). For the Pitman-Yor process this is given by (see [60], or [33, page 465])

$$p_\sigma(N_{n,1}, \ldots, N_{n,K_n}) = \frac{\prod_{i=1}^{K_n-1}(M + i\sigma)}{(M+1)^{[n-1]}} \prod_{j=1}^{K_n} (1-\sigma)^{[N_{n,j}-1]}. \tag{3.5}$$

Here $a^{[n]} = a(a+1)\cdots(a+n-1)$ is the ascending factorial, with $a^{[0]} = 1$ by convention. For the case that $M = 0$, it is shown in [23], that provided the partition is nontrivial $(1 < K_n < n)$, the maximiser $\hat\sigma_n$ of this likelihood exists. Moreover, if the true distribution P_0 is discrete, with atoms satisfying, for $\alpha_0(u) = \#\{x\colon 1/P_0\{x\} \le u\}$ and some $\sigma_0 \in (0,1)$,

$$\sup_{u>1} \frac{|\alpha_0(u) - Lu^{\sigma_0}|}{\sqrt{u^{\sigma_0}\log(eu)}} < \infty,$$

then [23] shows that the maximum likelihood estimator satisfies

$$\hat\sigma_n = \sigma_0 + O_P(n^{-\sigma_0/2}\sqrt{\log n}).$$

Thus the coefficient of regular variation σ_0 may be viewed as a true value of σ, identified by the maximum likelihood estimator.

For the following theorem we need only the consistency of $\hat\sigma_n$, which we prove in Section 3.4.3 for general M, under the condition that α_0 is regularly varying. We also consider the full Bayes approach, and show that the posterior distribution of σ concentrates asymptotically around the empirical likelihood estimator, and hence contracts to σ_0, under the same condition.

Theorem 3.2.6. *Let $P_0 = (1-\lambda)P_0^d + \lambda P_0^c$ where P_0^d is a discrete and P_0^c an atomless probability distribution. If $\hat\sigma_n$ are estimators based on $X_1, \ldots, X_n$ such that $\hat\sigma_n \to \sigma_0$ in probability, and P is the posterior Pitman-Yor process of Theorem 3.2.1, then the process*

$$\sqrt{n}\Big(P - \mathbb{P}_n - \frac{\hat\sigma_n K_n}{n}(G - \tilde{\mathbb{P}}_n)\Big)\Big|\, X_1, \ldots, X_n$$

tends to the same limit process as in Theorem 3.2.1 with σ replaced by σ_0, in probability. If P_0 is discrete with atoms such that α_0 given in (3.4) is regularly varying of exponent $\sigma_0 \in (0,1)$, then this is true for the maximum likelihood estimator $\hat\sigma_n$. Furthermore, in this case for Π_σ a prior distribution on σ with continuous positive density on $[0,1]$, the posterior distribution of P in the model $\sigma \sim \Pi_\sigma$, $P|\,\sigma \sim \mathrm{PY}(\sigma, M, G)$

and $X_1,\ldots,X_n|\, P,\sigma \sim P$ satisfies the assertion of Theorem 3.2.1, with σ in the left side also interpreted as a random posterior variable and σ in the right side replaced by σ_0. Finally if P_0 possesses a nontrivial atomless component (i.e. $\lambda > 0$), then $\hat{\sigma}_n \to \sigma_0 := 1$.

The proof of the theorem is deferred to Section 3.4.2. The final assertion of the theorem underlines again the deficiency of the Pitman-Yor process for distributions with a continuous component, which is not solved by estimating the type parameter . The type estimate tends to type 1 instead of the desired type 0 corresponding to the Dirichlet prior.

Besides the type parameter, the prior precision parameter M could be replaced by a data-dependent version. However, unlike the type parameter, this prior precision does not appear in the asymptotics of the posterior distribution of $\sqrt{n}(P-\mathbb{P}_n)$. Moreover, inspection of the proof of Theorems 3.2.1 and 3.2.6 shows that the convergence in these theorems is uniform in $M \ll \sqrt{n}$. Thus data-dependent M will not lead to new insights.

In the case of a discrete distribution P_0 for which the atoms decrease too slowly to ensure that $K_n/\sqrt{n}$ tends to zero, the bias term (3.3) could still tend to zero if $\tilde{\mathbb{P}}_n \to G$. However, we show below that $\tilde{\mathbb{P}}_n(A) \to 0$, for any set A that contains only finitely many atoms of P_0, and hence such convergence is false in any reasonable sense. Furthermore, the (in)consistency result (3.1) shows that in the case that P_0 possesses a continuous component, the center measure $G = P_0^c$ is the only choice for which the posterior distribution is even consistent. A data-dependent center measure might achieve this, but in the present context would come down to the original problem of estimating P_0. Hierarchical choices (and hence random) of the center measure are considered in [8, 7], but with the different aim of finding hierarchical structures in the data.

Lemma 3.2.7. *If P_0 is discrete with infinitely many support points, then $\tilde{\mathbb{P}}_n f \to 0$ in probability for any bounded function f with finite support. Furthermore, $\tilde{\mathbb{P}}_n f \to f_\infty$ in probability, for any bounded function f for which there exists a number f_∞ such that $\sup_{x:P_0\{x\}<\delta} |f(x) - f_\infty| \to 0$, as $\delta \downarrow 0$.*

Proof. Let x_j be the atoms of P_0, ordered by decreasing size $p_j := P_0\{x_j\}$ and set $f_j = f(x_j)$. Arguing as in the proof of Corollary 3.2.2, we can obtain (for the variance also see [42], formulas (39)–(40))

$$\begin{aligned}\mathbb{E}(K_n\tilde{\mathbb{P}}_n f) &= \sum_{j=1}^{\infty} f_j\bigl(1-(1-p_j)^n\bigr),\\ \operatorname{var}(K_n\tilde{\mathbb{P}}_n f) &= \sum_{j=1}^{\infty} f_j^2\bigl[(1-p_j)^n-(1-p_j)^{2n}\bigr]\\ &\qquad+\sum\sum_{i\neq j} f_i f_j\bigl[(1-p_i-p_j)^n-(1-p_i)^n(1-p_j)^n\bigr].\end{aligned}$$

For $f=1$ these expressions reduce to $\mathbb{E}K_n$ and $\operatorname{var} K_n$. As all terms of the series in $\mathbb{E}K_n$ tend to 1, it can be seen that $\mathbb{E}K_n \to \infty$. Furthermore, it can be seen that $\operatorname{var} K_n \le \mathbb{E}K_n$, as the terms in the second series in $\operatorname{var} K_n$ are negative and the terms of the first series are bounded above by the terms in $\mathbb{E}K_n$. Because the terms of the series tend to zero as $n\to\infty$, for fixed i,j, and $f_j \to f_\infty$, as $j\to\infty$, for general f as in the second assertion of the lemma, the expressions are asymptotically equivalent to $f_\infty \mathbb{E}K_n + o(1)$ and $f_\infty^2 \operatorname{var} K_n + o(1)$, as $n\to\infty$. It follows that

$$\operatorname{var}\Bigl(\frac{K_n\tilde{\mathbb{P}}_n f}{\mathbb{E}(K_n\tilde{\mathbb{P}}_n f)}-1\Bigr)=\frac{\operatorname{var}(K_n\tilde{\mathbb{P}}_n f)}{\bigl(\mathbb{E}(K_n\tilde{P}P_n f)\bigr)^2}=\frac{f_\infty^2\operatorname{var} K_n+o(1)}{\bigl(f_\infty\mathbb{E}K_n+o(1)\bigr)^2}.$$

Since $\operatorname{var} K_n \le \mathbb{E}K_n \to \infty$, the right side tends to zero if $f_\infty \neq 0$. Then it follows that $K_n\tilde{\mathbb{P}}_n f/\mathbb{E}(K_n\tilde{\mathbb{P}}_n f)\to 1$, in probability. Taking $f=1$, we see that $K_n/\mathbb{E}K_n \to 1$, in probability. Combining the preceding, we conclude that $\tilde{\mathbb{P}}_n f/f_\infty \to 1$, in probability.

If $f_\infty = 0$, then it follows that $\mathbb{E}(K_n\tilde{\mathbb{P}}_n f)\to 0$. Combination with the fact that $K_n\to\infty$, almost surely, gives that $\tilde{\mathbb{P}}_n f \to 0$, in probability.

If f has finite support, then the condition on f in the second part holds with $f_\infty = 0$ and hence $\tilde{\mathbb{P}}_n f\to 0$, in probability. ■

3.3 Numerical illustration

To illustrate that credible sets can be off, we carried out three simulation experiments, involving three discrete true probability distributions P_1, P_2, P_3 on $\mathbb{N}$. We focused on a credible interval for the probability of the set $[2,\infty)$. The measure P_1 is finitely discrete and given in Table 3.1, while P_2 and P_3 are given by the formulas

$$P_2\{k\}\propto\frac{1}{k^2},\qquad\qquad P_3\{k\}\propto\frac{1}{k^{1.5}}.$$

By the results of [42], as $n\to\infty$ the number K_n of distinct observations in a sample of size n from these distributions are asymptotically equal to 6, and proportional to $\sqrt{n}$ and to $n^{2/3}$, respectively, for P_1, P_2 and P_3. Thus $K_n/\sqrt{n}$ tends to 0, a positive

Table 3.1: Probability distribution P_1

k	1	2	3	4	5	6
$P_1(X=k)$	0.1	0.1	0.2	0.2	0.3	0.1

constant and ∞, respectively, and a bias is expected for P_2 and P_3, but not for P_1, where P_2 is a boundary case.

As prior parameters we used $\sigma = 1/2$ and $M = 1$ and G the normal distribution with mean and variance 1. The choice $M = 1$ means that the prior is not biased exceedingly against the true distribution.

The Pitman-Yor posterior distribution can be simulated using the explicit representation given by [58] (see Section 3.4). Following Algorithm 1 from [3], we truncated the infinite series in the representation at a finite value, ensuring that the total weight of the tail is smaller than $n^{-1/2}$ so that the approximation is accurate within our context. We simulated 10000 samples from each of P_1, P_2 and P_3 and for five different sample sizes: $n = 10, 10^2, 10^3, 10^4, 10^5$. For each sample we computed a 95% credible interval for $P[2,\infty)$ from its marginal posterior distribution, constructed using the 0.025 and the 0.975 posterior quantiles. We next computed coverage as the proportion of the 10000 replications that the true value, $P_1[2,\infty)$, $P_2[2,\infty)$ or $P_3[2,\infty)$, belonged to the interval. We did the same with the credible interval shifted by $(\sigma K_n/\sqrt{n})\big(G[2,\infty) - \tilde{\mathbb{P}}_n[2,\infty)\big)$, derived from (3.3).

Tables 3.2 and 3.3 summarise the results. For P_1 both the corrected uncorrected intervals perform satisfactorily, whereas for P_2 and P_3 the uncorrected intervals undercover, severely so for P_3, while the corrected intervals perform reasonably well, although not perfectly. The simulation results thus confirm the theoretical findings.

Table 3.2: Coverage of uncorrected posterior 95% credible intervals

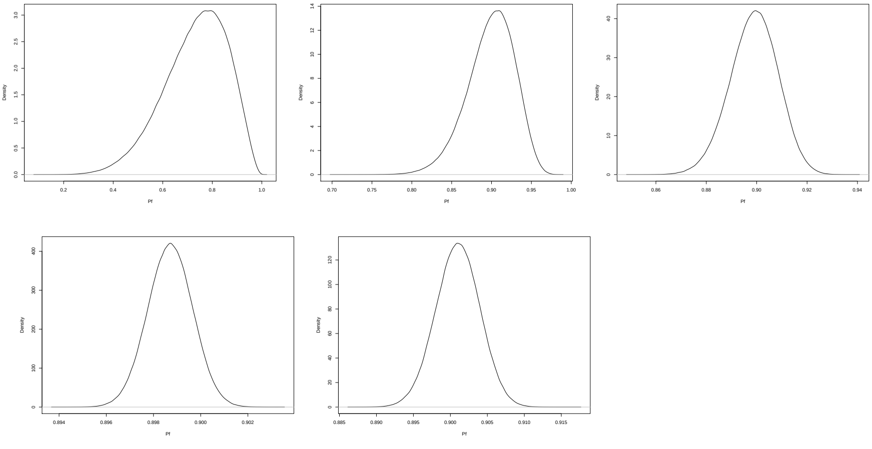
n	10	100	1000	10000	100000
P_1	0.660	0.940	0.957	0.940	0.947
P_2	0.707	0.772	0.790	0.845	0.838
P_3	0.559	0.231	0.035	0.0	0.0

Table 3.3: Coverage of corrected posterior 95% credible intervals

n	10	100	1000	10000	100000
P_1	0.990	0.967	0.958	0.942	0.945
P_2	0.814	0.941	0.958	0.971	0.971
P_3	0.884	0.956	0.959	0.985	0.949

To illustrate the asymptotic normality of the posterior distribution, Figure 3.1 shows density plots of the marginal posterior distribution of $P[2,\infty)$, given samples of vari-

Figure 3.1: Density of the marginal posterior distribution of $P[2,\infty)$ based on n observations from P_1, for $n = 10, 100, 1000$ (top row) and $n = 10^4, 10^5$. The true value of the parameter is $P_1[2,\infty) = 0.9$.



ous sizes from P_1. The plots were computed from the 100000 replicates, using the R "density" function. The normal approximation is satisfactory for $n = 1000$, but the posterior is visibly skewed for $n = 100$.

3.4 Proofs

Let $\tilde{X}_1, \ldots, \tilde{X}_{K_n}$ be the distinct values in $X_1, \ldots, X_n$, and let $N_{1,n}, \ldots, N_{K_n,n}$ be their multiplicities. By Corollary 20 in [60] (or see [33, Theorem 14.37]), the posterior distribution of the Pitman-Yor process can be characterised as the distribution of

$$\mathrm{PY}_n = R_n S_n + (1 - R_n) Q_n, \tag{3.6}$$

for

$$S_n = \sum_{i=1}^{K_n} W_{n,i} \delta_{\tilde{X}_i}, \tag{3.7}$$

and independent variables R_n, W_n, Q_n with, conditionally on $X_1, \ldots, X_n$, distributed according to:

- $R_n \sim \mathrm{B}(n - \sigma K_n, M + \sigma K_n)$,
- $Q_n \sim \mathrm{PY}(\sigma, M + \sigma K_n, G)$,
- $W_n = (W_{n,1}, \ldots, W_{n,K_n}) \sim \mathrm{Dir}(K_n; N_{n,1} - \sigma, \ldots, N_{n,K_n} - \sigma)$.

Here B and Dir refer to the beta and Dirichlet distributions, respectively. The number K_n will tend almost surely to the total number of atoms of P_0 in the case that P_0 is finitely discrete, and it will tend to infinity otherwise. In the latter case the rate of growth can have any order n^γ, for $0 < \gamma \leq 1$. (See Theorem 8 of [42], where it is shown that $K_n/\mathbb{E}K_n \to 1$, almost surely, where any rate can occur for $\mathbb{E}K_n$.) The proofs below use that K_n/n tends to the mass λ of the continuous part of P_0, and the limit of the related sequence $n^{-1}K_n\tilde{\mathbb{P}}_n f = n^{-1}\sum_{i=1}^{K_n} f(\tilde{X}_i)$.

Lemma 3.4.1. *The number K_n of distinct values among $X_1, \ldots, X_n \overset{iid}{\sim} \lambda P_0^c + (1-\lambda)P_0^d$ satisfies $K_n/n \to \lambda$, almost surely. The number K_n^d of those values that belong to the set S of atoms of P_0^d satisfies $K_n^d/n \to 0$, almost surely.*

Proof. The number of distinct values not in S is $K_n^c{:}= n\mathbb{P}_n(S^c)$ and hence $K_n^c/n \to P_0(S^c) = \lambda$, almost surely. If $S = \{x_1, x_2, \ldots\}$, then the number of distinct values in S is bounded above by $m + n\mathbb{P}_n\{x_{m+1}, x_{m+2}, \ldots\}$, for any m, and hence $K_n^d/n \leq m/n + \mathbb{P}_n\{x_{m+1}, x_{m+2}, \ldots\} \to P_0\{x_{m+1}, x_{m+2}, \ldots\}$, almost surely, for every m. ∎

A class $\mathcal{F}$ of measurable functions $f{:}\, \mathcal{X} \to \mathbb{R}$ is P_0-*Glivenko-Cantelli* if the uniform law of large numbers holds: $\sup_{f\in\mathcal{F}} |\mathbb{P}_n f - P_0 f| \to 0$, outer almost surely (see e.g. [78], Chapter 2.4; we write "outer" because the supremum may not be measurable; for standard examples this is superfluous). An *envelope function* of $\mathcal{F}$ is a measurable function $F{:}\, \mathcal{X} \to \mathbb{R}$ such that $|f| \leq F$, for every $f \in \mathcal{F}$.

Lemma 3.4.2. *Suppose $\mathcal{F}$ has an envelope function with $P_0F < \infty$. If S is the set of atoms of P_0^d, then $\sup_{f\in\mathcal{F}}\bigl|n^{-1}\sum_{i=1}^{K_n} f(\tilde{X}_i)1_{\tilde{X}_i\in S}\bigr| \to 0$, outer almost surely. Furthermore, if $\sigma_n \to \sigma \in [0,1]$, and $\mathcal{F}$ is a P_0-Glivenko-Cantelli class, then uniformly in $f \in \mathcal{F}$, outer almost surely,*

$$\mathbb{P}_n f + \frac{\sigma_n K_n}{n}\tilde{\mathbb{P}}_n f \to Tf{:}= (1-\lambda)P_0^d f + (1-\sigma)\lambda P_0^c f.$$

Proof. For any M, the supremum is bounded above by $n^{-1}K_n^d M + \mathbb{P}_n F 1_{F>M} \to P_0 F 1_{F>M}$, almost surely. The first term tends to zero by Lemma 3.4.1, for any M. The second term can be made arbitrarily small by choosing M large.

For the convergence in the display we write $K_n\tilde{\mathbb{P}}_n f = \sum_{i=1} f(X_i)1_{X_i\notin S} + \sum_{i=1}^{K_n} f(\tilde{X}_i)1_{\tilde{X}_i\in S}$. By the first assertion, the second sum divided by n tends to zero, uniformly in f. The first sum divided by n tends to $\lambda P_0^c f$, where the convergence is uniform in $f \in \mathcal{F}$ if $\mathcal{F}$ is a Glivenko-Cantelli class (which implies that the set of functions $x \mapsto f(x)1_{S^c}(x)$ is a Glivenko-Cantelli class, in view of [79]). Thus the left side of the display tends to $P_0 f + \sigma\lambda P_0^c f$, which is equal to Tf. ∎

3.4.1 Proof of Theorem 3.2.1

The left side $\sqrt{n}\big(\mathrm{PY}_n - \mathbb{P}_n + (\sigma K_n/n)(\tilde{\mathbb{P}}_n - G)\big)$ of the theorem can be decomposed as

$$\sqrt{n}\Big(R_n - 1 + \frac{\sigma K_n}{n}\Big)(S_n - Q_n) + \sqrt{n}\bigg(S_n\Big(1 - \frac{\sigma K_n}{n}\Big) - \mathbb{P}_n + \frac{\sigma K_n}{n}\tilde{\mathbb{P}}_n\bigg) \\ + \sigma\sqrt{K_n}(Q_n - G)\sqrt{\frac{K_n}{n}}. \tag{3.8}$$

We derive the limit distributions of these three terms in Lemmas 3.4.3–3.4.5 below. For later use it will be helpful to allow $\sigma \in (0,1)$ to depend on n. For this reason we give precise proofs of the first two lemmas, although they are very similar to results obtained in [41, 33]. The main novelty is in the third lemma. For simplicity we assume that $\sigma_n \in (0,1)$ converges to a limit, which we allow to be 0 or 1.

Lemma 3.4.3. *If $\sigma_n \to \sigma \in [0,1]$, then*

$$\sqrt{n}\left(R_n - 1 + \frac{\sigma_n K_n}{n}\right) \mid X_1, \ldots, X_n \rightsquigarrow N\big(0, (1-\sigma\lambda)\sigma\lambda\big), \qquad a.s. \tag{3.9}$$

Proof. We can represent the beta variable R_n as the quotient $R_n = U_n/(U_n + V_n)$, for independent gamma variables $U_n \sim \Gamma(u_n, 1)$ and $V_n \sim \Gamma(v_n, 1)$, for $u_n = n - \sigma_n K_n$ and $v_n = M + \sigma_n K_n$ the means, and also variances, of the latter variables. We can decompose

$$(U_n + V_n)\Big(R_n - \frac{u_n}{u_n + v_n}\Big) = \frac{v_n}{u_n + v_n}(U_n - u_n) - \frac{u_n}{u_n + v_n}(V_n - v_n).$$

Since $\sigma_n K_n/n \to \sigma\lambda \in [0,1]$, we have $v_n/(u_n + v_n) \to \sigma\lambda$ and $u_n/(u_n+v_n) \to 1 - \sigma\lambda$. Furthermore, $(U_n + V_n)/n \to 1$, almost surely, by the law of large numbers.

If $\sigma\lambda < 1$, then $n - \sigma_n K_n \to \infty$ and hence $(U_n - u_n)/\sqrt{u_n} \rightsquigarrow Z_1 \sim N(0,1)$, by the central limit theorem. It follows that $(U_n - u_n)/\sqrt{n} \rightsquigarrow Z_1\sqrt{1-\sigma\lambda}$. If $\sigma\lambda = 1$, then $\mathrm{var}(U_n/\sqrt{n}) = u_n/n \to 0$ and hence $(U_n - u_n)/\sqrt{n} \rightsquigarrow 0$, where the limit 0 is identical to $Z_1\sqrt{1-\sigma\lambda}$ in this case. Thus in all cases $(U_n - u_n)/\sqrt{n} \rightsquigarrow Z_1\sqrt{1-\sigma\lambda}$.

If $\sigma\lambda > 0$, then $\sigma_n K_n \to \infty$ and hence $(V_n - v_n)/\sqrt{v_n} \rightsquigarrow Z_2 \sim N(0,1)$, by the central limit theorem. It follows that $(V_n - vn)/\sqrt{n} \rightsquigarrow Z_2\sqrt{\sigma\lambda}$. If $\sigma\lambda = 1$, then $\mathrm{var}(V_n/\sqrt{n}) = v_n/n \to 0$ and hence $(V_n - v_n)/\sqrt{n} \rightsquigarrow 0$, where the limit 0 is identical to $Z_2\sqrt{\sigma\lambda}$ in this case. Thus in all cases $(V_n - v_n)/\sqrt{n} \rightsquigarrow Z_2\sqrt{\sigma\lambda}$.

Combining the preceding, we see that the sequence $\sqrt{n}\big(R_n - u_n/(u_n + v_n)\big)$ converges weakly to $\sigma\lambda Z_1\sqrt{1-\sigma\lambda} + (1-\sigma\lambda)Z_2\sqrt{\sigma\lambda}$. As the limit variable has variance $(1-\sigma\lambda)\sigma\lambda$ and $u_n/(u_n+v_n) = (1 - \sigma_n K_n/n)(1 + O(1/n))$, this concludes the proof. ■

Lemma 3.4.4. *If $\sigma_n \to \sigma \in [0,1]$ and $K_n \to \infty$ and $\mathcal{F}$ is a class of finitely many G-square-integrable functions, then in $\mathbb{R}^{\mathcal{F}}$,*

$$\sigma_n \sqrt{K_n}(Q_n - G)|\, X_1, \ldots, X_n \rightsquigarrow \sqrt{\sigma(1-\sigma)}\, \mathbb{G}_G. \qquad a.s. \qquad (3.10)$$

The convergence is also true in $\ell^\infty(\mathcal{F})$ if $\mathcal{F}$ possesses a G-square integrable envelope function and the Pitman-Yor process $\mathrm{PY}(\sigma, \sigma, G)$ satisfies the central limit theorem in this space.

Proof. The process $Q_n \sim \mathrm{PY}(\sigma_n, M + \sigma_n K_n, G)$ centered at mean zero can be represented as

$$Q_n - G \sim \sum_{i=0}^{K_n} W_{n,i}(P_i - G),$$

where $(W_{n,0}, \ldots, W_{n,K_n}) \sim \mathrm{Dir}(K_n + 1; M, \sigma_n, \ldots, \sigma_n)$ is independent of the independent processes $P_0 \sim \mathrm{PY}(\sigma_n, M, G)$ and $P_i \overset{\text{iid}}{\sim} \mathrm{PY}(\sigma_n, \sigma_n, G)$, for $i = 1, \ldots K_n$ (see e.g. Proposition 14.35 in [33]). The variable $W_{n,0}$ is $B(M, K_n\sigma_n)$-distributed, whence

$$\sigma_n \sqrt{K_n} \mathbb{E}\big|W_{n,0}(P_0 - G)f\big| = \frac{\sigma_n\sqrt{K_n}M}{M + K_n\sigma_n}\mathbb{E}|(P_0 - G)f| \le \frac{M}{\sqrt{K_n}}\sqrt{Gf^2},$$

where the moment of $(P_0 - G)f$ can be obtained from Proposition 14.34 in [33]. Next by the gamma representation of the Dirichlet distribution (e.g. Propositions G.2 and G.3 in [33]), we can represent

$$\sigma_n \sqrt{K_n} \sum_{i=1}^{K_n} W_{n,i}(P_i - G) \sim (1 - W_{n,0}) \frac{K_n^{-1/2} \sum_{i=1}^{K_n} V_{n,i}(P_i - G)}{K_n^{-1} \sum_{i=1}^{K_n} V_{n,i}/\sigma_n},$$

where the variables $V_{i,n} \overset{\text{iid}}{\sim} \Gamma(\sigma_n, 1)$ are independent of $W_{n,0}$ and the P_i. The triangular array of variables $V_{n,i}(P_i - G)$ are i.i.d. for every n with

$$\begin{aligned}
&\mathbb{E}V_{n,1}^2\big((P_1 - G)f\big)^2 = \sigma_n(1+\sigma_n)G(f - Gf)^2 \frac{1-\sigma_n}{1+\sigma_n},\\
&\mathbb{E}V_{n,1}^2\big((P_1 - G)f\big)^2 1_{|V_{n,1}(P_1 - G)f| \ge M_n} \to 0,
\end{aligned}$$

for any $M_n \to \infty$. The second claim is implied by the uniform integrability of the set of variables $W_\sigma := V_\sigma^2\,((P_\sigma - G)f)^2$, for $\sigma \in [0,1]$, where $V_\sigma \sim \Gamma(\sigma, 1)$ is independent of $P_\sigma \sim \mathrm{PY}(\sigma, \sigma, G)$, and W_0 and W_1 are defined to be degenerate at 0, in agreement with the first line of the preceding display. This itself is a consequence of the continuity of the map $\sigma \mapsto W_\sigma$ from $[0,1]$ to $L_2(\Omega)$ and the Dunford-Pettis theorem. The continuity follows from the norm continuity, $\mathrm{E}W_{\sigma_n}^2 \to \mathrm{E}W_\sigma^2$, if $\sigma_n \to \sigma$, by the first assertion in the display, combined with the continuity in distribution of $\sigma \mapsto W_\sigma$. Therefore, the sequence $K_n^{-1/2} \sum_{i=1}^{K_n} V_{n,i}(P_i - G)$ tends to a normal distribution with mean zero and variance $\sigma(1-\sigma)G(f - Gf)^2$, by the Lindeberg central limit theorem. The linearity of the process in f shows that as a process it tends marginally in distribution to the process $\sqrt{\sigma(1-\sigma)}\,\mathbb{G}_G$. Because $\mathrm{var}\bigl(K_n^{-1}\sum_{i=1}^{K_n} V_{n,i}/\sigma_n\bigr) = 1/(K_n\sigma_n)$, we have $K_n^{-1}\sum_{i=1}^{K_n} V_{n,i}/\sigma_n \to 1$, in probability, if $K_n\sigma_n \to \infty$. Since also $1 - W_{n,0} \to 1$, the proof is complete in the case that $K_n\sigma_n \to \infty$.

If $K_n\sigma_n$ remains bounded, then necessarily $\sigma_n \to 0$, as $K_n \to \infty$, by assumption. Then

$$\begin{aligned}\sigma_n^2 K_n \mathbb{E}\Bigl(\sum_{i=0}^{K_n} W_{n,i}(P_i - G)f\Bigr)^2 &= \sigma_n^2 K_n \sum_{i=0}^{K_n}\sum_{j=0}^{K_n} \mathbb{E} W_{n,i}W_{n,j}(P_i - G)f(P_j - G)f\\ &\leq \sigma_n^2 K_n \mathbb{E}\Bigl(\sum_{i=0}^{K_n} W_{n,i}\Bigr)^2 Gf^2 \leq \sigma_n^2 K_n Gf^2.\end{aligned}$$

Since this tends to zero, the lemma holds also in this case, with a limit process equal to 0, which is equal to $\sqrt{\sigma(1-\sigma)}\mathbb{G}_G$.

For the final assertion we note that the preceding argument gives the convergence of $\sup_{f\in\mathcal{F}} \sigma_n\sqrt{K_n}W_{n,0}(P_0 - G)f$ to zero for any class $\mathcal{F}$ with square-integrable envelope function. The convergence of $K_n^{-1/2}\sum_{i=1}^{K_n} V_{n,i}(P_i - G)$ in $\infty(\mathcal{F})$ follows from the convergence of $K_n^{-1/2}\sum_{i=1}^{K_n}(P_i - G)$ by the multiplier central limit theorem (e.g. Lemma 2.9.1 and Theorem 2.9.2 in [78]). ∎

Lemma 3.4.5. *If $\sigma_n \to \sigma \in [0,1]$, where $\sigma\lambda < 1$, then for any P_0-Donsker class with square-integrable envelope function*

$$\sqrt{n}\Bigl(S_n\Bigl(1 - \frac{\sigma_n K_n}{n}\Bigr) - \mathbb{P}_n + \frac{\sigma K_n}{n}\tilde{\mathbb{P}}_n\Bigr) \rightsquigarrow \mathbb{W} - \frac{1}{1-\sigma\lambda}\mathbb{W}1\,T, \qquad a.s., \tag{3.11}$$

in $\ell^\infty(\mathcal{F})$, where $\mathbb{W} = \sqrt{\lambda(1-\sigma)}\mathbb{G}_{P_0^c} + \sqrt{1-\lambda}\mathbb{G}_{P_0^d}$, for independent Brownian bridge processes $\mathbb{G}_{P_0^c}$ and $\mathbb{G}_{P_0^d}$, and T is the (deterministic) process defined in Lemma 3.4.2. The convergence is true in probability for any P_0-Donsker class. If $\sigma_n \to \sigma \in [0,1]$, where $\sigma\lambda = 1$, then the sequence of processes tends to the zero proces.

Proof. A gamma representation for the multinomial vector W_n in the definition of S_n is

$$W_{n,i} = \frac{U_{i,0} + \sum_{j=1}^{N_{n,i}-1} U_{i,j}}{\sum_{i=1}^{K_n}\Bigl(U_{i,0} + \sum_{j=1}^{N_{n,i}-1} U_{i,j}\Bigr)},$$

for all $U_{i,j}$ independent, $U_{i,0} \sim \Gamma(1-\sigma,1)$ and $U_{i,j} \sim \Gamma(1,1)$, for $j \geq 1$. Relabel the n variables $U_{i,j}$ as $\xi_{n,1},\ldots,\xi_{n,n}$, as follows. Let S be the set of all atoms of P_0. An observation X_i that is not contained in S appears exactly once in the set $\{X_1,\ldots,X_n\}$ of observations; set the variable $\xi_{n,i}$ with the corresponding i equal to $U_{i,0}$. Every X_i

that is contained in S appears $N_{n,i} \geq 1$ times among $X_1, \ldots, X_n$; set the $\xi_{n,j}$ with indices corresponding to these appearances equal to $U_{i,0}, U_{i,1}, \ldots, U_{i,N_{n,i}-1}$. Then

$$S_n = \sum_{i=1}^{K_n} W_{n,i} f(\tilde{X}_i) = \frac{n^{-1}\sum_{i=1}^n \xi_{n,i} f(X_i)}{n^{-1}\sum_{i=1}^n \xi_{n,i}} =: \frac{\overline{S}_n f}{\overline{S}_n 1}, \tag{3.12}$$

and the left side of the lemma can be decomposed as

$$\begin{aligned} S_n f\, \sqrt{n}\Big(1 - \frac{\sigma_n K_n}{n} - \overline{S}_n 1\Big) + \sqrt{n}\Big(\overline{S}_n f - \mathbb{P}_n f + \frac{\sigma_n K_n}{n}\tilde{\mathbb{P}}_n f\Big) \\ = -S_n f \sqrt{n}(\overline{S}_n 1 - T_n 1) + \sqrt{n}(\overline{S}_n f - T_n f), \end{aligned}$$

where $T_n f = \mathbb{P}_n f - (\sigma_n K_n/n)\tilde{\mathbb{P}}_n f$ tends to Tf, by Lemma 3.4.2. We shall show that $\sqrt{n}(\overline{S}_n - T_n) |\, X_1, \ldots, X_n \rightsquigarrow \mathbb{W}$. Then $S_n f \to Tf/T1 = Tf/(1 - \sigma\lambda)$, and the result follows in the case that $\sigma\lambda < 1$.

The variables $\xi_{n,1}, \ldots, \xi_{n,n}$ are independent. The K_n variables corresponding to the distinct values are $\Gamma(1-\sigma, 1)$-distributed; the others are $\Gamma(1,1)$-distributed. Thus the conditional mean and variance of $\overline{S}_n f$ are given by

$$\begin{aligned} \sum_{i=1}^n (\mathbb{E}\xi_{n,i}) f(X_i) &= \sum_{i=1}^n f(X_i) - \sigma \sum_{i=1}^{K_n} f(\tilde{X}_i) = T_n f, \\ \frac{1}{n}\sum_{i=1}^n (\operatorname{var} \xi_{n,i}) f^2(X_i) &= \frac{1}{n}\sum_{i=1}^n f^2(X_i) - \frac{\sigma}{n}\sum_{i=1}^{K_n} f^2(\tilde{X}_i) \to Tf^2, \qquad \text{a.s.,} \end{aligned}$$

by Lemma 3.4.2. The limit variance is equal to $\operatorname{var} \mathbb{W} f$. To complete the proof of the convergence $\sqrt{n}(\overline{S}_n - T_n) f |\, X_1, \ldots, X_n \rightsquigarrow \mathbb{W}$, it suffices to verify the Lindeberg-Feller condition. We have, for $\xi_n \sim \Gamma(1 - \sigma_n, 1)$ and $\bar{\xi}_n \sim \Gamma(1,1)$,

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\xi_{n,i}^2 f^2(X_i) 1_{|\xi_{n,i} f(X_i)| > \varepsilon\sqrt{n}} |\, X_1, \ldots, X_n\right) \\ &\qquad \leq \max\Big(\mathbb{E}\xi_n^2 1_{|\xi_n| \max_{1\leq i\leq n} |f(X_i)| > \varepsilon\sqrt{n}}, \mathbb{E}\bar{\xi}_n^2 1_{|\bar{\xi}_n| \max_{1\leq i \leq n} |f(X_i)| > \varepsilon\sqrt{n}}\Big)\, \mathbb{P}_n f^2. \end{aligned}$$

This tends to zero for every sequence $X_1, X_2, \ldots$ such that both $\mathbb{P}_n f^2 = O(1)$ and $\max_{1\leq i \leq n} |f(X_i)|/\sqrt{n} \to 0$, which is almost every sequence if $P_0 f^2 < \infty$.

By the Cramér-Wold device and linearity in f, the convergence is then implied for finite sets of f.

For convergence as processes in $\ell^\infty(\mathcal{F})$ for a general Donsker class, it suffices to prove asymptotic tightness (see e.g. Theorem 1.5.4 in [78]). The processes $n^{-1/2}\sum_{i=1}^n (\xi_{n,i} - \mathbb{E}\xi_{n,i}) f(X_i)$ are multiplier processes with mean zero, independent multipliers. Because the multipliers are not i.i.d., a direct application of the conditional multiplier central limit theorem (see Theorem 2.9.7 in [78]) is not possible. However, the multipliers have two forms $\Gamma(1-\sigma,1)$ and $\Gamma(1,1)$. By Jensen's inequality, for any collection $\mathcal{G}$ of functions,

$$\mathbb{E}_\xi \Big\|\sum_{i=1}^n(\xi_{n,i}-\mathbb{E}\xi_{n,i})f(X_i)\Big\|_{\mathcal{G}}^* \le \mathbb{E}_{\xi,\xi'} \Big\|\sum_{i=1}^n \big(\xi_{n,i}-\mathbb{E}\xi_{n,i}+\xi'_{n,i}-\mathbb{E}\xi'_{n,i}\big)\,f(X_i)\Big\|_{\mathcal{G}}^*,$$

for any random variables $\xi'_{n,i}$ independent of the $\xi_{n,i}$. We can choose these variables so that all $\xi_{n,i}+\xi_{n,i}\overset{\text{iid}}{\sim}\Gamma(1,1)$. The process in the right side then does have i.i.d. multipliers, and the asymptotic tightness follows from the i.i.d. case (as in [78]), Theorems 3.6.13, 2.9.6 and 2.9.7; also see Corollary 2.9.9; we apply the preceding inequality with $\mathcal{G}$ equal to the set of differences $f-g$ of functions $f,g\in\mathcal{F}$ with $L_2(P_0)$-norm of $f-P_0f-g+P_0g$ smaller than δ).

Finally if $\sigma\lambda=1$, then both $\sigma=1$ and $\lambda=1$. The second implies that $P_0=P_0^c$, $K_n=n$ and $\tilde{\mathbb{P}}_n=\mathbb{P}_n$. Thus in this case $S_n(1-\sigma_nK_n/n)=\sum_{i=1}^n W_{n,i}f(X_i)(1-\sigma_n)$, for $(W_{n,1},\ldots,W_{n,n})\sim\text{Dir}(n,1-\sigma_n,\ldots,1-\sigma_n)$, and $T_nf=\mathbb{P}_nf(1-\sigma_n)$.. We can now compute

$$\begin{aligned}\mathbb{E}\Big(\sum_{i=1}^n W_{n,i}f(X_i)|\,X_1,\ldots,X_n\Big)&=\mathbb{P}_nf,\\ \text{var}\Big(\sum_{i=1}^n W_{n,i}f(X_i)|\,X_1,\ldots,X_n\Big)&=\sum_{i=1}^n\sum_{j=1}^n\text{cov}(W_{n,i},W_{n,j})f(X_i)f(X_j)\\ &\le\sum_{i=1}^n\frac{(n-1)f^2(X_i)}{n^2(n(1-\sigma_n)+1)}\le\frac{\mathbb{P}_nf^2}{n(1-\sigma_n)},\end{aligned}$$

as the covariances between the $W_{n,i}$ are negative. This implies that the conditional mean and variance of $\sqrt{n}(S_n(1-\sigma_n)-T_nf)$ tend to zero, as $\sigma_n\to1$. ∎

We are ready to complete the proof of Theorem 3.2.1. If $K_n\to\infty$, then Lemmas 3.4.5–3.4.4 together with the convergence $K_n/n\to\lambda$ immediately give the convergence of the second and third terms in the decomposition (3.8). Furthermore, these lemmas give that $S_n-Q_n\to Tf/(1-\sigma\lambda)$, which combined with Lemma 3.4.3 gives the convergence of the first term in (3.8).

If K_n remains bounded, then Lemma 3.4.4 does not apply. However, since the process Q_n will run through finitely many different Pitman-Yor processes, we have $Q_n-G=O_P(1)$ and hence the third term in (3.8) is $O_P(1/\sqrt{n})$, still under the assumption that K_n is bounded. Lemma 3.4.5 is still valid, and hence the second term in (3.8) converges to a Gaussian process as before. We can divide this term by $1-\sigma K_n/n\to1$, to see that $S_n\to T$, in view of Lemma 3.4.2. The sequence K_n can remain bounded only if $\lambda=0$ and then the normal limit in Lemma 3.4.3 is degenerate, whence $\sqrt{n}(R_n-1)=-\sigma K_n/\sqrt{n}+o_P(1)=o_P(1)$, almost surely, again under the assumption that K_n is bounded. Combined this shows that the first term in (3.8) tends to zero.

3.4.2 Proof of Theorem 3.2.6

Make the dependence on σ of the Pitman-Yor posterior process and its limit explicit by writing $\mathrm{PY}_n(\sigma)$ and $\mathbb{G}(\sigma)$ for the process PY_n in (3.6) and the right side in Theorem 3.2.1, and set

$$\mathrm{CPY}_n(\sigma) = \sqrt{n}\big(\mathrm{PY}_n(\sigma) - \mathbb{P}_n - \frac{\sigma K_n}{n}(G - \tilde{\mathbb{P}}_n)\big).$$

Lemmas 3.4.3–3.4.5 give

$$\sup_{\sigma\in(0,1)}\sup_{h\in\mathrm{BL}_1}\Big|\mathbb{E}\Big(h\big(\mathrm{CPY}_n(\sigma)\big)|\,X_1,\ldots,X_n\Big) - \mathbb{E}h\big(\mathbb{G}(\sigma)\big)\Big| \to 0, \tag{3.13}$$

in probability. This immediately gives that for every data-dependent $\hat{\sigma}_n$ that take their values in the interval $(0,1)$,

$$\sup_{h\in\mathrm{BL}_1}\Big|\mathbb{E}\Big(h\big(\mathrm{CPY}_n(\hat{\sigma}_n)\big)|\,X_1,\ldots,X_n\Big) - \mathbb{E}h\big(\mathbb{G}(\hat{\sigma}_n)\big)\Big| \to 0,$$

in probability, where the second expectation is on the limit process $\mathbb{G}(\hat{\sigma}_n)$ for given, fixed $\hat{\sigma}_n$. The continuity of the limit process in σ shows that, for $\hat{\sigma}_n \to \sigma_0$ in probability,

$$\sup_{h\in\mathrm{BL}_1}\Big|\mathbb{E}h\big(\mathbb{G}(\hat{\sigma}_n)\big) - \mathbb{E}h\big(\mathbb{G}(\sigma_0)\big)\Big| \to 0,$$

in probability. Combined the two preceding displays give the first assertion of Theorem 3.2.6.

For discrete P_0 with regularly varying atoms, the convergence of the maximum likelihood estimator $\hat{\sigma}_n$ to its coefficient of regular variation $\sigma_0 \in (0,1)$ is shown in Theorem 3.4.7, and hence the preceding argument applies.

In a hierarchical Bayesian setup with a prior on σ and given σ the Pitman-Yor prior on P, the posterior distribution of P can be decomposed as

$$\begin{aligned}&\mathbb{E}\big(h\big(\mathrm{CPY}_n(\sigma)|\,X_1,\ldots,X_n\big)\\ &\qquad= \int \mathbb{E}\Big(h\big(\mathrm{CPY}_n(\sigma)\big)|\,\sigma, X_1,\ldots,X_n\Big)\,\Pi_n(d\sigma|\,X_1,\ldots,X_n),\end{aligned}$$

where $\Pi_n(d\sigma|\,X_1,\ldots,X_n)$ refers to the posterior distribution of σ given the observations $X_1,\ldots,X_n$, and $\mathrm{CPY}_n(\sigma)|\,\sigma, X_1,\ldots,X_n$ is the standardised Pitman-Yor posterior distribution for given σ, considered in Theorem 3.2.1. The uniformity (3.13) shows that the expectation in the integral on the right side can be replaced asymptotically by $\mathbb{E}h\big(\mathbb{G}(\sigma)\big)$, uniformly in $h \in \mathrm{BL}_1$, whenever the posterior distribution of σ concentrates with probability tending to one on the interval $(0,1)$. In particular, this is true if the posterior distribution of σ is consistent for some value $\sigma_0 \in (0,1)$, i.e. if it concentrates asymptotically within the interval $(\sigma_0 - \varepsilon, \sigma_0 + \varepsilon)$, for every $\varepsilon > 0$. This consistency is shown in the proposition below. Given posterior consistency, by the continuity of the limit process in σ, the expectation $\mathbb{E}h\bigl(\mathbb{G}(\sigma)\bigr)$ can in turn in the limit be replaced by $\mathbb{E}h\bigl(\mathbb{G}(\sigma_0)\bigr)$, uniformly in $h \in \mathrm{BL}_1$. This gives the second assertion of Theorem 3.2.6.

3.4.3 Estimating the type parameter

A measurable function $\alpha\colon [1,\infty) \to \mathbb{R}_+$ is said to be *regularly varying* (at ∞) of order γ if, for all $u > 0$, as $n \to \infty$,

$$\frac{\alpha(nu)}{\alpha(n)} \to u^\gamma. \tag{3.14}$$

It is known (see e.g. [6] or the appendix to [19]) that if the limit of the sequence of quotients on the left exists for every u, then it necessarily has the form u^γ, for some γ, as in (3.14). If we write $\alpha(u) = u^\gamma L(u)$, then L will be *slowly varying*: a function that is regularly varying of order 0. Then $\alpha(n) = n^\gamma L(n)$, and it can be shown that $n^{\gamma-\delta} \ll \alpha(n) \ll \alpha^{n+\delta}$, for every $\delta > 0$, so that the rate of growth of α is n^γ to "first order". (See Potter's theorem, [6], Theorem 1.5.6, or [19], Proposition B.1.9-5).

Example 3.4.6. *For the probability distribution $(p_j)_{j\in\mathbb{N}}$ with $p_j = C/j^\alpha$, for some $\alpha > 1$, the function $\alpha(u)\colon= \#(j\colon 1/p_j \leq u) = \lfloor (Cu)^{1/\alpha} \rfloor$ is regularly varying of order $\gamma = 1/\alpha$.*

We consider the empirical Bayes estimator $\hat\sigma_n$, the maximum likelihood estimator in the model $P|\,\sigma \sim \mathrm{PY}\,(\sigma, M, G)$ and $X_1,\ldots,X_n|\,P,\sigma \sim P$ given observations $X_1,\ldots,X_n$. We also consider the posterior distribution of σ given $X_1,\ldots,X_n$ in the model $\sigma \sim \Pi_\sigma$, $P|\,\sigma \sim \mathrm{PY}\,(\sigma, M, G)$ and $X_1,\ldots,X_n|\,P,\sigma \sim P$, for a given prior distribution Π_σ on $(0,1)$. In both cases the likelihood for observing $X_1,\ldots,X_n$ is proportional to (3.5). Hence $\hat\sigma_n$ is the point of maximum of this function and, by Bayes theorem, the posterior distribution has density relative to Π_σ proportional to (3.5).

In the following theorem we consider these objects under the assumption that $X_1,\ldots,X_n$ are an i.i.d. sample from a distribution P_0. Consistency of $\hat\sigma_n$ for σ_0 means that $\hat\sigma_n \to \sigma_0$ in probability. Contraction of the posterior distribution to σ_0 means that $\Pi_n(\sigma\colon |\sigma-\sigma_0| > \varepsilon|\,X_1,\ldots,X_n)$ tends to zero in probability, for every $\varepsilon > 0$.

Theorem 3.4.7. *If P_0 is discrete with atoms such that $\alpha_0(u)\colon= \#\{x\colon 1/P_0\{x\} \leq u\}$ is regularly varying of exponent $\sigma_0 \in (0,1)$, then the empirical Bayes estimator $\hat\sigma_n$ is consistent for σ_0. Furthermore, for a prior distribution Π_σ on σ with a density that is bounded away from zero and infinity, the posterior distribution of σ contracts to σ_0.*

Proof. Up to an additive constant the log likelihood can be written

$$\begin{aligned}\Lambda_n(\sigma) &= \sum_{l=1}^{K_n-1}\log(M+l\sigma)+\sum_{j=1:N_{n,j}\geq 2}^{K_n}\sum_{l=0}^{N_{n,j}-2}\log(1-\sigma+l)\\ &=\sum_{l=1}^{K_n-1}\log(M+l\sigma)+\sum_{l=1}^{n-1}\log(l-\sigma)Z_{n,l+1},\end{aligned}$$

where $Z_{n,l}=\#(1\leq j\leq K_n\colon N_{n,j}\geq l)$ is the number of distinct values of multiplicity at least l in the sample $X_1,\ldots,X_n$. (In the case that all observations are distinct and hence $N_{n,j}=1$ for every j, the second term of the likelihood is equal to 0.) The concavity of the logarithm shows that the log likelihood is a strictly concave function of σ. For $\sigma\downarrow 0$, it tends to a finite value, while for $\sigma\uparrow 1$ it tends to $-\infty$ if the term with $l=1$ is present in the second sum, i.e. if there is at least one tied observation. This happens with probability tending to 1 as $n\to\infty$. The derivative of the log likelihood is equal to

$$\Lambda_n'(\sigma)=\sum_{l=1}^{K_n-1}\frac{l}{M+l\sigma}-\sum_{l=1}^{n-1}\frac{1}{l-\sigma}Z_{n,l+1}. \tag{3.15}$$

The left limit at $\sigma=0$ is $\Lambda_n'(0)=\frac{1}{2}K_n(K_n-1)-\sum_{l=1}^{n-1}l^{-1}Z_{n,l+1}$. Since $Z_{n,l}\leq Z_{n,1}=K_n$, a crude bound on the sum is $K_n\log n$, which shows that the derivative at $\sigma=0$ tends to infinity if $K_n\gg\log n$. In that case the unique maximum of the log likelihood in $[0,1]$ is taken in the interior of the interval, and hence $\hat\sigma_n$ satisfies $\Lambda_n'(\hat\sigma_n)=0$.

Under the condition that α_0 is regularly varying of exponent $\sigma_0\in(0,1)$, the sequence $\alpha_n\colon=\alpha_0(n)$ is of the order n^{σ_0} up to slowly varying terms. By Theorems 9 and 1‘ of [42], the sequence K_n/α_n tends almost surely to $\Gamma(1-\sigma_0)$ and hence in particular $K_n\gg\log n$.

We show below that $\Lambda_n'(\sigma)/\alpha_n\to\lambda(\sigma)$ in probability, for every σ, and a strictly decreasing function λ with $\lambda(\sigma_0)=0$. It follows that $\Lambda_n'(\sigma_0-\varepsilon)>0$ and $\Lambda_n'(\sigma_0+\varepsilon)<0$ with probability tending to one, for every fixed $\varepsilon>0$. Then $\sigma_0-\varepsilon<\hat\sigma_n<\sigma_0+\varepsilon$ with probability tending to one, by the monotonicity of $\sigma\mapsto\Lambda_n'(\sigma)$, and hence the consistency of $\hat\sigma_n$ follows.

The monotonicity of Λ_n' and the fact that $\Lambda_n'(\hat\sigma_n)=0$, give that on the event $\sigma_0+\varepsilon>\hat\sigma_n$,

$$\begin{aligned}&\Lambda_n(\sigma)\geq\Lambda_n(\sigma_0+\varepsilon), && \text{if } \hat\sigma_n<\sigma<\sigma_0+\varepsilon,\\ &\Lambda_n(\sigma)\leq\Lambda_n(\sigma_0+\varepsilon)+\Lambda_n'(\sigma_0+\varepsilon)(\sigma-\sigma_0-\varepsilon), && \text{if } \sigma>\sigma_0+\varepsilon.\end{aligned}$$

It follows that on the event $\sigma_0 + \varepsilon > \hat{\sigma}_n$,

$$\begin{aligned}\Pi_n\big(\sigma > \sigma_0 + \varepsilon |\, X_1, \ldots, X_n\big) &= \frac{\int_{\sigma_0+\varepsilon}^1 e^{\Lambda_n(\sigma)}\, d\Pi_\sigma(\sigma)}{\int_0^1 e^{\Lambda_n(\sigma)}\, d\Pi_\sigma(\sigma)} \\ &\le \frac{\int_{\sigma_0+\varepsilon}^1 e^{\Lambda_n(\sigma_0+\varepsilon)+\Lambda_n'(\sigma_0+\varepsilon)(\sigma-\sigma_0-\varepsilon)}\, d\Pi_\sigma(\sigma)}{\int_{\hat{\sigma}_n}^{\sigma_0+\varepsilon} e^{\Lambda_n(\sigma_0+\varepsilon)}\, d\Pi_\sigma(\sigma)} \\ &\lesssim \frac{\int_0^\infty e^{\Lambda_n'(\sigma_0+\varepsilon)u}\, du}{\sigma_0+\varepsilon-\hat{\sigma}_n} = \frac{1}{-\Lambda_n'(\sigma_0+\varepsilon)(\sigma_0+\varepsilon-\hat{\sigma}_n)},\end{aligned}$$

where the proportionality constant depends on the density of Π_σ only. Since $-\Lambda_n'(\sigma_0 + \varepsilon)/\alpha_n \to -\lambda(\sigma_0+\varepsilon) > 0$ and $\sigma_0 + \varepsilon - \hat{\sigma}_n \to \varepsilon$ in probability, the right side tends to zero in probability. Combined with a similar argument on the left tail of the posterior distribution, this shows that the posterior distribution contracts to σ_0.

It remains to be shown that $\Lambda_n'(\sigma)/\alpha_n \to \lambda(\sigma)$, in probability for a strictly decreasing function λ with a unique zero at σ_0. The variables $Z_{n,l}$ can be written as $Z_{n,l} = \sum_{j=1}^\infty 1_{M_{n,j}\ge l}$, for $M_{n,j}$ the number of observations equal to x_j. As $K_n = Z_{n,1}$, the function Λ_n' can be written in the form

$$\Lambda_n'(\sigma) = \sum_{l=1}^{K_n-1} \frac{l}{M+l\sigma} - \sum_{l=1}^\infty \sum_{j=1}^\infty \frac{1_{M_{n,j}\ge l+1}}{l-\sigma} = \sum_{j=1}^\infty \Big[\frac{1_{M_{n,j}\ge 1}}{\sigma} - g_\sigma(M_{n,j})\Big] - \frac{h_\sigma(K_n)}{\sigma},$$

where $g_\sigma(0) = g_\sigma(1) = 0$ and $g_\sigma(m) = \sum_{l=1}^{m-1} \frac{1}{l-\sigma}$, for $m \ge 2$, and $h_\sigma(k) = 1 + \sum_{l=1}^{k-1} M/(M+l\sigma) \le 1+(M/\sigma)\log(1+k\sigma/M)$. It is shown in [42] (and repeated below) that $\mathrm{E}K_n/\alpha_n \to \Gamma(1-\sigma_0)$ and hence $\mathrm{E}h_\sigma(K_n) \le 1 + (M/\sigma)\log(1 + \mathrm{E}K_n\sigma/M) = O(\log n) = o(\alpha_n)$, so that the term on the far right is asymptotically negligible.

It is shown in Lemma 3.4.8 that

$$\mathbb{E}\frac{1}{\alpha_n}\sum_{j=1}^\infty \Big[\frac{1_{M_{n,j}\ge 1}}{\sigma} - g_\sigma(M_{n,j})\Big] \to \frac{\Gamma(1-\sigma_0)}{\sigma} - \sum_{m=1}^\infty \frac{\Gamma(m+1-\sigma_0)}{m!(m-\sigma)} =: \lambda(\sigma).$$

The limit function λ is strictly decreasing. The value of the series at $\sigma = \sigma_0$ is equal to

$$\sum_{m=1}^\infty \frac{\Gamma(m-\sigma_0)}{m!} = \int_0^\infty (e^x - 1)x^{-\sigma_0-1}e^{-x}\, dx = \int_0^\infty (1-e^{-x})x^{-\sigma_0-1}\, dx.$$

By partial integration, this can be further rewritten as $\int_0^\infty x^{-\sigma_0}/\sigma_0\, e^{-x}\, dx = \Gamma(1 - \sigma_0)/\sigma_0$. We conclude that $\lambda(\sigma_0) = 0$.

To complete the proof it suffices to show that the variance of the variables in the left side of the second last display tend to zero. For $i \ne j$, the conditional distribution of $M_{n,i}$ given $M_{n,j}=m$ is binomial with parameters $(n-m,p_i)$, which is stochastically smaller than the marginal binomial (n,p_i) distribution of $M_{n,i}$. It follows that $\mathrm{E}(h(M_{n,i})|\, M_{n,j})\le \mathrm{E}h(M_{n,i})$, for every nondecreasing function h, whence $h(M_{n,i})$ and $h(M_{n,j})$ are negatively correlated for every nonnegative, nondecreasing function h. Applying this with $h(m)=1_{m\ge 1}$ and $h=g_\sigma$, we find that

$$\operatorname{var}\frac{1}{\alpha_n}\sum_{j=1}^\infty 1_{M_{n,j}\ge 1}\le \frac{1}{\alpha_n^2}\sum_{j=1}^\infty \operatorname{var}1_{M_{n,j}\ge 1}\le \frac{1}{\alpha_n^2}\sum_{j=1}^\infty \mathrm{E}1_{M_{n,j}\ge 1},$$
$$\operatorname{var}\frac{1}{\alpha_n}\sum_{j=1}^\infty g_\sigma(M_{n,j})\le \frac{1}{\alpha_n^2}\sum_{j=1}^\infty \operatorname{var}g_\sigma(M_{n,j})\le \frac{1}{\alpha_n^2}\sum_{j=1}^\infty \mathrm{E}g_\sigma^2(M_{n,j}).$$

By Lemma 3.4.8, both right sides are of the order $O(1/\alpha_n)$. This concludes the proof that $\Lambda_n'(\sigma)/\alpha_n\to\lambda(\sigma)$, in probability. ∎

Lemma 3.4.8. *Suppose that* $\alpha(u){:}=\#\{j{:}\,1/p_j\le u\}$ *is regularly varying at* ∞ *of order* $\gamma\in(0,1)$. *For any* $\sigma\in(0,1)$, *and* $g_\sigma(m)=\sum_{l=1}^{m-1}\frac{1}{l-\sigma}$, *for* $m\ge 2$, *and* $M_{n,j}\sim$ *Binomial*(n,p_j),

(i) $\frac{1}{\alpha(n)}\sum_{j=1}^\infty \mathrm{E}1_{M_{n,j}\ge 1}\to\Gamma(1-\gamma)$,

(ii) $\frac{1}{\alpha(n)}\sum_{j=1}^\infty \mathrm{E}g_\sigma(M_{n,j})\to\sum_{m=1}^\infty\frac{\Gamma(m+1-\gamma)}{m!(m-\sigma)}$,

(iii) $\frac{1}{\alpha(n)}\sum_{j=1}^\infty \mathrm{E}g_\sigma^2(M_{n,j})\to\sum_{k=1}^\infty\sum_{l=1}^\infty\frac{\Gamma(k\vee l+1-\gamma)}{(k-\sigma)(l-\sigma)(k\vee l)!}$.

Proof. Because $\mathrm{P}(M_{n,j}=0)=(1-p_j)^n$, the series in the left side of (i) is equal to

$$\sum_{j=1}^\infty\Bigl(1-(1-p_j)^n\Bigr)=\int_1^\infty\Bigl(1-\Bigl(1-\frac{1}{u}\Bigr)^n\Bigr)\,d\alpha(u)=n\int_0^1\alpha\Bigl(\frac{1}{s}\Bigr)(1-s)^{n-1}\,ds,$$

by Fubini's theorem, since $1-(1-1/u)^n=\int_0^{1/u}n(1-s)^{n-1}\,ds$. It follows that the left side of (i) can be written

$$\int_0^n\frac{\alpha(n/s)}{\alpha(n)}\Bigl(1-\frac{s}{n}\Bigr)^{n-1}\,ds.$$

By regular variation of α, the integrand tends pointwise to $s^{-\gamma}e^{-s}$, as $n\to\infty$. By Potter's theorem, the quotient $\alpha(n/s)/\alpha(n)$ is bounded above by a multiple of $(1/s)^{\gamma-\delta}\vee(1/s)^{\gamma+\delta}$, for any given $\delta>0$, while $(1-s/n)^{n-1}\le e^{-s(1-\delta)}$, by the inequality $1-x\le e^{-x}$, for $x\in\mathbb{R}$. Therefore, by the dominated convergence theorem the integral converges to $\int_0^\infty s^{-\gamma}e^{-s}\,ds=\Gamma(1-\gamma)$.

The series in the left side of (ii) is equal to

$$\sum_{j=1}^\infty\sum_{m=2}^n g_\sigma(m)\binom{n}{m}p_j^m(1-p_j)^{n-m}=\sum_{m=2}^n g_\sigma(m)\binom{n}{m}\int_1^\infty\Bigl(\frac{1}{u}\Bigr)^m\Bigl(1-\frac{1}{u}\Bigr)^{n-m}\,d\alpha(u).$$

Writing $(1/u)^m(1-1/u)^{n-m}=\int_0^{1/u}s^{m-1}(1-s)^{n-m-1}(m-ns)\,ds$ (for $m\geq 1$) and applying Fubini's theorem, we can rewrite this as

$$\begin{aligned}\sum_{m=2}^{n}g_\sigma(m)\binom{n}{m}\int_0^1\alpha\Big(\frac{1}{s}\Big)s^{m-1}(1-s)^{n-m-1}(m-ns)\,ds&\\ =\int_0^1\sum_{l=1}^{n-1}\frac{1}{l-\sigma}\sum_{m=l+1}^{n}\binom{n}{m}s^{m-1}(1-s)^{n-m-1}(m-ns)\,\alpha\Big(\frac{1}{s}\Big)\,ds&\\ =\int_0^1\sum_{l=1}^{n-1}\frac{n-l}{l-\sigma}\binom{n}{l}s^l(1-s)^{n-l-1}\,\alpha\Big(\frac{1}{s}\Big)\,ds=\sum_{l=1}^{n-1}\frac{1}{l-\sigma}\mathrm{E}\alpha\Big(\frac{1}{S_{l,n}}\Big),&\end{aligned}$$

for $S_{l,n}\sim\mathrm{Beta}(l+1,n-l)$, where the second last equality follows from Lemma 3.4.9. Representing $S_{l,n}$ as $\Gamma_l/(\Gamma_l+\Gamma_{n-l})$, for independent variables $\Gamma_l\sim\Gamma(l+1,1)$ and $\Gamma_{n-l}\sim\Gamma(n-l,1)$, we see that the left side of (ii) is equal to

$$\sum_{l=1}^{n-1}\frac{1}{l-\sigma}\mathbb{E}\frac{\alpha\big(1+\Gamma_{n-l}/\Gamma_l\big)}{\alpha(n)}=\sum_{l=1}^{n-1}\frac{1}{l-\sigma}\mathbb{E}\frac{\alpha\big((n^{-1}+n^{-1}\Gamma_{n-l}/\Gamma_l)n\big)}{\alpha(n)}.$$

The sequence $U_{l,n}\colon=(n^{-1}+n^{-1}\Gamma_{n-l}/\Gamma_l)$ tends almost surely to $1/\Gamma_l$, by the law of large numbers, as $n\to\infty$, for fixed l. Since the convergence in (3.14) is automatically uniform in compacta contained in $(0,\infty)$ (see [19], Theorem B.1.4), it follows that $\alpha(U_{l,n}n)/\alpha(n)\to(1/\Gamma_l)^\gamma$, almost surely. Furthermore, by Potter's theorem $\alpha(U_{l,n}n)/\alpha(n)\lesssim U_{l,n}^{\gamma+\delta}\vee U_{l,n}^{\gamma-\delta}$, where $U_{l,n}^\beta\leq 1+(n^{-1}\Gamma_{n-l})^\beta(1/\Gamma_l)^\beta$ is uniformly integrable for every $\beta<1$, since $n^{-1}\Gamma_{n-l}\to 1$ in L_1 and $\mathbb{E}(1/\Gamma_l)^\beta<\infty$, so that $n^{-1}\Gamma_{n-l}/\Gamma_l\to 1/\Gamma_l$ in L_1, in view of the independence of Γ_{n-l} and Γ_l. By dominated convergence we conclude that $\mathbb{E}\alpha(U_{l,n}n)/\alpha(n)\to\mathrm{E}(1/\Gamma_l)^\gamma=\Gamma(l+1-\gamma)/l!$. Since $\mathrm{E}U_{l,n}^{\gamma+\delta}\vee U_{l,n}^{\gamma-\delta}\lesssim\mathrm{E}(1/\Gamma_l)^{-\gamma+\delta}\lesssim l^{-\gamma+\delta}$, a second application of the dominated convergence theorem shows that the preceding display tends to $\sum_{l=1}^\infty(l-\sigma)^{-1}\Gamma(l+1-\gamma)/l!$.

For the proof of (iii) we write $g_\sigma^2(m)=\sum_{k=1}^{m-1}\sum_{l=1}^{m-1}(k-\sigma)^{-1}(l-\sigma)^{-1}$ and follow the same steps as in (ii) to write the left side of (iii) as

$$\sum_{k=1}^{n-1}\sum_{l=1}^{n-1}\frac{1}{k-\sigma}\frac{1}{l-\sigma}\mathbb{E}\frac{\alpha(1/S_{k\vee l,n})}{\alpha(n)}.$$

This is seen to converge to the limit as claimed by the same arguments as under (ii). ■

Lemma 3.4.9. *For every $p\in[0,1]$ and $l\in\mathbb{N}\cup\{0\}$ and $n\in\mathbb{N}$,*

$$\sum_{m=l+1}^{n}\binom{n}{m}p^{m-1}(1-p)^{n-m-1}(m-np)=(n-l)\binom{n}{l}p^l(1-p)^{n-l-1}.$$

Proof. For X_{n-1} and X_n the numbers of successes in the first $n-1$ and n independent Bernoulli trials with success probability p, we have $\{X_n \geq l+1\} \subset \{X_{n-1} \geq l\}$ and $\{X_{n-1} \geq l\} - \{X_n \geq l+1\} = \{X_{n-1} = l, B_n = 0\}$, for B_n the outcome of the nth trial. This gives the identity $\mathrm{P}(X_{n-1} \geq l) - \mathrm{P}(X_n \geq l+1) = \mathrm{P}(X_{n-1} = l)(1-p)$. We multiply this by $n/(1-p)$ to obtain the identity given by the lemma, which we first rewrite using that $m\binom{n}{m} = n\binom{n-1}{m-1}$ and $(n-l)\binom{n}{l} = n\binom{n-1}{l}$. ■

Finally consider the situation that P_0 possesses a nontrivial continuous component. In this case the empirical Bayes estimator tends to 1.

Theorem 3.4.10. *If $P_0 = (1-\lambda)P_0^d + \lambda P_0^c$ where P_0^d is a discrete and P_0^c an atomless probability distribution with $\lambda > 0$ and such that $\alpha_0(u) := \#\{x: 1/P_0\{x\} \leq u\}$ is regularly varying of exponent $\sigma_0 \in (0,1)$, then $\hat{\sigma}_n \to 1$ in probability.*

Proof. By Lemma 3.4.1 the sequence K_n/n tends to λ in probability. The second term in the derivative of the log likelihood (3.15) depends on tied observations only (through the variables $Z_{n,l}$ with $l \geq 2$), and the arguments from the proof of Theorem 3.4.7 show that this term retains the order $O_P(\alpha_0(n))$. Thus it follows that $\Lambda_n'(\sigma)/n \to \lambda/\sigma$ in probability, whence it is positive with probability tending to one and the likelihood increasing in σ. ■

3.5 Acknowledgements

The authors would like to thank Botond Szabó for his extensive feedback. This work has been presented several times and the ensuing discussions and remarks by the referees helped identify where the exposition could be improved.

3.6 Mean and variance of posterior distribution

In this appendix we derive explicit formulas for the mean and variance of the posterior distribution. The limit of the variances can be seen to be equal to variance of the limit variable in Theorem 3.2.1.

Lemma 3.6.1. *Let $P \sim \mathrm{PY}(\sigma, M, G)$ where $\sigma \geq 0$. Then the mean and variance of the posterior distribution of P based on observations $X_1, \ldots, X_n | P \overset{iid}{\sim} P$ are as* *follows*

$$\begin{aligned}
\mathbb{E}[Pf|X_1,\ldots,X_n] &= \sum_{j=1}^{K_n} \frac{N_{j,n}-\sigma}{n+M} f(\tilde{X}_j) + \frac{M+\sigma K_n}{n+M} Gf, \\
\operatorname{var}(Pf|X_1,\ldots,X_n) &= \Big[\sum_{j=1}^{K_n} \frac{N_{j,n}-\sigma}{n-K_n\sigma} f(\tilde{X}_j) - Gf\Big]^2 \frac{(n-\sigma K_n)(M+\sigma K_n)}{(n+M)^2(n+M+1)} \\
&\quad - \frac{\Big(\sum_{j=1}^{K_n}(N_{j,n}-\sigma) f(\tilde{X}_j)\Big)^2}{(n-\sigma K_n)(n+M)(n+M+1)} + \frac{\sum_{j=1}^{K_n}(N_{j,n}-\sigma) f(\tilde{X}_j)^2}{(n+M)(n+M+1)} \\
&\quad + \frac{(1-\sigma)(M+\sigma K_n+1)}{(n+M)(n+M+1)} Var_G(f).
\end{aligned}$$

Lemma 3.6.2. *Suppose* $X_1,\ldots,X_n \overset{iid}{\sim} P_0$, *where* $P_0 = (1-\lambda)P_0^d + \lambda P_0^c$. *If* P *follows a* PY (σ, M, G) *process, then the posterior distribution in the model* $X_1,\ldots,X_n|P \sim P$, P_0 *almost surely*

$$\begin{aligned}
\mathbb{E}[Pf|X_1,\ldots,X_n] &\to (1-\lambda)P_0^d + (1-\sigma)\lambda P_0^c + \lambda\sigma G \\
n\operatorname{var}(Pf|X_1,\ldots,X_n) &\to (1-\lambda)\, Var_{P_0^d}(f) + (1-\sigma)\lambda\, Var_{P_0^c}(f) \\
&\quad + (1-\sigma)\sigma\lambda\, Var_G(f) \\
&\quad + \frac{(1-\sigma)\lambda(1-\lambda)}{1-\sigma\lambda} \left(P_0^d(f) - P_0^c(f)\right)^2 \\
&\quad + (1-\sigma\lambda)\sigma\lambda \left(\frac{(1-\lambda)P_0^d(f) + (1-\sigma)\lambda P_0^c(f)}{1-\sigma\lambda} - Gf\right)^2.
\end{aligned}$$

Proof of Lemma 3.6.1. We begin by recalling the posterior distribution from Section 3.4. Note that we have the following results:

- $\mathbb{E}[R_n] = \frac{n-K_n\sigma}{n+M}$ and $\operatorname{Var}(R_n) = \frac{(n-K_n\sigma)(M+K_n\sigma)}{(n+M)^2(n+M+1)}$.
- $\mathbb{E}[Q_n(f)] = G(f)$, $\operatorname{Var}(Q_n(f)) = \frac{1-\sigma}{M+\sigma K_n}\operatorname{Var}_G(f)$.

The first two results are standard results for Beta distributed random variables, and the last two results are because Q_n is a Pitman-Yor process. Now we just need to compute the moments for the weights W_j. We use the following results from the Dirichlet distribution. If $\tilde{X} \sim \operatorname{Dir}(K_n, \alpha_1,\ldots,\alpha_{K_n})$, then

$$\mathbb{E}[\tilde{X}_i] = \frac{\alpha_i}{\sum_{k=1}^{K_n}\alpha_k},$$

$$\operatorname{var}(\tilde{X}_i) = \frac{\alpha_i(\sum_{k=1}^{K_n}\alpha_k - \alpha_i)}{(\sum_{k=1}^{K_n}\alpha_k)^2(1+\sum_{k=1}^{K_n}\alpha_k)},$$

and

$$\mathrm{Cov}(\tilde{X}_i, \tilde{X}_j) = \frac{-\alpha_i \alpha_j}{(\sum_{k=1}^{K_n} \alpha_k)^2 (1 + \sum_{k=1}^{K_n} \alpha_k)}.$$

In our case $\alpha_i = N_{i,n} - \sigma$, $K = K_n$ and $\sum_{k=1}^{K_n} \alpha_k = n - \sigma K_n$. Then a direct computation shows that

$$\mathbb{E}[\sum_{j=1}^{K_n} W_j f(\tilde{X}_j)] = \sum_{j=1}^{K_n} \frac{N_{j,n} - \sigma}{n - K_n \sigma} f(\tilde{X}_j).$$

For the variance we use that, for independent random variables, the variance of the sum is the sum of the covariances.

$$\begin{aligned}
\mathrm{var}(\sum_{i=1}^{K_n} W_j f(\tilde{X}_j) | X_1, \dots, X_n) &= \sum_{i \neq j} \mathrm{Cov}(W_i, W_j) f(\tilde{X}_i) f(\tilde{X}_j) + \sum_{i=1}^{K_n} \mathrm{Var}(W_i) f(\tilde{X}_i)^2 \\
&= \sum_{i \neq j} \frac{-(N_{i,n} - \sigma)(N_{j,n} - \sigma)}{(n - \sigma K_n)^2 (n - \sigma K_n + 1)} f(\tilde{X}_i) f(\tilde{X}_j) \\
&\quad + \sum_{i=1}^{K_n} \frac{(N_{i,n} - \sigma)(n - \sigma K_n - N_{i,n} + \sigma)}{(n - \sigma K_n)^2 (n - \sigma K_n + 1)} f(\tilde{X}_i)^2 \\
&= -\frac{\left(\sum_{j=1}^{K_n} (N_{j,n} - \sigma) f(\tilde{X}_j)\right)^2}{(n - \sigma K_n)^2 (n - \sigma K_n + 1)} + \frac{\sum_{j=1}^{K_n} (N_{j,n} - \sigma) f(\tilde{X}_j)^2}{(n - \sigma K_n)(n - \sigma K_n + 1)}.
\end{aligned}$$

Now we can compute the mean and variance. Using independence between R_n, W and Q_n and linearity we see that

$$\mathbb{E}[P(f) | X_1, \dots, X_n] = \sum_{j=1}^{K_n} \frac{N_{j,n} - \sigma}{n + M} f(\tilde{X}_j) + \frac{M + \sigma K_n}{n + M} G(f).$$

In order to compute the variance we apply the law of total variance. For any two random variables X, Y with finite second moment we have that

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}\,(X|Y)] + \mathrm{Var}\,(\mathbb{E}[X|Y])\,.$$

We split into conditioning on R_n and the rest, so we can use the independence between W and Q_n. We compute these piece by piece. First consider

First consider

$$\mathbb{E}\left[\mathrm{Var}\left(R_n \sum_{j=1}^{K_n} W_j f(\tilde{X}_j) + (1 - R_n) Q_n(f) | R_n\right)\right].$$

Due to the independence of W and Q_n given R_n

$$= \mathbb{E}\left[R_n^2 \mathrm{Var}\left(\sum_{j=1}^{K_n} W_j f(\tilde{X}_j)\right) + (1-R_n)^2 \mathrm{Var}\left(Q_n(f)\right)\right].$$

Simplifying the expression yields

$$= \mathbb{E}[R_n^2]\mathrm{Var}\left(\sum_{j=1}^{K_n} W_j f(\tilde{X}_j)\right) + \mathbb{E}[(1-R_n)^2]\mathrm{Var}\left(Q_n(f)\right)].$$

Filling in the known moments results in

$$= \frac{(n-\sigma K_n)(n+1-\sigma K_n)}{(n+M)(n+M+1)}\mathrm{Var}\left(\sum_{j=1}^{K_n} W_j f(\tilde{X}_j)\right) + \frac{(M+\sigma K_n)(M+\sigma K_n+1)}{(n+M)(n+M+1)}\mathrm{Var}\left(Q_n(f)\right).$$

Expanding the variance terms and simplifying gives

$$= -\frac{\left(\sum_{j=1}^{K_n}(N_{j,n}-\sigma)f(\tilde{X}_j)\right)^2}{(n-\sigma K_n)(n+M)(n+M+1)} + \frac{\sum_{j=1}^{K_n}(N_{j,n}-\sigma)f(\tilde{X}_j)^2}{(n+M)(n+M+1)} + \frac{(1-\sigma)(M+\sigma K_n+1)}{(n+M)(n+M+1)}\mathrm{Var}_G(f).$$

Next wel deal with

$$\mathrm{Var}\left(\mathbb{E}[R_n\sum_{j=1}^{K_n} W_j f(\tilde{X}_j) + (1-R_n)Q_n(f)|R_n]\right).$$

Computing the expected value gives

$$= \mathrm{Var}\left(R_n\sum_{j=1}^{K_n}\frac{N_{j,n}-\sigma}{n-K_n\sigma}f(\tilde{X}_j) + (1-R_n)G(f)\right).$$

Reorganising terms

$$= \mathrm{Var}\left(G(f) + R_n(\sum_{j=1}^{K_n}\frac{N_{j,n}-\sigma}{n-K_n\sigma}f(\tilde{X}_j) - G(f))\right).$$

The constant term does not contribute to the variance so can be , and then taking the square of the constant in front of R_n results in

$$= (\sum_{j=1}^{K_n}\frac{N_{j,n}-\sigma}{n-K_n\sigma}f(\tilde{X}_j) - G(f))^2\mathrm{Var}\left(R_n\right).$$

Computing the variance of R_n gives

$$= (\sum_{j=1}^{K_n} \frac{N_{j,n} - \sigma}{n - K_n\sigma} f(\tilde{X}_j) - G(f))^2 \frac{(n - \sigma K_n)(M + \sigma K_n)}{(n + M)^2(n + M + 1)}.$$

Therefore by the law of total variance we find the result. ■

Proof of Lemma 3.6.2. We begin with some basic results which we will apply in several places. We note the following two almost sure limits: $\frac{K_n}{n} \to \lambda$ P_0-almost surely and

$$\frac{\sum_{j=1}^{K_n}(N_{j,n} - \sigma)f(\tilde{X}_j)}{n} \to (1 - \lambda)P_0^d(f) + (1 - \sigma)\lambda P_0^c(f) \qquad P_0\text{-a.s.}$$

For the posterior mean we know the exact formula by Lemma 3.6.1 and therefore the following limit can be computed:

$$\begin{aligned}\mathbb{E}[P(f)|X_1, \ldots, X_n] &= \sum_{j=1}^{K_n} \frac{N_{j,n} - \sigma}{n + M} f(\tilde{X}_j) + \frac{M + \sigma K_n}{n + M} G(f) \\ &\to (1 - \lambda)P_0^d(f) + (1 - \sigma)\lambda P_0^c(f) + \lambda\sigma G(f) P_0\text{-a.s.}\end{aligned}$$

Recall from Lemma 3.6.1 the formula for the posterior variance. We analyse this term by term. They all follow directly from the remarks at the beginning of the the proof, and the limits hold P_0-almost surely.

First we find that

$$\sum_{j=1}^{K_n} \frac{N_{j,n} - \sigma}{n - K_n\sigma} f(\tilde{X}_j) \to \frac{(1 - \lambda)P_0^d(f) + (1 - \sigma)\lambda P_0^c(f)}{1 - \sigma\lambda}.$$

Secondly,

$$n\frac{(n - \sigma K_n)(M + \sigma K_n)}{(n + M)^2(n + M + 1)} \to (1 - \sigma\lambda)\sigma\lambda.$$

Next,

$$-n\frac{\left(\sum_{j=1}^{K_n}(N_{j,n} - \sigma)f(\tilde{X}_j)\right)^2}{(n - \sigma K_n)(n + M)(n + M + 1)} \to -\frac{\left((1 - \lambda)P_0^d(f) + (1 - \sigma)\lambda P_0^c(f)\right)^2}{1 - \sigma\lambda}.$$

Also,

$$n\frac{\sum_{j=1}^{K_n}(N_{j,n} - \sigma)f(\tilde{X}_j)^2}{(n + M)(n + M + 1)} \to (1 - \lambda)P_0^d(f^2) + (1 - \sigma)\lambda P_0^c(f^2).$$

And finally,

$$n\frac{(1-\sigma)(M+\sigma K_n+1)}{(n+M)(n+M+1)}\mathrm{Var}_G(f)\to(1-\sigma)\sigma\lambda\mathrm{Var}_G(f).$$

This means we now have computed the limit of the posterior variance. We will now add all the terms together, and by the continuous mapping theorem we find that,

$$\begin{aligned}
n\mathrm{Var}\left(P(f)|X_1,\ldots,X_n\right)\to{}&(1-\sigma\lambda)\sigma\lambda\\
&\left(\frac{(1-\lambda)P_0^d(f)+(1-\sigma)\lambda P_0^c(f)}{1-\sigma\lambda}-G(f)\right)^2\\
&-\frac{\left((1-\lambda)P_0^d(f)+(1-\sigma)\lambda P_0^c(f)\right)^2}{1-\sigma\lambda}\\
&+(1-\lambda)P_0^d(f^2)+(1-\sigma)\lambda P_0^c(f^2)\\
&+(1-\sigma)\sigma\lambda\mathrm{Var}_G(f)\qquad\text{a.s. }P_0.
\end{aligned}$$

Note that

$$\begin{aligned}
&-\frac{\left((1-\lambda)P_0^d(f)+(1-\sigma)\lambda P_0^c(f)\right)^2}{1-\sigma\lambda}\\
&+(1-\lambda)P_0^d(f^2)+(1-\sigma)\lambda P_0^c(f^2)\\
&=(1-\lambda)\mathrm{Var}_{P_0^d}(f)+(1-\sigma)\lambda\mathrm{Var}_{P_0^c}(f)\\
&+\frac{(1-\sigma)\lambda(1-\lambda)}{1-\sigma\lambda}\left(P_0^d(f)-P_0^c(f)\right)^2.
\end{aligned}$$

Combining everything yields the Lemma. ■

Chapter 4

Estimating species diversity

This chapter is an adaptation of a paper submitted as: S. Franssen, A. van der Vaart, "Empirical and Full Bayes estimation of the type of a Pitman-Yor process".

4.1 Introduction

The Pitman-Yor process [59, 54] is a random discrete probability distribution, which can be used as a model for the relative abundance of species. It is characterised by a *type* parameter σ. Our main aim is statistical inference on this type parameter.

The Pitman-Yor process of type $\sigma = 0$ is the Dirichlet process [24], which is well understood, while negative types correspond to finitely discrete distributions and were considered in [17]. In this paper we concentrate on Pitman-Yor processes of positive type. The Pitman-Yor process is also known as the two-parameter Poisson-Dirichlet Process, is an example of a Poisson-Kingman process [57], and a species sampling process of Gibbs type [18].

The easiest definition is through *stick-breaking* ([54, 40]), as follows. The family of nonnegative Pitman-Yor processes is given by three parameters: a number $\sigma \in [0, 1)$, a number $M > -\sigma$ and an atomless probability distribution G on some measurable space $(\mathcal{X}, \mathcal{A})$. We say that a random probability measure P on $(\mathcal{X}, \mathcal{A})$ is a Pitman-Yor process (of nonnegative type), denoted $P \sim \mathrm{PY}(\sigma, M, G)$, if P can be represented as

$$P = \sum_{i=1}^{\infty} W_i \delta_{\theta_i},$$

where $W_i = V_i \prod_{j=1}^{i-1}(1 - V_j)$ for V_i independent variables with $V_i \sim \mathrm{B}(1 - \sigma, M + i\sigma)$, independent of $\theta_i \overset{\text{iid}}{\sim} G$.

It is clear from this definition that the realisations of P are discrete probability measures, with countably many atoms at random locations, with random weights. If one first draws $P \sim \mathrm{PY}(\sigma, M, G)$, and next given P a random sample $X_1, \ldots, X_n$ from P, then ties among the latter observations are likely. Each tie represents an observed species. Each species is identified by a label θ_i, but this does not play an important role in the present paper, except that labels are unique by the assumption that G is atomless. If the weights (W_i) correctly model the relative abundance of distinct species, then the (conditional) probability that a next observation X_{n+1} will be distinct from $X_1, \ldots, X_n$ is indicative of the likelihood of finding a new species given the past observations.

In a forensic setup one may consider two instances not present in a "database" $X_1, \ldots, X_n$ and compare the hypotheses that these two values are two independent draws X_{n+1}, X_{n+2} from the population or copies of a single draw X_{n+1}. If one of the new instances is a characteristic (e.g. DNA profile) found at a crime scene and the other a characteristic of a suspect, one can so derive the ratio of the probabilities that the characteristic at the crime scene originates from the perpetrator (two copies of a single draw X_{n+1}) or that the match is by chance (two draws X_{n+1}, X_{n+2} that happen to be the same). See [12, 13] (and Section 4.2.3) for further discussion.

As shown in [13] such probabilities are readily calculable under the Pitman-Yor model, but depend strongly on the type parameter σ. Thus it is desirable to estimate this parameter from the observed data. In this paper we consider both the empirical Bayes estimator (which is the maximum likelihood estimator in the Bayesian setup with the Pitman-Yor process viewed as a prior) and the full Bayes posterior of σ. We prove that the empirical Bayes estimator is asymptotically normal, and show that the posterior distribution satisfies a corresponding Bernstein-von Mises theorem. We apply these results to the forensic problem.

The Pitman-Yor process is characterised by a second parameter M, called the prior precision. As its name suggests, this is a prior modelling parameter, and perhaps is better not estimated. We show that the asymptotics of the type parameter are almost independent of the prior precision, and that the problem of estimating the prior precision degenerates as $n \to \infty$.

Other applications of the Pitman-Yor process, in genetics or topic modelling, can be found in [83, 76, 35, 3, 23]. The limiting case of the Pitman-Yor process with $M \downarrow 0$ is a (particular) Poisson-Kingman process (see [59], or Example 14.47 in [33]). Estimation of the type parameter for this process was considered in [23], who give conditions so that the maximum likelihood estimator converges to a limit at a certain rate. The present paper generalizes and refines these results in the case of a general M (by providing a sharp rate and a limit distribution), and adds a full Bayes analysis.

The posterior distribution when the Pitman-Yor process is used as a nonparametric prior on the distribution of the observations was studied in [41] for known type parameter and continuous observations, and for unknown type and general observations in [27]. In the latter paper it is shown that the posterior distribution is not very sensitive to the type parameter, and its asymptotics could be established knowing just consistency for this parameter, and hence without interception of the precise results of the present paper.

4.2 Main results

We consider statistical procedures derived from the Bayesian model in which a probability distribution P is drawn from the Pitman-Yor process, viewed as a prior over the set of probability measures, and next given P the observations $X_1, \ldots, X_n$ are an i.i.d. sample from P. To estimate the parameters σ (or (σ, M)) of the Pitman-Yor process, we consider the maximum likelihood estimator, based on the marginal distribution of $(X_1, \ldots, X_n)$ in this setup, and a full Bayes approach. Because the maximum likelihood estimator estimates a parameter of the prior, in this setup this estimator is also called an empirical Bayes estimator. The full Bayes approach adds a prior distribution over σ (or (σ, M)) to the Bayesian hierarchy and then uses the conditional distribution of σ given $X_1, \ldots, X_n$, the posterior distribution, for further inference.

While these two procedures are defined by the Bayesian setup, our theoretical results obtained below are frequentist Bayes: we study these two procedures, both functions of the observations $X_1, \ldots, X_n$, under the assumption that these observations are a random sample from a given distribution P_0.

Both statistical procedures are based on the Bayesian likelihood for $X_1, \ldots, X_n$. This can be conveniently obtained by considering the exchangeable random partition of the set $\{1, 2, \ldots, n\}$ generated by the sample through the equivalence relation

$$i \equiv j \qquad \text{if and only if} \qquad X_i = X_j.$$

An alternative way to generate $X_1, \ldots, X_n$ is to generate first the partition and next attach to each set in the partition a value generated independently from the center measure G (see e.g. [33], Lemma 14.11 for a precise statement), duplicating this as many times as there are indices in the set, in order to form the observations $X_1, \ldots, X_n$. Because the parameter σ enters only in creating the partition, the partition is a sufficient statistic for σ. Because of exchangeability, the vector $(K_n, N_{n,1}, \ldots, N_{n,K_n})$ of the number K_n of sets in the partition and the cardinalities $N_{n,i}$ of the partitioning sets (i.e. the multiplicity of X_i in $X_1, \ldots, X_n$) is already sufficient for (M, σ) and hence the empirical Bayes estimator and posterior distribution of (M, σ) are the same, whether based on observations $(X_1, \ldots, X_n)$ or on observations $(K_n, N_{n,1}, \ldots, N_{n,K_n})$.

The likelihood function for (M, σ) is therefore equal to the probability of a particular partition, called the *exchangeable partition probability functio*n (EPPF). For the

Pitman-Yor process this is given by (see [58], or [33, page 465])

$$p_\sigma(N_{n,1},\dots,N_{n,K_n}) = \frac{\prod_{i=1}^{K_n-1}(M+i\sigma)}{(M+1)^{[n-1]}}\prod_{j=1}^{K_n}(1-\sigma)^{[N_{n,j}-1]}. \tag{4.1}$$

Here $a^{[n]} = a(a+1)\cdots(a+n-1)$ is the ascending factorial, with $a^{[0]} = 1$ by convention.

Although we adopt the Pitman-Yor process as a prior for the distribution P of the observations, and then arrive at the likelihood (4.1), we shall investigate the maximum likelihood estimator and posterior distribution under the frequentist-Bayes assumption that in reality the observations $X_1,\dots,X_n$ are an i.i.d. sample from a given distribution P_0. It turns out that the asymptotic properties of the maximizer of (4.1) are then determined by the function $\alpha_0\colon(1,\infty)\to\mathbb{N}$ given by

$$\alpha_0(u){:}= \#\Bigl\{x\colon \frac{1}{P_0\{x\}} \le u\Bigr\}. \tag{4.2}$$

The function α_0 is nondecreasing and increases by jumps of size 1. Following [42], we shall assume that this function is regularly varying. The paper [42] derived (distributional) limits of characteristics such as K_n. In the present paper we use similar methods to analyse the likelihood function (4.1).

Recall that a measurable function $\alpha\colon(0,\infty)\to\mathbb{R}_+$ is said to be *regularly varying* (at ∞) of order γ if, for all $u>0$, as $n\to\infty$,

$$\frac{\alpha(nu)}{\alpha(n)} \to u^\gamma.$$

It is known (see e.g. [6] or the appendix to [19]) that if the limit of the sequence of quotients on the left exists for every u, then it necessarily has the form u^γ, for some γ. If we write $\alpha(u) = u^\gamma L(u)$, then L will be *slowly varying*: a function that is regularly varying of order 0. Then $\alpha(n) = n^\gamma L(n)$, and it can be shown that $n^{\gamma-\delta} \ll \alpha(n) \ll n^{\gamma+\delta}$, for every $\delta>0$, so that the rate of growth of α is n^γ to "first order".

Since the function α_0 in (4.2) increases by steps, it is discontinuous and so is the associated slowly varying function. We assume that there exists $\sigma_0\in(0,1)$ and a continuously differentiable slowly varying function L_0 such that, for every $u>1$,

$$\bigl|\alpha_0(u) - u^{\sigma_0}L_0(u)\bigr| \le Cu^{\beta_0}, \qquad \text{for some } \beta_0<\sigma_0 \text{ and } C>0, \tag{4.3}$$

$$\bigl|L_0'(u)\bigr| \le C_\delta\, u^{-1+\delta}, \qquad \text{for any } \delta>0, \text{ for some } C_\delta. \tag{4.4}$$

If (4.3) holds for a slowly varying function L_0, then α_0 is regularly varying of order σ_0 (see Lemma 4.3.10). The rationale for (4.4) is that a slowly varying function $u\mapsto L(u)$ will always grow slower than any power u^δ. It may not be differentiable, but if it is, then it is reasonable that the derivative grows slower than any power of $u^{-1+\delta}$. Condition (4.4) is satisfied, for instance, by powers $L_0(u) = (\log u)^r$ of the logarithmic function, for $r \in \mathbb{R}$, and the functions $L_0(u) = e^{(\log u)^r}$, for $r < 1$. Although we assume (4.3)-(4.4) for $u > 1$, the bounds are asymptotic in the sense that if they hold for $u > U$ and some U, then they are valid for $u > 1$ and possibly larger constants C and C_δ (and some extension of L_0), since the right sides are bounded away from zero on intervals $(1, U]$.

Example 4.2.1. *For the probability distribution* $(p_j)_{j\in\mathbb{N}}$ *with* $p_j = c/j^\alpha$, *for some* $\alpha > 1$, *we have* $\alpha_0(u){:}= \#(j{:}\, 1/p_j \leq u) = \lfloor (cu)^{1/\alpha} \rfloor$. *Then* $|\alpha_0(u) - (cu^{1/\alpha})| \leq 1$ *and hence* (4.3)-(4.4) *are satisfied with* $\sigma_0 = 1/\alpha$, $L_0(u) = c^{1/\alpha}$, $C = 1$, $\beta_0 = 0$, *and* $C_\delta = 0$.

Example 4.2.2. *In [23] it is assumed that* $|\alpha_0(u) - Lu^{\sigma_0}| \leq Cu^{\sigma_0/2}\sqrt{\log(eu)}$, *for every* $u > 1$ *and some constant* C. *This implies* (4.3)-(4.4) *with* L_0 *constant and any* β_0 *slightly bigger than* $\sigma_0/2$ *(and hence easily satisfying the restriction in* (4.3)*).*

4.2.1 Estimating the type parameter

The following theorem shows that the empirical likelihood estimator is asymptotically normal, after scaling by the rate $\sqrt{\alpha_0(n)}$ and centering at the zero $\sigma_{0,n}$ of the function

$$E_{0,n}(\sigma) = \int_0^n \alpha_0\Big(\frac{n}{s}\Big) e^{-s}\Big(\frac{1}{\sigma} - \sum_{m=1}^\infty \frac{s^m}{m!(m-\sigma)}\Big)\, ds. \tag{4.5}$$

In Lemma 4.3.4 these zeros are shown to converge to the exponent σ_0 of regular variation of α_0, at a rate depending on the function α_0.

For the moment we keep M fixed and let $\hat\sigma_n$ be the maximizer of the likelihood (4.1) with respect to σ, for given M. The following theorem shows that the asymptotic behaviour of $\hat\sigma_n$ is the same for every M.

Theorem 4.2.3. *Assume that* P_0 *is discrete with atoms such that* $\alpha_0(u){:}= \#\{x{:}\, 1/P_0\{x\} \leq u\}$ *satisfies* (4.3)-(4.4) *with exponent* $\sigma_0 \in (0, 1)$. *Then the empirical Bayes estimator* $\hat\sigma_n$, *the point of maximum of* (4.1), *satisfies* $\sqrt{\alpha_0(n)}(\hat\sigma_n - \sigma_{0,n}) \rightsquigarrow N(0, \tau_1^2/\tau_2^4)$, *where* $\sigma_{0,n}$ *are the roots of the functions* $E_{0,n}$ *in* (4.5) *and* τ_1 *is given in* (4.12) *and* $\tau_2^2 = -E_0'(\sigma_0)$, *for* E_0 *given in* (4.9).

The proof of the theorem is deferred to Section 4.3.1.

The theorem centers the estimators at the zeros $\sigma_{0,n}$ of the functions $E_{0,n}$ in (4.5). It is shown in Lemma 4.3.4 that these zeros tend to the coefficient of regular variation σ_0 of α_0, and hence $\hat\sigma_n \to \sigma_0$, in probability. However, the rate of this convergence may be too slow to replace $\sigma_{0,n}$ by σ_0 in the centering of $\hat\sigma_n$. For the case that the function L_0 in (4.3) can be taken constant, Lemma 4.3.4 shows that $\sigma_{0,n} - \sigma_0 = O(n^{-(\sigma_0 - \beta)})$, for β_0 as in (4.3), and hence if $\beta_0 < \sigma_0/2$, then $\sqrt{\alpha_0(n)}(\sigma_{0,n} - \sigma_0) \to 0$, and hence also $\sqrt{\alpha_0(n)}(\hat\sigma_n - \sigma_0) \rightsquigarrow N(0, \tau_1^2/\tau_2^4)$. If α_0 contains nontrivial slowly varying terms, then the rate of convergence $\sigma_{0,n} \to \sigma_0$ will typically be much slower than $\alpha_0(n)^{-1/2}$

and the latter result will fail. (Lemma 4.3.4 gives the rate $L_0'(n)n/L_0(n)$, and its proof shows that this is sharp, for instance: $1/\log n$ if $L_0(s) = \log s$.)

In general, we could say that the estimators $\hat\sigma_n$ roughly, but possibly not quite, estimate the degree of regular variation of α_0. If one believes in the likelihood, then this is an indication that the type parameter has a more subtle interpretation than the degree of regular variation, rendering it extra worth while to use principled methods for its estimation. (If interest were in the coefficient of regular variation σ_0, then direct approaches may be preferable.)

In the special case considered in Example 4.2.2, the rate of the estimator is $\sqrt{\alpha_n(n)} = n^{\sigma_0/2}$. This is a faster rate than obtained (in the case that $M = 0$) in [23], who showed that $\hat\sigma_n = \sigma_0 + O_P(n^{-\sigma_0/2}\sqrt{\log n})$ under the condition in Example 4.2.2. (Our improved rate centers at $\sigma_{0,n}$; for centering at σ_0, the condition of Example 4.2.2 must be slightly strengthened to have an exponent strictly smaller than $\sigma_0/2$ rather than $\sigma_0/2$.)

The asymptotic variance of the sequence $\sqrt{\alpha_0(n)}(\hat\sigma_n - \sigma_{0,n})$ is a one-dimensional form of the sandwich formula, which is clear if it is written as $\tau_2^{-2}\tau_1^2\tau_2^{-2}$. It appears that in general $\tau_1 \neq \tau_2$, which is explainable by the fact that the likelihood used to define $\hat\sigma_n$ is the Bayesian marginal likelihood, which is misspecified relative to the frequentist distribution of $X_1, \dots, X_n$: the likelihood behaves like a general contrast function rather than a likelihood.

Next consider the posterior distribution of σ given $X_1, \dots, X_n$ in the model $\sigma \sim \Pi_\sigma$, $P|\,\sigma \sim \mathrm{PY}\,(\sigma, M, G)$ and $X_1, \dots, X_n|\,P, \sigma \sim P$, for a given prior distribution Π_σ on $(0, 1)$. Since the likelihood for observing $X_1, \dots, X_n$ is proportional to (4.1), by Bayes rule the posterior distribution has density relative to Π_σ proportional to (4.1). We study the posterior distribution under the assumption that $X_1, \dots, X_n$ are an i.i.d. sample from a distribution P_0.

Theorem 4.2.4. *Assume that P_0 is a discrete distribution with atoms such that $\alpha_0(u){:}= \#\{x{:}\, 1/P_0\{x\} \leq u\}$ satisfies (4.3)-(4.4) with exponent $\sigma_0 \in (0, 1)$. For a prior distribution Π_σ on $\sigma \in (0, 1)$ with a bounded density that is positive and continuous at σ_0, the posterior distribution of σ satisfies*

$$\sup_B \Bigl|\Pi_n(\sigma \in B|\, X_1, \dots, X_n) - N\Bigl(\hat\sigma_n, \frac{1}{\alpha_0(n)\tau_2^2}\Bigr)(B)\Bigr| \to 0,$$

where the supremum is taken over all Borel sets B in $(0, 1)$, and $\tau_2^2 = -E_0'(\sigma_0)$, for E_0 given in (4.9). In particular, the posterior distribution for σ contracts to σ_0. Furthermore, the posterior mean $\tilde\sigma_n = \mathrm{E}(\sigma|\, X_1, \dots, X_n)$ satisfies $\sqrt{\alpha_0(n)}(\tilde\sigma_n - \hat\sigma_n) \to 0$, in probability.

The proof Theorem 4.2.4 is deferred to Section 4.3.2.

Apart from the unusual scaling rate $\alpha_0(n)$, Theorem 4.2.4 is of the Bernstein-von Mises type, for a misspecified model (see [44]). Misspecification arises, because the likelihood corresponds to the Bayesian model, but the observations are sampled from P_0.

4.2.2 Estimating the precision

The parameter M is commonly referred to as the *prior precision.* This name suggests that this is truly a prior modelling parameter and estimating it from the data may be illogical. Theorem 4.2.3 shows that the maximum likelihood estimator $\hat\sigma_{n,M}$ of σ, for a given M, satisfies that the sequence $\sqrt{\alpha_0(n)}(\hat\sigma_{n,M}-\sigma_{n,0})$ tends to a centered normal distribution, for every M, where the limit is independent of M. Inspection of the proof shows that $\hat\sigma_{n,M_n}$, for a sequence M_n, has the same behaviour, as long as $M_n \ll \sqrt{\alpha_0(n)}/\log n$. Furthermore, if M is equipped with a prior over a compact (or slowly increasing) interval, then the posterior distribution of σ still satisfies the assertion of Theorem 4.2.4, where the limit does not involve M. Thus for a very wide range of prior precisions, the estimators for the type parameter are asymptotically equivalent. This may again suggest that the parameter M plays a different role than the type parameter.

Nevertheless, we might use the likelihood function to obtain also a maximum likelihood or Bayes estimator for M. In the following theorem we consider the maximum likelihood estimator, where the parameter M is restricted to a compact set $[0,\bar M]$. (The proof extends to $\bar M=\bar M_n$ that increase not too fast to infinity.)

The limiting value M_0 of the maximum likelihood estimator $\hat M_n$ depends on the fine details of the regular variation of the function α_0 in (4.3)-(4.4), through the limit (assumed to exist)

$$K_0{:}=\lim_{n\to\infty}\Bigl[\frac{c_0L_0'(n)n\log n}{L_0(n)}+\log L_0(n)\Bigr],$$

where $c_0=\Gamma(1-\sigma_0)(1+\sigma_0)/(\sigma_0\tau_2^2)$. Define $M_0=0$ or $M_0=\bar M$ if the limit is $K_0=\infty$ or $K_0=-\infty$, respectively, and otherwise, set it equal to the maximizer in $[0,\bar M]$ of the function

$$M\mapsto\frac{M}{\sigma_0}\bigl(K_0+\log\Gamma(1-\sigma_0)\bigr)+\log\Gamma(1+M)-\log\Gamma\Bigl(1+\frac{M}{\sigma_0}\Bigr).$$

Theorem 4.2.5. *Assume that P_0 is a discrete distribution with atoms such that $\alpha_0(u){:}=\#\{x{:}\,1/P_0\{x\}\le u\}$ satisfies (4.3)-(4.4) with exponent $\sigma_0\in(0,1)$ and a function L_0 such that $u\mapsto L_0'(u)u$ is slowly varying. Then the joint maximum likelihood estimator $(\hat M_n,\hat\sigma_n)$ satisfies $\hat M_n\to M_0$ in probability and $\hat\sigma_n$ has identical behaviour as in Theorem 4.2.3.*

The proof of the theorem is deferred to Section 4.3.3.

Example 4.2.6. *If L_0 is constant, then $K_0=\log L_0$ is finite, and it can be arbitrary large or small, depending on the nonasymptotic properties of the sequence $P_0\{x_j\}$. For* *instance, the choices* $p_j = c/j^\alpha$ *for every* $j > J$, *and* $p_j = \eta$, *for* $j \leq J$, *are possible for every constants* $c > 0$, $\alpha > 1$, $J \in \mathbb{N}$, *and* $\eta \in (0, 1)$ *such that* $J\eta + \sum_{j>J} c/j^\alpha = 1$. *Then* $\alpha_0(u) = J + \lfloor (cu)^{1/\alpha} - J \rfloor = \lfloor (cu)^{1/\alpha} \rfloor$, *for* $u > 1/\eta$, *and hence* $L_0 = c^{1/\alpha}$, *as* (4.3) *is determined by* $u \to \infty$. *Depending on the constants* c *and* α, *the constant* M_0 *can be any value in* $[0, \bar{M}]$.

Example 4.2.7. *If* $L_0(u) = \log u$, *then* $K_0 = \infty$, *and hence* $M_0 = \bar{M}$. *If* $L_0(u) = 1/\log u$, *then* $K_0 = -\infty$, *and hence* $M_0 = 0$.

4.2.3 Forensic application

Consider again the Bayesian model in which $X_1, \ldots, X_n$ are drawn independently from a distribution P generated from the Pitman-Yor process. A next observation X_{n+1} will either be equal to one of the current observations $X_1, \ldots, X_n$ or constitute a new type. If $\tilde{X}_1, \ldots, \tilde{X}_{K_N}$ are the distinct values in $X_1, \ldots, X_n$ and $N_{n,1}, \ldots, N_{n,K_n}$ are the multiplicities of these values in the sample, then it is known that (see [56, 57], or [33], page 465)

$$\mathrm{P}(X_{n+1} = \tilde{X}_i |\, X_1, \ldots, X_n, \sigma, M) = \frac{N_{n,i} - \sigma}{M + n}, \qquad i = 1, \ldots, K_n. \tag{4.6}$$

The remaining mass $(M + K_n\sigma)/(M + n)$ is the probability of obtaining a new species, distinct from $\tilde{X}_1, \ldots, \tilde{X}_{K_N}$. The probability distribution defined by these numbers is known as the *prediction probability function.*

In the forensic setup discussed in [12, 13], the sample $X_1, \ldots, X_n$ represents a database of characteristics of individuals (say DNA profiles), and given is a new profile, not present in the database, that has been found both at the crime scene and on a defendant who has been charged with the crime. The prosecution argues that the defendant is the perpetrator who left her profile on the crime scene and hence only a single new observation X_{n+1} is involved. The defence argues that the perpetrator and the defendant are two different individuals, who happen to have the same profile, and hence two independent observations X_{n+1}, X_{n+2} are involved, which were observed to take the same value. The two hypotheses can be made precise in a Bayesian hierarchy describing a generative model for the database $X_1, \ldots, X_n$, the profile X_{n+1} found at the crime scene and the profile Y found on the suspect. According to the prosecution the generative model is:

(i) $(\sigma, M) \sim \Pi_{\sigma,M}$.

(ii) $P |\, (\sigma, M) \sim$ Pitman-Yor (σ, M).

(iii) $X_1, \ldots, X_{n+1} |\, P, \sigma, M \overset{\text{iid}}{\sim} P$.

(p) $Y |\, X_1, \ldots, X_{n+1}, P, \sigma, M \sim \delta_{X_{n+1}}$.

The fourth step (p) expresses that the profile Y found on the defendant is identical to the profile found on the crime scene, because it results from the same individual X_{n+1}

chosen from the population. The defence agrees with steps (i)-(iii) of the hierarchy, but replaces (p) by:

(d) $Y|\, X_1, \ldots, X_{n+1}, P, \sigma, M \sim P$.

This expresses that the defendant's profile is just another draw X_{n+2} from the population. In the observed data the value of this draw happens to be the same as the profile X_{n+1} at the crime scene.

To decide on the case we might evaluate the ratio of the likelihoods of the full evidence $X_1, \ldots, X_{n+1}, Y$ under the two hypotheses. Since the (marginal) likelihood of $X_1, \ldots, X_{n+1}$ is determined by (i)–(iii), it is the same under both hypotheses. Hence the relative likelihood is the ratio of the conditional likelihoods of Y given $X_1, \ldots, X_{n+1}$. For the prosecution Y depends deterministically on $X_1, \ldots, X_{n+1}$ (it must be equal to X_{n+1}) and hence the conditional probability of the observed value is equal to 1. For the defence the conditional likelihood of Y is the probability that an $(n+2)$th observation X_{n+2} happens to be of the same species as X_{n+1}. Since X_{n+1} has multiplicity $N_{n+1,K_{n+1}} = 1$ among $X_1, \ldots, X_{n+1}$ (by our assumption on the observed data), given (σ, M) this probability is $(1-\sigma)/(M+n+1)$, as determined by the prediction probability function (4.6). The unconditional probability is the integral of this relative to the posterior distribution of (σ, M), i.e. $\mathbb{E}\big((1-\sigma)/(M+n+1)|\, X_1, \ldots, X_{n+1}\big)$. The likelihood ratio of prosecution versus defence is therefore

$$1 \Big/ \ \mathbb{E}\Big(\frac{1-\sigma}{n+1+M}|\, X_1, \ldots, X_{n+1}\Big).$$

Theorem 4.2.8. *Under the assumptions of Theorem 4.2.4, the posterior distribution of $\phi = (1-\sigma)/(M+n+1)$ satisfies*

$$\sup_B \Big|\Pi_n(\phi \in B|\, X_1, \ldots, X_n) - N\Big(\frac{1-\hat\sigma_n}{n+M+1}, \frac{1}{(n+M+1)^2\alpha_0(n)\tau_2^2}\Big)(B)\Big| \to 0.$$

Moreover, the posterior mean $\tilde\phi_n = \mathbb{E}\big(\phi|\, X_1, \ldots, X_{n+1}\big)$ satisfies

$$\sqrt{\alpha_n(n)}\Big(\frac{1}{n\tilde\phi_n} - \frac{1}{1-\sigma_{n,0}}\Big) \rightsquigarrow N\Big(0, \frac{\tau_1^2}{(1-\sigma_0)^4\tau_2^4}\Big).$$

These assertions remain true if M is equipped with a prior supported on a compact interval in $[0, \infty)$.

Proof. The first assertion is immediate from Theorem 4.2.4 and the definition of ϕ. For the second assertion we note that

$$\frac{1}{n\tilde\phi_n} - \frac{1}{1-\sigma_{0,n}} == \frac{\tilde\sigma_n - \sigma_{0,n} + (1-\tilde\sigma_n)(M+1)/(M+n+1)}{(1-\tilde\sigma_n)/(M+n+1)}.$$

By Theorem 4.2.4 the sequence $\sqrt{\alpha_0(n)}(\tilde\sigma_n - \sigma_{n,0})$ is asymptotically equivalent to the sequence $\sqrt{\alpha_0(n)}(\hat\sigma_n - \sigma_{0,n})$, which tends to the $N(0, \tau_1^2/\tau_2^4)$-distribution, by Theorem 4.2.3. The second assertion follows by Slutzky's lemma. ■

4.3 Proofs

The logarithm of the likelihood (4.1) can be written

$$\begin{aligned}\Lambda_n(\sigma, M) &= \sum_{l=1}^{K_n-1} \log(M+l\sigma) + \sum_{j=1:N_{n,j}\geq 2}^{K_n} \sum_{l=1}^{N_{n,j}-1} \log(l-\sigma) - \sum_{i=1}^{n-1} \log(M+i)\\ &= \sum_{l=1}^{K_n-1} \log(M+l\sigma) + \sum_{l=1}^{n-1} \log(l-\sigma) Z_{n,l+1} - \sum_{i=1}^{n-1} \log(M+i), \qquad (4.7)\end{aligned}$$

where $Z_{n,l} = \#(1 \leq j \leq K_n \colon N_{n,j} \geq l)$ is the number of distinct values of multiplicity at least l in the sample $X_1, \ldots, X_n$. (In the case that all observations are distinct and hence $N_{n,j} = 1$ for every j, the second term of the likelihood is equal to 0.)

For the proofs of Theorems 4.2.3 and 4.2.4, we fix the argument M and drop it from the notation. The concavity of the logarithm shows that the log likelihood is a strictly concave function of σ. For $\sigma \downarrow 0$, it tends to a finite value if $M > 0$ and to $-\infty$ if $M = 0$, while for $\sigma \uparrow 1$ it tends to $-\infty$ if the term with $l = 1$ is present in the second sum, i.e. if there is at least one tied observation. This happens with probability tending to 1 as $n \to \infty$. The derivative of the log likelihood with respect to σ is equal to

$$\Lambda_n'(\sigma) = \sum_{l=1}^{K_n-1} \frac{l}{M+l\sigma} - \sum_{l=1}^{n-1} \frac{Z_{n,l+1}}{l-\sigma}. \qquad (4.8)$$

The left limit at $\sigma = 0$ is $\Lambda_n'(0) = \frac{1}{2} K_n(K_n-1)/M - \sum_{l=1}^{n-1} l^{-1} Z_{n,l+1}$. Since $Z_{n,l} \leq Z_{n,1} = K_n$, a crude bound on the sum is $K_n \log n$, which shows that the derivative at $\sigma = 0$ tends to infinity if $K_n \gg \log n$. In that case the unique maximum of the log likelihood in $[0,1]$ is taken in the interior of the interval, and hence $\hat\sigma_n$ satisfies $\Lambda_n'(\hat\sigma_n) = 0$.

Set $\alpha_n = \alpha_0(n)$. Under the condition that α_0 is regularly varying of exponent $\sigma_0 \in (0,1)$, the sequence α_n is of the order n^{σ_0} up to slowly varying terms. By Theorems 9 and 1‘ of [42], the sequence K_n/α_n tends almost surely to $\Gamma(1-\sigma_0)$ and hence in particular $K_n \gg \log n$, and the conclusion of the preceding paragraph that $\Lambda_n'(\hat\sigma_n) = 0$ pertains.

By Lemma 4.3.4, the functions $E_{0,n}/\alpha_n$, for $E_{0,n}$ defined in (4.5), converge to the function E_0 defined by

$$E_0(\sigma) = \frac{\Gamma(1-\sigma_0)}{\sigma} - \sum_{m=1}^{\infty} \frac{\Gamma(m+1-\sigma_0)}{m!(m-\sigma)}, \qquad (4.9)$$

and this function vanishes at $\sigma = \sigma_0$. By monotonicity of these functions, the zeros $\sigma_{0,n}$ of the functions $E_{0,n}$ tend to the zero $\sigma_0 \in (0,1)$ of the limit function.

4.3.1 Proof of Theorem 4.2.3

The monotonicity of Λ_n', the definition of $\hat\sigma_n$ and the fact that $-E_{0,n}(\sigma_{0,n}) = 0$ give that

$$\begin{aligned}\mathrm{P}\big(\sqrt{\alpha_n}(\hat\sigma_n - \sigma_{0,n}) \le x\big) &= \mathrm{P}\Big(\Lambda_n'\Big(\sigma_{0,n} + \frac{x}{\sqrt{\alpha_n}}\Big) \le 0\Big)\\ &= \mathrm{P}\Big(\frac{1}{\sqrt{\alpha_n}}\Big[\Lambda_n'\Big(\sigma_{0,n} + \frac{x}{\sqrt{\alpha_n}}\Big) - E_{0,n}\Big(\sigma_{0,n} + \frac{x}{\sqrt{\alpha_n}}\Big)\Big]\\ &\qquad\qquad \le -\frac{1}{\sqrt{\alpha_n}}\Big[E_{0,n}\Big(\sigma_{0,n} + \frac{x}{\sqrt{\alpha_n}}\Big) - E_{0,n}(\sigma_{0,n})\Big]\Big).\end{aligned}$$

The variables in the left side of the last probability are asymptotically normal by Lemma 4.3.1, while by the mean value theorem the numbers on the right side of the inequality are equal to $-xE_{0,n}'(\sigma_n)/\alpha_n$ for numbers σ_n between $\sigma_{0,n}$ and $\sigma_{0,n}+x/\sqrt{\alpha_n}$ and hence $\sigma_n \to \sigma_0$. Thus $-xE_{0,n}'(\sigma_n)/\alpha_n \to x\tau_2^2$, by Lemma 4.3.4. The asymptotic normality of $\hat\sigma_n$ follows.

4.3.2 Proof of Theorem 4.2.4

We first prove that the posterior distribution is $\sqrt{\alpha_n}$-consistent: $\Pi_n\big(\sqrt{\alpha_n}|\sigma - \sigma_{0,n}| > m_n|\, X_1, \ldots, X_n\big) \to 0$ in probability for any $m_n \to \infty$. By the monotonicity of Λ_n' and the fact that $\Lambda_n'(\hat\sigma_n) = 0$, for given $\sigma_n > \hat\sigma_n$,

$$\begin{aligned}&\Lambda_n(\sigma) \ge \Lambda_n(\sigma_n), &&\text{if } \hat\sigma_n < \sigma < \sigma_n,\\ &\Lambda_n(\sigma) \le \Lambda_n(\sigma_n) + \Lambda_n'(\sigma_n)(\sigma - \sigma_n), &&\text{if } \sigma > \sigma_n.\end{aligned}$$

It follows that

$$\begin{aligned}\Pi_n\big(\sigma > \sigma_n|\, X_1, \ldots, X_n\big) &= \frac{\int_{\sigma_n}^1 e^{\Lambda_n(\sigma)}\, d\Pi_\sigma(\sigma)}{\int_0^1 e^{\Lambda_n(\sigma)}\, d\Pi_\sigma(\sigma)}\\ &\le \frac{\int_{\sigma_n}^1 e^{\Lambda_n(\sigma_n)+\Lambda_n'(\sigma_n)(\sigma-\sigma_n)}\, d\Pi_\sigma(\sigma)}{\int_{\hat\sigma_n}^{\sigma_n} e^{\Lambda_n(\sigma_n)}\, d\Pi_\sigma(\sigma)}\\ &\lesssim \frac{\int_0^\infty e^{\Lambda_n'(\sigma_n)u}\, du}{\sigma_n - \hat\sigma_n} = \frac{1}{-\Lambda_n'(\sigma_n)(\sigma_n - \hat\sigma_n)},\end{aligned}$$

where the proportionality constant depends on the density of Π_σ only. If we choose $\sigma_n = \sigma_{0,n} + x/\sqrt{\alpha_n}$, then, by Lemmas 4.3.1 and 4.3.3, since $E_{0,n}(\sigma_{0,n}) = 0$,

$$\frac{\Lambda_n'(\sigma_n)}{\sqrt{\alpha_n}} = O_P(1) + \frac{E_{0,n}(\sigma_n)}{\sqrt{\alpha_n}} = O_P(1) + \frac{E_{0,n}'(\tilde\sigma_n)}{\alpha_n}x = O_P(1) - \tau_2^2 x.$$

Theorem 4.2.3 gives that $\sqrt{\alpha_n}(\sigma_n - \hat\sigma_n) = x + O_p(1)$, and hence the probability that $-\Lambda_n'(\sigma_n)(\sigma_n - \hat\sigma_n)$ is bigger than some fixed constant can be made arbitrarily large by choosing large enough x. This shows that the preceding display tends to zero in probability for $\sigma_n = \sigma_{0,n} + m_n/\sqrt{\alpha_n}$ and any $m_n \to \infty$. The probability of the events $\sigma_n > \hat{\sigma}_n$, on which the preceding displays are valid, then also tends to one. Combined with a similar argument on the left tail of the posterior distribution, this shows that the posterior contracts to $\sigma_{0,n}$ at rate $1/\sqrt{\alpha_n}$.

Since $\sqrt{\alpha_n}(\hat{\sigma}_n - \sigma_{0,n}) = O_p(1)$, by Theorem 4.2.3, there exists $m_n \to \infty$ so that the sets $C_n\colon= \{\sigma\colon \sqrt{\alpha_n}|\sigma - \hat{\sigma}_n| \le m_n\}$ have posterior probability tending to one. The total variation measure between the posterior measure $\Pi_n(\sigma \in \cdot\,|\, X_1, \ldots, X_n)$ and the conditioned posterior measure $\Pi_n(\sigma \in \cdot\,|\, X_1, \ldots, X_n, \sigma \in C_n)$ is bounded above by $2\Pi_n(\sigma \notin C_n|\, X_1, \ldots, X_n)$ and hence tends to zero in probability. Thus it suffices to prove the Gaussian approximation to the conditioned posterior measure.

By Lemma 4.3.2 and the fact that $\Lambda_n'(\hat{\sigma}_n) = 0$, a second order Taylor expansion gives

$$\sup_{\sigma \in C_n} \Bigl| \frac{\Lambda_n(\sigma) - \Lambda_n(\hat{\sigma}_n)}{\alpha_n(\sigma - \hat{\sigma}_n)^2} + \frac{1}{2}\tau_2^2 \Bigr| \overset{P}{\to} 0.$$

We conclude that there exist random variables ε_n that tend to zero in probability with, for every $\sigma \in C_n$,

$$-\frac{1}{2}(\sigma - \hat{\sigma}_n)^2 \alpha_n(\tau_2^2 + \varepsilon_n) \le \Lambda_n(\sigma) - \Lambda_n(\hat{\sigma}_n) \le -\frac{1}{2}(\sigma - \hat{\sigma}_n)^2 \alpha_n(\tau_2^2 - \varepsilon_n).$$

Then, for π_σ a density of the prior measure Π_σ,

$$\Pi_n\bigl(\sigma \in B|\, X_1, \ldots, X_n, \sigma \in C_n\bigr) \le \frac{\int_{B\cap C_n} e^{-(\sigma-\hat{\sigma}_n)^2\alpha_n(\tau_2^2-\varepsilon_n)/2}\, d\sigma}{\int_{C_n} e^{-(\sigma-\hat{\sigma}_n)^2\alpha_n(\tau_2^2+\varepsilon_n)/2}\, d\sigma} \frac{\sup_{\sigma\in C_n} \pi_\sigma(\sigma)}{\inf_{\sigma \in C_n} \pi_\sigma(\sigma)}$$

By changing variables we see that $\Pi_n\bigl(\sqrt{\alpha_n}(\sigma - \hat{\sigma}_n) \in B|\, X_1, \ldots, X_n, \sigma \in C_n\bigr)$ can be bounded above by

$$\begin{aligned}
&\frac{\int_{-m_n}^{m_n} 1_B(s) e^{-s^2(\tau_2^2-\varepsilon_n)/2}\, ds}{\int_{-m_n}^{m_n} e^{-s^2(\tau_2^2+\varepsilon_n)/2}\, ds}(1 + o_P(1))\\
&\qquad = \frac{\mathrm{P}\bigl(Z/\sqrt{\tau_2^2 - \varepsilon_n} \in B \cap (-m_n, m_n)\bigr)}{\mathrm{P}\bigl(Z/\sqrt{\tau_2^2 + \varepsilon_n} \in (-m_n, m_n)\bigr)}(1 + o_P(1)),
\end{aligned}$$

for Z a standard normal variable and the probabilities understood to refer to Z only. This tends in probability to $\mathrm{P}(Z/\tau_2 \in B)$, uniformly in B.

By the same method, switching + and − signs and sup and inf, we can derive the same expression as an asymptotic lower bound. This concludes the proof of the first assertion of Theorem 4.2.4.

Because convergence in total variation norm implies convergence of the expectations of bounded, measurable functions, to prove the convergence of the posterior mean, it suffices to show that

$$\int_{\sqrt{\alpha_n}(\sigma-\hat\sigma_n)>m_n} \sqrt{\alpha_n}|\sigma-\hat\sigma_n|\,d\Pi_n(\sigma|\,X_1,\ldots,X_n)\to 0,$$

in probability, for every $m_n\to\infty$, combined with a similar estimate on the left tail. By the argument at the beginning of the proof this expectation is bounded above by, for $\sigma_n=\hat\sigma_n+x/\sqrt{\alpha_n}$ with $x>0$,

$$\frac{\int_{\sqrt{\alpha_n}(\sigma-\hat\sigma_n)>m_n}\sqrt{\alpha_n}(\sigma-\hat\sigma_n)e^{\Lambda_n'(\sigma_n)(\sigma-\sigma_n)}\,d\Pi_\sigma(\sigma)}{\sigma_n-\hat\sigma_n}$$
$$=\frac{\int_{m_n-x}^\infty(u+x)e^{\Lambda_n'(\sigma_n)u/\sqrt{\alpha_n}}\,du}{x}$$

Because $\Lambda_n'(\sigma_n)/\sqrt{\alpha_n}=\sqrt{\alpha_n}(\sigma_n-\sigma_{0,n})\Lambda_n''(\tilde\sigma_n)/\alpha_n=(x+O_P(1))(-\tau_2^2+o_P(1))$, the exponential in the integrand is with probability arbitrarily close to 1 bounded above by e^{-cu}, for some $c>0$, if x is chosen large enough. On the event where this is the case the integral is bounded above by a multiple of $\int_{m_n-x}^\infty(u+x)e^{-cu}\,du\to 0$, as $n\to\infty$. It follows that the right side tends to zero in probability.

4.3.3 Proof of Theorem 4.2.5

The maximum likelihood estimator of M can be obtained in two steps: first we maximize the log likelihood (4.7) over σ for fixed M, yielding $\hat\sigma_{n,M}$, and next we maximize the "profile log likelihood" $M\mapsto\Lambda_n(\hat\sigma_{n,M},M)$. From Theorem 4.2.3 we know that $\hat\sigma_{n,M}$ will be contained in a neighbourhood of σ_0, with probability tending to 1. To proceed further, we first obtain a stochastic expansion of $\hat\sigma_{n,M}$ that refines this result.

The estimator $\hat\sigma_{n,M}$ solves $\Lambda_n'(\sigma,M)=0$, where the prime means the partial derivative with respect to σ. The derivative can be written

$$\Lambda_n'(\sigma,M)=\sum_{i=1}^{K_n-1}\frac{i}{M+i\sigma}-\sum_{l=1}^{n-1}\frac{1}{l-\sigma}Z_{n,l+1}\qquad=\frac{K_n}{\sigma}-G_n(\sigma)-\frac{h_{\sigma,M}(K_n)}{\sigma},$$

for $G_n(\sigma)=\sum_{l=1}^{n-1}\frac{1}{l-\sigma}Z_{n,l+1}$, and

$$h_{\sigma,M}(k)=1+\sum_{l=1}^{k-1}\frac{M}{M+l\sigma}\le 1+\frac{M}{\sigma}\log\Bigl(1+\frac{k\sigma}{M}\Bigr).\qquad(4.10)$$

Expansion of the equation $\Lambda_n'(\hat\sigma_{n,M},M)=0$ around the zero $\sigma_{0,n}$ of the function $E_{0,n}$ in (4.5) gives

$$\hat\sigma_{n,M}-\sigma_{0,n}=-\frac{\Lambda_n'(\sigma_{0,n},M)}{\Lambda_n''(\sigma_{0,n},M)+\Lambda_n'''(\tilde\sigma_{n,M},M)(\hat\sigma_{n,M}-\sigma_{0,n})/2}.$$

It is shown in Lemma 4.3.1 that $V_n\colon= \bigl(K_n/\sigma_{n,0} - G_n(\sigma_{n,0})\bigr)/\sqrt{\alpha_n}$, which is free of M, tends to a centered normal distribution, and it is shown in Lemma 4.3.2 that the sequence $\bigl(-K_n/\sigma_{0,n}^2 - G_n'(\sigma_{0,n})\bigr)/\alpha_n$ tends in probability to $-\tau_2^2\colon= E_0'(\sigma_0)$. It can similarly be seen that $Z_n\colon= \bigl(2K_n/\sigma_{0,n}^3 - G_n''(\sigma_{0,n})\bigr)/\alpha_n$ tends in probability to a constant z. Furthermore, it follows from the bound in (4.10) and the fact that K_n/α_n converges almost surely, that $h_{\sigma_{0,n},M}(K_n) = O_P(\log n)$, uniformly in M belonging to compacta, and from the definition of $h_{\sigma,M}(K_n)$ that its first and second partial derivatives relative to σ are of the same order. Therefore, uniformly in M,

$$\begin{aligned}\Lambda_n'(\sigma_{0,n},M) &= \frac{K_n}{\sigma_{0,n}} - G_n(\sigma_{0,n}) - \frac{h_{\sigma_{0,n},M}(K_n)}{\sigma_{0,n}} = \alpha_n^{1/2}V_n - \frac{h_{\sigma_{0,n},M}(K_n)}{\sigma_{0,n}},\\ \Lambda_n''(\sigma_{0,n},M) &= \frac{-K_n}{\sigma_{0,n}^2} - G_n'(\sigma_{0,n}) - \frac{d}{d\sigma}\Bigl[\frac{h_{\sigma,M}(K_n)}{\sigma}\Bigr]_{\sigma=\sigma_{0,n}} = -\alpha_n\tau_2^2 + O_P(\log n),\\ \Lambda_n'''(\tilde\sigma_{n,M},M) &= \frac{2K_n}{\sigma_{0,n}^3} - G_n''(\sigma_{0,n}) - \frac{d^2}{d\sigma^2}\Bigl[\frac{h_{\sigma,M}(K_n)}{\sigma}\Bigr]_{\sigma=\sigma_{0,n}} = \alpha_n Z_n + O_P(\log n).\end{aligned}$$

Substituting these expansions in the preceding display gives that

$$\begin{aligned}\hat\sigma_{n,M} - \sigma_{0,n} &= \frac{\alpha_n^{-1/2}V_n - \alpha_n^{-1}h_{\sigma_{0,n},M}(K_n)/\sigma_{0,n}}{\tau_2^2 - Z_n(\hat\sigma_{n,M} - \sigma_{0,n}) + O_P(\log n/\alpha_n)}\\ &= \Bigl(\frac{V_n}{\alpha_n^{1/2}\tau_2^2} - \frac{h_{\sigma_{0,n},M}(K_n)}{\alpha_n\sigma_{0,n}\tau_2^2}\Bigr)\Bigl(1 + \frac{Z_n(\hat\sigma_{n,M} - \sigma_{0,n})}{\tau_2^2} + O_P\Bigl(\frac{\log n}{\alpha_n}\Bigr)\Bigr).\end{aligned}$$

We can solve $\hat\sigma_{n,M} - \sigma_{0,n}$ from this as

$$(\hat\sigma_{n,M} - \sigma_{0,n})\Bigl(1 - \frac{V_nZ_n}{\alpha_n^{1/2}\tau_2^4}\Bigr) = \frac{V_n}{\alpha_n^{1/2}\tau_2^2} - \frac{h_{\sigma_{0,n},M}(K_n)}{\alpha_n\sigma_{0,n}\tau_2^2} + O_P\Bigl(\frac{\log n}{\alpha_n^{3/2}}\Bigr).$$

Hence, for $W_{n,M} = h_{\sigma_{0,n},M}(K_n)/(\sigma_{0,n}\tau_2^2\log n)$,

$$\begin{aligned}\hat\sigma_{n,M} &= \sigma_{0,n} + \frac{V_n}{\alpha_n^{1/2}\tau_2^2} + \frac{V_n^2Z_n}{\alpha_n\tau_2^6} - \frac{W_{n,M}\log n}{\alpha_n} + O_P\Bigl(\frac{\log n}{\alpha_n^{3/2}}\Bigr)\\ &=: \tilde\sigma_n - W_{n,M}\log n/\alpha_n + O_P(\log n/\alpha_n^{3/2}).\end{aligned}$$

The variables V_n, Z_n and $W_{n,M}$ are all bounded in probability. The quantity $\tilde\sigma_n$ is defined as the sum of the first three terms on the right in the preceding line, and does not depend on M.

We are now ready to expand the profile likelihood $\Lambda_n(\hat\sigma_{n,M},M)$. The first and third terms on the far right side of (4.7) can be expanded with the help of Lemma 4.3.12

as, uniformly in M in compacta,

$$\sum_{l=1}^{K_n-1} \log(M+l\hat\sigma_{n,M}) = K_n\log K_n + K_n\log(\hat\sigma_{n,M}/e) + \Bigl(\frac{M}{\hat\sigma_{n,M}}-\frac{1}{2}\Bigr)\log K_n$$

$$+\log\frac{\sqrt{2\pi}}{\hat\sigma_{n,M}} - \log\Gamma\Bigl(1+\frac{M}{\hat\sigma_{n,M}}\Bigr) + O_P\Bigl(\frac{1}{K_n}\Bigr),$$

$$\sum_{i=1}^{n-1}\log(M+i) = n\log n - n + \Bigl(M-\frac{1}{2}\Bigr)\log n +$$

$$+\log\sqrt{2\pi} - \log\Gamma(1+M) + O\Bigl(\frac{1}{n}\Bigr).$$

To find the point of maximum M, we can drop terms that depend on K_n and n only, and add terms that do not depend on M. Thus finding the maximizer of the profile likelihood is equivalent to finding the maximizer of

$$(K_n-1)\log\frac{\hat\sigma_{n,M}}{\tilde\sigma_n} + \sum_{l=1}^{n-1}\log\frac{l-\hat\sigma_{n,M}}{l-\tilde\sigma_n}Z_{n,l+1} + \frac{M\log K_n}{\hat\sigma_{n,M}} - M\log n$$

$$+\log\Gamma(1+M) - \log\Gamma\Bigl(1+\frac{M}{\hat\sigma_{n,M}}\Bigr) + O_P\Bigl(\frac{1}{K_n}\Bigr). \qquad (4.11)$$

The sum of the first two terms can be expanded as

$$(K_n-1)\log\Bigl(1-\frac{W_{n,M}\log n}{\alpha_n\tilde\sigma_n} + O_P\Bigl(\frac{\log n}{\alpha_n^{3/2}}\Bigr)\Bigr) - \sum_{l=1}^{n-1}\log\Bigl(1-\frac{\hat\sigma_{n,M}-\tilde\sigma_n}{l-\tilde\sigma_n}\Bigr)Z_{n,l+1}$$

$$= -\frac{K_nW_{n,M}\log n}{\alpha_n\tilde\sigma_n} - \sum_{l=1}^{n-1}\Bigl(\frac{\hat\sigma_{n,M}-\tilde\sigma_n}{l-\tilde\sigma_n}\Bigr)Z_{n,l+1} + O_P\Bigl(\frac{\log n}{\sqrt{\alpha}_n}\Bigr)$$

$$= -\frac{W_{n,M}\log n}{\alpha_n}\Bigl(\frac{K_n}{\tilde\sigma_n} - \sum_{l=1}^{n-1}\frac{Z_{n,l+1}}{l-\tilde\sigma_n}\Bigr) + O_P\Bigl(\frac{\log n}{\sqrt{\alpha}_n}\Bigr)$$

$$= -\frac{W_{n,M}\log n}{\alpha_n}\bigl(\sqrt{\alpha_n}\tilde V_n + E_{0,n}(\tilde\sigma_n)\bigr) + O_P\Bigl(\frac{\log n}{\sqrt{\alpha_n}}\Bigr),$$

where $\tilde V_n$ tends to a centered normal distribution by Lemma 4.3.1. The right side is of the order $O_P(\log n/\sqrt{\alpha_n})$.

Since $L_n = \sqrt{\alpha_n}\bigl(K_n/\alpha_n - \Gamma(1-\sigma_0)\bigr)$ tends to a centered normal distribution, we have $K_n = \alpha_n\Gamma(1-\sigma_0) + O_P(\sqrt{\alpha_n})$, so that $\log K_n = \log\alpha_n + \log\Gamma(1-\sigma_0) + O_P(1/\sqrt{\alpha_n})$. Therefore, if $\alpha_n = n^{\sigma_0}\bar L_0(n)$, then the sum of the third and fourth terms on the right of (4.11) are

$$M\Big[\frac{\sigma_0\log n+\log\bar L_0(n)+\log\Gamma(1-\sigma_0)+O_P(1/\sqrt{\alpha_n})}{\sigma_{0,n}}\Big]\Big[1+O_P\Big(\frac{1}{\sqrt{\alpha_n}}\Big)\Big]-M\log n$$
$$=M\Big(\frac{\sigma_0}{\sigma_{0,n}}-1\Big)\log n+\frac{M}{\sigma_{0,n}}\big(\log\bar L_0(n)+\log\Gamma(1-\sigma_0)\big)+O_P\Big(\frac{\log n}{\sqrt{\alpha_n}}\Big).$$

By Lemma 4.3.4, $\sigma_{0,n}-\sigma_0=O(n^{-\delta})+O\big(L_0'(n)n/L_0(n)\big)$, whence the right side is of the order $(\log n)L_0'(n)n/L_0(n)+\log L_0(n)+\log\Gamma(1-\sigma_0)$.

We conclude that up to terms that do not depend on M, the profile log likelihood (4.11) is equal to, for a constant $c>0$,

$$\frac{M}{\sigma_0}\Big[\frac{L_0'(n)nc\log n}{L_0(n)}+\log L_0(n)+\log\Gamma(1-\sigma_0)\Big]$$
$$+\log\Gamma(1+M)-\log\Gamma\Big(1+\frac{M}{\sigma_0}\Big)+o_P(1).$$

If the term within square brackets tends to infinity, then asymptotically this term dominates and the maximizing value $\hat M_n$ will tend to the end point of the interval. If this term tends to minus infinity, then it also dominates and the maximum value will tend to 0. If the term converges to a limit, then the whole expression tends to a function of the form $M\mapsto aM+\log\Gamma(1+M)-\log\Gamma(1+M/\sigma_0)$. This function is concave and tends to $-\infty$ as $M\to\infty$. Its derivative at zero may be both positive or negative, depending on a and σ_0, and the point of maximum of the function may be at zero or at some positive location, possibly the upper limit $\bar M$ of the parameter. The sequence $\hat M_n$ will tend to this point of maximum.

4.3.4 Lemmas

For $g_\sigma(m)=\sum_{l=1}^{m-1}\frac{1}{l-\sigma}$, set

$$\tau_1^2=\frac{(2^{\sigma_0}-1)\Gamma(1-\sigma_0)}{\sigma_0^2}+\sum_{m=1}^\infty\frac{\Gamma(m-\sigma_0)(g_{\sigma_0}(m+1)+g_{\sigma_0}(m))}{m!}\qquad(4.12)$$
$$-\sum_{k=2}^\infty\sum_{m=1}^\infty\frac{g_{\sigma_0}(k)\Gamma(k+m+1-\sigma_0)}{k!m!2^{k+m-\sigma_0}(m-\sigma_0)}-\sum_{m=2}^\infty\frac{g_{\sigma_0}(m)\Gamma(m-\sigma_0)}{m!\,2^{m-\sigma_0-1}}.$$

Lemma 4.3.1. *For any $\sigma_n\xrightarrow{P}\sigma_0\in(0,1)$ and $M_n=o\big(\sqrt{\alpha_0(n)}/\log n\big)$, we have $\alpha_0(n)^{-1/2}\big(\Lambda_n'(\sigma_n,M_n)-E_{0,n}(\sigma_n)\big)\rightsquigarrow N(0,\tau_1^2)$.*

Proof. We denote the true distribution by $P_0=\sum_{i=1}^\infty p_i\delta_{x_i}$. The variables $Z_{n,l}$ can be written as $Z_{n,l}=\sum_{j=1}^\infty 1_{M_{n,j}\geq l}$, for $M_{n,j}$ the number of observations equal to x_j.

As $K_n = Z_{n,1}$, the function Λ_n' can be written in the form

$$\Lambda_n'(\sigma, M) = \sum_{j=1}^{\infty} \Bigl[\frac{1_{M_{n,j}\geq 1}}{\sigma} - g_\sigma(M_{n,j})\Bigr] - \frac{h_{\sigma,M}(K_n)}{\sigma},$$

where $g_\sigma(0) = g_\sigma(1) = 0$ and $g_\sigma(m) = \sum_{l=1}^{m-1} \frac{1}{l-\sigma}$, for $m \geq 2$. It is shown in [42] (and repeated below) that $\mathrm{E}K_n/\alpha_0(n) \to \Gamma(1-\sigma_0)$ and hence Jensen's inequality and (4.10) give $\mathrm{E}h_{\sigma_n,M}(K_n) \leq 1 + (M/\sigma_n)\log(1 + \mathrm{E}K_n\sigma_n/M) = O(M\log n) = o(\alpha_0(n)^{1/2})$, so that the term on the far right is asymptotically negligible.

To prove the asymptotic normality of the infinite sum, after centering and scaling, we first consider the case that the sample size n is taken to be a random variable N_n with the Poisson distribution with mean n, independent of the original variables. The resulting variables $M_{N_n,j}$ are then Poisson distributed with means np_j and independent across j, and the asymptotic normality can be proved using the Lindeberg central limit theorem. (For reference, an appropriate infinite series version is formulated in Lemma 4.3.7.) As shown in the proof of (i) and (iii) of Lemma 4.3.3, the function $E_{0,n}(\sigma)$ defined in (4.5) is equal to the expectation of $V_{k,n} = \sum_{j=1}^{\infty}\bigl[1_{M_{n,j}\geq 1}/\sigma - g_\sigma(M_{n,j})\bigr]$. Because $(1_{M_{n,j}=0})g_\sigma(M_{n,j}) = 0$, by Lemma 4.3.3 (ii), (v), (iv) and (viii), the variance of $V_{k,n}$ is given by

$$\begin{aligned}\sum_{j=1}^{\infty} \mathrm{var}\Bigl(\frac{1_{M_{N_n,j}\geq 1}}{\sigma} - g_\sigma(M_{N_n,j})\Bigr) &= \sum_{j=1}^{\infty}\Bigl[\frac{e^{-np_j}(1-e^{-np_j})}{\sigma^2} \\ &\quad + \mathrm{var}\, g_\sigma(M_{N_n,j}) - 2\frac{e^{-np_j}}{\sigma}\mathrm{E}g_\sigma(M_{N_n,j})\Bigr] \sim \alpha_0(n)\tau_1^2, \qquad (4.13)\end{aligned}$$

where τ_1^2 is equal to $(ii)/\sigma^2 + (v) - (vi) - (2/\sigma)(vii)$, for $(ii), (v), (vi)$ and (vii) shorthand for the expressions on the right sides in Lemma 4.3.3 (ii), (v), (vi) and (vii). Furthermore, the variables $1_{M_{N_n,j}\geq 1}$ are bounded and hence trivially satisfy the Lindeberg condition, while the Lindeberg condition on the variables $g_\sigma(M_{N_n,j})$ follows from the boundedness of $\alpha_0(n)^{-1}\sum_j \mathrm{E}g_\sigma(M_{N_n,j})^3$, by Lemma 4.3.3 (viii). Thus the Poissonized sums are asymptotically normal, by Lemma 4.3.7.

The proof can be completed by applying Lemma 4.3.6 to the variables $V_{k,n} = \sum_{j=1}^{\infty}\bigl(1_{M_{k,j}\geq 1}/\sigma_n - g_{\sigma_n}(M_{k,j})\bigr)$, with $a_n = \alpha_0(n)^{-1/2}$ and $E_{0,n}$ as given. To show that $a_n(V_{k_n,n} - V_{n,n})$ tends to zero in probability, we split $V_{k,n}$ in $\sum_j 1_{M_{k,j}\geq 1}/\sigma_n$ and $\sum_j g_{\sigma_n}(M_{k,j})$ and handle the two parts separately. Because $a_{k_n}/a_n \to 1$, it is not a loss of generality to assume that $k_n \geq n$. Because the binomial distributions binomial(n, p_j) are stochastically ordered in n and the functions $m \mapsto 1_{m\geq 1}$ and $m \mapsto g_\sigma(m)$ are increasing, the variables $a_n\sum_j(1_{M_{k,j}\geq 1} - 1_{M_{n,j}\geq 1})/\sigma$ and $a_n\sum_j(g_{\sigma_n}(M_{k,j}) - g_{\sigma_n}(M_{n,j}))$ are nonnegative, and hence it suffices to show that their expectations tend to zero. This is shown in Lemma 4.3.5. ∎

Lemma 4.3.2. *For any $\tilde\sigma_n \xrightarrow{P} \sigma_0 \in (0,1)$ and $M_n = o\bigl(\alpha_0(n)/\log n\bigr)$, we have $\alpha_0(n)^{-1}\Lambda_n''(\tilde\sigma_n, M_n) \to E_0'(\sigma_0)$, in probability.*

Proof. The second derivative is given by

$$\begin{aligned}\Lambda_n''(\sigma) &= -\sum_{l=1}^{K_n-1}\frac{l^2}{(M+l\sigma)^2} - \sum_{l=1}^{n-1}\frac{1}{(l-\sigma)^2}Z_{n,l+1}\\ &= -\frac{K_n-1}{\sigma^2} + \frac{1}{\sigma^2}\sum_{l=1}^{K_n-1}\Big[\frac{2M}{M+\sigma l} - \frac{M^2}{(M+\sigma l)^2}\Big] - \sum_{l=1}^{n-1}\frac{1}{(l-\sigma)^2}Z_{n,l+1}.\end{aligned}$$

It is shown in [42] (see his formula (66), or see the proof of Lemma 4.3.1), that $K_n/\alpha_0(n) \to \Gamma(1-\sigma_0)$ and $Z_{n,l}/\alpha_0(n) \to \Gamma(l-\sigma_0)/(l-1)!$, for every $l \geq 1$, in probability and in mean. We use this to infer the convergence of the first and last terms on the right divided by $\alpha_0(n)$. That the limit is equal to $E_0'(\sigma_0)$ follows by inspection of its form and Lemma 4.3.4. The second term is bounded above in absolute value by a multiple of $M \log K_n$ and divided by $\alpha_0(n)$ tends to zero. ■

4.3.5 Technical lemmas

In the next lemmas $(p_j)_{j=1}^{\infty}$ is a given infinite probability vector and α is the cumulative distribution function of the counting measure on the points $1/p_j$, for $j \in \mathbb{N}_+$. Furthermore, the function $g_\sigma\colon \mathbb{N} \to \mathbb{N}$ is given by $g_\sigma(1) = g_\sigma(2) = 0$ and

$$g_\sigma(m) = \sum_{l=1}^{m-1}\frac{1}{l-\sigma}, \qquad m \geq 2.$$

Lemma 4.3.3. *Suppose that $\alpha(u) := \#\{j: 1/p_j \leq u\}$ is regularly varying at ∞ of order $\gamma \in (0,1)$. Then, for any $\sigma_n \to \sigma \in (0,1)$, and independent $M_{n,j} \sim$ Poisson(np_j),*

(i) $\frac{1}{\alpha(n)}\sum_{j=1}^{\infty} \mathrm{E}1_{M_{n,j}\geq 1} \to \Gamma(1-\gamma)$,

(ii) $\frac{1}{\alpha(n)}\sum_{j=1}^{\infty} \mathrm{var}\,1_{M_{n,j}\geq 1} \to (2^\gamma - 1)\Gamma(1-\gamma)$,

(iii) $\frac{1}{\alpha(n)}\sum_{j=1}^{\infty} \mathrm{E}g_{\sigma_n}(M_{n,j}) \to \sum_{m=1}^{\infty}\frac{\Gamma(m+1-\gamma)}{m!(m-\sigma)}$,

(iv) $\frac{1}{\alpha(n)}\sum_{j=1}^{\infty} \mathrm{E}\frac{\partial g_{\sigma_n}}{\partial\sigma}(M_{n,j}) \to \sum_{m=1}^{\infty}\frac{\Gamma(m+1-\gamma)}{m!(m-\sigma)^2}$,

(v) $\frac{1}{\alpha(n)}\sum_{j=1}^{\infty} \mathrm{E}g_{\sigma_n}^2(M_{n,j}) \to \sum_{m=1}^{\infty}\frac{\Gamma(m+1-\gamma)(g_\sigma(m+1)+g_\sigma(m))}{m!(m-\sigma)}$,

(vi) $\frac{1}{\alpha(n)}\sum_{j=1}^{\infty} \big(\mathrm{E}g_{\sigma_n}(M_{n,j})\big)^2 \to \sum_{k=2}^{\infty}\sum_{m=1}^{\infty}\frac{g_\sigma(k)\Gamma(k+m+1-\gamma)}{k!m!2^{k+m-\gamma}(m-\sigma)}$.

(vii) $\frac{1}{\alpha(n)}\sum_{j=1}^{\infty} e^{-np_j}\mathrm{E}g_{\sigma_n}(M_{n,j}) \to \sum_{m=2}^{\infty}\frac{g_\sigma(m)\gamma\Gamma(m-\gamma)}{m!\,2^{m-\gamma}}$.

(viii) $\frac{1}{\alpha(n)}\sum_{j=1}^{\infty} \mathrm{E}g_{\sigma_n}^3(M_{n,j}) \to \sum_{m=1}^{\infty}\frac{\Gamma(m+1-\gamma)(g_\sigma^2(m+1)+g_\sigma(m+1)g_\sigma(m)+g_\sigma^2(m))}{m!(m-\sigma)}$,

All limits on the right sides are finite.

Proof. Assertions (i) and (ii) were stated in [42]; we include proofs for completeness.

The series in the left side of (i) is

$$\sum_{j=1}^{\infty} \mathrm{P}(M_{n,j} \geq 1) = \sum_{j=1}^{\infty}(1 - e^{-np_j}) = \int_1^{\infty}(1 - e^{-n/u})\, d\alpha(u) = \int_0^n \alpha\Big(\frac{n}{s}\Big)e^{-s}\, ds,$$

by Fubini's theorem (or partial integration), since $1 - e^{-n/u} = \int_0^{n/u} e^{-s}\, ds$. By the definition of regular variation $\alpha(n/s)/\alpha(n) \to s^{-\gamma}$, for every s, as $n \to \infty$. By Potter's theorem ([6], Theorem 1.5.6), for every $\delta > 0$ there exists $M > 1$ such that $\alpha(n/s)/\alpha(n) \leq s^{-\gamma-\delta} \vee s^{-\gamma+\delta}$, for every $s < n/M$. We can choose δ so that $\gamma + \delta < 1$, and then $\int_0^{\infty}(s^{-\gamma-\delta} \vee s^{-\gamma+\delta})e^{-s}\, ds < \infty$. For the corresponding M, we then have $\int_0^{n/M} \alpha(n/s)/\alpha(n)\, e^{-s}\, ds \to \int_0^{\infty} s^{-\gamma}e^{-s}\, ds = \Gamma(1-\gamma)$, by the dominated convergence theorem. For $s \geq n/M$, we have $\alpha(n/s) \leq \alpha(M)$ and hence $\int_{n/M}^n \alpha(n/s)e^{-s}\, ds \leq \alpha(M)e^{-n/M} = o(\alpha(n))$, as $n \to \infty$.

The series in (ii) is $\sum_{j=1}^{\infty} e^{-np_j}(1-e^{-np_j}) = \sum_{j=1}^{\infty}(1-e^{-2np_j}) - \sum_{j=1}^{\infty}(1-e^{-np_j})$. By the first paragraph this is asymptotic to $\big(\alpha(2n)-\alpha(n)\big)\Gamma(1-\gamma) \sim (2^{\gamma}-1)\alpha(n)\Gamma(1-\gamma)$, by regular variation of α.

For (iii) we write

$$\sum_{j=1}^{\infty} \mathrm{E}g_{\sigma}(M_{n,j}) = \sum_{j=1}^{\infty}\sum_{m=2}^{\infty} g_{\sigma}(m)\frac{e^{-np_j}(np_j)^m}{m!} = \sum_{m=2}^{\infty} \frac{g_{\sigma}(m)}{m!}\int_1^{\infty} e^{-n/u}\Big(\frac{n}{u}\Big)^m d\alpha(u).$$

Substituting $e^{-n/u}(n/u)^m = \int_0^{n/u} e^{-s}s^{m-1}(m-s)\, ds$ (valid for $m > 0$) and using Fubini's theorem, we can rewrite the right side as

$$\begin{aligned}
&\sum_{m=2}^{\infty} \frac{g_{\sigma}(m)}{m!}\int_0^n \alpha\Big(\frac{n}{s}\Big)e^{-s}s^{m-1}(m-s)\, ds \\
&= g_{\sigma}(2)\int_0^n \alpha\Big(\frac{n}{s}\Big)e^{-s}s\, ds + \sum_{m=2}^{\infty} \frac{g_{\sigma}(m+1) - g_{\sigma}(m)}{m!}\int_0^n \alpha\Big(\frac{n}{s}\Big)e^{-s}s^m\, ds \\
&= \int_0^n \alpha\Big(\frac{n}{s}\Big)e^{-s}\Big(\sum_{m=1}^{\infty} \frac{s^m}{m!(m-\sigma)}\Big)\, ds.
\end{aligned}$$

As before, regular variation and Potter's theorem give for $s \leq n/M$ the bound $\alpha(n/s)/\alpha(n) \lesssim s^{-\gamma-\delta} \vee s^{-\gamma+\delta}$, and then

$$\frac{\alpha(n/s)}{\alpha(n)}\Big(\sum_{m=1}^{\infty} \frac{s^m}{m!(m-\sigma)}\Big) \lesssim (s^{-\gamma-\delta} \vee 1)(e^s - 1 - s)\frac{1}{s}, \qquad s \leq n/M.$$

Furthermore, the left side tends pointwise to $s^{-\gamma}\sum_{m=1}^\infty s^m/(m!(m-\sigma))$. By the dominated convergence theorem,

$$\int_0^{n/M}\frac{\alpha(n/s)}{\alpha(n)}e^{-s}\Bigl(\sum_{m=1}^\infty\frac{s^m}{m!(m-\sigma)}\Bigr)\,ds\to\int_0^\infty s^{-\gamma}e^{-s}\sum_{m=1}^\infty\frac{s^m}{m!(m-\sigma)}\,ds.$$

The right side is the limit as given. Since $\sum_j p_j=1$, we have that $\alpha(u)=\#(p_j\geq 1/u)\leq u$. Therefore

$$\int_{n/M}^n\alpha\Bigl(\frac{n}{s}\Bigr)e^{-s}\Bigl(\sum_{m=1}^\infty\frac{s^m}{m!(m-\sigma)}\Bigr)\,ds\leq\int_{n/M}^\infty\frac{n}{s}\frac{1}{s}\,ds\leq M.$$

This is of lower order than $\alpha(n)$ and hence the proof of the third assertion is complete.

For (iv) we follow the same approach as in (iii), replacing g_σ by $\dot g_\sigma=\partial/\partial\sigma g_s$, and then at the end substitute $\dot g_s(m+1)-\dot g_\sigma(m)=1/(m-\sigma)^2$.

For (v) again we follow the same approach as under (iii), now replacing g_σ by g_σ^2. At the end we write the difference $g_\sigma^2(m+1)-g_\sigma^2(m)$ as $(m-\sigma)^{-1}\bigl(g_\sigma(m+1)+g_\sigma(m)\bigr)$ and complete the argument as before, where we can bound $g_\sigma(m+1)+g_\sigma(m)$ by a multiple of $\log m$, for large m, and use that $\sum_m s^m\log m/m!\lesssim e^s(s^\delta\vee 1)$, for every s, by Lemma 4.3.11, with a sufficiently small $\delta>0$.

The series in (vi) is equal to

$$\begin{aligned}\sum_{j=1}^\infty\sum_{k=2}^\infty\sum_{m=2}^\infty g_\sigma(k)g_\sigma(m)&\frac{e^{-2np_j}(np_j)^{k+m}}{k!m!}\\&=\sum_{k=2}^\infty\sum_{m=2}^\infty\frac{g_\sigma(k)g_\sigma(m)}{k!\,m!}\int_1^\infty e^{-2n/u}\Bigl(\frac{n}{u}\Bigr)^{k+m}\,d\alpha(u).\end{aligned}$$

Substituting $e^{-2n/u}(2n/u)^{k+m}=\int_0^{2n/u}e^{-s}s^{k+m-1}(k+m-s)\,ds$ and using Fubini's theorem, we can rewrite the right side as

$$\begin{aligned}\sum_{k=2}^\infty\sum_{m=2}^\infty&\frac{g_\sigma(k)g_\sigma(m)}{k!\,m!\,2^{k+m}}\int_0^{2n}\alpha\Bigl(\frac{2n}{s}\Bigr)e^{-s}s^{k+m-1}(k+m-s)\,ds\\&=\int_0^{2n}\alpha\Bigl(\frac{2n}{s}\Bigr)\frac{d}{ds}\Bigl[\Bigl(\sum_{k=2}^\infty\frac{g_\sigma(k)s^k}{k!\,2^k}\Bigr)^2e^{-s}\Bigr]\,ds\\&=\int_0^{2n}\alpha\Bigl(\frac{2n}{s}\Bigr)\Bigl(\sum_{k=2}^\infty\frac{g_\sigma(k)s^k}{k!\,2^k}\Bigr)e^{-s}\Bigl[\sum_{k=1}^\infty\frac{\bigl(g_\sigma(k+1)-g_\sigma(k)\bigr)s^k}{k!\,2^k}\Bigr]\,ds.\end{aligned}$$

In view of Lemma 4.3.11 and because $g_\sigma(k+1)-g_\sigma(k)=1/(k-\sigma)$, the integrand is bounded above by a multiple of $\alpha(2n/s)(e^{s/2}-1)e^{-s}s^{-1}(e^{s/2}-1)$. Using the dominated convergence theorem and arguments as before, we see that the right side divided by $\alpha(n)$ is asymptotic to the right side of (vi).

The extra factor e^{-np_j} in (vii) relative to (iii) leads to the same expression as in (iii), except that $e^{-n/u}$ must be replaced by $e^{-2n/u}$. Following the same argument, we find that the series in (vii) is equal to

$$\sum_{m=2}^{\infty}\frac{g_\sigma(m)}{m!\,2^m}\int_0^{2n}\alpha\Bigl(\frac{2n}{s}\Bigr)e^{-s}s^{m-1}(m-s)\,ds.$$

The functions $\sum_{m=2}^{\infty}\frac{g_\sigma(m)}{m!2^m}e^{-s}s^{m-1}|m-s|$ are uniformly integrable (thanks to the factors 2^m that are extra relative to (iii)) . Therefore, by arguments as before the display is asymptotically equivalent to the expression obtained by replacing $\alpha(2n/s)$ by $\alpha(n)(2/s)^\gamma$. Finally we can use that $m\Gamma(m-\gamma)-\Gamma(m+1-\gamma)=\gamma\Gamma(m-\gamma)$.

For (viii) we follow the same approach as under (iii), replacing g_σ by g_σ^3, where at the end we write the difference $g_\sigma^3(m+1)-g_\sigma^3(m)$ as $(m-\sigma)^{-1}(g_\sigma^2(m+1)+g_\sigma(m+1)g_\sigma(m)+g_\sigma^2(m))$.

The finiteness of the limitss can be proved with the help of Lemma 4.3.9 by comparison with standard series. ∎

For $\alpha(u){:=}\#\{j{:}\,1/p_j\le u\}$ as in the preceding, define a function E_n by

$$E_n(\sigma)=\int_0^n\alpha\Bigl(\frac{n}{s}\Bigr)e^{-s}\Bigl(\frac{1}{\sigma}-\sum_{m=1}^{\infty}\frac{s^m}{m!(m-\sigma)}\Bigr)\,ds. \tag{4.14}$$

Lemma 4.3.4. *If the function α is regularly varying at ∞ of order $\gamma\in(0,1)$, then the functions E_n in* (4.14) *satisfy, as $n\to\infty$, for $\sigma_n\to\sigma\in(0,1)$,*

$$\begin{aligned}\frac{E_n(\sigma_n)}{\alpha(n)}&\to E(\sigma){:=}\frac{\Gamma(1-\gamma)}{\sigma}-\sum_{m=1}^{\infty}\frac{\Gamma(m+1-\gamma)}{m!(m-\sigma)}, \qquad (4.15)\\ \frac{E_n'(\sigma_n)}{\alpha(n)}&\to E'(\sigma).\end{aligned}$$

The limit function E vanishes at $\sigma=\gamma$ and the zeros $\sigma_{0,n}$ of E_n satisfy $\sigma_{0,n}\to\gamma$. Furthermore, if there exists a continuously differentiable function $L{:}\,[1,\infty)\to\mathbb{R}$ such that $|\alpha(u)-u^\gamma L(u)|\le Cu^\beta$, for every $u>1$, and some $C>0$ and $\beta<\gamma$, and such that $s\mapsto L'(s)s$ is slowly varying at ∞, then, as $n\to\infty$,

$$\sigma_{0,n}-\gamma=O\Bigl(\frac{n^{\beta-\gamma}}{L(n)}\Bigr)-\frac{\Gamma(1-\gamma)(1+\gamma)}{\gamma^2E'(\gamma)}\Bigl(\frac{L'(n)n}{L(n)}\Bigr)\bigl(1+o(1)\bigr).$$

In particular, if L can be taken constant, then $\sigma_{0,n}-\gamma=O(n^{\beta-\gamma})$.

Proof. As shown in the proof of Lemma 4.3.3 (i) and (iii), the function $E_n(\sigma)$ is equal to $\sum_{j=1}^\infty \bigl(\mathrm{E}1_{M_{n,j}\geq 1}/\sigma - \mathrm{E}g_\sigma(M_{n,j})\bigr)$, with the notation as in the lemma. Therefore, by the lemma the limit function E is equal to the limit in (i) divided by σ minus the limit in (iii), i.e. the right side of (4.15). The limit in (iii) at $\sigma=\gamma$ can be written

$$\sum_{m=1}^\infty \frac{\Gamma(m-\gamma)}{m!} = \sum_{m=1}^\infty \int_0^\infty \frac{s^{m-\gamma-1}}{m!} e^{-s}\,ds = \int_0^\infty (1-e^{-s})s^{-\gamma-1}\,ds.$$

By partial integration, this can be further rewritten as $\int_0^\infty x^{-\gamma}/\gamma\, e^{-x}\,dx = \Gamma(1-\gamma)/\gamma$. Thus $E(\gamma) = \Gamma(1-\gamma)/\gamma - \Gamma(1-\gamma)/\gamma = 0$.

The limit of $E_n'(\sigma_n)/\alpha(n)$ is obtained similarly from (i) and (iv) of Lemma 4.3.3, and it is seen to be equal to the derivative $E'(\sigma)$.

The functions E_n and E are monotonely decreasing and $E_n(\gamma-\varepsilon)/\alpha(n)\to E(\gamma-\varepsilon) > 0$ and $E_n(\gamma+\varepsilon)/\alpha(n)\to E(\gamma+\varepsilon)<0$, for every $\varepsilon>0$, by (4.15). This shows that $\sigma_{0,n}\in(\gamma-\varepsilon,\gamma+\varepsilon)$ eventually and hence $\sigma_{0,n}\to\gamma$.

By the mean value theorem, $E_n(\sigma_{0,n}) - E_n(\gamma) = E_n'(\tilde\sigma_n)(\sigma_{0,n}-\gamma)$, for some $\tilde\sigma_n\to\gamma$. Since $E_n(\sigma_{0,n}) = 0$ and $E_n'(\tilde\sigma_n)/\alpha(n)\to E'(\gamma)<0$, it follows that $\sigma_{0,n}-\gamma = -\bigl(E_n(\gamma)/\alpha(n)\bigr)/\bigl(E'(\gamma)+o(1)\bigr)$. This shows that $\sigma_{0,n}\to\gamma$ at the same rate of convergence as $E_n(\gamma)/\alpha(n)\to E(\gamma)=0$.

To investigate the latter rate, define functions H_n by

$$H_n(\sigma) = \int_0^n L\Bigl(\frac{n}{s}\Bigr) s^{-\gamma} e^{-s}\Bigl(\frac{1}{\sigma} - \sum_{m=1}^\infty \frac{s^m}{m!(m-\sigma)}\Bigr)\,ds.$$

The functions H_n are monotonely decreasing, with, by the assumption on α,

$$\begin{aligned}\bigl|E_n(\sigma) - n^\gamma H_n(\sigma)\bigr| &\leq \int_0^n \Bigl(\frac{n}{s}\Bigr)^\beta e^{-s}\Bigl|\frac{1}{\sigma} - \sum_{m=1}^\infty \frac{s^m}{m!(m-\sigma)}\Bigr|\,ds\\ &\lesssim n^\beta\Bigl(\frac{1}{\sigma} + \frac{1}{1-\sigma}\int_0^n \frac{1}{s^{1+\beta}} e^{-s}(e^s-1-s)\,ds\Bigr).\end{aligned}$$

The right side is $O(n^\beta)$, if σ is bounded away from 0 and 1.

Let $\bar H_n$ be defined as H_n, but with the integral extended from $(0,n)$ to the full half line $(0,\infty)$. The assumption on α does not specify $L(u)$ for $u\leq 1$, and hence does not specify $L(n/s)$ for $s\geq n$, but we can extend the function L to a continuously differentiable function on $(0,\infty)$ in such a way that it vanishes on a neighbourhood of 0 and hence is uniformly bounded on $[0,1]$. Then $|H_n(\sigma) - \bar H_n(\sigma)| \lesssim \int_n^\infty s^{-\gamma}e^{-s}(1+s^{-1}(e^s-1))\,ds = O(n^{-\gamma})$, if σ is bounded away from 0 and 1.

Splitting the two parts of the integrand in $\bar{H}_n$ and performing partial integration on the second part we find

$$\begin{aligned}\bar{H}_n(\gamma) &= \int_0^\infty L\Bigl(\frac{n}{s}\Bigr)\frac{s^{-\gamma}}{\gamma}e^{-s}\,ds - \int_0^\infty L\Bigl(\frac{n}{s}\Bigr)\sum_{m=1}^\infty \frac{s^{m-\gamma-1}}{m!}\,e^{-s}\,ds \\ &\qquad\qquad + \int_0^\infty L'\Bigl(\frac{n}{s}\Bigr)\frac{n}{s^2}\sum_{m=1}^\infty \frac{s^{m-\gamma}}{m!(m-\gamma)}\,e^{-s}\,ds \\ &= \int_0^\infty L\Bigl(\frac{n}{s}\Bigr)\frac{s^{-\gamma}}{\gamma}e^{-s}\,ds + \int_0^\infty L\Bigl(\frac{n}{s}\Bigr)(1-e^{-s})\,d\Bigl(\frac{s^{-\gamma}}{\gamma}\Bigr) \\ &\qquad\qquad + \int_0^\infty L'\Bigl(\frac{n}{s}\Bigr)\frac{n}{s^2}\sum_{m=1}^\infty \frac{s^{m-\gamma}}{m!(m-\gamma)}\,e^{-s}\,ds \\ &= 0 + \int_0^\infty L'\Bigl(\frac{n}{s}\Bigr)\frac{n}{s^2}\Bigl[(1-e^{-s})\frac{s^{-\gamma}}{\gamma} + \sum_{m=1}^\infty \frac{s^{m-\gamma}e^{-s}}{m!(m-\gamma)}\Bigr]\,ds \\ &\sim L'(n)n\int_0^\infty \frac{1}{s}\Bigl[(1-e^{-s})\frac{s^{-\gamma}}{\gamma} + \sum_{m=1}^\infty \frac{s^{m-\gamma}e^{-s}}{m!(m-\gamma)}\Bigr]\,ds,\end{aligned}$$

since the function $s \mapsto L'(s)s$ is slowly varying at infinity, so that $L'(n/s)n/s \sim L'(n)n$ as $n \to \infty$, for every $s > 0$. To justify the last step we can use Potter's theorem ([6], Theorem 1.5.6) as before, to infer for every $\delta > 0$ the existence of a constant $M > 1$ such that $L'(n/s)n/s/(L'(n)n) \lesssim s^\delta \vee s^{-\delta}$, for $s < n/M$ and $n \geq M$, so that on this interval the integrand is dominated by a multiple of $(s^\delta \vee s^{-\delta})\bigl[s^{-1}(1-e^{-s})s^{-\gamma} + \sum_{m=1} s^{m-\gamma}e^{-s}/(m+1)!\bigr] \lesssim s^{-\gamma-\delta} \wedge s^{-1-\gamma+\delta}$, which is integrable for sufficiently small $\delta > 0$. Furthermore, for $s > n/M$, the function $|L'(n/s)n/s|$ is uniformly bounded, whence the integral over the interval $[n/M, \infty)$ is bounded above by a multiple of $\int_{n/M}^\infty s^{-1-\gamma}\,ds \lesssim n^{-\gamma} \ll L'(n)n$.

By the identities obtained in the beginning of the proof, the integral in the right of the preceding display is identical to $\Gamma(1-\gamma)/\gamma^2 + \Gamma(1-\gamma)/\gamma = \Gamma(1-\gamma)(1+\gamma)/\gamma^2$.

Combining the preceding, we find that $E_n(\gamma) = n^\gamma\bigl(\bar{H}_n(\gamma) + O(n^{-\gamma})\bigr) + O(n^\beta) = n^\gamma L'(n)n + O(n^\beta)$ and hence $E_n(\gamma)/\alpha(n) = O(n^\gamma L'(n)n/\alpha(n)) + O(n^\beta/\alpha(n))$. ■

Lemma 4.3.5. *Suppose that $\alpha(u) := \#\{j : 1/p_j \leq u\}$ is regularly varying at ∞ of order $\gamma \in (0,1)$. Then for any $\sigma_n \to \sigma \in (0,1)$, and independent $M_{n,j} \sim$ Binomial(n, p_j), and $k_n \geq n$ with $k_n - n = O(\sqrt{n})$,*

$$\sum_{j=1}^\infty \mathrm{E}(1_{M_{k_n,j}\geq 1} - 1_{M_{n,j}\geq 1}) = o\bigl(\alpha(n)^{1/2}\bigr),$$

Furthermore, if there exists a continuously differentiable function $L : [1,\infty) \to \mathbb{R}$ such that $|\alpha(u) - u^\gamma L(u)| \leq Cu^\beta$, for every $u > 1$, and some $C > 0$ and $\beta < \gamma$, and

$|L'(u)| \leq C_\delta u^{-1+\delta}$, *for every* $u > 1$ *and* $\delta > 0$ *and some* $C_\delta > 0$, *then*

$$\sum_{j=1}^{\infty} \mathrm{E}\big(g_{\sigma_n}(M_{k_n,j}) - g_{\sigma_n}(M_{n,j})\big) = o\big(\alpha(n)^{1/2}\big).$$

Proof. Because $\mathrm{P}(M_{n,j} = 0) = (1-p_j)^n$, the left side of the first assertion is equal to

$$\sum_{j=1}^{\infty}\big((1-p_j)^n - (1-p_j)^{k_n}\big) = \int_1^\infty \Big(1 - \frac{1}{u}\Big)^n \Big(1 - \Big(1 - \frac{1}{u}\Big)^{k_n - n}\Big)\, d\alpha(u).$$

By the inequalities $1 - x \leq e^{-x}$, for $x \in \mathbb{R}$, and $1 - (1-x)^r \leq rx$, for $x \in [0,1]$ and $r \in \mathbb{N}$, this is bounded above by

$$\int_1^\infty e^{-n/u}(k_n - n)\frac{1}{u}\, d\alpha(u) = \frac{k_n - n}{n} \int_0^n \alpha\Big(\frac{n}{s}\Big) e^{-s}(1-s)\, ds,$$

by Fubini's theorem, since $e^{-n/u}(n/u) = \int_0^{n/u} e^{-s}(1-s)\, ds$. As in the proof of Lemma 4.3.3, the integral is $\alpha(n)\big(\Gamma(1-\gamma) - \Gamma(2-\gamma)\big)(1 + o(1))$. Therefore, the preceding display divided by $\alpha(n)^{1/2}$ is of the order $\alpha(n)^{1/2}(k_n - n)/n \sim \alpha(n)^{1/2} n^{-1/2}$. This tends to zero, as for every $\delta > 0$ we have that $\alpha(n) \leq n^{\gamma+\delta}$ eventually, by Potter's theorem, where $\gamma < 1$ by assumption.

To prove the second assertion we first write

$$\begin{aligned}\sum_{j=1}^{\infty} \mathrm{E} g_{\sigma_n}(M_{n,j}) &= \sum_{j=1}^{\infty}\sum_{m=2}^{n} g_\sigma(m)\binom{n}{m} p_j^m (1-p_j)^{n-m} \\ &= \sum_{m=2}^{n} g_\sigma(m)\binom{n}{m} \int_1^\infty \Big(\frac{1}{u}\Big)^m \Big(1 - \frac{1}{u}\Big)^{n-m} d\alpha(u).\end{aligned}$$

Writing $(1/u)^m(1-1/u)^{n-m} = \int_0^{1/u} s^{m-1}(1-s)^{n-m-1}(m - ns)\, ds$ (for $m > 0$) and applying Fubini's theorem, we can rewrite this as

$$\begin{aligned}&\sum_{m=2}^{n} g_\sigma(m)\binom{n}{m} \int_0^1 \alpha\Big(\frac{1}{s}\Big) s^{m-1}(1-s)^{n-m-1}(m - ns)\, ds \\ &\quad = \int_0^1 \sum_{l=1}^{n-1} \frac{1}{l-\sigma} \sum_{m=l+1}^{n} \binom{n}{m} s^{m-1}(1-s)^{n-m-1}(m-ns)\, \alpha\Big(\frac{1}{s}\Big)\, ds \\ &\quad = \int_0^1 \sum_{l=1}^{n-1} \frac{n-l}{l-\sigma}\binom{n}{l} s^l (1-s)^{n-l-1}\, \alpha\Big(\frac{1}{s}\Big)\, ds,\end{aligned}$$

by Lemma 4.3.8. Thus the left side of the second assertion can be written in the form, with $k = k_n$,

$$\int_0^1 \sum_{l=1}^{n-1} \frac{s^l(1-s)^{n-l-1}}{l-\sigma}\Big[(k-l)\binom{k}{l}(1-s)^{k-n} - (n-l)\binom{n}{l}\Big]\,\alpha\Big(\frac{1}{s}\Big)\,ds \\ + \int_0^1 \sum_{l=n+1}^{k} \frac{k-l}{l-\sigma}\binom{k}{l} s^l(1-s)^{k-l-1}\,\alpha\Big(\frac{1}{s}\Big)\,ds. \tag{4.16}$$

Because $\alpha(1/s) \leq 1/s$, the second term is bounded above by $\sum_{l>n}(k-l)/(l-\sigma)\binom{k}{l}B(l,k-l)$, for B the beta function. This is further bounded by $k\sum_{l>n} 1/((l-\sigma)l) \lesssim k/n \lesssim 1$.

By the assumption that $|\alpha(u) - u^\gamma L(u)| \leq Cu^\beta$, if in the first term, we replace $\alpha(1/s)$ by $s^{-\gamma}L(1/s)$, the error is bounded above by a multiple of

$$\int_0^1 \sum_{l=1}^{n-1} \frac{s^l(1-s)^{n-l-1}}{l-\sigma}\Big|(k-l)\binom{k}{l}(1-s)^{k-n} - (n-l)\binom{n}{l}\Big|\,s^{-\beta}\,ds.$$

The sum of the terms with $l > \sqrt{n}$ is bounded above by $a_{k,n} + a_{n,n}$, for

$$a_{k,n} = \sum_{l>\sqrt{n}} \frac{k-l}{l-\sigma}\binom{k}{l}B(l-\beta+1,k-l) \lesssim \sum_{l>\sqrt{n}} \frac{\Gamma(l-\beta+1)}{(l-\sigma)l!}\frac{k!}{\Gamma(k-\beta+1)}.$$

In view of Lemma 4.3.9, $a_{k,n}$ is bounded above by a multiple of $\sqrt{n}^{-\beta}k^\beta = O(n^{\beta/2}) = o\big(\alpha(n)^{1/2}\big)$. In the sum of the terms with $l \leq \sqrt{n}$, we decompose $k-l = (k-n)+(n-l)$ and $\binom{k}{l} = \sum_i \binom{n}{l-i}\binom{k-n}{i}$, and bound

$$\Big|(k-l)\binom{k}{l}(1-s)^{k-n} - (n-l)\binom{n}{l}\Big| \leq (k-n)\binom{k}{l}(1-s)^{k-n} \\ + (n-l)\Big[\binom{n}{l}\big(1-(1-s)^{k-n}\big) + \sum_{i\geq 1}\binom{n}{l-i}\binom{k-n}{i}\Big].$$

The middle term in the right is bounded above by $(n-l)\binom{n}{l}(k-n)s$. Thus the sum of the terms with $l \leq \sqrt{n}$ is bounded above by the sum of the three integrals

$$\begin{aligned}
&\int_0^1 \sum_{l\leq\sqrt{n}} \frac{s^l(1-s)^{n-l-1}}{l-\sigma}(k-n)\binom{k}{l}(1-s)^{k-n}s^{-\beta}ds\\
&= \sum_{l\leq\sqrt{n}} \frac{B(l+1-\beta,k-l)}{l-\sigma}(k-n)\binom{k}{l} = \sum_{l\leq\sqrt{n}} \frac{\Gamma(l+1-\beta)}{(l-\sigma)l!}\frac{\Gamma(k-l)}{(k-l)!}\frac{(k-n)k!}{\Gamma(k+1-\beta)}\\
&\qquad\lesssim \sum_{l\leq\sqrt{n}} \frac{1}{l^{1+\beta}}\frac{k-n}{k-l}k^{\beta} \lesssim n^{\beta-1/2} \leq n^{\beta/2},\\
&\int_0^s \sum_{l\leq\sqrt{n}} \frac{s^l(1-s)^{n-l-1}}{l-\sigma}(n-l)\binom{n}{l}(k-n)ss^{-\beta}ds\\
&= \sum_{l\leq\sqrt{n}} \frac{B(l+2-\beta,n-l)}{l-\sigma}(k-n)(n-l)\binom{n}{l} = \sum_{l\leq\sqrt{n}} \frac{\Gamma(l+2-\beta)}{(l-\sigma)l!}\frac{(k-n)n!}{\Gamma(n+2-\beta)}\\
&\qquad\lesssim \sqrt{n}^{1-\beta}(k-n)n^{\beta-1} \lesssim n^{\beta/2},\\
&\int_0^1 \sum_{l\leq\sqrt{n}} \frac{s^l(1-s)^{n-l-1}}{l-\sigma}(n-l)\sum_i \binom{n}{l-i}\binom{k-n}{i}s^{-\beta}ds\\
&= \sum_{l\leq\sqrt{n}}\sum_{i\geq 1} \frac{B(l+1-\beta,n-l)}{l-\sigma}(n-l)\binom{n}{l-i}\binom{k-n}{i}\\
&\qquad= \sum_{l\leq\sqrt{n}}\sum_{i\geq 1} \frac{\Gamma(l+1-\beta)}{(l-\sigma)(l-i)!}\frac{(n-l)!}{(n-l+i)!}\frac{n!}{\Gamma(n+1-\beta)}\binom{k-n}{i}\\
&\qquad\lesssim \sum_{l\leq\sqrt{n}}\sum_{i\geq 1} l^{i-1-\beta}\frac{1}{(n/2)^i}n^{\beta}\binom{k-n}{i} \leq \sum_{i\geq 1}\frac{\sqrt{n}^{i-\beta}}{(n/2)^i}n^{\beta}\binom{k-n}{i}\\
&\qquad\leq \Bigl(1+\frac{2}{\sqrt{n}}\Bigr)^{k-n} n^{\beta/2} \lesssim n^{\beta/2}.
\end{aligned}$$

We conclude that replacing $\alpha(1/s)$ by $s^{-\gamma}L(1/s)$ in the first part of (4.16) makes a difference of at most of the order $n^{\beta/2} = o\bigl(\alpha(n)^{1/2}\bigr)$. Finally, we consider the expression

$$\begin{aligned}
&\int_0^1 \sum_{l=1}^{n-1} \frac{s^l(1-s)^{n-l-1}}{l-\sigma}\Bigl[(k-l)\binom{k}{l}(1-s)^{k-n} - (n-l)\binom{n}{l}\Bigr]\, s^{-\gamma}L\Bigl(\frac{1}{s}\Bigr)\, ds\\
&\qquad= \sum_{l=1}^{n-1} \frac{\Gamma(l+1-\gamma)}{(l-\sigma)l!}\Bigl[\frac{k!}{\Gamma(k+1-\gamma)}\mathrm{E}L(1/S_{l,k}) - \frac{n!}{\Gamma(n+1-\gamma)}\mathrm{E}L(1/S_{l,n})\Bigr],
\end{aligned}$$

where $S_{l,k}$ is a random variable with the beta distribution with parameters $l+1-\gamma$ and $k-l$. The bound $|L'(u)| \leq C_\delta u^{-1+\delta}$ gives that $L(u) \lesssim u^{\delta}$, and hence $|\mathrm{E}L(1/S_{l,k})| \lesssim$

$\mathrm{E}S_{l,k}^{-\delta} = B(l+1-\gamma-\delta, k-l)/B(l+1-\gamma, k-l) \lesssim k^\delta$. Therefore, after bounding the difference with the help of the triangle inequality, the sum of the terms with $l > m$, can be bounded by $b_{k,m} + b_{n,m}$, for

$$b_{k,m} = \sum_{l>m} \frac{\Gamma(l+1-\gamma)}{(l-\sigma)l!} \frac{k!}{\Gamma(k+1-\gamma)} k^\delta \lesssim \Big(\frac{k}{m}\Big)^\gamma k^\delta.$$

For $m = n^{1/2+\delta/\gamma+\eta}$, the right side is of the order $n^{\gamma/2-\eta\gamma} = o\big(\alpha(n)^{1/2}\big)$, for any $\eta > 0$. The sum of the terms with $l \le m$ is bounded above by

$$\begin{aligned}\sum_{l=1}^m \frac{\Gamma(l+1-\gamma)}{(l-\sigma)l!} &\Big[\frac{k!}{\Gamma(k+1-\gamma)} - \frac{n!}{\Gamma(n+1-\gamma)}\Big]\mathrm{E}L(1/S_{l,k}) \\ &\qquad\qquad + \frac{n!}{\Gamma(n+1-\gamma)}\Big[\mathrm{E}L(1/S_{l,k}) - \mathrm{E}L(1/S_{l,n})\Big] \\ &\lesssim \sum_{l=1}^m \frac{1}{l^{1+\gamma}}\Big[k^\gamma - n^\gamma + O\Big(\frac{1}{n}\Big)\Big]k^\delta + \sum_{l=1}^m \frac{n^\gamma}{l^{1+\gamma}}\Big|\mathrm{E}L(1/S_{l,k}) - \mathrm{E}L(1/S_{l,n})\Big|.\end{aligned}$$

Here $k^\gamma - n^\gamma = n^\gamma\big((1+(k-n)/n)^\gamma - 1\big) \lesssim n^{\gamma-1/2}$, so that the first term is of the order $n^{\gamma-1/2}k^\delta = o\big(\alpha(n)^{1/2}\big)$, for sufficiently small $\delta > 0$, and hence is asymptotically negligible. For the second term we represent $S_{l,k}$ and $S_{l,n}$ using independent, gamma variables $\bar\Gamma_l$, Γ_{n-l} and Γ_{k-n} with shape parameters $l+1-\gamma$, $n-l$ and $k-n$, and write $|\mathrm{E}L(1/S_{l,k}) - \mathrm{E}L(1/S_{l,n})|$ as

$$\begin{aligned}\Big|\mathrm{E}L\Big(&\frac{\bar\Gamma_l + \Gamma_{n-l} + \Gamma_{k-n}}{\bar\Gamma_l}\Big) - \mathrm{E}L\Big(\frac{\bar\Gamma_l + \Gamma_{n-l}}{\bar\Gamma_l}\Big)\Big| \\ &= \Big|\mathrm{E}\int_0^{\Gamma_{k-n}} L'\Big(\frac{\bar\Gamma_l + \Gamma_{n-l} + u}{\bar\Gamma_l}\Big)\frac{1}{\bar\Gamma_l}\,du\Big| \\ &\lesssim \int_0^\infty \mathrm{P}(\Gamma_{k-n} > u)\mathrm{E}\frac{1}{(\bar\Gamma_l + \Gamma_{n-l} + u)^{1-\delta}}\frac{1}{\bar\Gamma_l^\delta}\,du \le \mathrm{E}\Gamma_{k-n}\mathrm{E}\frac{1}{\Gamma_{n-l}^{1-\delta}}\mathrm{E}\frac{1}{\bar\Gamma_l^\delta}.\end{aligned}$$

The three expecations can be computed explicitly in terms of the gamma function. Substituting the resulting expressions in the second sum of the second last display, we see that this is bounded above by

$$\sum_{l=1}^m \frac{n^\gamma(k-n)}{l^{1+\gamma}} \frac{\Gamma(n-l-1+\delta)}{\Gamma(n-l)} \frac{\Gamma(l+1-\gamma-\delta)}{\Gamma(l+1-\gamma)} \lesssim n^\gamma(k-n)\sum_{l=1}^m \frac{1}{l^{1+\gamma+\delta}(n-l)^{1-\delta}}.$$

Since the summation indices satisfy $l \le m \ll n$, so that $n - l \gtrsim n/2$, this is of the order $n^{-1/2+\gamma+\delta} = o\big(\alpha(n)^{1/2}\big)$, for sufficiently small δ. ∎

4.3.6 Supporting lemmas

Lemma 4.3.6. *Suppose that $V_{k,n}$, for $k, n \in \mathbb{N}$, are random variables independently of random variables $N_n \sim$ Poisson(n) so that $a_n(V_{N_n,n} - E_n) \rightsquigarrow N(0,\tau^2)$ and $a_n(V_{k_n,n} -$*

$V_{n,n}) \to 0$ in probability for every k_n with $|k_n - n| = O(\sqrt{n})$, for $n \to \infty$ and given numbers a_n and E_n. Then $a_n(V_{n,n} - E_n) \rightsquigarrow N(0,\tau^2)$.

Proof. For any Lipschitz function $h\colon \mathbb{R} \to [0,1]$ and k_n as given, as $n \to \infty$,

$$\Big|\mathrm{E}h\big(a_n(V_{k_n,n} - E_n)\big) - \mathrm{E}h\big(a_n(V_{n,n} - E_n)\big)\Big| \le \mathrm{E}a_n|V_{k_n,n} - V_{n,n}| \wedge 1 \to 0.$$

By the central limit theorem the probability $\mathrm{P}\big(|N_n - n| > \sqrt{n}M\big)$ can be made arbitrarily small uniformly in n by choosing sufficiently large M. Then $\mathrm{E}h\big(a_n(V_{n,n} - E_n)\big)$ is arbitrarily close to

$$\begin{aligned}\mathrm{E}h\big(a_n(V_{n,n} - E_n)\big) &\sum_{k:|k-n|\le\sqrt{n}M} \mathrm{P}(N_n = k)\\ &= \sum_{k:|k-n|\le\sqrt{n}M} \mathrm{E}h\big(a_n(V_{k,n} - E_n)\big)\mathrm{P}(N_n = k) + o(1),\end{aligned}$$

by the preceding display, as $n \to \infty$, for every fixed M. The sum in the right side is arbitrarily close to $\mathrm{E}h\big(a_n(V_{N_n,n} - E_n)\big)$ uniformly in n, if M is chosen sufficiently large, which tends to $\mathrm{E}h(\tau Z)$, for $Z \sim N(0,1)$ as $n \to \infty$, by assumption. We conclude that $\mathrm{E}h\big(a_n(V_{n,n} - E_n)\big)$ is arbitrarily close to $\mathrm{E}h(\tau Z)$, as $n \to \infty$. ∎

Lemma 4.3.7. *If $X_{n,1}, X_{n,2}, \ldots$ are independent random variables with $s_n^2 := \sum_{j=1}^\infty \operatorname{var} X_{n,j} < \infty$ and $s_n^{-2}\sum_{j=1}^\infty \mathrm{E}X_{n,j}^2 1_{|X_{n,j}|>\varepsilon s_n} \to 0$, for every $\varepsilon > 0$, then $\sum_{j=1}^\infty (X_{n,j} - \mathrm{E}X_{n,j})/s_n \rightsquigarrow N(0,1)$.*

Proof. The variables $Y_{n,j} = (X_{n,j} - \mathrm{E}X_{n,j})/s_n$ have mean zero and $\sum_j \mathrm{E}Y_{n,j}^2 = 1$. Choose integers $k_n \uparrow \infty$ such that $\sum_{j\le k_n} \mathrm{E}Y_{n,j}^2 \uparrow 1$. Then the sequence $\sum_{j>k_n} Y_{n,j}$ tends to zero in second mean and hence it suffices to show that $\sum_{j\le k_n} Y_{n,j} \rightsquigarrow N(0,1)$. The latter follows from the Lindeberg central limit theorem provided that $\sum_{j\le k_n} \mathrm{E}Y_{n,j}^2 1_{|Y_{n,j}|>\varepsilon} \to 0$, for every $\varepsilon > 0$. To see that this is satisfied, first note that the Lindeberg condition implies that $\max_j \mathrm{E}X_{n,j}^2/s_n^2 \to 0$ and hence both $s_n^{-2}\mathrm{E}\sum_j |\mathrm{E}X_{n,j}|^2 1_{|X_{n,j}|>\varepsilon s_n} \le o(1)\sum_j \mathrm{P}(|X_{n,j}| > \varepsilon s_n) \to 0$ and $|\mathrm{E}X_{n,j}| \le \varepsilon s_n$, for every j eventually, for every fixed $\varepsilon > 0$. This shows that the variables $X_{n,j}$ also satisfy the centered "infinite Lindeberg condition" $\sum_j \mathrm{E}Y_{n,j}^2 1_{|Y_{n,j}|>\varepsilon} \to 0$, for every $\varepsilon > 0$, which implies the Lindeberg condition for the finite array $Y_{n,1}, \ldots, Y_{n,k_n}$. ∎

Lemma 4.3.8. *For every $p \in [0,1]$ and $l \in \mathbb{N} \cup \{0\}$ and $n \in \mathbb{N}$,*

$$\sum_{m=l+1}^n \binom{n}{m} p^{m-1}(1-p)^{n-m-1}(m - np) = (n-l)\binom{n}{l} p^l (1-p)^{n-l-1}.$$

Proof. For X_{n-1} and X_n the numbers of successes in the first $n-1$ and n independent Bernoulli trials with success probability p, we have $\{X_n \geq l+1\} \subset \{X_{n-1} \geq l\}$ and $\{X_{n-1} \geq l\} - \{X_n \geq l+1\} = \{X_{n-1} = l, B_n = 0\}$, for B_n the outcome of the nth trial. This gives the identity $\mathrm{P}(X_{n-1} \geq l) - \mathrm{P}(X_n \geq l+1) = \mathrm{P}(X_{n-1} = l)(1-p)$. We multiply this by $n/(1-p)$ to obtain the identity given by the lemma, which we first rewrite using that $m\binom{n}{m} = n\binom{n-1}{m-1}$ and $(n-l)\binom{n}{l} = n\binom{n-1}{l}$. ∎

Lemma 4.3.9. *For every $\gamma \in (0,1)$, as $n \to \infty$,*

$$\frac{\Gamma(n-\gamma)n^\gamma}{\Gamma(n)} = 1 + O\Big(\frac{1}{n}\Big).$$

Proof. By Stirling's approximation, the quotient is

$$\begin{aligned}\frac{\sqrt{2\pi/(n-\gamma)}\,\left(\frac{n-\gamma}{e}\right)^{n-\gamma}\big(1+O(1/n)\big)n^\gamma}{\sqrt{2\pi/n}\,\left(\frac{n}{e}\right)^{n}\big(1+O(1/n)\big)} &= \Big(\frac{n-\gamma}{n}\Big)^n e^\gamma\Big(1+O\Big(\frac{1}{n}\Big)\Big)\\ &= \Big(e^{-\gamma}+O\Big(\frac{1}{n}\Big)\Big)e^\gamma\Big(1+O\Big(\frac{1}{n}\Big)\Big).\end{aligned}$$

∎

Lemma 4.3.10. *If $\alpha\colon (1,\infty) \to \mathbb{R}$ satisfies $|\alpha(u) - u^\gamma L(u)| \leq Cu^\beta$, for $u > 1$ and $\beta < \gamma$ and a slowly varying function L, then α is regularly varying of order γ.*

Proof. Since L is slowly varying, we have $L(n) \gtrsim n^{-\delta}$, for every $\delta > 0$, and hence $n^\gamma L(n) \gtrsim n^{\gamma-\delta} \gg n^\beta$, for sufficiently small $\delta > 0$. Since $|\alpha(n) - n^\gamma L(n)| \lesssim n^\beta$ by assumption, it follows that $\alpha(n) \gg n^\beta$ and $n^\gamma L(n)/\alpha(n) \to 1$, as $n \to \infty$. By the assumption we have $\alpha(nu) = (nu)^\gamma L(nu) + O(n^\beta) = (nu)^\gamma L(n)(1+o(1)) + O(n^\beta)$, for every u as $n \to \infty$, since L is slowly varying, and hence $\alpha(nu)/\alpha(n) = u^\gamma\big(n^\gamma L(n)/\alpha(n)\big)(1+o(1)) + O(n^\beta/\alpha(n)) \to u^\gamma$. ∎

Lemma 4.3.11. *For every $s \geq 0$ and $\delta \in (0,1]$:*

(i) $\sum_{m=1}^\infty s^m m^\delta/m! \leq s^\delta e^s$,

(ii) $\sum_{m=2}^\infty s^m m^\delta/m! \leq s^\delta(e^s-1)$.

Proof. Since $s^m m^\delta/m! = (s^m m/m!)^\delta (s^m/m!)^{1-\delta}$, Hölder's inequality with $p = 1/\delta$ and $q = 1/(1-\delta)$ gives that the sums on the left are bounded above by $\big(\sum_m (s^m m/m!)\big)^\delta \big(\sum_m (s^m/m!)\big)^{1-\delta}$, where the summation starts at $m = 1$ for (i) and at $m = 2$ for (ii). In the case of (i) the first series is bounded above by se^s and the second by $e^s - 1$, while in the case of (ii) the bounds $s(e^s-1)$ and $e^s - 1 - s$ pertain. ∎

Lemma 4.3.12. *For $K \to \infty$, we have* $\sum_{i=1}^{K-1} \log(M + i\sigma) = K \log K + K \log(\sigma/e) + (M/\sigma - 1/2) \log K + \log(\sqrt{2\pi}/\sigma) - \log \Gamma(1 + M/\sigma) + O(1/K)$*, where the remainder is bounded above by a universal multiple of* $(M/\sigma + 1)^2/K$*, for all* $M \geq 0, \sigma > 0$.

Proof. The sum is equal to $(K-1)\log\sigma + \log\Gamma(K + M/\sigma) - \log\Gamma(1+M/\sigma)$. By the expansion for the log Gamma function, the middle term can be expanded as

$$\log\Gamma\Big(K + \frac{M}{\sigma}\Big) = \Big(K + \frac{M}{\sigma} - \frac{1}{2}\Big)\log\Big(K + \frac{M}{\sigma}\Big) - \Big(K + \frac{M}{\sigma}\Big) + \log\sqrt{2\pi} + O\Big(\frac{1}{K}\Big),$$

where the remainder term is uniform in M and σ. Next expand $\log(K + M/\sigma)$ as $\log K + M/(\sigma K) + O((M/\sigma)^2/K^2)$. Finally we collect terms. ∎

Chapter 5

Prerequisite theory for the Deep learning paper

In this chapter, we quickly introduce some of the relevant theory we use surrounding Deep Neural networks.

5.1 Deep neural networks

Deep neural networks are inspired by the brain. The brain has many layers of neurons. Each of these takes some input, usually collected from other neurons. They combine these inputs to get a certain response, this is a nonlinear function from the inputs. This function we will call the activation function. In a simplified schematic representation, it looks like Figure 5.1.

To transform this into a mathematical set-up we work as follows. We start with the input X. These will be the covariates in our regression. Our goal is to estimate a function f, which takes as an input X and outputs the object we try to predict. If our activation function is denoted by σ we can build a deep neural network as follows. X is assumed to live in $\mathbb{R}^d$ (or some subspace thereoff). We will define $H^0(X) = X$. Next, we extend the definition inductively. Then we can apply a linear map W^i to H^{i-1}, this gives a new vector $W^i H^{i-1}(X)$. We add some bias term b^i. Then we take our activation function σ and apply it to each component of our vector $W^i H^{i-1}(X) + b^i$. This produces $H^i(X) = \sigma(W^i H^{i-1}(X)+b^i)$. We produce a function $f_{W,b}(X)$. Adding in the bias term b can be circumvented by modelling the data with a vector $(1, X)$ instead, so it is not strictly necessary.

This defines a deep neural network (DNN). Fitting a DNN is finding the best parameters W, b. Because finding the best parameters is a nonconvex optimisation problem

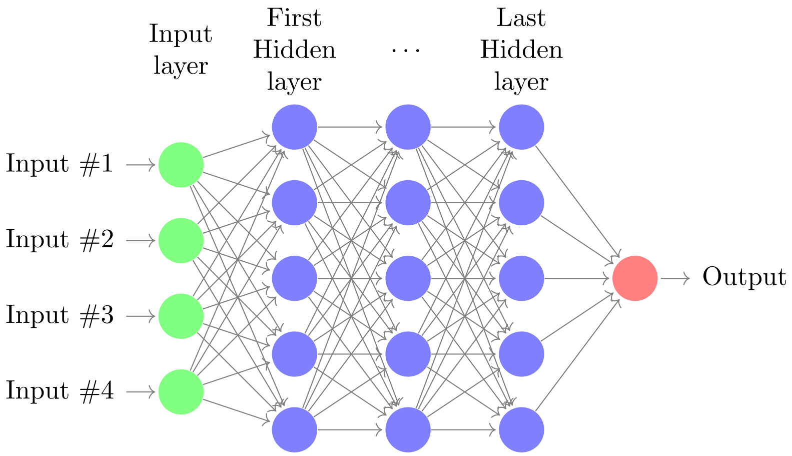


Figure 5.1: Neural Network with $L-1$ hidden layer.

we usually do not solve for the global optimum. Instead, we try to grudually improve our solution. This optimisation is usually done with the backpropagation algorithm. This algorithm is implemented in various software packages. We used Keras [14] and TensorFlow [1] to fit our neural networks. Because all these algorithms solve a lot of linear algebra these networks can be fit very fast on graphical processing units (GPU).

Our goal is to give theoretical guarantees for uncertainty quantification coming from DNN. We do this by using results found in [70, 73] for estimation using neural networks. Their approach is split into 2 parts that then come together. In the first part, they construct a small, sparse neural network that can approximate smooth functions well. In the second part, they show that small, sparse neural networks have small entropy. By putting these stepts together they can show that small, sparse neural networks have good statistical performance. We will give a brief overview of what they have done. For precise statements see the referenced papers.

5.2 Statistical properties of DNN

Step 1 As a first step, we approximate arbitrary smooth functions using deep neural networks. We want to show that DNN can approximate arbitrary smooth functions. To do that, we first construct small neural networks that will be useful later. We use these neural networks to construct products of terms.

We will pick as an activation function the ReLU function $\sigma(x) = \max(0, x)$. Then we can define functions T^k as follows:

$$T^k(x) := \sigma(x/2) - \sigma(x - 2^{1-2k}).$$

Next we define R^k by setting $R^1 = T^1$ and

$$R^k = T^k \circ R^{k-1}.$$

Finally, we define

$$S^m = \sum_{k=1}^{m} R^k(x).$$

Note that we can build this neural network with just $6m$ neurons.

The next step is to show that S^m and the function g given by $g(x) = x(1-x)$ are close together.

Lemma 5.2.1 (Lemma A.1, [70])**.**

$$\|g(x) - S^m(X)\| \leq 2^{-m}$$

Next, observe that

$$g(\frac{x-y+1}{2}) - g(\frac{x+y}{2}) + \frac{x+y}{2} - \frac{1}{4} = xy.$$

We can compute all these terms exactly using deep neural networks except g. And we can approximate g very precisely. Hence the function $(x, y) \mapsto xy$ can be appropriately well approximated using Neural networks. The next lemma is Lemma A.2 from [70].

Lemma 5.2.2. *For any positive integer m, there exists a sparse neural network $Mult_m$, such that $Mult_m(x, y) \in [0, 1]$ and*

$$\|Mult_m(x, y) - xy\| \leq 2^{-m}, \qquad \textit{for all } x, y \in [0, 1].$$

Moreover, $Mult_m(0, x) = Mult_m(x, 0) = 0$.

By using these products, one can build local Taylor polynomials. This construction is what is considered in [70]. However, we will follow [73] by using a spline basis instead. This spline basis has nice properties for the proofs we want to do later in our work. In [73, Lemma 1], they approximate the cardinal B spline basis using DNN. If we work in dimension d we can form the d-dimensional cardinal B-spline by tensoring the 1 dimensional cardinal B-splines. Denote $M^d_{0,0}$ the d dimensional cardinal B spline of degree m. They show that a small sparse neural network can approximate this spline. For full details, see [73, Lemma 1].

Lemma 5.2.3. *For all $\epsilon > 0$ there exists a small sparse neural network Spline that satisfies*

$$\|M^d_{0,0}(x) - \textit{Spline}(x)\| \leq \epsilon \quad \forall x \in [0, m+1]^d$$

and $\tilde{M}(x) = 0$ for all $x \notin [0, m+1]^d$.

We can then use this to give approximation guarantees for all functions in a Sobolev space $W^\beta([0, 1]^d)$.

Theorem 5.2.4. *For all $\epsilon > 0$ and all β-Sobolev smooth functions f there exists a small sparse neural network $\hat{f}_{DNN}$ that satisfies*

$$\|\hat{f}_{DNN}(x) - f(x)\| < \epsilon \quad \forall x \in [0,1]^d.$$

For the precise details on uniform sparse bounds see [73, Theorem 1].

Step 2 Controlling the Metric entropy.

The other ingredient that we need is the metric entropy of the class of sparse DNN. This is [70, Lemma 5]. This Lemma states that small and sparse deep neural networks have a small metric entropy.

Step 3 Putting it all together.

We can put everything together by using [70, Lemma 4]. This Lemma gives precise generalisation error bounds in terms of approximation error and metric entropy.

In the end, we can collect our results as in [70, Theorem 1] or [73, Theorem 2] for Hölder and Sobolev smooth functions respectively.

Chapter 6

Deep learning

This chapter is an adaption of a paper submitted as: S. Franssen, B. Szabó, "Uncertainty Quantification for nonparametric regression using Empirical Bayesian neural networks".

6.1 Introduction

Deep learning has received a lot of attention over the recent years due to its excellent performance in various applications, including personalized medicine [15], self driving cars [69, 64], financial institutions [39] and estimating power usage in the electrical grid [47, 43], just to mention a few. By now it is considered the state-of-the-art technique for image classification [45] or speech recognition [38].

Despite the huge popularity of deep learning, its theoretical underpinning is still limited, see for instance the monograph [2] for an overview. In our work we focus on the mathematical statistical aspects of how well feed-forward, multilayer artificial neural networks can recover the underlying signal in the noisy data. When fitting a neural network an activation function has to be selected. The most commonly used activation functions include the sigmoid, hyperbolic tangent, rectified linear unit (ReLU) and their variants. Due to computational advantages and available theoretical guarantees we consider the ReLU activation function in our work. The approximation properties of neural network with ReLU activation function has been investigated by several authors recently. In [50, 61] it was shown that deep networks with a smoothed version of ReLU can reduce sample complexity and the number of training parameters compared to shallow networks while reaching the same approximation accuracy. In the discussion paper [70] oracle risk bounds were derived for sparse neural networks in context of the multivariate nonparametric regression model. This in turn implies for Hölder regular classes (up to a logarithmic factor) rate optimal concentration rates and under additional structural assumptions (e.g. generalized additive models, sparse tensor decomposition) faster rates preventing the curse of dimensionality. The results of [70] were extended in different aspects by several authors. In [73] the more general Besov regularity classes were considered and adaptive estimation rates to these smoothness classes were derived. In [63] Bayesian sparse neural networks were proposed, where sparsity was induced by a spike-and-slab prior, and rate adaptive posterior contraction rates were derived. Finally, in [51] it was shown that the sparsity assumption on the neural network is not essential for the theoretical guarantees and similar results to [70] were derived for dense deep neural networks as well.

Most of the theoretical results focus on the recovery of the underlying signal of interest. However, it is at least as important to quantify how much we can rely on the procedure by providing reliable uncertainty statements. In statistics confidence regions are used to quantify the accuracy and remaining uncertainty of the method in a noisy model. Several approaches have already been proposed for statistical uncertainty quantification for neural networks, including bootstrap methods [53] or ensemble methods [46]. These methods are typically computationally very demanding especially for large neural networks. Bayesian methods are becoming also increasingly popular, since beside providing a natural way for incorporating expert information into the model via the prior they also provide built-in uncertainty quantification. The Bayesian counterpart of confidence regions are called credible regions which are the sets accumulating a prescribed, large fraction of the posterior mass. For neural networks various fully Bayesian methods were proposed, see for example [63, 81], however they quickly become computationally infeasible as the model size increases. To speed up the computations variational alternatives were proposed, see for instance [4]. An extended overview of machine learning methods for uncertainty quantification can be found in the survey [31].

Bayesian credible sets substantially depend on the choice of the prior and it is not guaranteed that they have confidence guarantees in the classical, frequentist sense. In fact it is known that credible sets do not always give valid uncertainty quantification in context of high-dimensional and nonparametric models, see for instance [21, 16] and hence their use for universally acceptable uncertainty quantification is not supported in general. In recent years frequentist coverage properties of Bayesian credible sets were investigated in a range of high-dimensional and nonparametric models and theoretical guarantees were derived on their reliability under (from various aspects) mild assumptions, see for instance [74, 10, 84, 65, 5, 67, 52] and references therein. However, we have only very limited understanding of the reliability of Bayesian uncertainty quantification in context of deep neural networks. To the best of our knowledge only (semi-)parametric aspects of the problem were studied so far [81], but these results do not provide uncertainty quantification on the whole functional parameter of interest.

In our work we propose a novel, empirical Bayesian approach with (relatively) fast computational time and derive theoretical, confidence guarantees for the resulting uncertainty statements. As a first step, we split the data into two parts and use the first part to train a deep neural network. We then use this empirical (i.e. data dependent) network to define the prior distribution used in our Bayesian procedure. We cut of the last layer of this neural network and take the linear combinations of the output of the previous layer with weights endowed by prior distributions, see the schematic representation of the prior in Section 6.2.1 below. The second part of the data is used to compute the corresponding posterior distribution, which will be used for inference. We study the performance of this method in the nonparametric random design regression model, but in principle our approach is applicable more widely. We derive optimal, minimax L_2-convergence rates for recovering the underlying functional parameter of interest and frequentist coverage guarantees for the slightly inflated credible sets. We also demonstrate the practical applicability of our method in a simulation study and verify empirically the asymptotic theoretical guarantees.

The rest of the paper is organized as follows. We present our main results in Section 6.2. After formally introducing the regression model we describe our Empirical Bayes Deep Neural Network (EBDNN) procedure in Section 6.2.1, list the set of assumptions under which our theoretical results hold in Section 6.2.2 and provide the guarantees for the uncertainty quantification in Section 6.2.3. In Section 6.3 we present a numerical analysis underlining our theoretical findings and providing a fast and easily implementable algorithm. The proofs are deferred to the Appendix. The proofs for the optimal posterior contraction rates and the frequentist coverage of the credible sets are given in Section 6.4. The approximation of the last layer of the neural network with B-splines is discussed in Section 6.5 and some relevant properties of B-splines are collected and verified in Section 6.6. Finally, general contraction and coverage results, on which we base the proofs in Section 6.4, are given in Section 6.7.

6.2 Main results

We consider in our analysis the multivariate random design regression model, where we observe pairs of random variables (X_1, Y_1),..., (X_n, Y_n) satisfying

$$Y_i = f_0(X_i) + Z_i, \qquad Z_i \overset{iid}{\sim} N(0, \sigma^2),\, X_i \overset{iid}{\sim} U([0,1]^d), \quad i = 1, ..., n,$$

for some unknown function $f_0 \in L^2([0,1]^d)$. We assume that the underlying function f_0 belongs to a β-smooth Sobolev ball $f_0 \in S_d^\beta(M)$ with known model hyperparameters $\beta, M, \sigma^2 > 0$. It is well known that the corresponding minimax L_2-estimation rate of f_0 is of order $\varepsilon_n = n^{-\beta/(d+2\beta)}$.

We will investigate the behaviour of multilayer neural networks in context of this nonparametric regression model. We propose an empirical Bayes type of approach, which recovers the underlying functional parameter with the (up to a logarithmic factor) minimax rate and provides reliable uncertainty quantification for the procedure.

6.2.1 Empirical Bayes Deep Neural Network (EBDNN)

We start by formally describing deep neural networks and then present our two-step Empirical Bayes approach. A deep neural network of depth $L > 0$ and width $p = (p_0, \dots, p_L)$ is a collection of weights $W = \{W^i | W^i \in \mathbb{R}^{p_i \times p_{i-1}}, i = 1, \dots, L\}$, shifts (or biases) $b = \{b^i | b^i \in \mathbb{R}^{p_i}, i = 1, \dots L-1\}$ and an activation function σ. There is a natural correspondence between deep neural networks with this architecture and functions $f_{W,b}(x)$: $\mathbb{R}^{p_0} \to \mathbb{R}^{p_L}$, with recursive formulation $f_{W,b}(x) = W^L H^{L-1}(x)$, where $H_j^0(x) = x_j$ and $H_j^i(x) = \sigma\left((W^i H^{i-1}(x))_j + b_j^i\right)$, for $j = 1, \dots, p_i$, $i = 1, ..., L$. Note that the activation function σ is not applied in the final iteration. Different types of activation functions are considered in the literature, including sigmoid, hyperbolic tangent, ReLU, ReLU square. In this work we focus on ReLU activation functions, i.e. we take $\sigma(x) = \max(x, 0)$.

Neural networks are very-high dimensional objects, with total number of parameters given by $\sum_{i=1}^L (p_{i-1} + 1)p_i$. Therefore, from a statistical perspective it is natural to introduce some additional structure in the form of sparsity by setting most of the model parameters W_{jk}^i $i = 1, ..., L$, $j = 1, ..., p_i$, $k = 1, ..., p_{i-1}$ and b_j^i, $i = 1, ..., L$, $j = 1, ..., p_i$ to zero. Such networks are called sparse, see the formal definition below.

Definition 6.2.1. *We call a deep neural network s-sparse if the weights W_{jk}^i and the biases b_j^i take values in $[-1, 1]$, and at most s of them are nonzero.*

Neural networks without sparsity assumptions are called dense networks and are more commonly used in practice. In our analysis we focus mainly on sparse networks but our method is flexible and can be easily extended to dense networks as well, which direction we briefly discuss in a subsequent section. Furthermore, we introduce boundedness on the neural network mainly for analytical, but also for practical reasons. We assume that $\|f_{W,b}\|_\infty < F$ for a fixed constant $F > 0$.

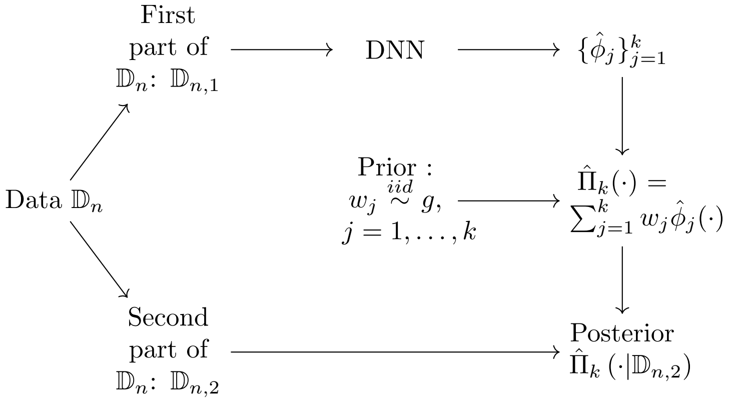
Next, note that in the last iteration of the recursive formulation $f_{W,b}(x) = W^L H^{L-1}(x)$ we take the linear combination of the functions $H_j^{L-1}(x)$, $j = 1, ..., p_{L-1}$. These functions take the role of data generated basis functions of the neural network and will play a crucial role in our method.

Definition 6.2.2. *We call the collection of functions $\hat{\phi}_j = H_j^{L-1}$, $j = 1, ..., p_{L-1}$ the DNN basis functions generated by the neural network.*

We propose a two stage, Empirical Bayes type of procedure. We start by splitting the dataset $\mathbb{D}_n = \big((X_1, Y_1), ...(X_n, Y_n)\big)$ into two (not necessarily equal) partition $\mathbb{D}_{n,1}$ and $\mathbb{D}_{n,2}$. We use the first dataset $\mathbb{D}_{n,1}$ to train the deep neural network. Then we build a prior distribution on the so constructed neural network and use the second dataset $\mathbb{D}_{n,2}$ to derive the corresponding posterior. More concretely, we cut-off the last layer of the neural network and take the (data driven) DNN basis functions $\hat{\phi}_j(x) = H_j^{L-1}(x)$, $j = 1, ..., p_{L-1}$ defined by the nodes in the $(L-1)$th layer. For convenience we use the notation $k = p_{L-1}$ for the number of DNN basis functions. We construct our prior distribution on the regression function by taking the linear combination of the so constructed basis functions and endowing the corresponding coefficients with prior distributions, i.e.

$$\hat{\Pi}_k(\cdot) = \sum_{j=1}^{k} w_j \hat{\phi}_j(\cdot), \qquad w_j \overset{iid}{\sim} g, \; j = 1, ..., k, \tag{6.1}$$

for some distribution g. Then the corresponding posterior is derived as the conditional distribution of the functional parameter given the second part of the data set $\mathbb{D}_{n,2}$. Please find below the schematic representation of our Empirical Bayes DNN prior and the corresponding posterior.


First part of $\mathbb{D}_n$: $\mathbb{D}_{n,1}$
DNN
$\{\hat{\phi}_j\}_{j=1}^k$
Data $\mathbb{D}_n$
Prior : $w_j \overset{iid}{\sim} g$, $j = 1, \ldots, k$
$\hat{\Pi}_k(\cdot) = \sum_{j=1}^k w_j \hat{\phi}_j(\cdot)$
Second part of $\mathbb{D}_n$: $\mathbb{D}_{n,2}$
Posterior $\hat{\Pi}_k(\cdot|\mathbb{D}_{n,2})$

We note, that often a pre-trained deep neural network is available corresponding to the regression problem of interest. In this case one can simply use that in stage one and compute the posterior based on the whole data set $\mathbb{D}_n$.

6.2.2 Assumptions on the EBDNN prior

We start by discussing the deep neural network produced in step one using the first dataset $\mathbb{D}_{n,1}$. As mentioned earlier we consider sparse neural networks following [70], but our results can be naturally extended to dense network as well. In [70, 73] optimal minimax concentration rates were derived for sparse neural networks under the assumptions that the networks are $s = k \log(n)$ sparse and have width $p = (d, 6k, \ldots, 6k, k, 1)$, with $k = k_n = n^{d/(d+2\beta)}$. We also apply these assumptions in our approach. However, since uncertainty quantification is a more complex task than estimation we need to introduce some additional structural requirements to our neural network framework.

One of the big advantage of deep neural networks is that they can learn the best fitting basis functions to the underlying structure of the functional parameter f_0 of interest, often resulting sharper recovery rates than using standard, fixed bases, see for instance [70, 73]. However, neural networks in general are highly flexible due to

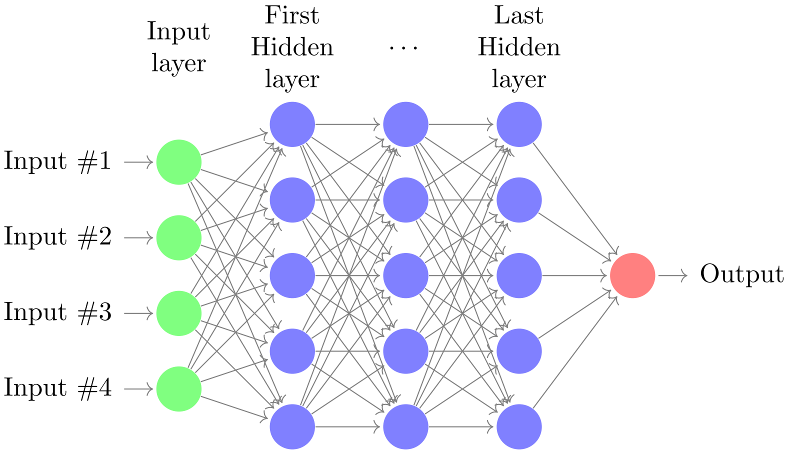


Figure 6.1: Neural Network with $L-1$ hidden layer.

the high-dimensional structure and do not provide a unique, for our goals appropriate representation. For instance, let us consider a neural network with ReLU activation function, see Figure 6.1 for schematic representation. Then let us include an additional layer in the network before the output layer consisting only one node, see Figure 6.2. This node takes the place of the output layer in the original network and since there is only one node in the so constructed last layer the output of the new network is the same as the original one (given that the output function is non-negative). Using the second neural network for our empirical Bayes prior is clearly sub-optimal as we end up with a one dimensional parametric prior for a nonparametric problem, resulting in overly confident uncertainty quantification. Therefore to avoid such pathological cases we introduce some additional structure to our neural network. We assume that the neural network produces nearly orthogonal basis functions, see the precise definition below.

Definition 6.2.3. *We say a neural network produces near orthogonal basis if the Gram matrix Σ_k given by $(\Sigma_k)_{i,j} = \langle \hat{\phi}_i, \hat{\phi}_j \rangle_2$ satisfies that $c_1 \mathbb{1}_k \leq \Sigma_k \leq c_2 \mathbb{1}_k$ for some $0 < c_1 < c_2 < \infty$ and for all $1 \leq i, j \leq k$.*

The requirement of near orthogonality is essential for our analysis to appropriately control the small ball probabilities of the prior distribution which is of key importance in Bayesian nonparametrics. Nevertheless, in view of the simulation study, given in Section 6.3 it seems that this assumption can be relaxed. In the numerical analysis section we do not impose this requirement on the algorithm and still get in our examples accurate recovery and reliable uncertainty quantification. We summarize the above assumptions below.

Assumption 6.2.4. *Let us take $k = k_n = n^{d/(d+2\beta)}$ and assume that the neural network $\hat{f}_n$ constructed in step 1 is*

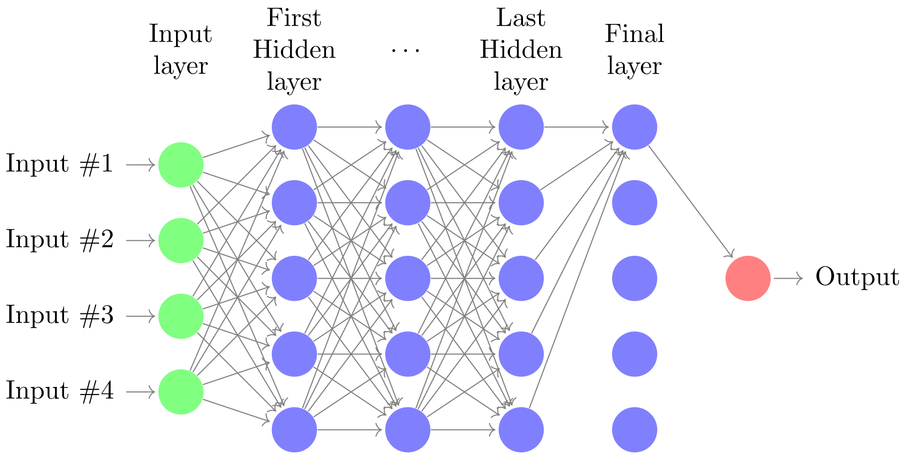


Figure 6.2: The modified Neural Network of Figure 6.1 by adding an additional layer before the output layer. This deep neural network provides the same output (given that the output function is non-negative), but in the last hidden layer it has only one node hence provides only one basis function.

- *bounded in supnorm* $\|\hat{f}_n\|_\infty \leq F$*,*
- $s = k\log(n)$ *sparse,*
- *has depth* $L = \log(n)\left\lceil \log_2(\max(4\beta, 4d))\right\rceil$
- *has width* $p = (n, 6k, \dots, 6k, k, 1)$*,*
- *there exists a* $C > 0$ *independent of* n *such that the DNN basis functions* $\hat{\phi} = (\hat{\phi}_1, ..., \hat{\phi}_k)$ *satisfy* $\|\theta^T\hat{\phi}\|_\infty \leq C\sqrt{k}\|\theta\|_\infty$*,*
- *and the corresponding Gram matrix* Σ_k*, given by* $(\Sigma_k)_{i,j} = \langle\hat{\phi}_i, \hat{\phi}_j\rangle_2$*, is nearly orthogonal for some* $0 < c_1 < c_2 < \infty$*.*

The class of deep neural networks satisfying Assumption 6.2.4 is denoted by $\mathcal{F}(L, p, s, F, C, c_1, c_2)$. In Appendix 6.5 we show that such kind of DNN basis $\hat{\phi}_1, ..., \hat{\phi}_k$ can be constructed. Next assume, similarly to [70, 73], that a near minimizer of the neural network can be obtained. This assumption is required to derive guarantees on the generalisation error of the network produced in step 1.

Assumption 6.2.5. *We assume that the network trained in step 1 is a near minimizer in expectation. Let* $f_0 \in S_d^\beta(M) \cap L_\infty(M)$ *and* $\epsilon_n = n^{-\beta/(2\beta+d)}$*, then the estimator* $\hat{f}_n$ *resulting from the neural network satisfies that*

$$\mathbb{E}_{f_0}\left[\frac{1}{n}\sum_{i=1}^n \left(y_i - \hat{f}_n(X_i)\right)^2 - \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n (Y_i - f(X_i))^2\right] \leq \epsilon_n \log(n)^3,$$

where $\mathcal{F} = \mathcal{F}(L, p, s, F, C, c_1, c_2)$*.*

It remained to discuss the choice of the prior distribution g on the coefficients of the DNN basis functions. In general we have a lot of flexibility in choosing g, but for analytical convenience we assume to work with continuous positive densities.

Assumption 6.2.6. *Assume that the density g in the prior* (6.1) *is continuous and positive.*

Remark 6.2.7. *We note that our proof requires only that the density is bounded away from zero and infinity on a small neighborhood of (the projection of) the true function f_0, hence it is sufficient to require that the density g is bounded away from zero and infinity on a large enough compact interval. Since one can construct a neural network with weights between -1 and 1, approximating the true function well enough, we can further relax our assumption and consider densities g supported on on* $[-1, 1]$*.*

6.2.3 Uncertainty quantification with EBDNN

Our main goal is to provide reliable uncertainty quantification for the outcome of the neural network. Our two-step Empirical Bayes approach gives a probabilistic solution to the problem which in turn can be automatically used to quantify the remaining uncertainty of the procedure. First we show that the corresponding posterior distribution recovers the underlying functional parameter of interest f_0 with the minimax contraction rate $\epsilon_n = n^{-\beta/(2\beta+d)}$ up to a logarithmic factor.

Theorem 6.2.8. *Let $\beta, M > 0$ and assume that the EBDNN prior $\hat{\Pi}_k$, given in* (6.1)*, satisfies Assumptions 6.2.4, 6.2.5 and 6.2.6. Then the corresponding posterior distribution contracts around the true function $f_0 \in S_d^\beta(M)$ at the near minimax rate, i.e.*

$$\limsup_{n\to\infty} \sup_{f_0 \in S_d^\beta(M) \cap L_\infty(M)} \mathbb{E}_{f_0}\Big(\hat{\Pi}_k\left(f\colon \|f - f_0\|_2 \geq M_n \log^3(n)\epsilon_n \middle| \mathbb{D}_{n,2}\right)\Big) = 0,$$

for all $M_n \to \infty$.

The proof of the theorem is given in Section 6.4.1. We note that one can easily construct estimators from the posterior inheriting the same concentration rate as the posterior contraction rate. For instance one can take the center of the smallest ball accumulating at least half of the posterior mass, see Theorem 2.5 of [32]. Furthermore, under not too restrictive conditions, it can be proved that the posterior mean achieves the same near optimal concentration rate as the whole posterior, see for instance page 507 of [32] or Theorem 2.3. of [36].

Our main focus is, however, on uncertainty quantification. The posterior is typically visualized and summarized by plotting the credible region C_α accumulating $1-\alpha$ fraction (typically one takes $\alpha = 0.05$) of the posterior mass. In our analysis we consider L_2-balls centered around an estimator $\hat{f}$ (typically the posterior mean or maximum a posteriori estimator), i.e.

$$C_\alpha = \{f\colon \|f-\hat{f}\|_2 \leq r_\alpha\} \qquad \text{satisfying} \qquad \hat{\Pi}_k(f\in C_\alpha|\mathbb{D}_{n,2}) = 1-\alpha.$$

More precisely, in case the posterior distribution is not continuous, then the radius r_α is taken to be the smallest such that $\hat{\Pi}_k(f\in C_\alpha|\mathbb{D}_{n,2}) \geq 1-\alpha$ holds.

However, Bayesian credible sets are not automatically confidence sets. To use them from a frequentist perspective reliable uncertainty quantification we have to show that they have good frequentist coverage, i.e.

$$\inf_{f_0\in S_d^\beta(M)} \mathbb{P}_{f_0}(f_0\in C_\alpha) \geq 1-\alpha.$$

In our analysis we introduce some additional flexibility by allowing the credible sets to be blown up by a factor L_n, i.e. we consider sets of the form

$$C_\alpha(L_n) = \{f\colon \|f-\hat{f}\|_2 \leq L_n r_\alpha\} \quad \text{with} \quad \hat{\Pi}_k(f\in C_\alpha(1)|\mathbb{D}_{n,2}) = 1-\alpha. \tag{6.2}$$

This additional blow up factor L_n is required as the available theoretical results in the literature on the concentration properties of the neural network are sharp only up to a logarithmic multiplicative term and we compensate for this lack of sharpness by introducing this additional flexibility. Furthermore, in view of our simulation study, it seems that a logarithmic blow up is indeed necessary to provide from a frequentist perspective reliable uncertainty statements, see Section 6.3.

The centering point of the credible sets can be chosen flexibly, depending on the problem of interest. In practice usually the posterior mean or mode is considered for computational and practical simplicity. Our results hold for general centering points under some mild conditions. We only require that the centering point attains nearly the optimal concentration rate. We formalize this requirement below.

Let us denote by $f^* = f_{\theta^*} = (\theta^*)^T\hat{\Phi}_k$, with $\hat{\Phi}_k = \big(\hat{\phi}_1, ..., \hat{\phi}_k(x)\big)$ the DNN basis, the Kullback-Leibler (KL) projection of f_0 onto our model $\Theta_k = \{\sum_{j=1}^k \theta_j\hat{\phi}_j\colon \theta\in\mathbb{R}^k\}$, i.e. let $\theta^*\in\mathbb{R}^k$ denote the minimizer of the function $\theta\mapsto KL(f_0, \theta^T\hat{\Phi}_k)$. We note that the KL projection is equivalent with the L_2-projection of f_0 to Θ_k in the regression model with Gaussian noise. We assume that the centering point of the credible set is close to f^*.

Assumption 6.2.9. *The centering point $\hat{\theta}$ (i.e. $\hat{f} = f_{\hat{\theta}} = \hat{\theta}^T\hat{\Phi}_k$) satisfies that for all $\delta > 0$ there exists $M_\delta > 0$ such that*

$$\sup_{f_0\in S_d^\beta(M)\cap L_\infty(M)} \mathbb{P}_{f_0}\Big(d_n(f^*,\hat{f}) \leq M_\delta n^{-\beta/(2\beta+d)}\Big) \geq 1-\delta. \tag{6.3}$$

This assumption on the centering point is mild. For instance considering the centering point of the smallest ball accumulating a large fraction (e.g. half) of the posterior mass as the center of the credible ball satisfies this assumption. The posterior mean is another good candidate for appropriately chosen priors.

Theorem 6.2.10. *Let $\beta > \frac{d}{2}$, $M > 0$ and assume that the EBDNN prior $\hat{\Pi}_k$, given in* (6.1)*, satisfies Assumptions 6.2.4, 6.2.5 and 6.2.6, and the centering point $\hat{f}$ satisfies Assumption 6.2.9. Then the EBDNN credible balls with inflating factor $L_{\delta,\alpha}\log^3(n)$ have uniform frequentist coverage and near optimal size, i.e. for arbitrary $\delta, \alpha > 0$ there exists $L_{\delta,\alpha} > 0$ such that*

$$\liminf_n \inf_{f_0 \in S_d^\beta(M) \cap L_\infty(M)} \mathbb{P}_{f_0}(f_0 \in C_\alpha(L_{\delta,\alpha}\log^3 n)) \geq 1 - \delta, \tag{6.4}$$

$$\liminf_n \inf_{f_0 \in S_d^\beta(M) \cap L_\infty(M)} \mathbb{P}_{f_0}(r_\alpha \leq C n^{-\beta/(2\beta+d)}) \geq 1 - \delta, \tag{6.5}$$

for some large enough $C > 0$.

We defer the proof of the theorem to Section 6.4.2.

6.3 Numerical Analysis

So far we have studied the EBDNN methodology from a theoretical, asymptotic perspective. In this section we investigate the finite sample behaviour of the procedure. First note that the theoretical bounds in [70, 73] are not known to be tight. Sharper bounds would result in more accurate procedure with smaller adjustments for the credible sets. For these reasons we study the performance of the EBDNN methodology in synthetic data sets, where the estimation and coverage properties can be empirically evaluated.

In our implementation we deviate for practical reasons in three points from the theoretical assumptions considered in the previous sections. First, in practice sparse deep neural networks are rarely used, as they are typically computationally too involved to train. Instead, dense deep neural networks are applied routinely which we will also adopt in our simulation study. Moreover, the global optima typically can not be retrieved when training a neural network. The common practice is to use gradient descent and aim for attaining good local minimizer. Finally, the softwares used to train deep neural networks do not necessarily return a near orthogonal deep neural network. Even worse, some of the produced basis functions can be collinear or even constantly zero. To guarantee that the produced basis functions are nearly orthogonal one can either apply the Gram-Schmidt procedure or introduce a penalty for collinearity. We do not pursue this direction in our numerical analysis, but use standard softwares and investigate the robustness of our procedure with respect to these aspects. So instead of studying our EBDNN methodology under our restrictive assumptions we investigate its performance in more realistic scenarios. We fit a dense deep neural network using standard gradient descent and do not induce sparsity or near orthogonality to our network.

6.3.1 Implementation details

We have implemented our EBDNN method in Python. We use Keras [14] and tensorflow [1] to fit a deep neural network using the first half of the data. We use gradient decent to fit a dense neural network with $L = \lceil \log_2(\beta) \log_2(n) \rceil$ layers. Each of the first $L-1$ hidden layers has width $6k_n$, with $k_n = n^{d/(2\beta+d)}$, and we apply the ReLU activation function on them. The last layer has width k_n and the identify map is taken as activation function on it, that is, we take the weighted linear combination of these basis functions. Then we extract the basis functions by removing the last layer and endow the corresponding weights by independent and identically distributed standard normal random variables, to exploit conjugacy and speed up the computations. We derive credible regions by sampling from the posterior using Numpy [37] and empirically computing the quantiles and the posterior mean used as the centering point. The corresponding code is available at [25].

6.3.2 Results of the numerical simulations

We consider two different regression functions in our analysis.

$$f_1 = \sum_{i=1}^{\infty} \frac{\sin(i)\cos(\pi(i-0.5)x)}{i^{1.5}}, \qquad f_2 = \sum_{i=1}^{\infty} \frac{\sin(i^2)\cos(\pi(i-0.5)x)}{i^{1.5}}.$$

Note that both of them belong to a Sobolev class with regularity 1. In the implementation we have considered a sufficiently large cut-off of their Fourier series expansion.

In our simulation study we investigate beyond the L_2-credible balls also L_∞ credible balls as well. In our theoretical studies we have derived good frequentist coverage after inflating the credible balls by a $\log^3(n)$ factor. In our numerical analysis we observe that (at least on a range of examples) a $\log(n)$ blow-up factor is sufficient, while a $\sqrt{\log(n)}$ blow-up is not enough.

We considered sample sizes $n = 1000$, 5000, 10000, and 50000, and repeated each of the experiments 1000 times. We report in Table 6.1 the average L_2-distance and the corresponding standard deviation between the posterior mean and the true functional parameter of interest.

n	1000	5000	10000	50000
f_1	213.33 ± 116.06	120.36 ± 29.29	96.84 ± 18.58	48.94 ± 9.19
f_2	221.20 ± 113.48	140.04 ± 46.69	108.82 ± 9.73	75.74 ± 3.60

Table 6.1: Average L_2-distance between the posterior mean and the true function based on 1000 repetitions. The sample sizes range from 1000 to 50000.

Furthermore, we investigate the frequentist coverage properties of the EBDNN credible sets by reporting the fraction of times the (inflated) credible balls contain the true function out of the 1000 runs in Table 6.2. One can observe that in case of function

f_2 a $\sqrt{\log n}$ blow up factor is not sufficient and the more conservative $\log n$ inflation has to be applied, which provides reliable uncertainty quantification in both cases.

function	blow-up	1000	5000	10000	50000
	none	0.0	0.0	0.0	0.0
f_1	$\sqrt{\log(n)}$	0.893	0.896	0.848	0.928
	$\log(n)$	0.951	0.997	1.0	1.0
	none	0.0	0.0	0.0	0.0
f_2	$\sqrt{\log(n)}$	0.834	0.693	0.736	0.003
	$\log(n)$	0.934	0.986	1.0	1.0

Table 6.2: Frequentist coverage of the inflated L_2-credible balls based on 1000 runs of the algorithm. Sample size is ranging between 1000 and 50000 and we considered multiplicative inflation factors between 1 and $\log n$.

We also report the size of the L_2 credible balls in Table 6.3. One can observe that the radius of the credible balls are substantially smaller than the average Euclidean distance between the posterior mean the true functions f_1 and f_2 of interest, respectively. This explains the necessity of the inflation factor applied to derive reliable uncertainty quantification from the Bayesian procedures. We illustrate the method in Figure 6.3. Note that the true function is inside of the region defined by the convex hull of the 95% closest posterior draws to the posterior mean.

n	1000	5000	10000	50000
f_1	101.91 ± 15.69	52.27 ± 3.60	38.45 ± 2.33	19.47 ± 1.03
f_2	91.26 ± 16.50	49.61 ± 3.76	37.46 ± 2.03	19.69 ± 0.85

Table 6.3: L_2 The average diameter and the corresponding standard deviation of the (non-inflated) credible balls based on 1000 runs of the algorithm.

Next we investigate the point wise and L_∞ credible regions. Compared to the L_2-credible balls we note that the L_∞ credible bands are roughly a factor 2 wider, see Table 6.4.

n	1000	5000	10000	50000
f_1	175.93 ± 28.33	101.10 ± 10.66	77.08 ± 6.50	46.38 ± 3.83
f_2	145.60 ± 30.59	94.65 ± 9.93	78.41 ± 5.91	51.26 ± 3.84

Table 6.4: The average supremum diameters and the corresponding standard deviation of the (non-inflated) credible balls based on 1000 runs of the algorithm.

At the same time, the distance between the posterior mean and the true regression function is roughly a factor 6 larger compared to the situation in the L_2 norm, see Table 6.5.

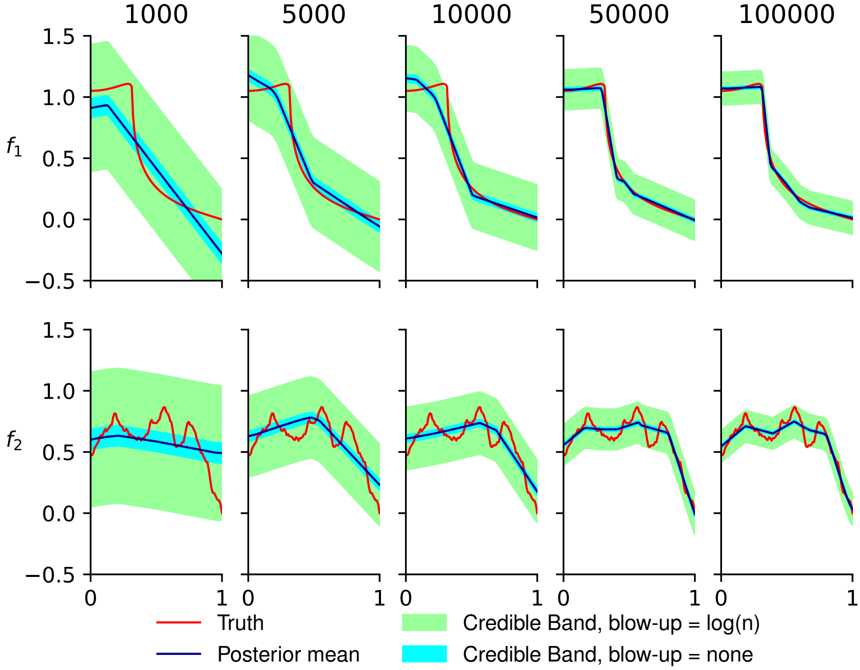


Figure 6.3: EBDNN L_2-credible balls illustrated by the region covered by the 95% closest draw from the posterior to the posterior mean in L_2-distance. Sample size increases from 1000 to 100000. The original credible sets are plotted by light blue and the inflated credible sets by green.

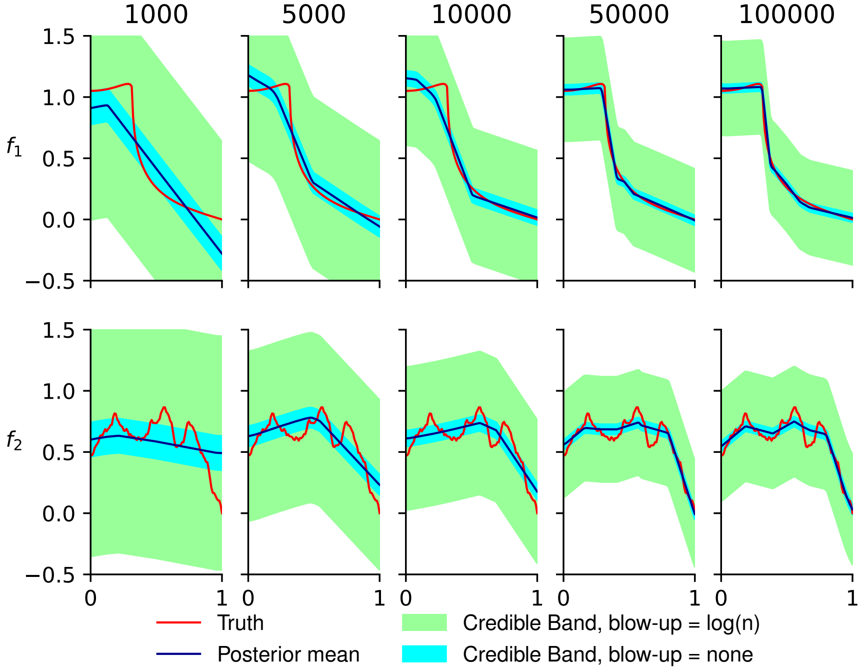


Figure 6.4: EBDNN L_∞-credible regions. The bands are formed by keeping the 95% closest draws from the posterior to the posterior mean in L_∞-distance. Sample size increases from 1000 to 100000. The original credible bands are plotted by light blue and the inflated credible bands by green.

n	1000	5000	10000	50000
f_1	442.45 ± 181.02	355.40 ± 55.67	325.34 ± 42.26	199.95 ± 31.32
f_2	569.47 ± 119.33	360.34 ± 81.54	255.59 ± 36.78	169.60 ± 11.41

Table 6.5: Average supremum norm-distances between the posterior mean and the true function based on 1000 repetitions.

This results in worse coverage results than in the L_2 case, although inflating the credible bands by a $\log(n)$ factor still results in reliable uncertainty quantification on our simulated data, see Table 6.6 and Figure 6.4.

function	blow-up	1000	5000	10000	50000
	none	0.0	0.0	0.0	0.0
f_1	$\sqrt{\log(n)}$	0.779	0.12	0.015	0.081
	$\log(n)$	0.957	0.997	1.0	1.0
	none	0.0	0.0	0.0	0.0
f_2	$\sqrt{\log(n)}$	0.101	0.181	0.384	0.503
	$\log(n)$	0.946	0.986	1.0	1.0

Table 6.6: Frequentist coverage of the inflated L_∞-credible balls based on 1000 runs of the algorithm. Sample size is ranging between 1000 and 50000 and we considered multiplicative inflation factors between 1 and $\log n$.

6.4 Proof of the main results

Before providing the proof of our main theorems we recall a few notations used throughout the section. We denote by $\mathbb{D}_{n,1}$ and $\mathbb{D}_{n,2}$ the first and second half of the data respectively, i.e.

$$\mathbb{D}_{n,1} = ((X_1, Y_1), \dots, (X_{\lfloor \frac{n}{2} \rfloor}, Y_{\lfloor \frac{n}{2} \rfloor})),$$
$$\mathbb{D}_{n,2} = ((X_{\lfloor \frac{n}{2} \rfloor+1}, Y_{\lfloor \frac{n}{2} \rfloor+1}), \dots, (X_n, Y_n)).$$

Furthermore, we denote by f^* the L_2-projection of f_0 into the linear space spanned by the DNN basis based on the first data set $\mathbb{D}_{n,1}$, as it was defined above Assumption 6.2.9. Next we give the proofs for our main theorems.

6.4.1 Proof of Theorem 6.2.8

First note that by triangle inequality $\|f - f_0\|_2 \leq \|f - f^*\|_2 + \|f^* - f_0\|_2$. We deal with the two terms on the right hand side separately.

In view of Lemma 6.7.3 (with $\mathbb{D}_n := \mathbb{D}_{n,2}$, $k = n^{d/(2\beta+d)}$, Σ_k defined by the DNN basis functions $\hat{\phi}_1$,..., $\hat{\phi}_k$, $\Pi_k = \hat{\Pi}_k$ and with respect to the conditional distribution given the first data set $\mathbb{P}_{f_0}^{|\mathbb{D}_{n,1}}$) we get that for every $\delta > 0$ there exists $M_\delta < \infty$,

$$\sup_{f_0 \in S_d^\beta(M)} \mathbb{E}_{f_0}^{|\mathbb{D}_{n,1}} \hat{\Pi}_k \Big(f \colon \|f - f^*\|_2 \geq M_\delta n^{-\beta/(2\beta+d)} | \mathbb{D}_{n,2} \Big) \leq \delta, \tag{6.6}$$

with $\mathbb{P}_{f_0}$-probability tending to one. Hence, it remained to deal with the term $\|f^* - f_0\|_2$. We introduce the event

$$A_n = \{\|f^* - f_0\|_2 \leq (M_n/2) n^{-\beta/(2\beta+d)} \log^3(n)\},$$

which is independent from the second half of the data $\mathbb{D}_{n,2}$, hence

$$\begin{aligned}&\mathbb{E}_{f_0}\hat{\Pi}_k\Big(\theta\colon \|f_\theta - f_0\|_2 \geq M_n n^{-\frac{\beta}{2\beta+d}}\log^3(n)\Big|\mathbb{D}_{n,2}\Big)\\&\quad\leq \mathbb{E}_{f_0}\Big(\mathbb{1}_{A_n}\mathbb{E}_{f_0}^{|\mathbb{D}_{n,1}}\hat{\Pi}_k(\theta\colon \|f_\theta - f^*\|_2 \geq M_n n^{-\frac{\beta}{2\beta+d}}|\mathbb{D}_{n,2})\Big) + \mathbb{E}_{f_0}\mathbb{1}_{A_n^c}.\end{aligned}$$

The first term is bounded by δ in view of assertion (6.6), while the second term tends to zero in view of Theorem 4 of [73] combined with Markov inequality.

6.4.2 Proof of Theorem 6.2.10

Let $L_n = L_{\epsilon,\alpha}\log(n)^3$, $\varepsilon_n = n^{-\beta/(2\beta+d)}$ and $k = n^{d/(2\beta+d)}$. Then by triangle inequality we get

$$\begin{aligned}\mathbb{P}_{f_0}\left(f_0 \in C_\alpha(L_n)\right) &= \mathbb{P}_{f_0}\Big(\|f_0 - \hat{f}\|_2 \leq L_n r_\alpha\Big)\\&\geq \mathbb{P}_{f_0}\Big(\hat{\Pi}_k(\theta\colon \|f_\theta - \hat{f}\|_2 \leq \|\hat{f} - f_0\|_2/L_n|\mathbb{D}_{n,2}) < 1-\alpha\Big).\end{aligned} \tag{6.7}$$

Furthermore, let us introduce the event $A_n = \{\|\hat{f} - f_0\|_2 \leq M_\varepsilon \log^3(n)\epsilon_n\}$. Note that by triangle inequality and in view of Assumption 6.2.9 (for large enough choice of M_ε) and Theorem 4 of [73] (with f^* denoting the L_2-projection of f_0 to the linear space spanned by the DNN basis) combined with Markov's inequality,

$$\begin{aligned}\mathbb{P}_{f_0}(A_n^c) \leq \mathbb{P}_{f_0}&\big(\|\hat{f} - f^*\|_2 \geq M_\varepsilon \epsilon_n/2\big)\\&+\mathbb{P}_{f_0}\big(\|f_0 - f^*\|_2 \geq M_\varepsilon \log^3(n)\epsilon_n/2\big)\\\leq \varepsilon/3 &+ \varepsilon/3 = (2/3)\varepsilon.\end{aligned}$$

Hence, the probability on the right hand side of (6.7) is lower bounded by

$$\mathbb{P}_{f_0}\Big(\hat{\Pi}_k(\theta\colon \|f_\theta - \hat{f}\|_2 \leq \frac{M_\varepsilon \epsilon_n}{L_{\epsilon,\alpha}}|\mathbb{D}_{n,2}) < 1-\alpha\Big) - (2/3)\varepsilon.$$

We finish the proof by showing that the first term in the preceding display is bounded from below by $1-\varepsilon/3$. Since by assumption the DNN basis is nearly orthogonal with $\mathbb{P}_{f_0}$-probability tending to one, we get in view of Lemma 6.7.2 below (applied with probability measure $\mathbb{P}_{f_0}^{|\mathbb{D}_{n,1}}$, $k = n^{d/(2\beta+d)}$ and $\Pi_k = \hat{\Pi}_k$) that for all $\varepsilon > 0$ there exists $\delta_{\varepsilon,\alpha} > 0$ such that

$$\sup_{f_0 \in S_d^\beta(M)} \mathbb{E}_{f_0}^{|\mathbb{D}_{n,1}}\hat{\Pi}_k\Big(f_\theta\colon \|f_\theta - \hat{f}\|_2 \leq \delta_{\varepsilon,\alpha}\sqrt{k/n}|\mathbb{D}_{n,2}\Big) \leq \varepsilon(1-\alpha)/3,$$

with $\mathbb{P}_{f_0}$-probability tending to one. Let us take $L_{\epsilon,\alpha} \geq M_\epsilon/\delta_{\epsilon,\alpha}$ and combine the preceding display with Markov's inequality,

$$\mathbb{P}_{f_0}\Big(\hat{\Pi}_k\big(\theta\colon \|f_\theta - \hat{f}\|_2 \leq \frac{M_\epsilon \epsilon_n}{L_{\varepsilon,\alpha}}|\mathbb{D}_{n,2}\big) \geq 1-\alpha\Big)$$
$$\leq \frac{\mathbb{E}_{f_0}\Big(\hat{\Pi}_k\big(\theta\colon \|f_\theta - \hat{f}\|_2 \leq \delta_{\epsilon,\alpha}\epsilon_n|\mathbb{D}_{n,2}\big)\Big)}{1-\alpha} \leq \varepsilon/3 + o(1),$$

concluding the proof.

Remark 6.4.1. *We point out that the extra multiplicative term* $\log^3(n)$ *is the result of the lack of sharpness in the convergence rate of deep neural network estimator* f_n^**. Sharper bounds for this estimation would result in smaller blow up factor.*

6.5 Approximation of Splines using Deep neural networks

This section considers the construction of orthonormal basis $\phi_1, \ldots, \phi_k$ in d-dimension using neural networks. We first show that splines can be approximated well with neural networks and then we achieve near orthonormality by rescaling. In this and the following section we use results from different sources. Notably we combine the results from [73] and [33, 71]. In the former they define the splines using the divided differences definition. In the latter books they use the convolution definition and rescaling. By [71, theorem 4.23] these definitions are equivalent. We summarize the main results in the following lemma.

Lemma 6.5.1. *There exist DNN basisfunctions* $\phi_1, \ldots, \phi_k$ *with* $k = n^{d/(d+2\beta)}$ *such that*

- *For every* $f_0 \in S_d^\beta(M) \cap L_\infty(M)$*, with* $\beta > d/2$*, there exists* $\theta = (\theta_1, ..., \theta_k) \in \ell_\infty(1)$ *such that* $\|f_0 - \sum_{j=1}^k \theta_j\phi_j\|_2 < \epsilon_n$ *with* $\varepsilon_n = n^{-\beta/(d+2\beta)}$*.*
- *The rescaled DNN basis functions* $\sqrt{k}\phi_1, \ldots, \sqrt{k}\phi_k$ *are nearly orthonormal in the sense of Definition 6.2.3.*
- *The basis functions are bounded in supremum norm, i.e.* $\|\phi_j\|_\infty \leq 2$*,* $j = 1, ..., k$*.*

Proof. In view of Lemma 6.6.2 there exists $\theta = (\theta_1, ..., \theta_k) \in \mathbb{R}^k$ such that $\|f_0 - \sum_{j=1}^k \theta_j B_j\|_2 \leq n^{-\beta/(d+2\beta)}$, where B_j, $j = 1, ..., k$ denote the cardinal B-splines of order $q \geq \beta$, see (6.9) and teh remark below it about the single index representation. Moreover, if $\|f_0\|_\infty < M$, then one can choose $\theta \in \mathbb{R}^k$ such that $\|\theta\|_\infty < M$. Furthermore, in view of Proposition 1 of [73] one can construct a DNN basis

$$\phi_1, \ldots, \phi_k, \quad \text{such that} \quad \|B_j - \phi_j\|_\infty \leq C/n, \quad j = 1, ..., k, \tag{6.8}$$

for some universal constant $C > 0$. Therefore, by triangle inequality

$$\begin{aligned}\|f_0 - \sum_{j=1}^{k}\theta_j\phi_j\|_2 &\leq \|f_0 - \sum_{j=1}^{k}\theta_j B_j\|_2 + \|\sum_{j=1}^{k}\theta_j(B_j - \phi_j)\|_2 \\ &\lesssim n^{-\beta/(d+2\beta)} + M\sqrt{k}/n \lesssim n^{-\beta/(d+2\beta)}.\end{aligned}$$

Then in Lemma 6.5.2 below we show that the above DNN basis $(\phi_j)_{j=1,..,k}$ inherits the near orthogonality of B-splines, which is verified for dimension d in Lemma 6.6.3. The boundedness of the B-splines, will be also inherited by the above DNN basis $(\phi_j)_{j=1,..,k}$ in view of Lemma 6.5.3. Moreover, basis can be rescaled in such a way that the coefficients are in the interval $[-1, 1]$. ■

We provide below the two lemmas used in the proof of the previous statement.

Lemma 6.5.2. *The rescaled DNN basis* $\sqrt{k}\phi = \left(\sqrt{k}\phi_1, \ldots, \sqrt{k}\phi_k\right)$ *given in* (6.8) *is nearly orthonormal in the sense of Definition 6.2.3.*

Proof. In view of Lemma 1 of [73] the above DNN basis has the same support as the B-splines of order $q = \lceil\beta\rceil$. Let us define the matrices $Q_k, R_k \in \mathbb{R}^{k\times k}$ as

$$\begin{aligned}(Q_k)_{i,j} &= \left\langle\sqrt{k}B_i, \sqrt{k}B_j\right\rangle = k\int_0^1 B_i(x)B_j(x)\,\mathrm{d}\,x, \\ (R_k)_{i,j} &= \left\langle\sqrt{k}\phi_i, \sqrt{k}\phi_j\right\rangle - \left\langle\sqrt{k}B_i, \sqrt{k}B_j\right\rangle = k\int_0^1 \phi_i(x)\phi_j(x) - B_i(x)B_j(x)dx,\end{aligned}$$

for $i, j \in \{1, \ldots, k\}$. Then $Q_k + R_k$ is the matrix consisting of the innerproducts in the constructed basis. Note that in view of (6.8) there exists a constant $C' > 0$ such that $|(R_k)_{i,j}| < C'k/n$. Furthermore, we note that a B-spline basis function of order q has intersecting support with at most $(2q)^d$ other B-spline basis functions. In view of Lemma 1 of [73], the same holds for the ϕ_j, $j = 1, ..., k$ basis. This means that there are at most $(2q)^d$ non-zero terms in every row or column and hence in total we have at most $(2q)^d k$ nonzero cells in the matrix.

Define $(M_k)_{i,j} = |(R_k)_{i,j}|$. Then the spectral radius of M_k is an upper bound of the spectral radius of R_k by Wielandt's theorem [82]. Since M_k is a nonnegative matrix, in view of the Perron-Frobenius theorem [55, 30], the largest eigenvalue in absolute value is bounded by constant times k^2/n. Next note that both Q_k and R_k are symmetric real matrices. Therefore, in view of the Weyl inequalities (see equation (1.54) of [75]), the eigenvalues of $Q_k + R_k$ can differ at most by constant times $k^2/n = o(1)$ from the eigenvalues of Q_k. We conclude the proof by noting that in view of Lemma 6.6.3 the eigenvalues of Q_k are bounded from below by c and from above by C for n large enough, hence the Gram matrix $Q_k + R_k$ also satisfies

$$\frac{1}{2}c\mathbb{1}_k \leq (Q_k + R_k) \leq 2C\mathbb{1}_k.$$

This means that the rescaled basis $\sqrt{k}\phi$ satisfies the near orthogonality requirement, see Assumption 6.2.3. ■

Lemma 6.5.3. *The DNN basis given in* (6.8) *satisfies that* $\|\phi_j\|_\infty \leq 2$.

Proof. In view of Lemma 6.6.3 the cardinal B-splines are bounded in supnorm by 1. This implies our statement by (6.8) and applying the triangle inequality. ■

6.6 Cardinal B-splines

One of the key steps in the proof of Lemma 6.5.1 is to use approximation of B-splines with deep neural networks, derived in [73]. In this chapter we collect properties of the cardinal B-splines used in our analysis. More specifically we show that they can be used to approximate functions in Besov spaces and we verify that they form a bounded, near orthogonal basis.

We start by defining cardinal B-splines of order q in $[0,1]$ and then extend the definition with tensors to the d-dimensional unit cube. Given $J+1$ knots $0 = t_0 < t_1 < ... < t_J = 1$, the function $f\colon [0,1] \mapsto \mathbb{R}$ is a spline of order q if its restriction to the interval $[t_i, t_{i+1}]$, $i = 0, ..., J-1$ is a polynomial of degree at most $q-1$ and $f \in C^{q-2}[0,1]$ (provided that $q \geq 2$). For simplicity we will consider equidistant knots, i.e. $t_i = i/J$, $i \in \{0, ..., J\}$, but our results can be extended to a more general knot structure as well.

Splines form a linear space and a convenient basis for this space are given by B-splines $B_{1,q}, ..., B_{J,q}$. B-splines are defined recursively in the following way. First let us introduce additional knots at the boundary $t_{-q+1} = ... = t_{-1} = t_0 = 0$ and $t_J = t_{J+1} = ... = t_{J+q-1}$. Then we define the first order B-spline basis as $B_{j,1}(x) = 1_{t_j \leq x < t_{j+1}}$, $j = 0, ..., J-1$. For higher order basis we use the recursive formula

$$B_{j,q}(x) = \frac{x - t_j}{t_{j+q-1} - t_j} B_{j,q-1}(x) + \frac{t_{j+q} - x}{t_{j+q} - t_{j+1}} B_{j+1,q-1}(x), \quad j = -q+1, ..., J-1.$$

From now on for simplicity we omit the order q of the B-splines from the notation, writing $B_1, ..., B_J$. We extend B-splines to dimension d by tensorisation. For $x \in [0,1]^d$ the d-dimensional cardinal B-splines are formed by taking the product of one dimensional B-splines, i.e. for $j \in \{1, ..., J\}^d$ and $x \in [0,1]^d$ we define

$$B_j(x) = \prod_{\ell=1}^{d} B_{j_\ell}(x_\ell). \tag{6.9}$$

Remark 6.6.1. *We note that the d-dimensional index $j \in \{1, ..., J\}^d$ can be replaced by a single index running from* 1 *to $J^d =: k$. In this section for convenience we work with the multi-index formulation, but in the rest of the paper we consider the single index formulation.*

Next we list a few key properties of d-dimensional cardinal B-splines used in our proofs. In view of Chapter 12 of [71] (see Definition 12.3 and Theorems 12.4-12.8) and Lemma E.7 of [33], the cardinal B-splines have optimal approximation properties in the following sense.

Lemma 6.6.2. *Let S be the space spanned by the cardinal B-splines of order $q \geq \beta$. Then there exists a constant $C > 0$ such that for all $f \in S_d^\beta(M)$ and all integers $\alpha \leq \beta$*

$$d(f, S) = \inf_{s \in S} \|f - s\|_2 \leq C k^{-\beta/d} \sum_{l=1}^{d} \left\| \frac{\partial^\alpha f}{\partial x_l^\alpha} \right\|_2,$$

with $k = J^d$. Moreover, if $\|f\|_\infty < F$, then one can pick s such that $\|s\|_\infty < F$.

Next we show that cardinal B-splines are near orthogonal. The one dimensional case was considered in Lemma E.6 of [33]. Here we extend these results to dimension d. Note that by tensorisation we will have $k = J^d$ spline basis functions.

Lemma 6.6.3. *Let us denote by $B = (B_j)_{j \in \{1,\dots,J\}^d}$ the collection of B-splines and by $\theta = (\theta_j)_{j \in \{1,\dots,J\}^d}$ the corresponding coefficients. Let $k = J^d$. Then there exists constant $c \in (0,1)$ such that*

$$c\|\theta\|_\infty \leq \|\theta^T B\|_\infty \leq \|\theta\|_\infty,$$
$$c\|\theta\|_2 \leq \sqrt{k}\|\theta^T B\|_2 \leq \|\theta\|_2.$$

Proof. The bounds for the supremum norm follow from Lemma 2.2 of [20], hence it remained to deal with the bounds for the L_2-norm.

Let I_i, $i \in \{1, ..., J\}^d$, denote the hypercube $\prod_{\ell=1}^d [(i_\ell - 1)/J, i_\ell/J]$ and C_i, $i \in \{1, ..., J\}^d$ the collection of B-splines B_j, $j \in \{1, ..., J\}^d$ which attain a nonzero value on the corresponding hypercube I_i. Then

$$\|\theta^T B\|_2^2 = \int_{[0,1]^d} (\theta^T B(x))^2 \,\mathrm{d}\, x = \sum_{i \in \{1,\dots,J\}^d} \int_{I_i} \left(\sum_{j \in C_i} \theta_j B_j(x) \right)^2 \mathrm{d}\, x.$$

Note that for a one-dimensional cardinal B-spline of degree q we can distinguish $(2q-1)$ different cases, i.e. if $q \leq i_\ell \leq J - q$, the 1 dimensional splines are just translations of each other. Since the d-dimensional B-splines are defined as a tensor product of d one-dimensional B-splines the number of distinct cases is $(2q-1)^d$.

Define the translation map $T_i \colon \mathbb{R}^d \mapsto \mathbb{R}^d$, $i \in \{1, ..., J\}^d$, to be the map given by $T_i(x) = (\frac{x_1 - i_1 + 1}{J}, \dots, \frac{x_d - i_d + 1}{J})$, then $\det T_i = J^{-d}$. This maps into the same space of polynomials regardless of i. This means

$$\int_{I_i} \left(\sum_{j \in C_i} \theta_j B_j(x) \right)^2 \mathrm{d}\, x = J^{-d} \int_{[0,1]^d} \left(\sum_{j \in C_i} \theta_j B_j(T_i(x)) \right)^2 \mathrm{d}\, x.$$

We argue per case now. On each of these hypercubes I_i, $i \in \{1,...,J\}^d$, the splines are locally polynomials. Then the inverse of T_i defines a linear map between the polynomials spanned by the splines and the space of polynomials P of order q. The splines define q^d basis functions on our cubes I_i. Observe that each of the linear maps T_i map the B-spline basis functions B_j, $j \in \{1,...,J\}^d$ to the same space of polynomials. Note that by [71, Theorem 4.5] and the rescaling property the 1 dimensional splines restricted to the interval $[\frac{i}{j}, \frac{i+1}{J}]$ are linearly-independent. By tensorisation it follows that the B-spline basis restricted to the hypercube I_i provides a linearly indepedent polynomial basis, hence $\sum_{j\in C_i} \theta_j^2$ defines a squared norm of the functions $x \mapsto \sum_{j\in C_i} \theta_j B_j(x)$, $x \in I_i$. Since in finite dimensional real vector spaces all norms are equivalent this results in

$$\begin{aligned}\sum_{j\in C_i} \theta_j^2 &\asymp \int_{[0,1]^d} \left(\sum_{j\in C_i} \theta_j B_j(T_i(x)) \right)^2 \mathrm{d}\,x \\ &= J^d \int_{I_i} \left(\sum_{j\in C_i} \theta_j B_j(x) \right)^2 \mathrm{d}\,x.\end{aligned}$$

In view of the argument above, there are at most $(2q-1)^d$ different groups of hypercubes, hence the above result can be extended to the whole interval $[0,1]^d$ as well (by taking the worst case scenario constants in the above inequality out of the finitely many one), i.e.

$$\begin{aligned}\sum_{i\in\{1,....,J\}^d} \sum_{j\in C_i} \theta_j^2 \asymp J^d &\sum_{i\in\{1,....,J\}^d} \int_{I_i} \left(\sum_{j\in C_i} \theta_j B_j(x) \right)^2 \mathrm{d}\,x \\ &= J^d \int_{[0,1]^d} \left(\sum_{j\in\{1,....,J\}^d} \theta_j B_j(x) \right)^2 \mathrm{d}\,x.\end{aligned}$$

Since every j on the left hand side occurs at most $(2q-1)^d$ many times in the sum, this leads us to

$$\|\theta\|_2^2 \asymp J^d \|\theta^T B\|_2^2,$$

for some universal constants, concluding the proof of our statement. ■

6.7 Concentration rates and uncertainty quantification of the posterior distribution

In this section we provide posterior contraction rates and lower bounds for the radius of the credible balls under general conditions. These results are then applied for the Empirical Bayes Deep Neural Network method in Section 6.4.

6.7.1 Coverage theorem - general form

In this section we first provide a general theorem on the size of credible sets based on sieve type of priors. This result can be used beyond the nonparametric regression model and is basically the adaptation of Lemma 4 of [67] to the non-adaptive setting with fixed sieve dimension k, not chosen by the empirical Bayes method as in [67]. This theorem is of separate interest, as it can be used for instance for extending our results to other models, including nonparametric classification.

We start by introducing the framework under which our results hold. We consider a general statistical model, i.e. we assume that our data $\mathbb{D}_n$ is generated from a distribution $\mathbb{P}_{f_0}$ indexed by an unknown functional parameter of interest f_0 belonging to some class of functions $\mathcal{F}$. Let us consider $k = k_n$ (not necessarily orthogonal) basis functions $\phi_1, ..., \phi_k \in \mathcal{F}$ and use the notation $f_\theta(x) = \sum_{i=1}^k \theta_i \phi_i(x)$. Then we define the class $\Theta_k = \{\sum_{i=1}^k \theta_i \phi_i, \theta_i \in \mathbb{R}, i = 1, ..., k\} \subset \mathcal{F}$ and equivalently we also refer to the elements of this class using the coefficients $(\theta_1, ..., \theta_k)$. We note that f_0 doesn't necessarily belong to the sub-class Θ_k.

Furthermore, let us consider a pseudometric d_n: $\mathcal{F} \times \mathcal{F} \mapsto \mathbb{R}$ and take $\theta^o = \arg\inf_{\theta \in \Theta_k} d_n(f_\theta, f_0)$, i.e. the projection of f_0 to the space Θ_k is f_{θ^o} with $\theta^o = (\theta_1^o, ..., \theta_k^o)^T \in \Theta_k$ denoting the corresponding coefficient vector. Let us also consider a metric d: $\Theta_k \times \Theta_k \mapsto \mathbb{R}$ on the k-dimensional parameter space Θ_k. Finally, we introduce the notation $B_k(\bar{\theta}, \varepsilon, d) = \{\theta \in \Theta_k\colon d(\theta, \bar{\theta}) \leq \varepsilon\}$ for the ε-radius d-ball in Θ_k centered at $\bar{\theta} \in \Theta_k$ and $B(\tilde{f}, \varepsilon, d_n) = \{f \in \mathcal{F}\colon d_n(\tilde{f}, f) \leq \varepsilon\}$ for the ε-radius d_n-ball centered at $\tilde{f} \in \mathcal{F}$.

The next theorem provides lower bound for the radius $r_{n,\alpha}$ of the credible balls

$$B(f_{\hat{\theta}}, \delta_\varepsilon \varepsilon_n, d_n) = \{f \in \mathcal{F}\colon d_n(f_\theta, f_{\hat{\theta}}) \leq r_{n,\alpha}\}$$

centered around an estimator $f_{\hat{\theta}}$. The radius $r_{n,\alpha}$ is defined as

$$\Pi_k\Big(\theta \in \Theta_k\colon f_\theta \in B(f_{\hat{\theta}}, r_{n,\alpha}, d_n) | \mathbb{D}_n\Big) = 1 - \alpha.$$

Before stating the theorem we introduce some assumptions.

A1 The centering point $f_{\hat{\theta}} \in \mathcal{F}$ satisfies that for all $\epsilon > 0$ there exists $M_\epsilon > 0$

$$\sup_{f_0 \in \mathcal{F}} \mathbb{P}_{f_0}\Big(d_n(f_{\theta^o}, f_{\hat{\theta}}) \leq M_\epsilon \sqrt{k/n}\Big) \geq 1 - \varepsilon.$$

A2 Assume that there exists $C_m > 0$ such that for all $\theta, \theta' \in \Theta_k$

$$C_m^{-1} d(\theta, \theta') \leq d_n(f_\theta, f_{\theta'}) \leq C_m d(\theta, \theta').$$

A3 Assume that for all $M, \epsilon > 0$ there exist constants $c_1, c_2, c_3, c_4, \delta_0, B_\epsilon > 0$ and $r \geq 2$ such that the following conditions hold

A3.i

$$B_k(\theta^o, \sqrt{k/n}, d) \subset S_n(k, c_1, c_2, r),$$

where $S_n(k, c_1, c_2, r) =$

$$\Big\{\theta \in \Theta_k\colon \mathbb{E}_{f_0} \log \frac{p_{\theta^o}}{p_\theta} \leq c_1 k, \mathbb{E}_{f_0}\Big(\log \frac{p_{\theta^o}}{p_\theta} - \mathbb{E}_{f_0} \log \frac{p_{\theta^o}}{p_\theta}\Big)^r \leq c_2 k^{r/2}\Big\}.$$

A3.ii Let $\bar{B}_k = \Theta_k \cap B(f_{\theta^o}, M_\epsilon \sqrt{k/n}, d_n)$. Then for every $f_0 \in \mathcal{F}$

$$\mathbb{P}_{f_0}\Big(\sup_{f_\theta \in \bar{B}_k} \ell_n(\theta) - \ell_n(\theta^o) \leq B_\varepsilon k\Big) \geq 1 - \varepsilon,$$

where $\ell_n(\theta)$ denotes the log-likelihood corresponding to the functional parameter $f_\theta \in \Theta_k$.

A3.iii For every δ_0 small enough

$$\frac{\sup_{\theta \in \bar{B}_k} \Pi_k\big(B_k(\theta, \delta_0 \sqrt{k/n}, d)\big)}{\Pi_k\big(B_k(\theta^o, \sqrt{k/n}, d)\big)} \leq c_4 e^{c_3 k \log(\delta_0)}.$$

Theorem 6.7.1. *Assume that conditions **A1**- **A3** hold. Then for every $\varepsilon > 0$ there exists a small enough $\delta_\varepsilon > 0$ such that*

$$\sup_{f_0 \in \mathcal{F}} \mathbb{E}_{f_0} \Pi_k\left(\theta \in \Theta_k\colon d_n(f_\theta, f_{\hat{\theta}}) \leq \delta_\varepsilon \sqrt{k/n} | \mathbb{D}_n\right) \leq \varepsilon.$$

Proof. First note that since Π_k is supported on Θ_k,

$$\Pi_k\left(\theta\colon d_n(f_\theta, f_{\hat{\theta}}) \leq \delta_\varepsilon \sqrt{k/n} | \mathbb{D}_n\right) = \frac{\int_{B(f_{\hat{\theta}}, \delta_\varepsilon \sqrt{k/n}, d_n) \cap \Theta_k} e^{\ell_n(\theta) - \ell_n(\theta^o)} d\Pi_k(\theta)}{\int_{\Theta_k} e^{\ell_n(\theta) - \ell_n(\theta^o)} d\Pi_k(\theta)}. \quad (6.10)$$

Next let us introduce the notations

$$\Omega_n(C) = \left\{ e^{Ck} \frac{\int_{\Theta_k} e^{\ell_n(\theta) - \ell_n(\theta^o)} d\Pi_k(\theta)}{\Pi_k\big(B_k(\theta^o, \sqrt{k/n}, d)\big)} \geq 1 \right\}, \quad (6.11)$$

$$\Gamma_n(B) = \sup_{B(f_{\hat{\theta}}, \delta_\varepsilon \sqrt{k/n}, d_n) \cap \Theta_k} \ell_n(\theta) - \ell_n(\theta^o) < Bk. \quad (6.12)$$

Note that in view of Assumptions **A3.ii** and **A1** we have that $\inf_{f_0} \mathbb{P}_{f_0}(\Gamma_n(B_\varepsilon)) \geq 1 - 2\epsilon$ for some large enough constant $B_\varepsilon > 0$ and in view of Assumption **A3.i** by using the standard technique for lower bound for the likelihood ratio ([33, Lemma

8.37]) we have with $\mathbb{P}_{f_0}$-probability bounded from below by $1-\varepsilon$ that there exists $c_0>0$ such that

$$\int_{\Theta_k} e^{\ell_n(\theta)-\ell_n(\theta^o)} d\Pi_k(\theta) \geq e^{-(c_0+1/\sqrt{\epsilon})k}\Pi_k\big(S_n(k,c_1,c_2,r)\big)$$
$$\geq e^{-(c_0+1/\sqrt{\epsilon})k_n}\Pi_k\big(B_k(\theta^o,\sqrt{k/n},d)\big),$$

hence $\mathbb{P}_{f_0}(\Omega_n(c_0+1/\sqrt{\epsilon}))\geq 1-\varepsilon$.

Therefore, in view of assumption **A2**, the right hand side of (6.10) is bounded from above on $A_n=\Omega_n(c_0+1/\sqrt{\epsilon})\cap\Gamma_n(B_\varepsilon)\cap\{d_n(f_{\theta^o},f_{\hat\theta})\leq M_\epsilon\sqrt{k/n}\}$ by

$$e^{(B_\varepsilon+c_0+1/\sqrt{\epsilon})k}\frac{\Pi_k\Big(\Theta_k\cap B(f_{\hat\theta},\delta_\varepsilon\sqrt{k/n},d_n)\Big)}{\Pi_k\Big(B_k(\theta^o,\sqrt{k/n},d)\Big)}$$
$$\leq e^{(c_0+B_\varepsilon+1/\sqrt{\epsilon})k}\frac{\Pi_k\big(B_k(\hat\theta,C_m\delta_\varepsilon\sqrt{k/n},d)\big)}{\Pi_k\Big(B_k(\theta^o,\sqrt{k/n},d)\Big)}$$
$$\leq Ce^{\big(c_0+B_\varepsilon+1/\sqrt{\epsilon}+c_3\log(C_m\delta_\varepsilon)\big)k}\leq\varepsilon,$$

for small enough choice of $\delta_\varepsilon>0$, where the last line follows from assumption **A3.iii** (with $\delta_0=C_m\delta_\varepsilon$). Furthermore, note that

$$\mathbb{P}_{f_0}(A_n^c)\leq\mathbb{P}_{f_0}\big(\Omega_n(c_0+1/\sqrt{\epsilon})\big)+\mathbb{P}_{f_0}\big(\Gamma_n(B_\varepsilon)\big)+\mathbb{P}_{f_0}\big(d_n(f_{\theta^o},f_{\hat\theta})\geq M_\epsilon\sqrt{k/n}\big)\leq 4\varepsilon,$$

where for the last term we used Assumption **A1**. Hence the $\mathbb{E}_{f_0}$-expected value of the first term on the right hand side of (6.10) is bounded from above by 5ε. ■

6.7.2 Coverage in nonparametric regression

We apply Theorem 6.7.1 in context of the uniform random design nonparametric regression model, i.e. we observe pairs of random variables $\mathbb{D}_n=\{(X_1,Y_1),...,(X_n,Y_n)\}$ satisfying that

$$Y_i=f_0(X_i)+\varepsilon_i,\qquad X_i\overset{iid}{\sim}\text{Unif}([0,1]^d),\quad \varepsilon_i\overset{iid}{\sim}N(0,1),\quad i=1,...,n,\qquad(6.13)$$

for some unknown functional parameter $f_0\in\mathcal{F}=L_2([0,1]^d,M)$. Let us denote by X the collection of design points, i.e. $X=(X_1,...,X_n)$.

Let us consider $k=k_n$ (not necessarily orthogonal) basis functions $\phi_1,...,\phi_k\in\mathcal{F}$ and use the notation $f_\theta(x)=\sum_{i=1}^k\theta_i\phi_i(x)$. We denote by $\Phi_{n,k}$ the empirical basis matrix consisting the basis functions $\phi_1,...,\phi_k$ evaluated at the design points $X_1,...,X_n$, i.e.

$$\Phi_{n,k}=\begin{pmatrix}\phi_1(X_1)&\phi_2(X_1)&\cdots&\phi_k(X_1)\\ \phi_1(X_2)&\phi_2(X_2)&\cdots&\phi_k(X_2)\\ \vdots&\vdots&\ddots&\vdots\\ \phi_1(X_n)&\phi_2(X_n)&\cdots&\phi_k(X_n)\end{pmatrix}.$$

Furthermore, let us denote the Gram matrix of the basis functions $\phi_1, ..., \phi_k$ with respect to the L_2 inner product $\langle f, g\rangle = \int_{[0,1]^d} f(x)g(x)dx$ by

$$\Sigma_k = \begin{pmatrix} \langle\phi_1,\phi_1\rangle & \langle\phi_1,\phi_2\rangle & \cdots & \langle\phi_1,\phi_k\rangle \\ \langle\phi_2,\phi_1\rangle & \langle\phi_2,\phi_2\rangle & \cdots & \langle\phi_2,\phi_k\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle\phi_k,\phi_1\rangle & \langle\phi_k,\phi_2\rangle & \cdots & \langle\phi_k,\phi_k\rangle \end{pmatrix}. \tag{6.14}$$

Finally, we need to impose the following near orthogonality assumption on the basis functions $\phi_1, ..., \phi_k$.

B1 Assume that there exists a constant $c_m \geq 1$ such that

$$c_m^{-1} I_k \leq \Sigma_k \leq c_m I_k.$$

In our analysis we consider a prior Π_k supported on functions of the form $f_\theta = \sum_{j=1}^k \theta_j\phi_j$. We take priors of the product form, i.e.

$$d\Pi_k(\theta) = \prod_{j=1}^k g(\theta_j)d\theta,$$

for a one dimensional density g, satisfying for every $M' > 0$ that there exists constants $\underline{c}, \overline{c} > 0$ such that

$$\underline{c} \leq g(x) \leq \overline{c}, \qquad x \in [-M', M'] \tag{6.15}$$

Lemma 6.7.2. *Consider the nonparametric regression model* (6.13) *and a prior* Π_k *satisfying assumption* (6.15)*. Let* $\beta > \frac{d}{2}$ *and let* $\mathcal{F} \subset S_d^\beta(M)$*. Denote the basis functions by* ϕ_j*,* $j = 1, \dots, k$*. Assume that the basis functions are bounded in supnorm by* Ck*. We assume that the Gram matrix* Σ_k *given in* (6.14)*, for the basis functions* ϕ_j*, satisfies* ***B1*** *and that* $\sum_{j=1}^k \phi_j(x)^2 \leq Ck^2$*, for all* $x \in [0,1]^d$*. Furthermore, assume that the centering point of the credible set satisfies assumption* ***A1****. Then for every* $\varepsilon > 0$ *there exists a small enough* $\delta_\varepsilon > 0$ *such that*

$$\sup_{f_0\in\mathcal{F}} \mathbb{E}_{f_0}\Pi_k\left(\theta \in \mathbb{R}^k \colon \|f_\theta - f_{\hat\theta}\|_2 \leq \delta_\varepsilon\sqrt{k/n}|\mathbb{D}_n\right) \leq \varepsilon.$$

Proof. We show below that the conditions of Theorem 6.7.1 hold in this model for the conditional probability given the design points $\mathbb{P}_{f_0}^{|X}(\cdot) = \mathbb{P}_{f_0}(\cdot|X)$, on an event $A_n \subset \mathcal{X}^n$, where $\mathcal{X} = [0,1]^d$, satisfying $\mathbb{P}_{f_0}(A_n^c) \leq \varepsilon$, taking d_n to be the empirical L_2 semi-metric, i.e. $d_n(f,g)^2 = \|f-g\|_n^2 = \sum_{i=1}^n \big(f(X_i) - g(X_i)\big)^2$ and d the ℓ_2-metric in $\Theta_k = \mathbb{R}^k$ i.e. $d(\theta,\theta') = \|\theta - \theta'\|_2$. Hence in view of Theorem 6.7.1, on the event A_n for every $\varepsilon > 0$ there exists a small enough $\delta_\varepsilon > 0$ such that

$$\sup_{f_0\in\mathcal{F}} \mathbb{E}_{f_0}^{|X}\Pi_k\left(\theta \in \mathbb{R}^k \colon \|f_\theta - f_{\hat\theta}\|_n \leq 2\delta_\varepsilon\sqrt{k/n}|\mathbb{D}_n\right) \leq \varepsilon.$$

Next note that in view of assertion (6.16), see below, we get on an event B_n, with $\mathbb{P}_{f_0}(B_n^c) \leq \varepsilon$, that

$$\|f_\theta - f_{\hat\theta}\|_n/2 \leq \|f_\theta - f_{\hat\theta}\|_2,$$

resulting in

$$\begin{aligned}
&\sup_{f_0\in\mathcal{F}} \mathbb{E}_{f_0}\Pi_k\left(\theta\colon \|f_\theta - f_{\hat\theta}\|_2 \leq \delta_\varepsilon\sqrt{k/n}|\mathbb{D}_n\right)\\
&\qquad\leq \sup_{f_0\in\mathcal{F}} \mathbb{E}_{f_0}\mathbb{E}_{f_0}^{|X}\left(\Pi_k\big(\theta\colon \|f_\theta - f_{\hat\theta}\|_n \leq 2\delta_\varepsilon\sqrt{k/n}|\mathbb{D}_n\big)\right)\\
&\qquad\leq \sup_{f_0\in\mathcal{F}} \mathbb{P}_{f_0}(A_n^c) + \mathbb{P}_{f_0}(B_n^c) + \varepsilon \leq 3\varepsilon.
\end{aligned}$$

It remained to prove that the conditions of Theorem 6.7.1 hold.

Condition A1. Follows by the choice of the centering point.

Condition A2. First note that $\Sigma_{n,k} = n^{-1}\Phi_{n,k}^T\Phi_{n,k}$ has mean Σ_k. Then by the modified version of Rudelson's inequality [68] we get that

$$\mathbb{E}_{f_0}\|\Sigma_{n,k} - \Sigma_k\|_2 \leq C\sqrt{\frac{\log k}{n}}\mathbb{E}_{f_0}(\|\boldsymbol{\phi}(X_1)\|_2^{\log n})^{1/\log n},$$

with $\boldsymbol{\phi}(X_1) = \big(\phi_1(X_1), ..., \phi_k(X_1)\big)^T$. Note that by the boundedness assumption $\sum_{j=1}^k \phi_j(x)^2 \leq Ck^2$, $x \in [0,1]^d$, the right hand side of the preceding display is bounded from above by constant times $\sqrt{k^2\log(k)/n}$ on an event B_n with $\mathbb{P}_{f_0}(B_n)$ tending to one. This upperbound is $o(1)$ if $\beta > \frac{d}{2}$. Therefore,

$$\Big|\|f_\theta\|_2^2 - \|f_\theta\|_n^2\Big| = \Big|\theta^T(\Sigma_k - \Sigma_{n,k})\theta\Big| \leq \|\Sigma_k - \Sigma_{n,k}\|_2\|\theta\|_2^2 = o_{\mathbb{P}_{f_0}}(\|\theta\|_2^2).$$

Furthermore, in view of Assumption **B1**

$$c_m^{-1}\|\theta\|_2^2 \leq \|f_\theta\|_2^2 = \theta^T\Sigma_k\theta \leq c_m\|\theta\|_2^2, \qquad \text{for all } \theta \in \mathbb{R}^k,$$

which in turn implies that on B_n

$$(2c_m)^{-1}\|\theta\|_2^2 \leq \|f_\theta\|_2^2/2 \leq \|f_\theta\|_n^2 \leq 2\|f_\theta\|_2^2 \leq 2c_m\|\theta\|_2^2, \tag{6.16}$$

holds for all $\theta \in \mathbb{R}^k$.

Condition A3.i. First note that for arbitrary $\theta \in \mathbb{R}^k$

$$\begin{aligned}\ell_n(\theta) - \ell_n(\theta^o) &= \frac{1}{2}\sum_{i=1}^n (Y_i - f_{\theta^o}(X_i))^2 - \frac{1}{2}\sum_{i=1}^n (Y_i - f_\theta(X_i))^2\\ &= -\sum_{i=1}^n \Big((f_\theta(X_i) - f_{\theta^o}(X_i))^2/2 - (Y_i - f_{\theta^o}(X_i))(f_\theta(X_i) - f_{\theta^o}(X_i))\Big)\\ &= -\sum_{i=1}^n \big(f_\theta(X_i) - f_{\theta^o}(X_i)\big)^2/2 - \sum_{i=1}^n \varepsilon_i (f_\theta(X_i) - f_{\theta^o}(X_i))\\ &\quad - \sum_{i=1}^n (f_0(X_i) - f_{\theta^o}(X_i))(f_\theta(X_i) - f_{\theta^o}(X_i))\\ &= -\sum_{i=1}^n \big(f_\theta(X_i) - f_{\theta^o}(X_i)\big)^2/2 - \sum_{i=1}^n \varepsilon_i \big(f_\theta(X_i) - f_{\theta^o}(X_i)\big), \qquad (6.17)\end{aligned}$$

where in the last line we used that f_{θ^o} is the orthogonal projection of f_0 to $\Theta_k = \{\sum_{i=1}^k \theta_i \phi_i \colon \theta \in \mathbb{R}^k\}$ with respect to the empirical Euclidean norm d_n. Then by taking $\mathbb{E}_{f_0}^{|X}$-expectation on both sides of (6.17) we get that

$$\mathbb{E}_{f_0}^{|X}\big(\ell_n(\theta) - \ell_n(\theta^o)\big) = n\|f_\theta - f_{\theta^o}\|_n^2/2 \leq c_m n \|\theta - \theta^o\|_2^2.$$

Similarly $\mathbb{E}_{f_0}^{|X}\big[\ell_n(\theta) - \ell_n(\theta^o) - \mathbb{E}_{f_0}^{|X}(\ell_n(\theta) - \ell_n(\theta^o))\big]^2 \leq 2c_m n \|f_\theta - f_{\theta^o}\|_2^2$, hence $B_k(\theta^o, \sqrt{k/n}, \|\cdot\|_2) \subset \mathcal{S}_n(k, c_m, 2c_m, 2)$.

Condition A3.ii. In view of assertion (6.17) and using Cauchy-Schwarz inequality (as in inequality (A.3) of the supplementary material of [67] we arrive at

$$\begin{aligned}\ell_n(\theta) - \ell_n(\theta^o) &= -n\|f_\theta - f_{\theta^o}\|_n^2/2 - \varepsilon^T \Phi_{n,k}(\theta - \theta^o)\\ &\leq -n\|f_\theta - f_{\theta^o}\|_n^2/2 + \|\varepsilon^T \Phi_{n,k}\|_2 \|\theta^o - \theta\|_2.\end{aligned}$$

We show below that with $\mathbb{P}_{f_0}^{|X}$-probability tending to one

$$\|\varepsilon^T \Phi_{n,k}\|_2^2 \leq Ckn. \qquad (6.18)$$

Hence on the same event we get that

$$\begin{aligned}\ell_n(\theta) - \ell_n(\theta^o) &\leq \sqrt{n}\|f_\theta - f_{\theta^o}\|_n \Big(C\sqrt{c_m} M_\varepsilon \sqrt{k} - \sqrt{n}\|f_\theta - f_{\theta^o}\|_n/2\Big)\\ &\leq 2c_m C^2 M_\varepsilon^2 k.\end{aligned}$$

It remained to prove that (6.18) holds with probability tending to one. Note that in view of assertion (6.16) on an event B_n, with $\mathbb{P}_{f_0}(B_n) \to 1$ we get for $\varepsilon \sim N_k(0, I_k)$

$$\|\varepsilon^T \Phi_{n,k}\|_2^2 = n\varepsilon^T \Sigma_{n,k} \varepsilon \leq 2c_m n \|\varepsilon\|_2^2.$$

Then by the properties of the χ^2_k distribution the right hand side of the preceding display is bounded from above by $4c_m^2 nk$ with probability tending to one as k tends to infinity.

Condition A3.iii In view of assertion (6.16) on the event B_n we get that

$$B(f_\theta, \delta_0\sqrt{k/n}, \|\cdot\|_n) \subset B_k(\theta, \sqrt{2c_m}\delta_0\sqrt{k/n}, \|\cdot\|_2),$$
$$B(f_{\theta^o}, \sqrt{k/n}, \|\cdot\|_n) \supset B_k(\theta^o, \sqrt{k/n}/\sqrt{2c_m}, \|\cdot\|_2).$$

Next note that in view of (6.16) on an event B_n with $\mathbb{P}_{f_0}(B_n) \to 1$ we have

$$(2c_m)^{-1}\|\theta^o\|_2^2 \leq \|f_{\theta^o}\|_n^2 \leq \|f_0\|_n^2 \leq M, \tag{6.19}$$

resulting in $|\theta_j^o| \leq \sqrt{2c_m M} < M'$. Therefore the density of the prior is bounded away from zero and infinity on a neighbourhood of θ^o which in turn implies that

$$\sup_{\theta\in B(f_{\theta^o}, M, \|\cdot\|_n)} \frac{\Pi_k\big(B(f_\theta, \delta_0\sqrt{k/n}, \|\cdot\|_n)\big)}{\Pi_k\big(B(f_{\theta^o}, \sqrt{k/n}, \|\cdot\|_n)\big)} \lesssim \frac{\mathrm{Vol}\big(B_k(\theta, \sqrt{2c_m}\delta_0\sqrt{k/n}, \|\cdot\|_2)\big)}{\mathrm{Vol}\big(B_k(\theta^o, \sqrt{k/n}/\sqrt{2c_m}, \|\cdot\|_2)\big)}$$
$$\lesssim e^{ck\log(2c_m\delta_0)}.$$

■

6.7.3 Misspecified contraction rates

Finally, we derive a contraction rate result for the posterior in our misspecified setting. We assume that our true model parameter is f_0, which however, does not necessarily belong to our model $\Theta_k = \{f_\theta = \sum_{i=1}^k \theta_i\phi_i\colon \theta \in \mathbb{R}^k\}$. Let us denote by $f^* = f_{\theta^*}$ the L_2-projection of f_0 into the subspace Θ_k. We show below that the posterior contracts with the L_2-rate $\sqrt{k/n}$ around f^* in the regression model.

Theorem 6.7.3. *Consider the random design nonparametric regression model with observations* $\mathbb{D}_n = \big((X_1, Y_1), ..., (X_n, Y_n)\big)$ *and assume that the Gram matrix* Σ_k *given in* (6.14) *satisfies Assumption* ***B1*** *and that the prior* Π_k *satisfies* (6.15)*. Then*

$$\limsup_{n\to\infty} \sup_{f_0\in S_d^\beta(M)\cap L_\infty(M)} \mathbb{E}_{f_0}\Pi_k\big(\theta\in\mathbb{R}^k\colon \|f_\theta - f^*\|_2 \geq M_n\sqrt{k/n}|\mathbb{D}_n\big) = 0$$

for all $M_n \to \infty$.

Proof. For ease of notation we set $\varepsilon_n = \sqrt{k/n}$. Then we show below that the following two inequalities hold for some constant $J > 0$,

$$\frac{\Pi_k\left(\theta\colon j\varepsilon_n \leq \|f_\theta - f^*\|_2 \leq 2j\varepsilon_n\right)}{\Pi_k\left(B(f^*, \varepsilon_n, \|\cdot\|_2)\right)} \leq e^{j^2k/8}, \qquad \text{for all } j \geq J, \tag{6.20}$$

$$\log N(\epsilon, \{f_\theta\colon \epsilon < \|f^* - f_\theta\|_2 \leq 2\epsilon\}, \|\cdot\|_2) \leq k, \quad \text{for all } \epsilon > 0. \tag{6.21}$$

The function class $\mathcal{F} = S_d^\beta(M) \cap L_\infty(M)$ is closed, convex, and uniformly bounded. Furthermore, since the Gaussian noise satisfies $\mathbb{E}_{f_0} e^{M|\varepsilon_i|} < \infty$ for all $M > 0$, in view of Lemma 8.41 of [33] condition (8.52) of [33] holds. Therefore, in view of Lemma 8.38 of [33] the logarithm of the covering number for testing under misspecification is bounded from above by $\log N(\epsilon, \{f_\theta \colon \epsilon < \|f^* - f_\theta\|_2 \leq 2\epsilon\}, \|\cdot\|_2)$, which in turn is bounded by k following from (6.21). Then our statement follows by applying Theorem 8.36 of [33] (with $\varepsilon_n = \bar{\varepsilon}_n = k/n$, $d = \|\cdot\|_2$, $\mathcal{P}_{n,1} = \mathcal{F}$, $\mathcal{P}_{n,2} = \emptyset$).

It remained to verify conditions (6.20) and (6.21).

Proof of (6.20). First note that in view of condition **B1**

$$\frac{\Pi_k\big(\theta \colon j\varepsilon_n \leq \|f_\theta - f^*\|_2 \leq 2j\varepsilon_n\big)}{\Pi_k\big(B(f^*, \varepsilon_n, \|\cdot\|_2)\big)} \leq \frac{\Pi_k\big(B_k(\theta^*, 2j\sqrt{c_m}\varepsilon_n, \|\cdot\|_2)\big)}{\Pi_k\left(B_k(\theta^*, \varepsilon_n/\sqrt{c_m}, \|\cdot\|_2)\right)}.$$

Furthermore, note that the prior density is bounded from above and below by $\overline{c}^k$ and $\underline{c}^k$, respectively, in a neigbourhood of f_{θ^*} following from assumption (6.15) and by similar argument as in (6.19). Therefore the prior probability of a given set A can be upper and lower bounded by the Euclidean volume of A times $\overline{c}^k$ and $\underline{c}^k$, respectively. This implies that the preceding display can be further bounded from above by

$$\Big(\frac{\overline{c}}{\underline{c}}\Big)^k \frac{\mathrm{Vol}\Big(B_k(\theta^*, 2j\sqrt{c_m}\varepsilon_n, \|\cdot\|_2)\Big)}{\mathrm{Vol}\Big(B_k(\theta^*, \varepsilon_n/\sqrt{c_m}, \|\cdot\|_2)\Big)} \leq \Big(\frac{2c_m j\overline{c}}{\underline{c}}\Big)^k \leq e^{j^2k/8},$$

for all $j \geq J$, for J large enough.

Proof of (6.21). First note that by Assumption **B1**

$$\begin{aligned}&\log N(\epsilon, \{f_\theta \colon \epsilon < \|f_\theta - f^*\|_2 \leq 2\epsilon\}, \|\cdot\|_2)\\ &\qquad\leq \log N(\epsilon c_m^{-1/2}, \{\theta \colon \epsilon c_m^{-1/2} < \|\theta - \theta^*\|_2 \leq 2\epsilon c_m^{1/2}\}, \|\cdot\|_2).\end{aligned}$$

Then in view of the standard bound on the local entropies of k-dimensional Euclidean balls the right hand side is further bounded by constant times k, finishing the proof of our statement.

■

Bibliography

[1] Martín Abadi et al. *TensorFlow: Large-scale Machine Learning on Heterogeneous Systems.* 2015. URL: `https://www.tensorflow.org/`.

[2] Martin Anthony, Peter L Bartlett, Peter L Bartlett, et al. *Neural Network Learning: Theoretical Foundations.* Vol. 9. cambridge university press Cambridge, 1999.

[3] Julyan Arbel, Pierpaolo De Blasi, and Igor Prünster. "Stochastic Approximations to the Pitman-Yor Process". In: *Bayesian Analysis* (2018). DOI: `10.1214/18-BA1127`. URL: `https://doi.org/10.1214/18-BA1127`.

[4] Jincheng Bai, Qifan Song, and Guang Cheng. "Efficient Variational Inference for Sparse Deep Learning with Theoretical Guarantee". In: *Neurips* abs/2011.07439 (2020).

[5] Eduard Belitser. "On Coverage and Local Radial Rates of Credible Sets". In: *The Annals of Statistics* 45.3 (2017), pp. 1124–1151. DOI: `10.1214/16-AOS1477`. URL: `https://doi.org/10.1214/16-AOS1477`.

[6] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation.* Vol. 27. Encyclopedia of Mathematics and Its Applications. Cambridge University Press, Cambridge, 1989, pp. xx+494. ISBN: 0-521-37943-1.

[7] Federico Camerlenghi et al. "Distribution Theory for Hierarchical Processes". In: *The Annals of Statistics* 47.1 (Feb. 1, 2019), pp. 67–92. DOI: `10.1214/17-AOS1678`. URL: `https://doi.org/10.1214/17-AOS1678`.

[8] Federico Camerlenghi et al. "Latent Nested Nonparametric Priors (with Discussion)". In: *Bayesian Analysis* 14.4 (2019), pp. 1303–1356. ISSN: 1936-0975. DOI: `10.1214/19-BA1169`. URL: `https://doi-org.ezproxy.leidenuniv.nl/10.1214/19-BA1169`.

[9] Ismaël Castillo. "A Semiparametric Bernstein–von Mises Theorem for Gaussian Process Priors". In: *Probability Theory and Related Fields* 152.1-2 (2012), pp. 53–99. ISSN: 0178-8051. DOI: `10.1007/s00440-010-0316-5`. URL: `https://doi-org.tudelft.idm.oclc.org/10.1007/s00440-010-0316-5`.

[10] Ismaël Castillo and Richard Nickl. "On the Bernstein-von Mises Phenomenon for Nonparametric Bayes Procedures". In: *The Annals of Statistics* 42.5 (Oct. 1, 2014). ISSN: 0090-5364. DOI: `10.1214/14-AOS1246`. arXiv: `1310.2484`. URL: `http://arxiv.org/abs/1310.2484` (visited on 10/01/2021).

[11] Ismaël Castillo and Judith Rousseau. "A Bernstein–von Mises Theorem for Smooth Functionals in Semiparametric Models". In: *The Annals of Statistics* 43.6 (2015), pp. 2353–2383. ISSN: 0090-5364. DOI: 10.1214/15-AOS1336. URL: https://doi-org.tudelft.idm.oclc.org/10.1214/15-AOS1336.

[12] Giulia Cereda. "Current Challenges in Statistical DNA Evidence Evaluation". PhD thesis. Leiden University, 2017.

[13] Giulia Cereda and Richard D. Gill. *A Nonparametric Bayesian Approach to the Rare Type Match Problem.* 2019. arXiv: 1908.02954 [stat.AP].

[14] Francois Chollet et al. *Keras.* 2015. URL: https://github.com/fchollet/keras.

[15] Davide Cirillo and Alfonso Valencia. "Big Data Analytics for Personalized Medicine". In: *Systems Biology • Nanobiotechnology* 58 (Aug. 1, 2019), pp. 161–167. ISSN: 0958-1669. DOI: 10.1016/j.copbio.2019.03.004. URL: https://www.sciencedirect.com/science/article/pii/S0958166918301903.

[16] David Freedman. "Wald Lecture: On the Bernstein-von Mises Theorem with Infinite-Dimensional Parameters". In: *The Annals of Statistics* 27.4 (Aug. 1, 1999), pp. 1119–1141. DOI: 10.1214/aos/1017938917. URL: https://doi.org/10.1214/aos/1017938917.

[17] P. De Blasi, A. Lijoi, and I. Prünster. "An Asymptotic Analysis of a Class of Discrete Nonparametric Priors". In: *Statistica Sinica* 23.3 (2013), pp. 1299–1321.

[18] P. De Blasi et al. "Are Gibbs-type Priors the Most Natural Generalization of the Dirichlet Process?" In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.2 (Feb. 2015), pp. 212–229. ISSN: 0162-8828, 2160-9292. DOI: 10.1109/TPAMI.2013.217. arXiv: 1503.00163 [math, stat]. URL: http://arxiv.org/abs/1503.00163 (visited on 10/20/2022).

[19] Laurens de Haan and Ana Ferreira. *Extreme Value Theory.* Springer Series in Operations Research and Financial Engineering. Springer, New York, 2006, pp. xviii+417. ISBN: 978-0-387-23946-0. DOI: 10.1007/0-387-34471-3. URL: https://doi-org.tudelft.idm.oclc.org/10.1007/0-387-34471-3.

[20] R. de Jonge and J.H. van Zanten. "Adaptive Estimation of Multivariate Functions Using Conditionally Gaussian Tensor-Product Spline Priors". In: *Electronic Journal of Statistics* 6 (2012), pp. 1984–2001. DOI: 10.1214/12-EJS735. URL: https://doi.org/10.1214/12-EJS735.

[21] Dennis D. Cox. "An Analysis of Bayesian Inference for Nonparametric Regression". In: *The Annals of Statistics* 21.2 (June 1, 1993), pp. 903–923. DOI: 10.1214/aos/1176349157. URL: https://doi.org/10.1214/aos/1176349157.

[22] Monroe D. Donsker. "Justification and Extension of Doob's Heuristic Approach to the Komogorov-Smirnov Theorems". In: 23 (1952), pp. 277–281. ISSN: 0003-4851.

[23] Stefano Favaro and Zacharie Naulet. "Near-Optimal Estimation of the Unseen under Regularly Varying Tail Populations". Apr. 7, 2021. arXiv: 2104.03251 [math, stat]. URL: http://arxiv.org/abs/2104.03251 (visited on 05/26/2021).

[24] T. Ferguson. "Prior Distributions on Spaces of Probability Measures". In: *The Annals of Statistics* 2 (1974), pp. 615–629. ISSN: 0090-5364.

[25] S. E. M. P. Franssen. *Uncertainty Quantification Using Empirical Bayesian Deep Neural Networks.* URL: `sempf1992.github.io`.

[26] S. E. M. P. Franssen and Botond Szabo. "Frequentist Coverage Guarantees of Empirical Bayesian Uncertainty Quantification Using Deep Neural Network Regression".

[27] S. E. M. P. Franssen and A. W. van der Vaart. "Bernstein-von Mises Theorem for the Pitman-Yor Process of Nonnegative Type". In: *Electronic Journal of Statistics* 16.2 (Jan. 2022), pp. 5779–5811. ISSN: 1935-7524, 1935-7524. DOI: `10.1214/22-EJS2077`. URL: `https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-16/issue-2/Bernstein-von-Mises-theorem-for-the-Pitman-Yor-process-of/10.1214/22-EJS2077.full` (visited on 11/09/2022).

[28] S. E. M. P. Franssen and A. W. van der Vaart. *Empirical and Full Bayes Estimation of the Type of a Pitman-Yor Process.* Aug. 30, 2022. DOI: `10.48550/arXiv.2208.14255`. arXiv: `2208.14255 [math, stat]`. URL: `http://arxiv.org/abs/2208.14255` (visited on 11/09/2022).

[29] David A. Freedman. "On the Asymptotic Behavior of Bayes Estimates in the Discrete Case II". In: *The Annals of Mathematical Statistics* 36.2 (Apr. 1, 1965), pp. 454–456. DOI: `10.1214/aoms/1177700155`. URL: `https://doi.org/10.1214/aoms/1177700155`.

[30] Georg Frobenius. "Ueber Matrizen Aus Nicht Negativen Elementen". In: *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften* (1912), pp. 456–477.

[31] Jakob Gawlikowski et al. "A Survey of Uncertainty in Deep Neural Networks". July 7, 2021. arXiv: `2107.03342 [cs, stat]`. URL: `http://arxiv.org/abs/2107.03342` (visited on 08/16/2021).

[32] Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. "Convergence Rates of Posterior Distributions". In: *Annals of Statistics* 28.2 (Apr. 2000), pp. 500–531. DOI: `10.1214/aos/1016218228`. URL: `https://doi.org/10.1214/aos/1016218228`.

[33] Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference.* Vol. 44. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2017, pp. xxiv+646. ISBN: 978-0-521-87826-5. DOI: `10.1017/9781139029834`. URL: `https://doi-org.tudelft.idm.oclc.org/10.1017/9781139029834`.

[34] Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models.* Cambridge Series in Statistical and Probabilistic Mathematics. New York, NY: Cambridge University Press, 2016. 690 pp. ISBN: 978-1-107-04316-9.

[35] Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. "Interpolating Between Types and Tokens by Estimating Power-Law Generators". In: Advances in Neural Information Processing Systems. 2005.

[36] Qiyang Han. "Oracle Posterior Contraction Rates under Hierarchical Priors". In: *Electronic Journal of Statistics* 15.1 (2021), pp. 1085–1153.

[37] Charles R. Harris et al. "Array Programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[38] Geoffrey Hinton et al. "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups". In: *IEEE Signal processing magazine* 29.6 (2012), pp. 82–97.

[39] Jian Huang, Junyi Chai, and Stella Cho. "Deep Learning in Finance and Banking: A Literature Review and Classification". In: *Frontiers of Business Research in China* 14.1 (June 8, 2020), p. 13. ISSN: 1673-7431. DOI: 10.1186/s11782-020-00082-6. URL: https://doi.org/10.1186/s11782-020-00082-6.

[40] Hemant Ishwaran and Lancelot F James. "Gibbs Sampling Methods for Stick-Breaking Priors". In: *Journal of the American Statistical Association* 96.453 (2001), pp. 161–173. DOI: 10.1198/016214501750332758.

[41] Lancelot James. "Large Sample Asymptotics for the Two-Parameter Poisson-Dirichlet Process". In: *Pushing the limits of contemporary Statistics: Contributions in Honor of Jayanta k. Ghosh* 3 (2008). DOI: htpps://doi.org/10.1214/074921708000000147.

[42] Samuel Karlin. "Central Limit Theorems for Certain Infinite Urn Schemes". In: *Journal of Mathematics and Mechanics* 17.4 (1967). DOI: 10.1512/iumj.1968.17.17020.

[43] Jin-Young Kim and Sung-Bae Cho. "Electric Energy Consumption Prediction by Deep Learning with State Explainable Autoencoder". In: *Energies* 12 (Feb. 2019), p. 739. DOI: 10.3390/en12040739.

[44] Bas J. K. Kleijn and Aad W. van der Vaart. "The Bernstein-Von-Mises Theorem under Misspecification". In: *ResearchGate* (2012). DOI: 10.1214/12-EJS675. URL: https://www.researchgate.net/publication/254212684_The_Bernstein-Von-Mises_theorem_under_misspecification (visited on 05/25/2022).

[45] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet Classification with Deep Convolutional Neural Networks". In: *Advances in neural information processing systems* 25 (2012), pp. 1097–1105.

[46] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles". In: *Advances in Neural Information Processing Systems.* Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

[47] Fan Liang et al. "Deep Learning-Based Power Usage Forecast Modeling and Evaluation". In: *Proceedings of the 9th International Conference of Information and Communication Technology [ICICT-2019] Nanning, Guangxi, China January 11-13, 2019* 154 (Jan. 1, 2019), pp. 102–108. ISSN: 1877-0509. DOI: 10.1016/j.procs.2019.06.016. URL: https://www.sciencedirect.com/science/article/pii/S1877050919307859.

[48] A. Lo. "A Remark on the Limiting Posterior Distribution of the Multiparameter Dirichlet Process". In: *Sankhyā Ser. A* 48.2 (1986), pp. 247–249. ISSN: 0581-572X.

[49] Albert Y. Lo. "Weak Convergence for Dirichlet Processes". In: *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 45.1 (1983), pp. 105–111. ISSN: 0581-572X. JSTOR: 25050418.

[50] Hrushikesh Mhaskar, Qianli Liao, and Tomaso Poggio. "When and Why Are Deep Networks Better than Shallow Ones?" In: *Proceedings of the AAAI Conference on Artificial Intelligence.* Vol. 31. 1. 2017.

[51] Michael Kohler and Sophie Langer. "On the Rate of Convergence of Fully Connected Deep Neural Network Regression Estimates". In: *The Annals of Statistics* 49.4 (Aug. 1, 2021), pp. 2231–2249. DOI: 10.1214/20-AOS2034. URL: https://doi.org/10.1214/20-AOS2034.

[52] Francois Monard, Richard Nickl, and Gabriel P. Paternain. "Statistical Guarantees for Bayesian Uncertainty Quantification in Nonlinear Inverse Problems with Gaussian Process Priors". In: *The Annals of Statistics* 49.6 (2021), pp. 3255–3298. DOI: 10.1214/21-AOS2082. URL: https://doi.org/10.1214/21-AOS2082.

[53] Ian Osband et al. "Deep Exploration via Bootstrapped DQN". In: *Advances in Neural Information Processing Systems.* Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper/2016/file/8d8818c8e140c64c743113f563cf750f-Paper.pdf.

[54] M. Perman, J. Pitman, and M. Yor. "Size-Biased Sampling of Poisson Point Processes and Excursions". In: *Probability Theory and Related Fields* 92.1 (1992), pp. 21–39. ISSN: 0178-8051. DOI: 10.1007/BF01205234. URL: http://dx.doi.org.prox.lib.ncsu.edu/10.1007/BF01205234.

[55] Oskar Perron. "Zur Theorie Der Matrices". In: *Mathematische Annalen* 64.2 (June 1, 1907), pp. 248–263. ISSN: 1432-1807. DOI: 10.1007/BF01449896. URL: https://doi.org/10.1007/BF01449896.

[56] J. Pitman. "Exchangeable and Partially Exchangeable Random Partitions". In: *Probability Theory and Related Fields* 102.2 (1995), pp. 145–158. ISSN: 0178-8051. DOI: 10.1007/BF01213386. URL: http://dx.doi.org.prox.lib.ncsu.edu/10.1007/BF01213386.

[57] J. Pitman. "Poisson-Kingman Partitions". In: *Statistics and Science: A Festschrift for Terry Speed.* Vol. 40. IMS Lecture Notes Monogr. Ser. Beachwood, OH: Inst. Math. Statist., 2003, pp. 1–34. DOI: 10.1214/lnms/1215091133. URL: http://dx.doi.org.prox.lib.ncsu.edu/10.1214/lnms/1215091133.

[58] J. Pitman. "Some Developments of the Blackwell-MacQueen Urn Scheme". In: *Statistics, Probability and Game Theory.* Vol. 30. IMS Lecture Notes Monogr. Ser. Hayward, CA: Inst. Math. Statist., 1996, pp. 245–267. DOI: 10.1214/lnms/1215453576. URL: http://dx.doi.org.prox.lib.ncsu.edu/10.1214/lnms/1215453576.

[59] J. Pitman and M. Yor. "The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator". In: *Annals of Probability* 25.2 (1997),

pp. 855–900. ISSN: 0091-1798. DOI: 10.1214/aop/1024404422. URL: http://dx.doi.org.prox.lib.ncsu.edu/10.1214/aop/1024404422.

[60] Jim Pitman. “Random Discrete Distributions Invariant Under Size-biased Permutation”. In: *Advances in Applied Probability* 28.2 (1996), pp. 525–539. DOI: 10.2307/1428070.

[61] Tomaso Poggio et al. “Why and When Can Deep-but Not Shallow-Networks Avoid the Curse of Dimensionality: A Review”. In: *International Journal of Automation and Computing* 14.5 (2017), pp. 503–519.

[62] D. Pollard. *Convergence of Stochastic Processes.* Springer Series in Statistics. New York: Springer-Verlag, 1984, pp. xiv+215. ISBN: 0-387-90990-7. DOI: 10.1007/978-1-4612-5254-2. URL: http://dx.doi.org.prox.lib.ncsu.edu/10.1007/978-1-4612-5254-2.

[63] Nicholas G. Polson and Veronika Ročková. “Posterior Concentration for Sparse Deep Learning”. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems.* NIPS’18. Red Hook, NY, USA: Curran Associates Inc., Dec. 3, 2018, pp. 938–949.

[64] Qing Rao and Jelena Frtunikj. “Deep Learning for Self-Driving Cars: Chances and Challenges”. In: *Proceedings of the 1st International Workshop on Software Engineering for AI in Autonomous Systems.* SEFAIS ’18. New York, NY, USA: Association for Computing Machinery, May 28, 2018, pp. 35–38. ISBN: 978-1-4503-5739-5. DOI: 10.1145/3194085.3194087. URL: https://doi.org/10.1145/3194085.3194087 (visited on 12/15/2021).

[65] Kolyan Ray. “Adaptive Bernstein–von Mises Theorems in Gaussian White Noise”. In: *The Annals of Statistics* 45.6 (2017), pp. 2511–2536. DOI: 10.1214/16-AOS1533. URL: https://doi.org/10.1214/16-AOS1533.

[66] Kolyan Ray and Aad van der Vaart. “Semiparametric Bayesian Causal Inference”. In: *The Annals of Statistics* 48.5 (2020), pp. 2999–3020. ISSN: 0090-5364. DOI: 10.1214/19-AOS1919. URL: https://doi-org.tudelft.idm.oclc.org/10.1214/19-AOS1919.

[67] Judith Rousseau and Botond Szabo. “Asymptotic Frequentist Coverage Properties of Bayesian Credible Sets for Sieve Priors”. In: *The Annals of Statistics* 48.4 (Aug. 2020), pp. 2155–2179. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/19-AOS1881. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-48/issue-4/Asymptotic-frequentist-coverage-properties-of-Bayesian-credible-sets-for-sieve/10.1214/19-AOS1881.full (visited on 06/21/2021).

[68] M. Rudelson. “Random Vectors in the Isotropic Position”. In: *Journal of Functional Analysis* 164.1 (May 10, 1999), pp. 60–72. ISSN: 0022-1236. DOI: 10.1006/jfan.1998.3384. URL: https://www.sciencedirect.com/science/article/pii/S0022123698933845.

[69] S. Ramos et al. “Detecting Unexpected Obstacles for Self-Driving Cars: Fusing Deep Learning and Geometric Modeling”. In: *2017 IEEE Intelligent Vehicles Symposium (IV).* 2017 IEEE Intelligent Vehicles Symposium (IV). June 11–14, 2017, pp. 1025–1032. DOI: 10.1109/IVS.2017.7995849.

[70] Johannes Schmidt-Hieber. "Nonparametric Regression Using Deep Neural Networks with ReLU Activation Function". In: *The Annals of Statistics* (Aug. 2017). DOI: 10.1214/19-AOS1875. URL: https://www.researchgate.net/publication/319235694_Nonparametric_regression_using_deep_neural_networks_with_ReLU_activation_function (visited on 09/22/2021).

[71] Larry Schumaker. *Spline Functions: Basic Theory*. 3rd ed. Cambridge Mathematical Library. Cambridge: Cambridge University Press, 2007. ISBN: 978-0-521-70512-7. DOI: 10.1017/CBO9780511618994. URL: https://www.cambridge.org/core/books/spline-functions-basic-theory/843475201223F90091FFBDDCBF210BFB.

[72] Helmut Strasser. *Mathematical Theory of Statistics: Statistical Experiments and Asymptotic Decision Theory*. De Gruyter, Apr. 20, 2011. ISBN: 978-3-11-085082-6. DOI: 10.1515/9783110850826. URL: https://www.degruyter.com/document/doi/10.1515/9783110850826/html (visited on 05/25/2022).

[73] Taiji Suzuki. "Adaptivity of Deep ReLU Network for Learning in Besov and Mixed Smooth Besov Spaces: Optimal Rate and Curse of Dimensionality". In: (Oct. 2018). URL: https://arxiv.org/abs/1810.08033.

[74] Botond Szabo, A. W. van der Vaart, and J. H. van Zanten. "Frequentist Coverage of Adaptive Nonparametric Bayesian Credible Sets". In: *The Annals of Statistics* 43.4 (2015), pp. 1391–1428. DOI: 10.1214/14-AOS1270. URL: https://doi.org/10.1214/14-AOS1270.

[75] Terrence Tao. *Topics in Random Matrix Theory*. URL: https://terrytao.files.wordpress.com/2011/02/matrix-book.pdf.

[76] Yee Whye Teh. "A Hierarchical Bayesian Language Model Based on Pitman-Yor Processes". In: pp. 985–992.

[77] A. W. van der Vaart. *Asymptotic Statistics*. Vol. 3. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998, pp. xvi+443. ISBN: 978-0-521-78450-4. DOI: 10.1017/CBO9780511802256. URL: https://doi.org/10.1017/CBO9780511802256.

[78] Aad van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, 1996.

[79] Aad van der Vaart and Jon A. Wellner. "Preservation Theorems for Glivenko-Cantelli and Uniform Glivenko-Cantelli Classes". In: *High Dimensional Probability, II (Seattle, WA, 1999)*. Vol. 47. Progr. Probab. Birkhäuser Boston, Boston, MA, 2000, pp. 115–133.

[80] J.H. van Zanten. *High and Infinite Dimensional Models Lecture Notes*. 2018.

[81] Yuexi Wang and Veronika Rockova. "Uncertainty Quantification for Sparse Deep Learning". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, Aug. 26–28, 2020, pp. 298–308. URL: https://proceedings.mlr.press/v108/wang20b.html.

[82] Eric W. Weisstein. *Wielandt's Theorem*. In: *MathWorld*. URL: https://mathworld.wolfram.com/WielandtsTheorem.html.

[83] Frank Wood et al. "A Stochastic Memoizer for Sequence Data". In: *Proceedings of the 26th Annual International Conference on Machine Learning.* ICML '09. New York, NY, USA: Association for Computing Machinery, 2009, pp. 1129–1136. ISBN: 978-1-60558-516-1. DOI: `10.1145/1553374.1553518`. URL: `https://doi.org/10.1145/1553374.1553518`.

[84] William Weimin Yoo and Subhashis Ghosal. "Supremum Norm Posterior Contraction and Credible Sets for Nonparametric Multivariate Regression". In: *The Annals of Statistics* 44.3 (2016), pp. 1069–1102. DOI: `10.1214/15-AOS1398`. URL: `https://doi.org/10.1214/15-AOS1398`.

Summary

In this thesis, we investigate the properties of Bayesian methods. In particular, we want to give frequentist guarantees for Bayesian methods. A Bayesian starts with specifying their apriori belief as a probability distribution, the *prior* distribution. The prior is their inherently subjective beliefs. After a Bayesian has specified their prior, they collect data and compute the *posterior* distribution. For a Bayesian, this posterior distribution encodes their new beliefs on the world. However, this prior was subjective. Thus the posterior is also subjective. So we can wonder, will this posterior distribution give a better representation of reality? Will it be more accurate? The posterior distribution quantifies a subjective belief of uncertainty. How reliable is this quantification of uncertainty?

These questions lie at the foundation of this thesis. They have been answered for certain classes of prior distributions. However, they have not been fully answered for all distributions in use. In this thesis, in the introduction, we explain the foundational statistical theory to study these questions. In particular, we show how to apply *Schwartz* theorem and the *Bernstein-von Mises* theorems to study posterior distributions. We then turn to novel research.

In our research, we investigated the *Pitman-Yor* prior and the behaviour of hyperpriors on its parameters. To do this, we formulated and proved the Bernstein-von Mises theorems in this context. Together with the tools introduced in the introduction, we can provide frequentist guarantees for these methods. In these works, we show that the Pitman-Yor process induces a bias in the posterior. This bias means that when you use the Pitman-Yor process for distribution estimation you must use a bias correction. We identify the explicit bias and derive the asymptotic distribution. However, the asymptotic distribution shows that the credible sets will be unreliable if the true distribution has a continuous component. This result shows that we cannot trust the uncertainty quantification in this case. It was already known that the posterior distribution is inconsistent in this case. Hence, the Pitman-Yor process should be avoided in this situation.

For the hyperprior, we studied the estimation of the type of the true distribution. This hyperprior comes up in various scenarios, for example, when trying to estimate

the number of distinct species in a group. It also appears in various applications, for example, in interpreting DNA evidence in case of a rare genotype. We again derive a Bernstein-von Mises theorem to investigate the posterior distribution. Moreover, we study frequentist estimators for the hyperpriors and give frequentist guarantees.

Moreover, we proposed a novel Bayesian methodology for uncertainty quantification in deep learning. Using a variant of Schwartz's theorem, we can directly provide frequentist guarantees for our method. Our proposed method has two advantages. First, it is much faster than competing methods for quantifying uncertainty. Moreover, it has theoretical guarantees, which competing methods often lack.

Samenvatting

In dit proefschrift bestuderen we the eigenschappen van Bayesiaanse methoden. In het bijzonder willen we frequentistische garanties geven voor Bayesiaanse methoden. Een Bayesiaan begint met het geven van hun a priori geloof in de vorm van een kansverdeling, de *prior* verdeling. De prior verdeling is hun, inherent subjectieve, geloof. Nadat de Bayesiaan hun prior verdeling heeft bepaald, verzamelen ze data en berekenen ze de *posteriori* verdeling. Voor de Bayesiaan geeft deze posteriori verdeling hun nieuwe geloof over de werkelijkheid. Hun prior was echter subjectief. Hierdoor is hun posteroor ook subjectief. Dus kunnen we ons afvragen hoe goed deze posteriori verdeling is. Geeft deze posteriori verdeling een betere representatie van de werkelijkheid? Verder geeft een posteriori verdeling een subjectieve kwantificatie van onzekerheid. Hoe betrouwbaar is deze kwantificatie van onzekerheid?

Deze vragen liggen ten grondslag van dit proefschrift. Ze zijn voor bepaalde klassen van priors beantwoord. Echter zijn ze niet voor alle klassen van priors volledig beantwoord. In deze thesis, in de inleiding, beschrijven we eerst de statistische theorie die nodig is om deze vragen te formuleren en beantwoorden. In het bijzonder zullen we zien hoe we de stelling van Schwartz en de Bernstein-von Mises stellingen kunnen gebruiken om de posteriori te bestuderen. Daarna gaan we kijken naar het gedane onderzoek.

In ons onderzoek hebben we de *Pitman-Yor* prior en de hyperpriors op de parameters bestudeerd. Om dit te kunnen doen moesten we Bernstein-von Mises stellingen formuleren en bewijzen. Samen met de technieken uit de inleiding van dit proefschrift kunnen we frequentistische garanties geven voor deze methoden. In het bijzonder tonen we aan dat het Pitman-Yor process een statistische vertekening heeft. We identificeren precies wat deze vertekening is en laten zien hoe je hiervoor kan corrigeren. We tonen echter ook aan dat als de ware verdeling een continue component heeft er meer mis is met de posteriori. De asymptotische verdeling van de posteriori is in dat geval niet de juiste verdeling. Hierdoor is de onzekerheid kwantificatie onbetrouwbaar. Het was echter al bekend dat de posteriori verdeling in dit geval ook *inconsistent* is, het vind de ware verdeling sowieso niet. Dus kun je beter het Pitman-Yor process as prior vermijden als je vermoed dat er een continue component is in de ware verdeling.

Voor de hyperprior hebben we het schatten van het type van de ware verdeling bestudeerd. Dit is een waarde die terug komt in verschillende statistische toepassingen. Een voorbeeld hiervan is het schatten van het aantal verschillen soorten in een groep. Dit is weer relevant bij de interpretatie van de sterkte van DNA bewijs in het geval van een zeldzaam genotype. We doen dit door middel van een Bernstein-von Mises stelling voor de hyperpriors. Verder bestuderen we puntschatters voor de hyperpriors en geven we frequentistische garanties.

Tenslotte hebben we een nieuwe Bayesiaanse methode ontwikkeld voor het kwantificeren van onzekerheid in *deep learning*. Door middel van een een variant van de stelling van Schwartz kunnen we direct frequentistische garanties geven. Onze methode heeft twee voordelen ten opzichte van de concurrentie. Ten eerste is onze methode veel sneller voor het kwantificeren van de onzekerheid. Ten tweede heeft onze methode theoretische garanties, welke vaak ontbreken bij de concurrenten.

Acknowledgements

The acknowledgements is the chapter that many PhDs dread the most to write. It is easy to forget to acknowledge someone when trying to thank everyone you interacted with over the span of 4+ years. This dissertation is not just a product of me working alone. It came to be due to my interactions with others.

First, I thank Botond Szabó and Aad van der Vaart. I learned a lot from them. I had my first course on Bayesian statistics by Aad and Botond, never stopped teaching me new things. Not only technical knowledge. They taught me soft skills like how to present my work. If I had questions, I could always rely on their help. We have discussed many topics, which have guided my research direction and presentation. In particular, Aads insight into the technical presentation of the results was helpful and inspiring.

I would also like to thank my coauthors: my supervisors and Jeanne Nguyen. Without their help, this work would be more limited in scope and insight.

I thank the reading committee for their time reading my manuscript.

I thank my colleagues, both from Delft and Leiden. We did not only meet for work but also for board games, drinks and celebrations. Without the interaction with others, it would have been four bland years.

I thank the members of my reading group. They have been with me on my whims of reading. Together we learned a lot of statistical theory, and I hope we can continue this in the future. There was a great atmosphere in the group. This atmosphere led to space for learning.

A special mention goes to my friend Sophie, who read an earlier version of this dissertation and helped me improve it. Her help was valuable for teaching me insights from someone outside of statistics. Her questions and confusion indicated that I went too quick in my explanation, skipping too many steps.

Moreover, I would like to thank all my friends and family. They supported me during this journey.

Finally, I would like to thank you, the reader, for reading this dissertation.

Curriculum Vitae

Stefan Esther Mariëlle Patrick Franssen was born on 30 July, 1992 in Maastricht, The Netherlands. After attending high school in Maastricht, he started econometrics at Tilburg University before switching to Mathematics at Utrecht University. He completed his Bachelor's degree in 2015 with a bachelor's thesis on topological K-Theory, supervised by Marius Crainic. He pursued a master's in Mathematical Sciences at Utrecht University. He wrote his master's thesis "Topics in Bayesian nonparametrics" under the supervision of Aad van der Vaart.

In 2018 he started a PhD in Leiden under the supervision of Aad van der Vaart, with Botond Szabó acting as the co-promotor. During his PhD, Stefan was a teaching assistant and lecturer for different statistics and calculus courses. He was the supervisor of Maxime Casera's master's thesis. Stefan is the founder and organiser of the reading group on mathematical statistics. Moreover, he presented his research in Bayesian nonparametrics at numerous international conferences.

Publications

Published

[27] S. E. M. P. Franssen and A. W. van der Vaart. "Bernstein-von Mises Theorem for the Pitman-Yor Process of Nonnegative Type". In: *Electronic Journal of Statistics* 16.2 (Jan. 2022), pp. 5779–5811. ISSN: 1935-7524, 1935-7524. DOI: 10.1214/22-EJS2077. URL: https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-16/issue-2/Bernstein-von-Mises-theorem-for-the-Pitman-Yor-process-of/10.1214/22-EJS2077.full (visited on 11/09/2022)

Submitted

[28] S. E. M. P. Franssen and A. W. van der Vaart. *Empirical and Full Bayes Estimation of the Type of a Pitman-Yor Process.* Aug. 30, 2022. DOI: 10.48550/arXiv.2208.14255. arXiv: 2208.14255 [math, stat]. URL: http://arxiv.org/abs/2208.14255 (visited on 11/09/2022)

[26] S. E. M. P. Franssen and Botond Szabo. "Frequentist Coverage Guarantees of Empirical Bayesian Uncertainty Quantification Using Deep Neural Network Regression"