mathematics

MDPI

Article

Asymptotic Efficiency of Point Estimators in Bayesian Predictive Inference

Emanuele Dolera [1,2]

[1] Department of Mathematics, University of Pavia, Via Adolfo Ferrata 5, 27100 Pavia, Italy; emanuele.dolera@unipv.it
[2] Collegio Carlo Alberto, Piazza V. Arbarello 8, 10134 Torino, Italy

Abstract: The point estimation problems that emerge in Bayesian predictive inference are concerned with random quantities which depend on both observable and non-observable variables. Intuition suggests splitting such problems into two phases, the former relying on estimation of the random parameter of the model, the latter concerning estimation of the original quantity from the distinguished element of the statistical model obtained by plug-in of the estimated parameter in the place of the random parameter. This paper discusses both phases within a decision theoretic framework. As a main result, a non-standard loss function on the space of parameters, given in terms of a Wasserstein distance, is proposed to carry out the first phase. Finally, the asymptotic efficiency of the entire procedure is discussed.

Keywords: asymptotic efficiency; bayesian predictive inference; compatibility equations; decision theory; de Finetti's representation theorem; exchangeability; Wasserstein distance

MSC: 62A01; 62C10; 62C12; 60F17

Citation: Dolera, E. Asymptotic Efficiency of Point Estimators in Bayesian Predictive Inference. *Mathematics* **2022**, *10*, 1136. https://doi.org/10.3390/math10071136

Academic Editor: Jiancang Zhuang

Received: 1 March 2022
Accepted: 29 March 2022
Published: 1 April 2022

1. Introduction

This paper carries on a project—conceived by Eugenio Regazzini some years ago, and partially developed in collaboration with Donato M. Cifarelli—which aims at proving why and how some classical, frequentist algorithms from the theory of point estimation can be justified, under some regularity assumptions, within the Bayesian framework. See [1–4]. This project was inspired, in turn, by the works and the thoughts of Bruno de Finetti about the foundation of statistical inference, substantially based on the following principles.

1. De Finetti's vision of statistics is grounded on the irrefutable fact that the Bayesian standpoint—intended as the use of basic tools of probability theory and, especially, of conditional distributions—becomes a necessity for those who intend statistical inference as the utilization of observed data to update their original beliefs about other quantities of interest, not yet observed. See [5,6].
2. Rigorous notions of point estimation and optimality of an estimator can be achieved only within a decision-theoretic framework (see, e.g., [7]), at least if we admit all estimators into competition and disregard distinguished restrictions such as unbiasedness or equivariance. In turn, decision theory proves to be genuinely Bayesian, thanks to a well-known result by Abraham Wald. See [8] [Chapter 4].
3. At least from a mathematical stance, the existence of the prior distribution can be drawn from various representation theorems which, by pertaining to the more basic act of modeling incoming information, stand before the problem of point estimation. The most luminous example is the celebrated de Finetti representation theorem for exchangeable observations. See [6,9] and, for a predictive approach [10,11].

Indeed, these principles do not force the assessment of a specific prior distribution, but just lead the statistician to take cognizance that some prior has, in any case, to exist.

This fact agrees with de Finetti's indication to keep the concepts of "Bayesian standpoint" and "Bayesian techniques" as distinguished. See also [12].

Despite their robust logical coherence, orthodox Bayesian solutions to inferential problems suffer two main drawbacks on the practical, operational side, which may limit their use. On the one hand, it is rarely the case that a prior distribution is fully specified due to a lack of prior information, this phenomenon even being amplified by the choice of complex statistical models (e.g., of nonparametric type). On the other hand, the numerical tractability of the Bayesian solutions often proves to be a serious hurdle, especially in the presence of large datasets. For example, it suffices to mention those algorithms from Bayesian nonparametrics that involve tools from combinatorics (like permutations or set/integer partitions) having exponential algorithmic complexity. See, e.g., [13]. Finally, the implicit nature of the notion of *Bayesian estimator*, although conceptually useful, makes it hard to employ in practical problems, especially in combination with non-quadratic loss functions, even if noteworthy progress has been achieved from the numerical side in the last decade. All these issues still pervade modern statistical literature while, historically, they have paved the way firstly to the "Fisherian revolution" and then to more recent techniques such as empirical Bayes and objective Bayes methods. The ultimate result has been a proliferation of many *ad hoc* algorithms, often of limited conceptual value, that provide focused and operational solutions to very specific problems.

Aware of this trend, Eugenio Regazzini conceived his project with the aims of: reframing the algorithms of modern statistics—especially those obtained by frequentist techniques—within the Bayesian theory as summarized in points 1–3 above, showing whether they can be re-interpreted as good approximations of Bayesian algorithms. The rationale is that orthodox Bayesian theory could be open to accept even non-Bayesian solutions (hence, suboptimal ones if seen "through the glass of the prior") as long as such solutions prove to be more operational than the Bayesian ones and, above all, *asymptotically almost efficient*, in the Bayesian sense. This concept means that, for a fixed prior, the Bayesian risk function evaluated at the non-Bayesian estimator is approximately equal to the overall minimum of such risk function (achieved when evaluated at the Bayesian estimator), the error of approximation going to zero as the sample size increases. Of course, these goals can be carried out after providing quantitative estimates for the risk function, as done, for example, in some decision-theoretic work on the empirical Bayes approach to inference. See, e.g., the seminal work [14]. Indeed, Regazzini's project has much in common with the empirical Bayes theory, although the former strictly remains on the "orthodox Bayesian main way" whilst the latter mixes Bayesian and frequentist techniques. As to more practical results, an archetype of Regazzini's line of reasoning can be found in a previous statement from [15] [Section 5] which proves that the maximum likelihood estimator (MLE)—obtained in the classical context of n i.i.d. observations, driven by a regular parametric model—has the same Bayesian efficiency (coinciding with the *mean square error*, in this case) as the Bayesian estimator up to $O(1/n)$-terms, provided that the prior is smooth enough. Another example can be found in [16] where the authors, while dealing with species sampling problems, rediscover the so-called Good–Turing estimator for the probability of finding a new species (which is obtained via empirical Bayes arguments) within the Bayesian nonparametric setting described in [17]. Other examples are contained in [2,4]. In any case, Regazzini's project is not only a matter of "rigorously justifying" a given algorithm, but rather of logically conceiving an estimation problem from the beginning to the end by quantifying coherent degrees of approximation in terms of the Bayesian risk or, more generally, in terms of *speed of shrinkage of the posterior distribution* with respect to distances on the space of probability measures, these goals being proved *uniformly with respect to an entire class of priors*. Hence, this plan of action is conceptually antipodal to that of (nowadays called) "Bayesian consistency", i.e., to justify a Bayesian algorithm from the point of view of classical statistics.

1.1. Main Contributions and General Strategy

In this paper, we pursue Regazzini's project by considering some *predictive problems* where the quantity $U_{n,m}$ to be estimated depends explicitly on new (hitherto unobserved) variables $X_{n+1}, \ldots, X_{n+m}$, possibly besides the original sample variables $X_1, \ldots, X_n$ and an unobservable parameter T. Thus, $U_{n,m} = u_{n,m}(X_{n+1}, \ldots, X_{n+m}; X_1, \ldots, X_n; T)$. For simplicity, we confine ourselves to the simplest case in which both $(X_1, \ldots, X_n)$ and $(X_{n+1}, \ldots, X_{n+m})$ are segments of a whole sequence $\{X_i\}_{i \geq 1}$ of exchangeable $\mathbb{X}$-valued random variables, while T is a random parameter that makes the X_i's conditionally i.i.d. with a common distribution depending on T, in accordance with de Finetti's representation theorem. From the statistical point of view, the exchangeability assumption just reveals a supposed homogeneity between the observable quantities while, from a mathematical point of view, it simply states that the joint distribution of any k-subset of the X_i's depends only on k and not on the specific k-subset, for any $k \in \mathbb{N}$. Thus, we are setting our estimation problem within an orthodox Bayesian framework where, independently of the fact that we are able or not to precisely assess the prior distribution, such a prior has to exist for mere mathematical reasons. This solid theoretical background provides all the elements to logically formulate the original predictive estimation problem as the following decision-theoretic question: find

$$\hat{U}_{n,m} = \text{Argmin}_Z \mathbb{E}[\mathcal{L}_{\mathbb{U}}(U_{n,m}, Z)] \,, \tag{1}$$

where: $\mathcal{L}_{\mathbb{U}}$ is a suitable loss function on the space $\mathbb{U}$ in which $U_{n,m}$ takes its values; Z runs over the space of all $\mathbb{U}$-valued, $\sigma(X_1, \ldots, X_n)$-measurable random variables; the expectation is taken with respect to the joint distribution of $(X_1, \ldots, X_{n+m})$ and T. It is remarkable that the same estimation problem would have been meaningless in classical (Fisherian) statistics, which can solely consider the estimation of (a function of) the parameter, and not of random quantities. Now, the solution displayed in (1) depends of course on the prior and it is the optimal one when seen, in terms of the Bayesian risk, "with the glass of that prior". However, the above-mentioned difficulties about the assessment of a specific prior can diminish the practical (but not the conceptual) value of this solution, in the sense that it could prove to be non-operational in the case of a lack of prior information. Sometimes, when the prior is known up to further unknown parameters, another estimation problem is needed.

Our research is then focused on formalizing a general strategy aimed at producing, under regularity conditions, alternative estimators $U^*_{n,m}$ which prove to be asymptotically nearly optimal (as specified above), uniformly with respect to any prior in some class. More precisely, for any fixed prior in that class, we aim at proving the validity of the asymptotic expansions (as $n \to +\infty$),

$$\mathbb{E}\big[\mathcal{L}_{\mathbb{U}}(U_{n,m}, \hat{U}_{n,m})\big] = \hat{R}_{0,m} + \frac{1}{n}\hat{R}_{1,m} + o\left(\frac{1}{n}\right) \tag{2}$$

$$\mathbb{E}\big[\mathcal{L}_{\mathbb{U}}(U_{n,m}, U^*_{n,m})\big] = R^*_{0,m} + \frac{1}{n}R^*_{1,m} + o\left(\frac{1}{n}\right), \tag{3}$$

along with $\hat{R}_{i,m} = R^*_{i,m}$ for $i = 0, 1$, where $\hat{U}_{n,m}$ is the same as in (1). This is exactly the content of Theorem 5.1 and Corollary 5.1 in [15], which deal with the case where: $U_{n,m} = T$ (estimation of the parameter of the model), so that $\mathbb{U}$ coincides with the parameter space $\Theta \subseteq \mathbb{R}$; $\mathcal{L}_{\mathbb{U}}$ is the quadratic loss function, so that the risk function coincides with the mean square error; $\hat{U}_{n,m} = \mathbb{E}[T \mid X_1, \ldots, X_n]$ is the Bayesian estimator with respect to $\mathcal{L}_{\mathbb{U}}$; $U^*_{n,m}$ coincides with the MLE; $\hat{R}_{0,m} = R^*_{0,m} = 0$ and $\hat{R}_{1,m} = R^*_{1,m} = \int_\Theta [\mathrm{I}(\theta)]^{-1} \pi(\mathrm{d}\theta)$, I denoting the Fisher information of the model and π being any prior on Θ with positive and sufficiently smooth density (with respect to the Lebesgue measure). Moving to truly predictive problems, the main operational solutions come from the empirical Bayes theory, which shares Equation (1) with the approach we are going to present. However, the empirical Bayes theory very soon leaves the "Bayesian main way" by bringing some sort of

Law of Large Numbers into the game, in order to replace the unknown quantities (usually, the prior itself). Here, on the contrary, we pursue Regazzini's project by proposing a new method that remains on the Bayesian main way. It consists of the following six steps.

Step 1. Reformulate problem (1) into another (orthodox Bayesian) estimation problem about T, the random parameter of the model. Roughly speaking, start from the following de Finetti representation:

$$\mathbb{P}[X_1 \in A_1, \ldots, X_k \in A_k \mid T = \theta] = \mu^{\otimes_k}(A_1 \times \ldots \times A_k \mid \theta) := \prod_{i=1}^{k} \mu(A_i \mid \theta)\,, \tag{4}$$

valid for all $k \in \mathbb{N}$, Borel sets $A_1, \ldots, A_k$, $\theta \in \Theta$, and some probability kernel $\mu(\cdot|\cdot)$, which coincides with the statistical model for the single observation. Then, consider the following estimation problem: find

$$\hat{T}_{n,m} = \mathrm{Argmin}_W \mathbb{E}\Big[\mathcal{L}_{\Theta,(X_1,\ldots,X_n)}(T, W)\Big]\,, \tag{5}$$

where: $\mathcal{L}_{\Theta,(X_1,\ldots,X_n)}$ is a suitable loss function on Θ; W runs over the space of all Θ-valued, $\sigma(X_1, \ldots, X_n)$-measurable random variables; the expectation is taken with respect to the joint distribution of $(X_1, \ldots, X_n)$ and T. The explicit definition of $\mathcal{L}_{\Theta,(X_1,\ldots,X_n)}$ is given in terms of a Wasserstein distance, as follows:

$$\mathcal{L}_{\Theta,(x_1,\ldots,x_n)}(\theta, \tau) = \inf_{\Gamma} \int_{\mathbb{U}^2} \mathcal{L}_{\mathbb{U}}(u, v)\Gamma(\mathrm{d}u\mathrm{d}v), \tag{6}$$

where Γ runs over the Fréchet class of all probability measures on $\mathbb{U}^2$ with marginals $\gamma_{\theta,(x_1,\ldots,x_n)}$ and $\gamma_{\tau,(x_1,\ldots,x_n)}$, respectively, and $\gamma_{\theta,(x_1,\ldots,x_n)}$ stands for the pull-back measure $\mu^{\otimes_m}(\cdot|\theta) \circ u_{n,m}(\cdot; x_1, \ldots, x_n; \theta)^{-1}$ on $\mathbb{U}$.

Step 2. After getting the estimator $\hat{T}_{n,m}$ from (5), consider estimators $U^*_{n,m}$ that satisfy the following approximated version of problem (1): find

$$U^*_{n,m} = \mathrm{Argmin}_Z \int_{\mathbb{X}^m} \mathcal{L}_{\mathbb{U}}\Big(u_{n,m}(y_1, \ldots, y_m; X_1, \ldots, X_n; \hat{T}_{n,m}), Z\Big)\mu^{\otimes_m}(\mathrm{d}y_1 \ldots \mathrm{d}y_m \mid \hat{T}_{n,m}), \tag{7}$$

where Z runs over the space of all $\mathbb{U}$-valued, $\sigma(X_1, \ldots, X_n)$-measurable random variables.

Step 3. For the estimators $\hat{U}_{n,m}$ and $U^*_{n,m}$ that solve (1) and (7) respectively, prove that (2) and (3) hold along with $\hat{R}_{i,m} = R^*_{i,m}$ for $i = 0, 1$. This entails the asymptotic almost efficiency of $U^*_{n,m}$, which it is still a prior-dependent estimator. In any case, this step is crucial to show that the loss function $\mathcal{L}_{\Theta,(x_1,\ldots,x_n)}$ given in (6) is "Bayesianly well-conceived", that is, in harmony with the original aim displayed in (1).

Step 4. Identities (2) and (3) provides conditions on the statistical model $\mu(\cdot|\cdot)$ that possibly allows the existence of some *prior-free* estimator $\tilde{T}_{n,m}$ of T which turns out to be asymptotically almost efficient, with respect to the same risk function as that displayed on the right-hand side of (5). More precisely, this fact consists of proving the validity of the following identities (as $n \to +\infty$)

$$\mathbb{E}\Big[\mathcal{L}_{\Theta,(X_1,\ldots,X_n)}(T, \hat{T}_{n,m})\Big] = \hat{\rho}_{0,m} + \frac{1}{n}\hat{\rho}_{1,m} + o\left(\frac{1}{n}\right) \tag{8}$$

$$\mathbb{E}\Big[\mathcal{L}_{\Theta,(X_1,\ldots,X_n)}(T, \tilde{T}_{n,m})\Big] = \tilde{\rho}_{0,m} + \frac{1}{n}\tilde{\rho}_{1,m} + o\left(\frac{1}{n}\right), \tag{9}$$

along with $\hat{\rho}_{i,m} = \tilde{\rho}_{i,m}$ for $i = 0, 1$, where $\hat{T}_{n,m}$ is the same as in (5), for all prior distributions in a given class.

Step 5. After getting estimators $\tilde{T}_{n,m}$ as in **Step 4**, consider the *prior-free* estimators $\tilde{U}_{n,m}$ satisfying the analogous minimization problem as in (7), with $\hat{T}_{n,m}$ replaced by $\tilde{T}_{n,m}$.

Step 6. For any estimator $\tilde{U}_{n,m}$ found as in **Step 5**, prove the validity of the following identity (as $n \to +\infty$):

$$\mathbb{E}\left[\mathcal{L}_{\mathbb{U}}(U_{n,m}, \tilde{U}_{n,m})\right] = \tilde{R}_{0,m} + \frac{1}{n}\tilde{R}_{1,m} + o\left(\frac{1}{n}\right), \tag{10}$$

along with $\hat{R}_{i,m} = \tilde{R}_{i,m}$ for $i = 0, 1$, where the $\hat{R}_{i,m}$'s are the same as in (2), for all prior distributions in the same class as specified in **Step 4**. This last step shows why and how the frequentist (i.e., prior-free) estimator $\tilde{U}_{n,m}$ can be used, within the orthodox Bayesian framework, as a good approximation of the Bayesian estimator $\hat{U}_{n,m}$. This is particularly remarkable at least in two cases that do not exclude each other: when the estimator $\tilde{T}_{n,m}$ obtained from **Step 4** is much simpler and numerically manageable than $\hat{T}_{n,m}$; when prior information is sufficient to characterize only a class of priors, but not a specific element of it.

This plan of action obeys the following principles:

(A) The loss function $\mathcal{L}_{\Theta,(x_1,\ldots,x_n)}$ on Θ is harmoniously coordinated with the original choice of the loss function $\mathcal{L}_{\mathbb{U}}$ on $\mathbb{U}$. This principle is much aligned with de Finetti's thought (see [18]), since it remarks on the more concrete nature of the space $\mathbb{U}$ compared with the space Θ which is, in principle, only a set of labels. Hence, it is much more reasonable to firstly metrize the space $\mathbb{U}$ and then the space Θ accordingly (as in (6)), rather than directly metrize Θ—even without taking account of the original predictive aim.

(B) The Bayesian risk function associated with both $U^*_{n,m}$ and $\tilde{U}_{n,m}$ can be bounded from above by the sum of two quantities: the former taking account of the error in estimating T, the latter reflecting the fact that we are estimating both $U^*_{n,m}$ and $\tilde{U}_{n,m}$ from an "estimated distribution".

The former principle, whose formalization constitutes the main novelty of this work, is concerned with the geometrical structure of the space of the parameters Θ. This is what we call a *relativistic principle* in point estimation theory: the goal of estimating a random quantity that depends on the observations (possibly besides the parameter) yields a modification of the geometry of Θ, to be now thought of as a curved space according to a non-trivial geometry. Of course, this modified geometry entails a coordinated notion of mean square error, now referred to the Riemannian geodesic distance. The term relativistic principle just hints at the original main principle of General Relativity Theory according to which the presence of a massive body modifies the geometry of the physical surrounding space, by means of the well-known Einstein tensor equations. These equations formalize a sort of compatibility between the physical and the geometric structures of the space. Thus, the identities (43) and (44), as stated in Section 3 to properly characterize the (Riemannian) metric on Θ, we will call *compatibility equations*. Actually, the idea of metrizing the parameter space Θ in a non standard way is well-known since the pioneering paper [19] by Radhakrishna Rao, and has received so much attention in the statistical literature to give birth to a fertile branch called *Information Geometry*. See, e.g., [20]. In particular, the concepts of efficiency, unbiasedness, Cramér–Rao lower bounds, Rao–Blackwell and Lehmann–Scheffé theorems are by far best-understood in this non-standard (i.e., non-Euclidean) setting. See [21]. In any case, to the best of our knowledge, this is the first work which connects the use of a non-standard geometric setting on Θ with predictive estimation problems—even if some hints can be drawn from [22]. In our opinion, the lack of awareness about the aforesaid relativistic principle, combined with an abuse of the quadratic loss function on Θ, has produced a lot of actually sub-efficient algorithms, most of which focused on the estimation of certain probabilities, or of nonparametric objects. In these cases, the efficiency of the ensuing estimators is created artificially through a misuse of the quadratic loss, and it proves to be drastically downsized whenever these estimators are evaluated by means of other, more concrete loss functions which take account (as in (6)) of the natural geometry of the spaces of really observable quantities. To get an idea of this phenomenon, see the discussion about Robbins' estimators in Section 4.4 below.

1.2. Organization of the Paper

We conclude the introduction by summarizing the main results of the paper, which are threefold. The first block of results, including Theorem 1, Proposition 1 and Lemma 1 in Section 2.2, concerns some refinement of de Finetti's Law of Large Numbers for the log-likelihood process. The second block of theoretical results, developed in Section 3, contains:

(i) Proposition 2, which shows how to bound from above the Bayesian risk of any estimator of $U_{n,m}$ by using the Wasserstein distance;
(ii) Proposition 3, which explains how to use the Laplace method of the approximation of integrals to get asymptotic expansions of the Bayesian risk functions;
(iii) the formulation of the compatibility Equations (43) and (44);
(iv) the proof of the "asymptotic almost efficiency" of the estimator $U^*_{n,m}$ obtained in **Step 2**, via verification of identities (2) and (3);
(v) the successful completion of **Step 6**, that is, the proof of the "asymptotic almost efficiency" of estimators $\tilde{U}_{n,m}$ obtained in **Step 5**, via verification of identity (10).

The last block of results, contained in Section 4, consists of explicit verifications of the compatibility equations for some simple statistical models (Sections 4.1–4.3), and also the adaptation of our plan of action to the same Poisson-mixture model used by Herbert Robbins in [23] to illustrate his empirical Bayes approach to predictive inference (Section 4.4). Finally, all the proofs of the theoretical results are deferred to Section 5, while some conclusions and future developments are hinted at in Section 6.

2. Technical Preliminaries

We begin by rigorously fixing the mathematical setting, split into two subsections. The former will contain a very general framework which will serve to give a precise meaning to the questions presented in the Introduction and to state in full generality one of the main results, that is, Proposition 2 in Section 3. In fact, this statement will include some inequalities that, by carrying out the goal described in point (B) of the Introduction will constitute the starting point for all the results presented in Section 3. The second subsection will deal with a simplification of the original setting—essentially based on additional regularity conditions for the spaces $\mathbb{U}$ and Θ and for the statistical model $\mu(\cdot|\cdot)$—aimed at introducing the novel compatibility equations without too many technicalities.

2.1. The General Framework

Let $(\mathbb{X}, \mathscr{X})$ and $(\Theta, \mathscr{T})$ be standard Borel spaces called *sample space* (for any single observation) and *parameter space*, respectively. Consider a sequence $\{X_i\}_{i\geq 1}$ of $\mathbb{X}$-valued random variables (r.v.'s, from now on) along with another Θ-valued r.v. T, all the X_i's and T being defined on a suitable probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Assume that (4) holds for all $k \in \mathbb{N}$, $A_1, \ldots, A_k \in \mathscr{X}$ and $\theta \in \Theta$ with some given probability kernel $\mu(\cdot|\cdot) : \mathscr{X} \times \Theta \to [0,1]$, called *statistical model* (for any single observation). The validity of (4) entails that the X_i's are *exchangeable* and that

$$\mathbb{P}[X_1 \in A_1, \ldots, X_k \in A_k] = \int_\Theta \mu^{\otimes k}(A_1 \times \ldots \times A_k \mid \theta)\pi(\mathrm{d}\theta) =: \alpha_k(A_1 \times \ldots \times A_k) \quad (11)$$

holds for all $k \in \mathbb{N}$ and $A_1, \ldots, A_k \in \mathscr{X}$ with some given probability measure (p.m.) π on $(\Theta, \mathscr{T})$ called *prior distribution*. Identity (11) uniquely characterizes the p.m. α_k on $(\mathbb{X}^k, \mathscr{X}^k)$ for any $k \in \mathbb{N}$, this p.m. being called *law of k-observations*, where $\mathbb{X}^k$ ($\mathscr{X}^k$, respectively) denotes the k-fold cartesian product (σ-algebra product, respectively) of k copies of $\mathbb{X}$ ($\mathscr{X}$, respectively). Moreover, let

$$\pi_k(B \mid x_1, \ldots, x_k) := \mathbb{P}[T \in B \mid X_1 = x_1, \ldots, X_k = x_k]$$
$$\beta_k(A \mid x_1, \ldots, x_k) := \mathbb{P}[X_{k+1} \in A \mid X_1 = x_1, \ldots, X_k = x_k]$$

be two probability kernels, with $\pi_k(\cdot|\cdot) : \mathscr{T} \times \mathbb{X}^k \to [0,1]$ and $\beta_k(\cdot|\cdot) : \mathscr{X} \times \mathbb{X}^k \to [0,1]$, defined as respective solutions of the following disintegration problems

$$\mathbb{P}[X_1 \in A_1, \ldots, X_k \in A_k, T \in B] = \int_{A_1 \times \ldots \times A_k} \pi_k(B \mid x_1, \ldots, x_k) \alpha_k(\mathrm{d}x_1 \ldots \mathrm{d}x_k)$$
$$\mathbb{P}[X_1 \in A_1, \ldots, X_k \in A_k, X_{k+1} \in A] = \int_{A_1 \times \ldots \times A_k} \beta_k(A \mid x_1, \ldots, x_k) \alpha_k(\mathrm{d}x_1 \ldots \mathrm{d}x_k)$$

for any $k \in \mathbb{N}$, $A_1, \ldots, A_k, A \in \mathscr{X}$ and $B \in \mathscr{T}$. The probability kernels $\pi_k(\cdot|\cdot)$ and $\beta_k(\cdot|\cdot)$ are called *posterior distribution* and *predictive distribution*, respectively.

Let $(\mathbb{U}, \mathrm{d}_{\mathbb{U}})$ be a Polish metric space and, for fixed $n, m \in \mathbb{N}$, let $u_{n,m} : \mathbb{X}^m \times \mathbb{X}^n \times \Theta \to \mathbb{U}$ be a measurable map. Let $U_{n,m} := u_{n,m}(X_{n+1}, \ldots, X_{n+m}; X_1, \ldots, X_n; T)$ be the random quantity to be estimated with respect to the loss function $\mathcal{L}_{\mathbb{U}}(u,v) := \mathrm{d}^2_{\mathbb{U}}(u,v)$. Now, recall the notion of *barycenter* (also known as Fréchet mean) of a given p.m.. Let $(\mathbb{S}, \mathrm{d}_{\mathbb{S}})$ be a Polish metric space, endowed with its Borel σ-algebra $\mathscr{B}(\mathbb{S})$. Given a p.m. μ on $(\mathbb{S}, \mathscr{B}(\mathbb{S}))$, define

$$\mathrm{Bary}_{\mathbb{S}}[\mu; \mathrm{d}_{\mathbb{S}}] := \mathrm{Argmin}_{y \in \mathbb{S}} \int_{\mathbb{S}} \mathrm{d}^2_{\mathbb{S}}(x,y) \mu(\mathrm{d}x)$$

provided that μ has finite second moment ($\mu \in \mathcal{P}_2(\mathbb{S}, \mathrm{d}_{\mathbb{S}})$, in symbols) and that at least one minimum point exists. See [24–26] for results on existence, uniqueness and some characterizations of barycenters. Then, put

$$\rho_{n,m}(C \mid x_1, \ldots, x_n) := \mathbb{P}[U_{n,m} \in C \mid X_1 = x_1, \ldots, X_n = x_n],$$

meaning that $\rho_{n,m}(\cdot|\cdot) : \mathscr{B}(\mathbb{U}) \times \mathbb{X}^n \to [0,1]$ is a probability kernel that solves the disintegration problem

$$\mathbb{P}[X_1 \in A_1, \ldots, X_n \in A_n, U_{n,m} \in C] = \int_{A_1 \times \ldots \times A_n} \rho_{n,m}(C \mid x_1, \ldots, x_n) \alpha_k(\mathrm{d}x_1 \ldots \mathrm{d}x_k)$$

for any $A_1, \ldots, A_n \in \mathscr{X}$ and $C \in \mathscr{B}(\mathbb{U})$. If $\mathbb{E}[\mathrm{d}^2_{\mathbb{U}}(U_{n,m}, u_0)] < +\infty$ for some $u_0 \in \mathbb{U}$ and $\mathrm{Bary}_{\mathbb{U}}[\rho_{n,m}(\cdot \mid x_1, \ldots, x_n); \mathrm{d}_{\mathbb{U}}]$ exists uniquely for α_n-almost all $(x_1, \ldots, x_n)$, then

$$\hat{U}_{n,m} = \mathrm{Bary}_{\mathbb{U}}[\rho_{n,m}(\cdot \mid X_1, \ldots, X_n); \mathrm{d}_{\mathbb{U}}] \tag{12}$$

solves the minimization problem (1). To give an analogous formalization to the minimization problem (7), define

$$\gamma_{\theta,(x_1,\ldots,x_n)}(C) := \mu^{\otimes m}\Big(\{(y_1, \ldots, y_m) \in \mathbb{X}^m \mid u_{n,m}(y_1, \ldots, y_m; x_1, \ldots, x_n; \theta) \in C\} \,\Big|\, \theta\Big)$$

for any $\theta \in \Theta$, $(x_1, \ldots, x_n) \in \mathbb{X}^n$ and $C \in \mathscr{B}(\mathbb{U})$. Again, if $\gamma_{\theta,(x_1,\ldots,x_n)} \in \mathcal{P}_2(\mathbb{U}, \mathrm{d}_{\mathbb{U}})$ and $\mathrm{Bary}_{\mathbb{U}}[\gamma_{\theta,(x_1,\ldots,x_n)}(\cdot); \mathrm{d}_{\mathbb{U}}]$ exists uniquely for any $\theta \in \Theta$ and α_n-almost all $(x_1, \ldots, x_n)$, then

$$U^*_{n,m} = \mathrm{Bary}_{\mathbb{U}}[\gamma_{\hat{T}_{n,m},(X_1,\ldots,X_n)}; \mathrm{d}_{\mathbb{U}}] \tag{13}$$

solves the minimization problem (7). By the way, notice that a combination of de Finetti's representation theorem with basic properties of conditional distributions entails that

$$\rho_{n,m}(C \mid x_1, \ldots, x_n) = \int_{\Theta} \gamma_{\theta,(x_1,\ldots,x_n)}(C) \pi_n(\mathrm{d}\theta \mid x_1, \ldots, x_n) \tag{14}$$

for α_n-almost all $(x_1, \ldots, x_n)$. It remains to formalize the minimization problem (5). If $\gamma_{\theta,(x_1,\ldots,x_n)}, \gamma_{\tau,(x_1,\ldots,x_n)} \in \mathcal{P}_2(\mathbb{U}, \mathrm{d}_{\mathbb{U}})$, then the loss function in (6) satisfies

$$\mathcal{L}_{\Theta,(x_1,\ldots,x_n)}(\theta,\tau) = \mathcal{W}^2_{\mathbb{U}}\Big(\gamma_{\theta,(x_1,\ldots,x_n)}; \gamma_{\tau,(x_1,\ldots,x_n)}\Big),$$

where $\mathcal{W}_{\mathbb{U}}$ denotes the 2-Wasserstein distance on $\mathcal{P}_2(\mathbb{U}, \mathrm{d}_{\mathbb{U}})$. See [27] [Chapters 6–7] for more information on the Wasserstein distance. Therefore, if $\pi_n(\cdot \mid x_1, \dots, x_n) \in \mathcal{P}_2(\Theta, \mathcal{L}^{1/2}_{\Theta,(x_1,\dots,x_n)})$ and $\mathrm{Bary}_\Theta[\pi_n(\cdot \mid x_1, \dots, x_n); \mathcal{L}^{1/2}_{\Theta,(x_1,\dots,x_n)}]$ exists uniquely for α_n-almost all $(x_1, \dots, x_n)$, then

$$\hat{T}_{n,m} = \mathrm{Bary}_\Theta\Big[\pi_n(\cdot \mid X_1, \dots, X_n); \mathcal{L}^{1/2}_{\Theta,(X_1,\dots,X_n)}\Big] \tag{15}$$

solves the minimization problem (5).

To conclude, it remains to formalize the definition of various Bayesian risk functions, that will appear in the formulation of the main results. For any estimator $U^\dagger_{n,m} = u^\dagger_{n,m}(X_1, \dots, X_n)$ of $U_{n,m}$, obtained with a measurable $u^\dagger_{n,m} : \mathbb{X}^n \to \mathbb{U}$, put

$$\begin{aligned}\mathfrak{R}_{\mathbb{U}}[U^\dagger_{n,m}] &:= \mathbb{E}\Big[\mathcal{L}_{\mathbb{U}}(U_{n,m}, U^\dagger_{n,m})\Big] \\ &= \int_\Theta \int_{\mathbb{X}^{n+m}} \mathcal{L}_{\mathbb{U}}\Big(u_{n,m}(\mathbf{y};\mathbf{x};\theta), u^\dagger_{n,m}(\mathbf{x})\Big)\mu^{\otimes_{n+m}}(\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} \mid \theta)\pi(\mathrm{d}\theta) \\ &= \int_{\mathbb{X}^n} \int_\Theta \int_{\mathbb{X}^m} \mathcal{L}_{\mathbb{U}}\Big(u_{n,m}(\mathbf{y};\mathbf{x};\theta), u^\dagger_{n,m}(\mathbf{x})\Big)\mu^{\otimes_m}(\mathrm{d}\mathbf{y} \mid \theta)\pi_n(\mathrm{d}\theta \mid \mathbf{x})\alpha_n(\mathrm{d}\mathbf{x})\end{aligned} \tag{16}$$

provided that the integrals are finite. Here and throughout, the bold symbols $\mathbf{x}, \mathbf{y}$ are just short-hands to denote the vectors $(x_1, \dots, x_n)$ and $(y_1, \dots, y_m)$, respectively. Analogously, for any estimator $T^\dagger_{n,m} = t^\dagger_{n,m}(X_1, \dots, X_n)$ of T, obtained with a measurable $t^\dagger_{n,m} : \mathbb{X}^n \to \Theta$, put

$$\begin{aligned}\mathfrak{R}_\Theta[T^\dagger_{n,m}] &:= \mathbb{E}\Big[\mathcal{L}_{\Theta,(X_1,\dots,X_n)}(T, T^\dagger_{n,m})\Big] \\ &= \int_\Theta \int_{\mathbb{X}^n} \mathcal{L}_{\Theta,\mathbf{x}}\Big(\theta, t^\dagger_{n,m}(\mathbf{x})\Big)\mu^{\otimes_n}(\mathrm{d}\mathbf{x} \mid \theta)\pi(\mathrm{d}\theta) \\ &= \int_{\mathbb{X}^n} \int_\Theta \mathcal{L}_{\Theta,\mathbf{x}}\Big(\theta, t^\dagger_{n,m}(\mathbf{x})\Big)\pi_n(\mathrm{d}\theta \mid \mathbf{x})\alpha_n(\mathrm{d}\mathbf{x})\end{aligned} \tag{17}$$

provided that the integrals are finite.

2.2. The Simplified Framework

Start by assuming that $\mathbb{U} = \mathbb{R}$ and $\mathcal{L}_{\mathbb{U}}(u, v) = |u - v|^2$. Then, restrict the attention to those predictive problems in which the quantity to be estimated depends only on the new observations $X_{n+1}, \dots, X_{n+m}$ and on the random parameter T, but not on the observable variables $X_1, \dots, X_n$. This restriction is actually non-conceptual, and it is made only to diminish the mathematical complexity of the ensuing asymptotic expansions (valid as $n \to +\infty$), having this way fewer sources of dependence from the variable n. Thus, the quantity to be estimated has the form $u_m(X_{n+1}, \dots, X_{n+m}; T)$ for some measurable $u_m : \mathbb{X}^m \times \Theta \to \mathbb{R}$. From now on, it will be assumed that

$$\mathbb{E}\Big[\big(u_m(X_{n+1}, \dots, X_{n+m}; T)\big)^2\Big] < +\infty\,. \tag{18}$$

Whence, for the Bayesian estimator $\hat{U}_{n,m}$ in (12) existence and uniqueness are well-known: its explicit form is given by $\hat{U}_{n,m} = \hat{u}_{n,m}(X_1, \dots, X_n)$ with

$$\begin{aligned}\hat{u}_{n,m}(x_1, \dots, x_n) &= \mathbb{E}[u_m(X_{n+1}, \dots, X_{n+m}; T) \mid X_1 = x_1, \dots, X_n = x_n] \\ &= \int_\Theta \int_{\mathbb{X}^m} u_m(y_1, \dots, y_m; \theta)\mu^{\otimes_m}(\mathrm{d}y_1 \dots \mathrm{d}y_m \mid \theta)\pi_n(\mathrm{d}\theta \mid x_1, \dots, x_n),\end{aligned}$$

which is finite for α_n-almost all $(x_1, \ldots, x_n)$. The risk function $\mathfrak{R}_{\mathbb{U}}$ evaluated at $\hat{U}_{n,m}$ achieves its overall minimum value and, from (16), it takes the form:

$$\mathfrak{R}_{\mathbb{U}}[\hat{U}_{n,m}] = \int_{\mathbb{X}^n} \left\{ \int_{\Theta} \mathsf{v}(\theta)\pi_n(\mathrm{d}\theta \mid \mathbf{x}) + \int_{\Theta} [\mathsf{m}(\theta)]^2 \pi_n(\mathrm{d}\theta \mid \mathbf{x}) - \left(\int_{\Theta} \mathsf{m}(\theta)\pi_n(\mathrm{d}\theta \mid \mathbf{x}) \right)^2 \right\} \alpha_n(\mathrm{d}\mathbf{x}), \tag{19}$$

with

$$\mathsf{m}(\theta) := \int_{\mathbb{X}^m} u_m(y_1, \ldots, y_m; \theta)\mu^{\otimes_m}(\mathrm{d}y_1 \ldots \mathrm{d}y_m \mid \theta)$$
$$\mathsf{v}(\theta) := \int_{\mathbb{X}^m} [u_m(y_1, \ldots, y_m; \theta) - \mathsf{m}(\theta)]^2 \mu^{\otimes_m}(\mathrm{d}y_1 \ldots \mathrm{d}y_m \mid \theta)$$

thanks to the well-known "Law of Total Variance". See, e.g., [28] [Problem 34.10(b)]. As to the issue of estimating T, the first remarkable simplification induced by the above assumptions is that the p.m. $\gamma_{\theta,(x_1,\ldots,x_n)}$ is independent of $(x_1, \ldots, x_n)$. Whence,

$$\Delta(\theta, \tau) := [\mathcal{L}_{\Theta,(x_1,\ldots,x_n)}(\theta, \tau)]^{1/2} = \mathcal{W}_{\mathbb{U}}\left(\gamma_{\theta,(x_1,\ldots,x_n)}; \gamma_{\tau,(x_1,\ldots,x_n)}\right), \tag{20}$$

is, in turn, independent of $(x_1, \ldots, x_n)$ and *defines a distance* on Θ provided that

$$\gamma_{\theta,(x_1,\ldots,x_n)} = \gamma_{\tau,(x_1,\ldots,x_n)}$$

entails $\theta = \tau$. Thus, for any estimator $T^{\dagger}_{n,m} = t^{\dagger}_{n,m}(X_1, \ldots, X_n)$ of T, obtained with a measurable $t^{\dagger}_{n,m} : \mathbb{X}^n \to \Theta$, (17) becomes

$$\mathfrak{R}_{\Theta}[T^{\dagger}_{n,m}] = \int_{\mathbb{X}^n} \int_{\Theta} [\Delta(\theta, t^{\dagger}_{n,m}(\mathbf{x}))]^2 \pi_n(\mathrm{d}\theta \mid \mathbf{x})\alpha_n(\mathrm{d}\mathbf{x}) \,. \tag{21}$$

The last simplifications concern the basic object of the inference, i.e., the statistical model $\mu(\cdot|\cdot)$ and the prior π. First, assume that $\Theta = (a, b) \subseteq \mathbb{R}$ and that π has a density p (with respect to the Lebesgue measure). Even if this one-dimensionality assumption can seem a drastic simplification, it is again of a non-conceptual nature, and it is made to diminish the mathematical complexity of the ensuing statements. In fact, one of the goals of this work is to provide a Riamannian-like characterization of the metric space (Θ, Δ), and this is particularly simple in such a one-dimensional setting. The following arguments should be quite easily reproduced at least in a finite-dimensional setting (i.e., when $\Theta \subseteq \mathbb{R}^d$) by using basic tools of Riemannian geometry, such as local expansions of the geodesic distance. See, e.g., [29] [Chapter 5]. As to the statistical model $\mu(\cdot|\cdot)$, consider the following:

Assumption 1. *$\mu(\cdot|\cdot)$ is dominated by some σ-finite measure χ on $(\mathbb{X}, \mathscr{X})$ with a (distinguished version of the) density $f(\cdot|\theta)$ that satisfies:*

(i) *$f(x|\,\theta) > 0$ for all $x \in \mathbb{X}$ and $\theta \in \Theta$;*
(ii) *for any fixed $x \in \mathbb{X}$, $\theta \mapsto f(x|\,\theta)$ belongs to $\mathrm{C}^4(\Theta)$;*
(iii) *there exists a separable Hilbert space $\mathcal{H}$ for which $\log f(x|\,\cdot) \in \mathcal{H}$ for all $x \in \mathbb{X}$, and such that, for any open Θ' whose closure is compact in Θ ($\Theta' \Subset \Theta$, in symbols), the restriction operators $\mathcal{R}_{\Theta'} : h \mapsto h_{|\Theta'}$ are continuous from $\mathcal{H}$ to $\mathrm{C}^0(\overline{\Theta'})$;*
(iv) *$\int_{\mathbb{X}} |\log f(x|\,\theta)|^2 \mu(\mathrm{d}x \mid \theta) < +\infty$ for π-a.e. θ, and the Kullback-Leibler divergence*

$$\mathsf{K}(t \parallel \theta) := \int_{\mathbb{X}} \left(\frac{\log f(x|\,t)}{\log f(x|\,\theta)} \right) \mu(\mathrm{d}x \mid t) \tag{22}$$

is well-defined.

A canonical choice for the Hilbert space $\mathcal{H}$ is in the form of a *weighted Sobolev space* $H^r(\Theta;\pi)$ for some $r \geq 1$. See, e.g., [30,31] for definition and further properties of weighted Sobolev spaces, such as embedding theorems. By the way, it is worth remarking that such assumptions are made to easily state the following results. It is plausible they could be relaxed in future works.

In this regularity setting, introduce the sequence $\{H_n\}_{n\geq 1}$, where $H_n : \Omega \to \mathcal{H}$ represents the (normalized) *log-likelihood process*, that is

$$H_n := \frac{1}{n}\ell_n(\cdot; X_1, \ldots, X_n) := \frac{1}{n}\sum_{i=1}^{n} \log f(X_i|\,\cdot) = \int_{\mathbb{X}} \log f(\xi|\,\cdot)\mathfrak{e}_n^{(X_1,\ldots,X_n)}(\mathrm{d}\xi) \tag{23}$$

the symbol $\mathfrak{e}_n^{(X_1,\ldots,X_n)}$ standing for the *empirical measure* based on $(X_1, \ldots, X_n)$, i.e.,

$$\mathfrak{e}_n^{(X_1,\ldots,X_n)} := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}\,.$$

For completeness, any notation like $\ell_n(\cdot; X_1, \ldots, X_n)$ is just a short-hand to denote the entire function $\theta \mapsto \ell_n(\theta; X_1, \ldots, X_n)$. First of all, observe that H_n is a *sufficient statistics* in both classical and Bayesian sense. See [11]. Then, a version of de Finetti's Law of Large Numbers (see [9,32]) for the log-likelihood process can be stated as follows:

Theorem 1. *Under Assumption 1, define the following $\mathcal{H}$-valued r.v.*

$$H := \int_{\mathbb{X}} \log f(z|\,\cdot)\mu(\mathrm{d}z \mid T) = -\mathsf{K}(T \parallel \cdot) + \int_{\mathbb{X}} \log f(z|\,T)\mu(\mathrm{d}z \mid T)$$

along with $\nu_n(D) := \mathbb{P}[H_n \in D]$ and $\nu(D) := \mathbb{P}[H \in D]$, for any $D \in \mathscr{B}(\mathcal{H})$. Then, it holds that

$$H_n \xrightarrow{L^2} H \tag{24}$$

which, in turn, yields that $\nu_n \Rightarrow \nu$, where $\Rightarrow$ denotes weak convergence of p.m.'s on $(\mathcal{H}, \mathscr{B}(\mathcal{H}))$.

Then, to carry out the objectives mentioned in the Introduction, a quantitative refinement of the thesis $\nu_n \Rightarrow \nu$ is needed, as stated in the following proposition.

Proposition 1. *Let $C_b^2(\mathcal{H})$ denote the space of bounded, C^2 functionals on $\mathcal{H}$. Besides Assumption 1, suppose there exists a function $\Gamma(\cdot;\mu,\pi) : \mathcal{H} \to \mathbb{R}$ such that*

$$\frac{1}{2}\mathbb{E}\big[\mathrm{Hess}[\Psi]_H \otimes \mathrm{Cov}_T[\log f(X_i|\,\cdot)]\big] = \mathbb{E}[\Psi(H)\Gamma(H;\mu,\pi)] \tag{25}$$

holds for all functional $\Psi \in \mathrm{C}_b^2(\mathcal{H})$, where $\mathrm{Hess}[\Psi]_h$ denotes the Hessian of Ψ at $h \in \mathcal{H}$, $\otimes$ is the tensor product between quadratic forms (operators) and $\mathrm{Cov}_t[\log f(X_i|\,\cdot)]$ stands for the covariance operator of the $\mathcal{H}$-valued r.v.'s $\log f(X_i|\,\cdot)$ with respect to the p.m. $\mu(\cdot \mid t)$. Then,

$$\int_{\mathcal{H}} \Psi(h)\nu_n(\mathrm{d}h) = \int_{\mathcal{H}} \Psi(h)\nu(\mathrm{d}h) + \frac{1}{n}\int_{\mathcal{H}} \Psi(h)\Gamma(h;\mu,\pi)\nu(\mathrm{d}h) + o\left(\frac{1}{n}\right) \tag{26}$$

holds as $n \to +\infty$ for all continuous $\Psi : \mathcal{H} \to \mathbb{R}$ for which the above integrals are convergent.

For further information on second-order differentiability in Hilbert/Banach spaces, see [33,34]. By the way, the above identity (26) is a quantitative strengthening of de Finetti's theorem similar to the identities stated in Theorem 1.1 of [8] [Chapter 6], valid in a finite-dimensional setting. Later on, we will resort to *uniform* versions of (26), meaning that the $o(\frac{1}{n})$-term is uniformly bounded with respect to h. However, such a kind of results—much more in the spirit of the Central Limit Theorem—are very difficult to prove and, to the best of the author's knowledge, there are no known results in infinite-dimension. Examples in

finite-dimensional settings are given in [35,36], which prove Berry–Esseen like inequalities in the very specific context of Bernoulli r.v.s. See also [37]. Anyway, since one merit of [35] is to show how to use the classical Central Limit Theorem to prove an expansion as in (26), one could hope to follow that very same line of reasoning by resorting to some version of the central limit theorem for Banach spaces, such as that stated in [38]. Research on this is ongoing.

Now, to make the above Proposition 1 a bit more concrete, it is worth noticing the case in which $f(\cdot|\theta)$ is in exponential form. In fact, in this case, the identity (26) can be rewritten in a simpler form, condensed in the following statement.

Lemma 1. *Besides Assumption 1, suppose that* $f(x \mid \theta) = \exp\{\theta S(x) - M(\theta)\}$, *with some measurable* $S : \mathbb{X} \to \mathbb{R}$ *and* $M(\theta) := \log\left(\int_{\mathbb{X}} e^{\theta S(x)} \chi(\mathrm{d}x)\right) \in \mathbb{R}$ *for all* $\theta \in \Theta$. *Then,* (26) *holds with*

$$\nu(D) := \mathbb{P}\left[\left(\theta \mapsto \theta M'(T) - M(\theta)\right) \in D\right]$$

and

$$\Gamma\left((\theta \mapsto \theta M'(t) - M(\theta)); \mu, \pi\right) = \frac{M''(t)}{p(t)} \frac{\mathrm{d}^2}{\mathrm{d}y^2} \left[M''(V(y)) p(V(y)) V'(y)\right]_{\big|y=M'(t)},$$

where $V(M'(t)) = t$ *for any* $t \in \Theta$.

To conclude this subsection, consider the expressions (19)–(21) and notice that they depend explicitly on the posterior distribution $\pi_n(\cdot \mid x_1, \ldots, x_n)$. Now, thanks to Assumption 1, the mapping $t \mapsto \delta_t$ can be seen as defined on Θ and taking values in the dual space $\mathcal{H}^*$, with Riesz representative $\mathfrak{h}_t \in \mathcal{H}$. More formally, for any $h \in \mathcal{H}$ and $t \in \Theta$, it holds that $h(t) = {}_{\mathcal{H}}\langle h, \delta_t \rangle_{\mathcal{H}^*} = \langle h, \mathfrak{h}_t \rangle$, where $\langle \cdot, \cdot \rangle$ stands for the scalar product on $\mathcal{H}$ while ${}_{\mathcal{H}}\langle \cdot, \cdot \rangle_{\mathcal{H}^*}$ denotes the pairing between $\mathcal{H}$ and $\mathcal{H}^*$. In this notation, the posterior distribution can be rewritten in exponential form as:

$$\pi_n(B \mid X_1, \ldots, X_n) = \frac{\int_B \exp\{n\langle H_n, \mathfrak{h}_\theta \rangle\} \pi(\mathrm{d}\theta)}{\int_\Theta \exp\{n\langle H_n, \mathfrak{h}_\theta \rangle\} \pi(\mathrm{d}\theta)} = \pi_n^*(B \mid H_n) \tag{27}$$

for any $B \in \mathscr{T}$, the probability kernel $\pi_n^*(\cdot|\cdot) : \mathscr{T} \times \mathcal{H} \to [0,1]$ being defined by

$$\pi_n^*(B \mid h) := \frac{\int_B \exp\{n\langle h, \mathfrak{h}_\theta \rangle\} \pi(\mathrm{d}\theta)}{\int_\Theta \exp\{n\langle h, \mathfrak{h}_\theta \rangle\} \pi(\mathrm{d}\theta)}\,. \tag{28}$$

This is particularly interesting because it shows that the posterior distribution can always be thought of, in the presence of a dominated statistical model characterized by strictly positive, smooth densities, as an element of an exponential family, even if the original statistical model $\mu(\cdot|\cdot)$ is not in exponential form. By utilizing the kernel π_n^* in combination with the p.m. ν_n, the following re-writings of (19)–(21) are valid:

$$\begin{aligned} \mathfrak{R}_{\mathbb{U}}[\hat{U}_{n,m}] = \int_{\mathcal{H}} \Bigg\{ & \int_\Theta \mathsf{v}(\theta) \pi_n^*(\mathrm{d}\theta \mid h) + \int_\Theta [\mathsf{m}(\theta)]^2 \pi_n^*(\mathrm{d}\theta \mid h) \\ & - \left(\int_\Theta \mathsf{m}(\theta) \pi_n^*(\mathrm{d}\theta \mid h) \right)^2 \Bigg\} \nu_n(\mathrm{d}h) \end{aligned} \tag{29}$$

$$\mathfrak{R}_\Theta[T_{n,m}^\dagger] = \int_{\mathcal{H}} \int_\Theta [\Delta(\theta, \mathfrak{T}_{n,m}^\dagger(h))]^2 \pi_n^*(\mathrm{d}\theta \mid h) \nu_n(\mathrm{d}h), \tag{30}$$

where the mapping $\mathfrak{T}_{n,m}^\dagger$ is such that $\mathfrak{T}_{n,m}^\dagger(H_n) = t_{n,m}^\dagger(X_1, \ldots, X_n)$ holds $\mathbb{P}$-a.s.

3. Main Results

The first result establishes a relationship between the Bayesian risk functions $\mathfrak{R}_{\mathbb{U}}$ and $\mathfrak{R}_{\Theta}$ defined in (16) and (17), respectively. Due to the central role of this relationship, it will be formulated within the general framework described in Section 2.1.

Proposition 2. *Consider any estimator $U^{\dagger}_{n,m} = u^{\dagger}_{n,m}(X_1, \ldots, X_n)$ of $U_{n,m}$ and any estimator $T^{\dagger}_{n,m} = t^{\dagger}_{n,m}(X_1, \ldots, X_n)$ of T such that $\mathbb{E}[\mathrm{d}^2_{\mathbb{U}}(U^{\dagger}_{n,m}, u_0)] < +\infty$ holds for some $u_0 \in \mathbb{U}$ along with $\mathbb{E}\Big[\mathcal{L}_{\Theta,(X_1,\ldots,X_n)}(T^{\dagger}_{n,m}, t_0)\Big] < +\infty$ for some $t_0 \in \Theta$. Then, it holds*

$$\begin{aligned}\mathfrak{R}_{\mathbb{U}}[U^{\dagger}_{n,m}] \leq \mathfrak{R}_{\Theta}[T^{\dagger}_{n,m}] &+ \mathbb{E}\left[\int_{\mathbb{U}} \mathrm{d}^2_{\mathbb{U}}(U^{\dagger}_{n,m}, u)\gamma_{T^{\dagger}_{n,m},(X_1,\ldots,X_n)}(\mathrm{d}u)\right] \\ &+ 2\mathbb{E}\left[\mathcal{L}^{1/2}_{\Theta,(X_1,\ldots,X_n)}(T, T^{\dagger}_{n,m})\left(\int_{\mathbb{U}} \mathrm{d}^2_{\mathbb{U}}(U^{\dagger}_{n,m}, u)\gamma_{T^{\dagger}_{n,m},(X_1,\ldots,X_n)}(\mathrm{d}u)\right)^{1/2}\right]. \qquad (31)\end{aligned}$$

In particular, if the Bayesian risk function $\mathfrak{R}_{\Theta}$ is optimized by choosing $T^{\dagger}_{n,m} = \hat{T}_{n,m}$, where $\hat{T}_{n,m}$ is as in (15), and $U^{\dagger}_{n,m}$ is chosen equal to $U^{*}_{n,m}$, where $U^{*}_{n,m}$ is as in (13), then (31) becomes

$$\begin{aligned}\mathfrak{R}_{\mathbb{U}}[U^{*}_{n,m}] &\leq \mathfrak{R}_{\Theta}[\hat{T}_{n,m}] + \mathbb{E}\left[\int_{\mathbb{U}} \mathrm{d}^2_{\mathbb{U}}(U^{*}_{n,m}, u)\gamma_{\hat{T}_{n,m},(X_1,\ldots,X_n)}(\mathrm{d}u)\right] \\ &\quad + 2\mathbb{E}\left[\mathcal{L}^{1/2}_{\Theta,(X_1,\ldots,X_n)}(T, \hat{T}_{n,m})\left(\int_{\mathbb{U}} \mathrm{d}^2_{\mathbb{U}}(U^{*}_{n,m}, u)\gamma_{\hat{T}_{n,m},(X_1,\ldots,X_n)}(\mathrm{d}u)\right)^{1/2}\right] \\ &= \inf_{T^{\dagger}_{n,m}} \mathfrak{R}_{\Theta}[T^{\dagger}_{n,m}] + \mathbb{E}\left[\inf_{U^{\dagger}_{n,m}} \int_{\mathbb{U}} \mathrm{d}^2_{\mathbb{U}}(U^{\dagger}_{n,m}, u)\gamma_{\hat{T}_{n,m},(X_1,\ldots,X_n)}(\mathrm{d}u)\right] \\ &\quad + 2\mathbb{E}\left[\mathcal{L}^{1/2}_{\Theta,(X_1,\ldots,X_n)}(T, \hat{T}_{n,m})\left(\int_{\mathbb{U}} \mathrm{d}^2_{\mathbb{U}}(U^{*}_{n,m}, u)\gamma_{\hat{T}_{n,m},(X_1,\ldots,X_n)}(\mathrm{d}u)\right)^{1/2}\right]. \qquad (32)\end{aligned}$$

As an immediate remark, notice that the last member of (32) is obtained by first optimizing the risk $\mathfrak{R}_{\Theta}$ with respect to the choice of $T^{\dagger}_{n,m}$ and then, after getting $\hat{T}_{n,m}$, the term $\mathbb{E}\big[\int_{\mathbb{U}} \mathrm{d}^2_{\mathbb{U}}(U^{\dagger}_{n,m}, u)\gamma_{\hat{T}_{n,m},(X_1,\ldots,X_n)}(\mathrm{d}u)\big]$ is optimized with respect to the choice of $U^{\dagger}_{n,m}$. Of course, it can be argued about the convenience of this procedure—and it is actually due—even if, in most problems, it seems that the strategy proposed in Proposition 2 proves indeed to be the simplest and the most feasible one, above all if computational issues are taken into account. In fact, the absolute best theoretical strategy—consisting of optimizing the right-hand side of (31) jointly with respect to the choice of $(U^{\dagger}_{n,m}, T^{\dagger}_{n,m})$—turns out to be very often too complex and onerous to carry out. Therefore, it seems reasonable to quantify, at least approximately, how far the strategy of Proposition 2 is from absolute optimality, in terms of efficiency. Finally, the additional term

$$2\mathbb{E}\left[\mathcal{L}^{1/2}_{\Theta,(X_1,\ldots,X_n)}(T, \hat{T}_{n,m})\left(\int_{\mathbb{U}} \mathrm{d}^2_{\mathbb{U}}(U^{*}_{n,m}, u)\gamma_{\hat{T}_{n,m},(X_1,\ldots,X_n)}(\mathrm{d}u)\right)^{1/2}\right] \qquad (33)$$

will be reconsidered in next statement, within the simplified setting of Section 2.2. Indeed, by arguing asymptotically, it will be shown that it is essentially negligible, proving in this way a sort of "Pythagorean inequality".

Henceforth, to make the above remark effective, we will formulate the subsequent results within the simplified setting introduced in Section 2.2. Indeed, **Steps 1–3** mentioned in the Introduction are worthy of being reconsidered in light of Proposition 2. On the one hand, **Steps 1** and **2** boil down to checking the existence and uniqueness of the barycenters appearing in (15) and (13), for instance by using the results contained in [24–26]. On the other hand, **Step 3** hinges on the validity of (2) and (3), which are somewhat related to inequality (32). More precisely, (2) will be proved directly by resorting to identity (29),

while (3) will be obtained by estimating the right-hand side of (32). Here is a precise statement.

Proposition 3. *Besides Assumptions 1 and* (18)*, suppose that* $p > 0$ *and* $p \in \mathrm{C}^1(\Theta)$, $\mathsf{m}, \mathsf{v} \in \mathrm{C}^2(\Theta)$, $\Delta^2 \in \mathrm{C}^2(\Theta^2)$, *and* κ_t *is any element of* $\mathcal{H} \cap \mathrm{C}^3(\Theta)$ *with a unique minimum point at* $t \in \Theta$. *Then, it holds*

$$\int_\Theta \left\{ \mathsf{v}(\theta) + [\mathsf{m}(\theta)]^2 \right\} \pi_n^*(\mathrm{d}\theta \mid -\kappa_t) - \left(\int_\Theta \mathsf{m}(\theta) \pi_n^*(\mathrm{d}\theta \mid -\kappa_t) \right)^2$$
$$= \mathsf{v}(t) + \frac{1}{n\kappa_t''(t)} \left\{ [\mathsf{m}'(t)]^2 + \frac{1}{2}\mathsf{v}''(t) + \mathsf{v}'(t) \left[\frac{p'(t)}{p(t)} - \frac{1}{2}\frac{\kappa_t'''(t)}{\kappa_t''(t)} \right] \right\} + o\left(\frac{1}{n}\right) \tag{34}$$

$$\int_\Theta \Delta^2(\theta, \tau) \pi_n^*(\mathrm{d}\theta \mid -\kappa_t)$$
$$= \Delta^2(t, \tau) + \frac{1}{n\kappa_t''(t)} \left\{ \frac{1}{2}\frac{\partial^2}{\partial\theta^2}\Delta^2(\theta, \tau)_{\big|\theta=t} + \frac{\partial}{\partial\theta}\Delta^2(\theta, \tau)_{\big|\theta=t} \left[\frac{p'(t)}{p(t)} - \frac{1}{2}\frac{\kappa_t'''(t)}{\kappa_t''(t)} \right] \right\} + o\left(\frac{1}{n}\right) \tag{35}$$

as $n \to +\infty$, *for any* $\tau \in \Theta$.

Here, it is worth noticing that the asymptotic expansions derived in the above proposition are obtained by means of the Laplace method, as first proposed in [39]. See also [40] [Chapter 20]. At this stage, we face the problem of optimizing the left-hand side of (35) with respect to τ. Since the explicit expression of $\Delta^2(t, \tau)$ will be hardly known in closed form, a reasonable strategy considers, for fixed $t \in \Theta$, the optimization of the right-hand side of (35) with respect to τ, disregarding the remainder term $o(1/n)$. If $\Delta^2 \in \mathrm{C}^3(\Theta^2)$, this attempt leads to considering the equation

$$\frac{\partial}{\partial\tau} \left[\Delta^2(t, \tau) + \frac{1}{n\kappa_t''(t)} \left\{ \frac{1}{2}\frac{\partial^2}{\partial\theta^2}\Delta^2(\theta, \tau)_{\big|\theta=t} + \frac{\partial}{\partial\theta}\Delta^2(\theta, \tau)_{\big|\theta=t} \left[\frac{p'(t)}{p(t)} - \frac{1}{2}\frac{\kappa_t'''(t)}{\kappa_t''(t)} \right] \right\} \right] = 0 \tag{36}$$

and, since

$$\frac{\partial}{\partial\tau}\Delta^2(t, \tau)_{\big|\tau=t} = 0\,, \tag{37}$$

we have that any solution of (36) is of the form $\hat{\tau}_n = t + \epsilon_n$, with some ϵ_n that goes to zero as $n \to +\infty$. For completeness, the validity of (37) could be obtained by using the explicit expression of the Wasserstein distance due to Dall'Aglio. See [41].

If $\Delta^2 \in \mathrm{C}^4(\Theta^2)$, we can plug the expression of $\hat{\tau}_n$ into (35), and expand further the right-hand side. Exploiting that

$$\Delta^2(t, \hat{\tau}_n) = \frac{1}{2}\frac{\partial^2}{\partial\tau^2}\Delta^2(t, \tau)_{\big|\tau=t} \cdot \epsilon_n^2 + o(\epsilon_n^2)$$
$$\frac{\partial}{\partial t}\Delta^2(t, \hat{\tau}_n) = \frac{\partial}{\partial\tau}\left[\frac{\partial}{\partial t}\Delta^2(t, \tau)\right]_{\big|\tau=t} \cdot \epsilon_n + \frac{1}{2}\frac{\partial^2}{\partial\tau^2}\left[\frac{\partial}{\partial t}\Delta^2(t, \tau)\right]_{\big|\tau=t} \cdot \epsilon_n^2 + o(\epsilon_n^2)$$
$$\frac{\partial^2}{\partial t^2}\Delta^2(t, \hat{\tau}_n) = \frac{\partial^2}{\partial t^2}\Delta^2(t, \tau)_{\big|\tau=t} + \frac{\partial}{\partial\tau}\left[\frac{\partial^2}{\partial t^2}\Delta^2(t, \tau)\right]_{\big|\tau=t} \cdot \epsilon_n$$
$$+ \frac{1}{2}\frac{\partial^2}{\partial\tau^2}\left[\frac{\partial^2}{\partial t^2}\Delta^2(t, \tau)\right]_{\big|\tau=t} \cdot \epsilon_n^2 + o(\epsilon_n^2),$$

we get

$$\begin{aligned}
&\int_\Theta \Delta^2(\theta,\hat\tau_n)\pi_n^*(\mathrm{d}\theta \mid -\kappa_t)\\
&= \frac{1}{2}\frac{\partial^2}{\partial\tau^2}\Delta^2(t,\tau)_{\big|\tau=t}\cdot\epsilon_n^2 + \frac{1}{n\kappa_t''(t)}\left\{\frac{1}{2}\frac{\partial^2}{\partial t^2}\Delta^2(t,\tau)_{\big|\tau=t} + \frac{1}{2}\frac{\partial}{\partial\tau}\left[\frac{\partial^2}{\partial t^2}\Delta^2(t,\tau)\right]_{\big|\tau=t}\cdot\epsilon_n\right.\\
&\quad + \frac{1}{4}\frac{\partial^2}{\partial\tau^2}\left[\frac{\partial^2}{\partial t^2}\Delta^2(t,\tau)\right]_{\big|\tau=t}\cdot\epsilon_n^2 + \left[\frac{p'(t)}{p(t)} - \frac{1}{2}\frac{\kappa_t'''(t)}{\kappa_t''(t)}\right]\cdot\frac{\partial}{\partial\tau}\left[\frac{\partial}{\partial t}\Delta^2(t,\tau)\right]_{\big|\tau=t}\cdot\epsilon_n\\
&\quad \left.+\frac{1}{2}\left[\frac{p'(t)}{p(t)} - \frac{1}{2}\frac{\kappa_t'''(t)}{\kappa_t''(t)}\right]\cdot\frac{\partial^2}{\partial\tau^2}\left[\frac{\partial}{\partial t}\Delta^2(t,\tau)\right]_{\big|\tau=t}\cdot\epsilon_n^2\right\} + o(\epsilon_n^2) + o\left(\frac{1}{n}\right). \qquad (38)
\end{aligned}$$

The right-hand side of this expression has the form

$$a\cdot\epsilon_n^2 + \frac{1}{n}\left[A\cdot\epsilon_n^2 + B\cdot\epsilon_n + C\right] + o(\epsilon_n^2) + o\left(\frac{1}{n}\right),$$

so that the choice

$$\epsilon_n = -\frac{B}{2na}\left(1+\frac{A}{na}\right) + o\left(\frac{1}{n^2}\right) = -\frac{B}{2na} + o\left(\frac{1}{n}\right)$$

optimizes its expression. Whence,

$$\begin{aligned}
\hat\tau_n = t &- \frac{1}{n\kappa_t''(t)}\left(\frac{\partial^2}{\partial\tau^2}\Delta^2(t,\tau)_{\big|\tau=t}\right)^{-1}\cdot\left\{\frac{1}{2}\frac{\partial}{\partial\tau}\left[\frac{\partial^2}{\partial t^2}\Delta^2(t,\tau)\right]_{\big|\tau=t}\right.\\
&\left.+\left[\frac{p'(t)}{p(t)} - \frac{1}{2}\frac{\kappa_t'''(t)}{\kappa_t''(t)}\right]\cdot\frac{\partial}{\partial\tau}\left[\frac{\partial}{\partial t}\Delta^2(t,\tau)\right]_{\big|\tau=t}\right\} + o\left(\frac{1}{n}\right) \qquad (39)
\end{aligned}$$

and consequently

$$\int_\Theta \Delta^2(\theta,\hat\tau_n)\pi_n^*(\mathrm{d}\theta \mid -\kappa_t) = \frac{1}{2n\kappa_t''(t)}\frac{\partial^2}{\partial t^2}\Delta^2(t,\tau)_{\big|\tau=t} + o\left(\frac{1}{n}\right). \qquad (40)$$

A first consequence of these computations is that the (Bayesian) estimator $\hat T_{n,m}$ in (15) has the same form as (39) with t and κ_t replaced by the MLE, denoted by $\hat\theta_n$, and $-H_n$, respectively. Of course, this fact has some relevance only in the case that $\hat\theta_n$ exists and is unique. Moreover, coming back to (32), it is worth noticing that

$$\inf_{U_{n,m}^\dagger}\int_{\mathbb{U}}\left|U_{n,m}^\dagger - u\right|^2\gamma_{\hat\tau_n}(\mathrm{d}u) = \mathsf{v}(\hat\tau_n) = \mathsf{v}(t) + \mathsf{v}'(t)\epsilon_n + o\left(\frac{1}{n}\right), \qquad (41)$$

where we have dropped the dependence on $(X_1,\dots,X_n)$ in the expression of $\gamma_{\hat\tau_n}$, in agreement with the simplified setting of Section 2.2 we are following. The last preliminary remark is about the additional term (33) that appears in the last member of (32). In fact, exploiting from the beginning that $\mathbb{U}=\mathbb{R}$ and $\mathcal{L}_{\mathbb{U}}(u,v) = |u-v|^2$, we find that it reduces to

$$\begin{aligned}
&2\mathbb{E}\Big[\int_\Theta\int_{\mathbb{X}^m}\Big(u_m(\mathbf{y},\theta) - u_m(\mathbf{y},\hat T_{n,m})\Big)\Big(u_m(\mathbf{y},\hat T_{n,m}) - \mathsf{m}(\hat T_{n,m})\Big)\times\\
&\times\mu^{\otimes m}(\mathrm{d}\mathbf{y}\mid\theta)\pi_n(\mathrm{d}\theta\mid X_1,\dots,X_n)\Big] \qquad (42)
\end{aligned}$$

by which we notice that it also involves "covariance terms". The way is now paved to state the following

Theorem 2. *Besides Assumptions 1 and (18), suppose that* $\mathsf{m}, \mathsf{v} \in \mathrm{C}^2(\Theta)$ *and* $\Delta^2 \in \mathrm{C}^4(\Theta^2)$. *Then, the identities*

$$\frac{\partial^2}{\partial \tau^2}\Delta^2(t,\tau)_{\big|\tau=t} = -\frac{\partial}{\partial \tau}\left[\frac{\partial}{\partial t}\Delta^2(t,\tau)\right]_{\big|\tau=t} \tag{43}$$

$$\frac{1}{2}\mathsf{v}''(t) + [\mathsf{m}'(t)]^2 = \frac{1}{2}\frac{\partial^2}{\partial t^2}\Delta^2(t,\tau)_{\big|\tau=t} - \frac{1}{2}\left(\frac{\partial^2}{\partial \tau^2}\Delta^2(t,\tau)_{\big|\tau=t}\right)^{-1}\frac{\partial}{\partial \tau}\left[\frac{\partial^2}{\partial t^2}\Delta^2(t,\tau)\right]_{\big|\tau=t}\mathsf{v}'(t) \tag{44}$$

entail that

$$\int_\Theta \left\{\mathsf{v}(\theta) + [\mathsf{m}(\theta)]^2\right\}\pi_n^*(\mathrm{d}\theta \mid -\kappa_t) - \left(\int_\Theta \mathsf{m}(\theta)\pi_n^*(\mathrm{d}\theta \mid -\kappa_t)\right)^2 = \int_\Theta \Delta^2(\theta, \hat{\tau}_n)\pi_n^*(\mathrm{d}\theta \mid -\kappa_t) + \mathsf{v}(t) + \mathsf{v}'(t)\epsilon_n + o\left(\frac{1}{n}\right) \tag{45}$$

for any $t \in \Theta$, *any* κ_t *in* $\mathcal{H} \cap \mathrm{C}^3(\Theta)$ *with a unique minimum point at* $t \in \Theta$, *and any* $p > 0$ *with* $p \in \mathrm{C}^1(\Theta)$, *provided that the term in (42) is of* $o(\frac{1}{n})$*-type. Thus, if either*

(A1) (26) holds uniformly with respect to some class $\mathfrak{F}$ *of continuous functionals* $\Psi : \mathcal{H} \to \mathbb{R}$, *in the sense that*

$$\sup_{\Psi \in \mathfrak{F}}\left|\int_{\mathcal{H}} \Psi(h)\nu_n(\mathrm{d}h) = \int_{\mathcal{H}} \Psi(h)\nu(\mathrm{d}h) + \frac{1}{n}\int_{\mathcal{H}} \Psi(h)\Gamma(h;\mu,\pi)\nu(\mathrm{d}h)\right| = o\left(\frac{1}{n}\right)$$

(A2) both the functionals $h \mapsto \int_\Theta \{\mathsf{v}(\theta) + [\mathsf{m}(\theta)]^2\}\pi_n^*(\mathrm{d}\theta \mid h) - \left(\int_\Theta \mathsf{m}(\theta)\pi_n^*(\mathrm{d}\theta \mid h)\right)^2$ *and* $h \mapsto \inf_{\mathfrak{T}_{n,m}^\dagger} \int_\Theta [\Delta(\theta, \mathfrak{T}_{n,m}^\dagger(h))]^2\pi_n^*(\mathrm{d}\theta \mid h)$ *belong to* $\mathfrak{F}$, *for all* $n \in \mathbb{N}$

or

(B1) (34) and (40) hold uniformly for all κ_t *belonging to a given subset* $\mathcal{D}$ *of* $\mathcal{H}$

(B2) $\nu_n(\mathcal{D}) = 1$ *for all* $n \in \mathbb{N}$

then (2)–(3) are in force with

$$\hat{R}_{0,m} = R_{0,m}^* = \int_\Theta \mathsf{v}(t)\pi(\mathrm{d}t) \tag{46}$$

$$\hat{R}_{1,m} = R_{1,m}^* = \int_\Theta \frac{1}{\overline{\kappa}_t''(t)}\left\{[\mathsf{m}'(t)]^2 + \frac{1}{2}\mathsf{v}''(t) + \mathsf{v}'(t)\left[\frac{p'(t)}{p(t)} - \frac{1}{2}\frac{\overline{\kappa}_t'''(t)}{\overline{\kappa}_t''(t)}\right]\right\}\pi(\mathrm{d}t) + \int_\Theta \mathsf{v}(t)\Gamma(\overline{\kappa}_t;\mu,\pi)\pi(\mathrm{d}t), \tag{47}$$

where $\overline{\kappa}_t(\theta) := \mathsf{K}(t \,||\, \theta)$, *for any* $p > 0$ *with* $p \in \mathrm{C}^1(\Theta)$.

As announced in the Introduction, here we have minted the term *compatibility equations* to refer to identities (43) and (44). They actually constitute two "compatibility conditions" that involve only the statistical model, without any mention to the prior. The dependence on the quantity to be estimated is indeed hidden in the expression of Δ^2. More deeply, these equations can be viewed as a check on the compatibility between the original estimation problem (1) and the fact that we have metrized the space of the parameters Θ as in (20). Actually, they could have a more general value if interpreted as relations aimed at *characterizing* Δ^2, rather than imposing that this distance is given in terms of the Wasserstein distance as in (20). However, for a distance that is characterized differently from (20), an analogous of inequality (32) should be checked in terms of this new distance on Θ. As to the concrete check of the compatibility equations, we notice that the former identity (43) is generally valid as a consequence of the representation formula or the Wasserstein distance

due to Dall'Aglio (see [41]), as long as the exchange between derivatives and integrals is allowed. For the other identity (44), we have instead collected in Section 4 some examples of simple statistical models for which its verification proves to be quite simple. Finally, the issue of extending these equations in a higher dimension, including the infinite dimension, is deferred to Section 6.

Apropos of the other assumptions, the verification that the term in (42) is of $o(\frac{1}{n})$-type is generally straightforward. For instance, such a term is even equal to zero if u_m is independent of θ. As to the two groups of assumptions which are needed to prove (46) and (47), the latter block, formed by (B1) and (B2), is certainly easier to check. However, (B1) and (B2) can prove to be rather strong since they require the existence of the MLE for any $n \in \mathbb{N}$. On the other hand, checking (A1) and (A2) is generally harder since it constitutes a strong reinforcement of de Finetti's Law of Large Number for the log-likelihood process, similar in its conception to those stated in [35,36]. Moreover, the check of (A2) is more or less equivalent to prove a uniform regularity of the mapping $h \mapsto \pi_n^*(\mathrm{d}\theta \mid h)$, as a map from $\mathcal{H}$ into the space of p.m.'s on $(\Theta, \mathscr{T})$ metrized with a Wasserstein distance. This theory is presented and developed in [42,43]. In any case, these lines of research deserve further investigations, to be deferred to a forthcoming paper.

Finally, we consider **Steps 4–6** mentioned in the Introduction, in light of the previous results. In fact, the compatibility Equations (43) and (44) suggest two new compatibility conditions, which are necessary to get (10) along with $\hat{R}_{i,m} = \tilde{R}_{i,m}$ for $i = 0, 1$. A formal statement reads as follows.

Theorem 3. *Besides Assumptions 1 and* (18)*, suppose that* $\mathsf{m}, \mathsf{v} \in \mathrm{C}^2(\Theta)$, $\Delta^2 \in \mathrm{C}^4(\Theta^2)$. *Assume also that either (A1) and (A2) or (B1) and (B2) of* Theorem 2 *are in force. Then, any solution* $\hat{\tau}_n$ *of the following equations:*

$$\mathsf{v}(\hat{\theta}_n) = \mathsf{v}(\hat{\tau}_n) + \Delta^2(\hat{\tau}_n, \hat{\theta}_n) + o\left(\frac{1}{n}\right) \tag{48}$$

$$\mathsf{v}'(\hat{\theta}_n) = \frac{\partial}{\partial t}\Delta^2(\hat{\tau}_n, t)_{\big|t=\hat{\theta}_n} + o\left(\frac{1}{n}\right) \tag{49}$$

$$\frac{1}{2}\mathsf{v}''(\hat{\theta}_n) + [\mathsf{m}'(\hat{\theta}_n)]^2 = \frac{1}{2}\frac{\partial^2}{\partial t^2}\Delta^2(\hat{\tau}_n, t)_{\big|t=\hat{\theta}_n} + o\left(\frac{1}{n}\right), \tag{50}$$

where $\hat{\theta}_n$ *stands for the MLE, yields a prior-free estimator* $\hat{T}_{n,m}$ *and, through* ***Step 5****, another prior-free estimator* $\tilde{U}_{n,m}$ *that satisfies* (10) *along with* $\hat{R}_{i,m} = \tilde{R}_{i,m}$ *for* $i = 0, 1$*, where* $\hat{R}_{0,m}$ *and* $\hat{R}_{1,m}$ *are as in* (46) *and* (47)*, respectively, provided that the term in* (42) *is of* $o(\frac{1}{n})$*-type.*

The derivation of new prior free-estimators via this procedure represents a novel line of research that we would like to pursue in forthcoming works.

4. Applications and Examples

This section is split into four subsections, and has two main purposes. In fact, Sections 4.1–4.3 just contain explicit examples of very simple statistical models for which the compatibility equations are satisfied. These models are the one-dimensional Gaussian, the exponential and the Pareto model. Section 4.4 has a different nature, since it is devoted to a more concrete application of our approach to the original Poisson-mixture setting used by Herbert Robbins to introduce his own approach to empirical Bayes theory. Finally, Section 4.5 carries on the discussion initiated in Section 4.4 by showing a concrete application relative to one year of claims data for an automobile insurance company.

4.1. The Gaussian Model

Here, we have $\mathbb{X} = \Theta = \mathbb{R}$ and

$$\mu(A \mid \theta) = \int_A \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\theta)^2\}\mathrm{d}x \qquad (A \in \mathscr{B}(\mathbb{R}))$$

for some known $\sigma^2 > 0$. For simplicity, we put $m = 1$ and $u_1(y, \theta) = y$, which is tantamount to saying that the original predictive aim was focused on the estimation of X_{n+1}. In this setting, it is very straightforward to check that $\mathsf{m}(\theta) = \theta$ and $\mathsf{v}(\theta) = \sigma^2$. Moreover, in view of well-know computations on the Wasserstein distance (see [44,45]), it is also straightforward to check that $\Delta^2(\theta, \tau) = |\theta - \tau|^2$. Therefore, (43) becomes $2 = 2$, while (44) reduces to $1 = 1$. Finally, it is also possible to check the validity of (48)–(50) with the simplest choice $\hat{\tau}_n = \hat{\theta}_n$.

The case of constant mean and unknown variance will not be dealt with here because its treatment is substantially included in the following subsection. Apropos of the multidimensional variant of this model, very important in many statistical applications, we just mention the interesting paper [46] which paves the way, mathematically speaking, to write down the multidimensional analogous of the compatibility equations in a full Riemannian context.

4.2. The Exponential Model

Here, we have $\mathbb{X} = \Theta = (0, \infty)$ and

$$\mu(A \mid \theta) = \int_A \theta e^{-\theta x} \mathrm{d}x \qquad (A \in \mathscr{B}(0, +\infty)) \,.$$

Again, for simplicity, we put $m = 1$ and $u_1(y, \theta) = y$, which is tantamount to saying that the original predictive aim was focused on the estimation of X_{n+1}. In this setting, it is very straightforward to check that $\mathsf{m}(\theta) = 1/\theta$ and $\mathsf{v}(\theta) = 1/\theta^2$. Moreover, by resorting to Dall'Aglio representation of the Wasserstein distance (see [41]), it is also straightforward to check that $\Delta^2(\theta, \tau) = 2|1/\theta - 1/\tau|^2$. Although very simple, this is a very interesting example of non-Euclidean distance on $\Theta = (0, \infty)$. As to the validity of the compatibility equations, we easily see that (43) yields $4/t^4 = 4/t^4$, while (44) becomes:

$$\frac{3}{t^4} + \left(\frac{1}{t^2}\right)^2 = \frac{1}{2} \cdot \frac{4}{t^4} - \frac{1}{2}\left(\frac{8}{t^5}\right) \cdot \left(\frac{4}{t^4}\right)^{-1} \cdot \left(-\frac{2}{t^3}\right) .$$

4.3. The Pareto Model

Here, we have $\mathbb{X} = \Theta = (0, \infty)$ and

$$\mu(A \mid \theta) = \int_{A \cap (\theta, +\infty)} \frac{\alpha \theta^\alpha}{x^{\alpha+1}} \mathrm{d}x \qquad (A \in \mathscr{B}(0, +\infty))$$

for some known $\alpha > 2$. Again, for simplicity, we put $m = 1$ and $u_1(y, \theta) = y$, which is tantamount to saying that the original predictive aim was focused on the estimation of X_{n+1}. In this setting, it is very straightforward to check that $\mathsf{m}(\theta) = \frac{\alpha}{\alpha-1}\theta$ and $\mathsf{v}(\theta) = \frac{\alpha}{(\alpha-2)(\alpha-1)^2}\theta^2$. Moreover, by resorting to the Dall'Aglio representation of the Wasserstein distance (see [41]), it is also straightforward to check that $\Delta^2(\theta, \tau) = \frac{\alpha}{\alpha-2}|\theta - \tau|^2$. Of course, this is not a regular model since the support of $\mu(\cdot|\theta)$ varies with θ. Anyway, it is interesting to notice that the compatibility equations are still also valid in this case. Therefore, the analysis of such non-regular models should motivate further investigations about their intrinsic value.

4.4. Robbins Approach to Empirical BAYES

In his seminal paper [23], Herbert Robbins introduced the following model to present his own approach to empirical Bayes theory. The problem that he considers is inspired by car insurance data analysis, and it is only slightly different from a "standard" predictive problem. We start by putting $\mathbb{X} = \mathbb{N}_0^2$ and $\mathbb{U} = \mathbb{N}_0$, and considering exchangeable random variables X_i's with $X_i = (\xi_i, \eta_i)$. The practical meaning is that ξ_i represents the number of accidents experienced by the i-th customer in the past year, while η_i represents the number of accidents that the same i-th customer will experience in the current year. Then, Robbins (in his own notation) attaches to each customer a random parameter, say $\lambda_i > 0$ to the i-th

customer, which represents the rate of a Poisson distribution for that customer. Moreover, he considers the λ_i's as i.i.d. and, conditionally on the λ_i's, the X_i's become independent, and in addition ξ_i and η_i become i.i.d. with distribution $\mathrm{Poi}(\lambda_i)$, for all $i \in \mathbb{N}$. Robbins calls G the common distribution of the λ_i's and interpret it as a "prior distribution". However, if we strictly follow the Bayesian main way, we should call this distribution θ to avoid confusion, and just realize that we have, this way, defined the statistical model, that is

$$\mu(\{(k,h)\} \mid \theta) = \int_0^{+\infty} \frac{e^{-z}z^k}{k!}\frac{e^{-z}z^h}{h!}\theta(\mathrm{d}z) \qquad ((k,h) \in \mathbb{N}_0^2)\,. \tag{51}$$

Thus, the actual prior (Bayesianly speaking) is some p.m. π on the space of all p.m.'s on $((0,+\infty),\mathscr{B}(0,+\infty))$, while the random parameter T considered in the present paper is some *random probability measure*. Here, the objective—actually very practical and intuitively logic—is to estimate η_1 on the basis of the sample $(\xi_1,\ldots,\xi_n)$. Thus, our $U_{n,m}$ coincides with η_1 and the loss function is just, as usual, the quadratic loss. Throughout his paper, Robbins works under the conditioning to $T = \theta$ (that his under a fixed prior, in his own terminology). Hence, his "theoretical estimator" reads

$$\mathbb{E}_\theta[\eta_1 \mid (\xi_1,\ldots,\xi_n)] = \mathbb{E}_\theta[\eta_1 \mid \xi_1] = (\xi_1+1)\frac{p_\theta(\xi_1+1)}{p_\theta(\xi_1)}, \tag{52}$$

where $p_\theta(k) := \mu(\{k\}\times\mathbb{N}_0 \mid \theta)$. To get rid of the unobservable θ, Robbins exploits that $\theta = \mathbb{E}_\theta[\xi_1] = \sum_{k=0}^{+\infty} k p_\theta(k)$ to bring the Strong Law of Large Numbers into the game. Indeed, since

$$\hat{p}(k) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\xi_i = k\} \overset{\mathbb{P}_\theta - a.s.}{\longrightarrow} p_\theta(k)$$

holds for any θ, then it could be worth considering the (prior-free) estimator:

$$\tilde{U}_{n,m} = (\xi_1+1)\frac{\hat{p}(\xi_1+1)}{\hat{p}(\xi_1)}\,. \tag{53}$$

At this stage, if we want to maintain the Bayesian main way, we should make three basic considerations. First, given the statistical model (51), independently of the estimation problem, the assumption of exchangeability of the X_i's entails the existence of some prior distribution π, by de Finetti's representation theorem. Second, given the quadratic loss function on $\mathbb{U}$, the best (i.e., the most efficient) estimator is given by:

$$\hat{U}_{n,m} := \mathbb{E}[\eta_1 \mid (\xi_1,\ldots,\xi_n)],$$

where the expectation $\mathbb{E}$ depends of course on the prior π. Third, if we consider the above estimator as useless, because of an effective ignorance about the prior π, we are justified to consider the above $\tilde{U}_{n,m}$ as a possible approximation of $\hat{U}_{n,m}$, in the sense expressed by the joint validity of (2) and (10), with $\hat{R}_{i,m} = \tilde{R}_{i,m}$ for $i = 0,1$, uniformly with respect to a whole (possible very large) class of priors π. Unfortunately, it is not the case. Or rather, we could actually achieve this goal, in the presence of distinguished choices of π. Therefore, if there is ignorance on π, we can only consider the Robbins estimator as efficient "at zero-level", and not also "at $O(\frac{1}{n})$-level". If we follow the approach presented in this paper, the natural choice for an estimator is given by:

$$U^*_{n,m} = (\xi_1+1)\frac{\displaystyle\int_0^{+\infty}\left(\frac{e^{-z}z^{\xi_1+1}}{(\xi_1+1)!}\right)\hat{T}_{n,m}(\mathrm{d}z)}{\displaystyle\int_0^{+\infty}\left(\frac{e^{-z}z^{\xi_1}}{\xi_1!}\right)\hat{T}_{n,m}(\mathrm{d}z)}, \tag{54}$$

where the estimator $\hat{T}_{n,m}$ belongs to the effective space of the parameters Θ, that is the space of all p.m.'s on $((0,+\infty),\mathscr{B}(0,+\infty))$, and is identified as:

$$\hat{T}_{n,m} = \text{Argmin}_{\tau} \int_{\Theta} \mathcal{W}_2^2(\mu_{\xi_1,\theta}, \mu_{\xi_1,\tau}) \pi_n(\mathrm{d}\theta \mid \xi_1, \ldots, \xi_n), \tag{55}$$

with

$$\mu_{k,\theta}(A) := \frac{\displaystyle\int_A \left(\frac{e^{-z} z^k}{k!}\right) \theta(\mathrm{d}z)}{\displaystyle\int_0^{+\infty} \left(\frac{e^{-z} z^k}{k!}\right) \theta(\mathrm{d}z)} \qquad (A \in \mathscr{B}(0,+\infty)) \,.$$

The proof of the fact that our estimator is more efficient than Robbins estimator—at least asymptotically and uniformly with respect to a whole class of priors—will be given in a forthcoming paper. Indeed, such a proof will constitute only a first step towards a complete vindication of our approach. The crowing achievement of the project would be represented by the production of some prior-free approximation of $\hat{T}_{n,m}$ that could lead, through (54), to an efficient estimator $\tilde{U}_{n,m}$ up to the "$O(\frac{1}{n})$-level". Research on this is ongoing.

4.5. An Example of Real Data Analysis

This subsection represents a continuation of the analysis of Robbins' approach to empirical Bayes theory, hinting at some concrete applications. We display below a Table 1 from [47] which is relative to one year of claims data for a European automobile insurance company. The original source of the data is the actuarial work [48].

Table 1. Table reporting, in the second line, the exact counts of claimed accidents. Third and fourth lines display estimated numbers of accidents.

Claims	0	1	2	3	4	5	6	7
Counts	7840	1317	239	42	14	4	4	1
Robbins estimator	0.168	0.363	0.527	1.33	1.43	6.00	1.25	0
Gamma MLE	0.164	0.398	0.633	0.87	1.10	1.34	1.57	0

Here, a population of 9461 automobile insurance policy holders is considered. Out of these, 7840 made no claims during the year; 1317 made a single claim; 239 made two claims each and so forth, continuing to the one person who made seven claims. The insurance company is concerned about the claims each policy holder will make in the next year. The third and the fourth lines provide estimations of such numbers by following the original Robbins method (based on (53)) and another compound model discussed in Section 6.1 of [47], respectively. In particular, the Robbins estimator predicts that the 7840 policy holders that made no claims during the year will contribute to an amount of $7840 \times 0.168 \approx 1317$ accidents, and so on. Analogously, the compound model predicts that the same 7840 policy holders will contribute to an amount of $7840 \times 0.164 \approx 1286$ accidents, and so on. Moreover, it is worth noticing that the original Robbins estimator suffers the lack of certain regularity properties, such as monotonicity, so that various smoothed versions of it have been provided by other authors. See [49]. See also [50] [Chapter 5] for a comprehensive treatment.

Here, we seize the opportunity to give the reader a taste of our approach, as explained in Section 4.4. A detailed treatment would prove, in any case, too complex to be thoroughly developed in this paper, due to the significant amount of numerical techniques which are necessary to carry out our strategy. Indeed, the big issue is concerned with the implementation of the infinite-dimensional minimization problem (55), which is still under investigation. However, we can simplify the treatment by restricting the attention

on prior distributions π that put the total unitary mass, for example, on the set $\mathcal{E}$ of exponential distributions, so that $\theta(\mathrm{d}z) = \beta e^{-\beta z}\mathrm{d}z$ for $z > 0$ and some hyper-parameter $\beta > 0$. Thus, given some hyper-prior ζ on the hyper-parameter β, we can easily see that (55) boils down to a simple, one-dimensional minimization problem. Its solution $\hat{T}_{n,m}$ is provided by the distribution

$$\hat{\beta}_n e^{-\hat{\beta}_n z}\mathrm{d}z$$

with $\hat{\beta}_n$ coinciding with the harmonic mean of the posterior distribution of the hyper-parameter β. On the basis of the theory developed in the paper, this solution will prove asymptotically nearly optimal uniformly with respect to the (narrow) class of prior distributions that put the total unitary mass on $\mathcal{E}$. Whence, the estimator $U^*_{n,m}$ in (54) assumes the form

$$U^*_{n,m} = \frac{\xi_1 + 1}{\hat{\beta}_n + 1}\,.$$

This last estimator is, of course, not prior-free, because $\hat{\beta}_n$ depends on the prior ζ. However, to get a quick result, we can approximate $\hat{\beta}_n$ by means of the Laplace methods again yielding

$$\frac{\xi_1 + 1}{\hat{\beta}_n + 1} \approx (\xi_1 + 1)\frac{S_n}{S_n + n} := \tilde{U}_{n,m},$$

where S_n represents the total amount of accidents. Since $n = 9461$ and $S_n = 2028$ in the dataset under consideration, we provide the following new Table 2,

Table 2. Table reporting, in the second line, the exact counts of claimed accidents. Third line displays estimated numbers of accidents.

Claims	0	1	2	3	4	5	6	7
Counts	7840	1317	239	42	14	4	4	1
Estimator $\tilde{U}_{n,m}$	0.176	0.353	0.53	0.706	0.882	1.06	1.23	1.41

which is indeed comparable with the previous one. To give an idea, the Robbins estimator predicts 2019 total accidents for the next year, while the estimator $\tilde{U}_{n,m}$ above predicts 2033 total accidents for the next year.

In any case, a thorough analysis of this specific example deserves more attention, and will be developed in a forthcoming new paper.

5. Proofs

Gathered here are the proofs of the results stated in the main body of the paper.

5.1. Theorem 1

First, by following the same line of reasoning as in [32], conclude that the sequence $\{H_n\}_{n\geq 1}$ is a Cauchy sequence in $L^2(\Omega;\mathcal{H}) := \{W : \Omega \to \mathcal{H} \mid \mathbb{E}[\|W\|^2_{\mathcal{H}}] < +\infty\}$. Thus, by completeness, there exists a random element H^* in $L^2(\Omega;\mathcal{H})$ such that $H_n \xrightarrow{L^2} H^*$. Now, exploit the continuous embedding $\mathcal{H} \subset \mathrm{C}^0(\overline{\Theta})$. By de Finetti's Strong Law of Large Numbers (see [9]), $H_n(\theta)$ converges $\mathbb{P}$-a.s. to $-\mathsf{K}(T \parallel \theta) + \int_{\mathbb{X}} \log f(z|\, T)\mu(\mathrm{d}z \mid T) = H(\theta)$, for any fixed $\theta \in \Theta$. Since $H \in \mathcal{H}$ by Assumption 1, then $H = H^*$ as elements of $\mathcal{H}$. At this stage the conclusion that $\nu_n \Rightarrow \nu$ follows by the standard implication that L^2-convergence implies convergence in distribution, which is still true for random elements taking values in a separable Hilbert space. See [51].

5.2. Proposition 1

Start by considering a functional Ψ in $\mathrm{C}^2_b(\mathcal{H})$. Notice that

$$\int_{\mathcal{H}} \Psi(h)\nu_n(\mathrm{d}h) = \mathbb{E}[\Psi(H_n)]$$

and then expand the term $\Psi(H_n - H + H)$ by the Taylor formula (see [33,34]) to get

$$\Psi(H_n) = \Psi(H) + \langle \nabla\Psi(H), H_n - H\rangle + \frac{1}{2}\langle \mathrm{Hess}[\Psi]_H(H_n - H), H_n - H\rangle + o(\|H_n - H\|^2)\,.$$

Observe that H is $\sigma(T)$-measurable while, by de Finetti's representation theorem, the distribution of $H_n - H$, given T, coincides with the distribution of a sum of n i.i.d. random elements. Whence, the tower property of the conditional expectation entails:

$$\mathbb{E}[\Psi(H_n)] = \mathbb{E}[\Psi(H)] + \frac{1}{2n}\mathbb{E}[\mathrm{Hess}[\Psi]_H \otimes \mathrm{Cov}_T[\log f(X_i|\,\cdot)]] + o\left(\frac{1}{n}\right)$$

since $\mathbb{E}[H_n - H \mid T] = 0$ and, then,

$$\mathbb{E}[\langle \nabla\Psi(H), H_n - H\rangle \mid T] = \langle \nabla\Psi(H), \mathbb{E}[H_n - H \mid T]\rangle = 0$$

the expression $\mathbb{E}[H_n - H \mid T]$ being intended as a Bochner integral. Thus, the main identity (26) follows immediately from (25), for any $\Psi \in C_b^2(\mathcal{H})$. Once (26) is established for regular Ψ's, one can extend its validity to more general continuous Ψ's by standard approximation arguments.

5.3. Lemma 1

First, observe that:

$$-\mathsf{K}(T \parallel \theta) + \int_{\mathbb{X}} \log f(z|\,T)\mu(\mathrm{d}z \mid T) = \theta M^{'}(T) - M(\theta)\,.$$

Notice also that:

$$\int_{\mathcal{H}} \Psi(h)\nu_n(\mathrm{d}h) = \mathbb{E}\left[\Psi\left(\theta \mapsto \frac{\theta}{n}\sum_{i=1}^{n} S(X_i) - M(\theta)\right)\right].$$

Then, repeat the same arguments as in the previous proof, getting

$$\int_{\mathcal{H}} \Psi(h)\nu_n(\mathrm{d}h) = \int_{\mathcal{H}} \Psi(h)\nu(\mathrm{d}h) + \frac{1}{2n}\int_{\Theta}\left[\frac{\mathrm{d}^2}{\mathrm{d}x^2}\Psi(\theta \mapsto x\theta - M(\theta))\right]_{\big|x=M'(t)} M^{''}(t)p(t)\mathrm{d}t + o\left(\frac{1}{n}\right).$$

For standard exponential families, the function M' is ono-to-one, with inverse function V. Whence, by indicating the range of M' as $Cod(M')$,

$$\begin{aligned}
&\int_{\Theta}\left[\frac{\mathrm{d}^2}{\mathrm{d}x^2}\Psi(\theta \mapsto x\theta - M(\theta))\right]_{\big|x=M'(t)} M^{''}(t)p(t)\mathrm{d}t \\
&= \int_{Cod(M')}\left[\frac{\mathrm{d}^2}{\mathrm{d}x^2}\Psi(\theta \mapsto x\theta - M(\theta))\right] M^{''}(V(x))p(V(x))V^{'}(x)\mathrm{d}x \\
&= \int_{Cod(M')}\Psi(\theta \mapsto x\theta - M(\theta))\left[\frac{\mathrm{d}^2}{\mathrm{d}x^2}[M^{''}(V(x))p(V(x))V^{'}(x)]\right]\mathrm{d}x,
\end{aligned}$$

where, for the last identity, a double integration-by-parts has been used. Finally, changing the variable according to $x = M^{'}(t)$ leads to the desired result.

5.4. Proposition 2

A disintegration argument shows that

$$\begin{aligned}\mathfrak{R}_{\mathbb{U}}[U^{\dagger}_{n,m}] &= \int_{\mathbb{X}^n}\int_{\Theta\times\mathbb{X}^m}\mathcal{L}_{\mathbb{U}}\Big(u_{n,m}(\mathbf{y};\mathbf{x};\theta),u^{\dagger}_{n,m}(\mathbf{x})\Big)\times\\ &\quad\times\mathbb{P}[(X_{n+1},\ldots,X_{n+m})\in \mathrm{d}\mathbf{y},T\in\mathrm{d}\theta\mid(X_1,\ldots,X_n)=\mathbf{x}]\alpha_n(\mathrm{d}\mathbf{x})\\ &= \int_{\mathbb{X}^n}\int_{\Theta\times\mathbb{X}^m}\mathcal{L}_{\mathbb{U}}\Big(u_{n,m}(\mathbf{y};\mathbf{x};\theta),u^{\dagger}_{n,m}(\mathbf{x})\Big)\mu^{\otimes_m}(\mathrm{d}\mathbf{y}\mid\theta)\pi_n(\mathrm{d}\theta\mid\mathbf{x})\alpha_n(\mathrm{d}\mathbf{x})\\ &= \int_{\mathbb{X}^n}\int_{\Theta}\mathcal{W}^2_{\mathbb{U}}\Big(\gamma_{\theta,\mathbf{x}};\delta_{u^{\dagger}_{n,m}(\mathbf{x})}\Big)\pi_n(\mathrm{d}\theta\mid\mathbf{x})\alpha_n(\mathrm{d}\mathbf{x})\,.\end{aligned}$$

Then, use the triangular inequality for the Wasserstein distance to obtain:

$$\mathcal{W}_{\mathbb{U}}\Big(\gamma_{\theta,\mathbf{x}};\delta_{u^{\dagger}_{n,m}(\mathbf{x})}\Big)\leq\mathcal{W}_{\mathbb{U}}\Big(\gamma_{\theta,\mathbf{x}};\gamma_{\tau,\mathbf{x}}\Big)+\mathcal{W}_{\mathbb{U}}\Big(\gamma_{\tau,\mathbf{x}};\delta_{u^{\dagger}_{n,m}(\mathbf{x})}\Big)$$

for any $\tau\in\Theta$. Take the square of both side and observe that:

$$\mathcal{W}_{\mathbb{U}}\Big(\gamma_{\tau,\mathbf{x}};\delta_{u^{\dagger}_{n,m}(\mathbf{x})}\Big)=\int_{\mathbb{U}}\mathrm{d}^2_{\mathbb{U}}(u,u^{\dagger}_{n,m}(\mathbf{x}))\gamma_{\tau,\mathbf{x}}(\mathrm{d}u)\,.$$

Now, (31) is proved by letting $\tau=T^{\dagger}_{n,m}$ after noticing that the latter summand in the above right-hand side is independent of θ.

Finally, (32) is obtained by first optimizing the risk $\mathfrak{R}_{\Theta}$ with respect to the choice of $T^{\dagger}_{n,m}$ and then, after getting $\hat{T}_{n,m}$, the term $\mathbb{E}\big[\int_{\mathbb{U}}\mathrm{d}^2_{\mathbb{U}}(U^{\dagger}_{n,m},u)\gamma_{\hat{T}_{n,m},(\mathrm{X}_1,\ldots,\mathrm{X}_n)}(\mathrm{d}u)\big]$ is optimized with respect to the choice of $U^{\dagger}_{n,m}$.

5.5. Proposition 3

Preliminarily, use Theorem 1 in [52] [Section II.1] to prove that:

$$\int_{\Theta}\varphi(\theta)e^{-\kappa_t(\theta)}\mathrm{d}\theta=\frac{2\sqrt{\pi}}{\sqrt{n}}e^{-\kappa_t(t)}\left[c_0+\frac{c_2}{2n}+o\left(\frac{1}{n}\right)\right]$$

holds for any $\varphi\in\mathrm{C}^2(\Theta)$ such that $\varphi(t)\neq0$, where

$$\begin{aligned}c_0&:=\frac{b_0}{2a_0^{1/2}}\\ c_2&:=\left\{\frac{b_2}{2}-\frac{3a_1b_1}{a_0}+[5a_1^2-4a_0a_2]\frac{3b_0}{16a_0^2}\right\}\times\frac{1}{a_0^{3/2}},\end{aligned}$$

with $a_0:=\frac{1}{2}\kappa_t^{''}(t)$, $a_1:=\frac{1}{3!}\kappa_t^{'''}(t)$, $a_2:=\frac{1}{4!}\kappa_t^{''''}(t)$, $b_0=\varphi(t)$, $b_1=\varphi^{'}(t)$ and $b_2=\frac{1}{2}\varphi^{''}(t)$. Moreover, from that very same theorem, it holds that:

$$\int_{\Theta}\varphi(\theta)e^{-\kappa_t(\theta)}\mathrm{d}\theta=\sqrt{\pi}e^{-\kappa_t(t)}\left[\frac{c_1}{n^{3/2}}+o\left(\frac{1}{n^{3/2}}\right)\right]$$

for any $\varphi\in\mathrm{C}^2(\Theta)$ with a zero of order 1 at t, where

$$c_1:=\left[\frac{b_1^*}{2}-\frac{a_1b_0^*}{2a_0}\right]\frac{1}{a_0}$$

with $b_0^*:=\varphi^{'}(t)$ and $b_1^*:=\frac{1}{2}\varphi^{''}(t)$. At this stage, application of this formulas gives:

$$\int_{\Theta}\mathrm{m}(\theta)\pi_n^*(\mathrm{d}\theta\mid-\kappa_t)=\mathrm{m}(t)+\frac{1}{na_0}\left[\frac{1}{4}\mathrm{m}^{''}(t)+\frac{1}{2}\mathrm{m}^{'}(t)\frac{p'(t)}{p(t)}-\frac{3}{4}\frac{a_1}{a_0}\mathrm{m}^{'}(t)\right]+o\left(\frac{1}{n}\right)$$

and

$$\int_{\Theta} \mathsf{m}^2(\theta)\pi_n^*(\mathrm{d}\theta \mid -\kappa_t) = \mathsf{m}^2(t) + \frac{1}{na_0}\left[\frac{1}{2}(\mathsf{m}'(t))^2 + \frac{1}{2}\mathsf{m}''(t)\mathsf{m}(t) + \mathsf{m}'(t)\mathsf{m}(t)\frac{p'(t)}{p(t)} - \frac{3}{2}\frac{a_1}{a_0}\mathsf{m}'(t)\mathsf{m}(t)\right] + o\left(\frac{1}{n}\right).$$

Then, in addition,

$$\int_{\Theta} [\mathsf{v}(\theta) - \mathsf{v}(t)]\pi_n^*(\mathrm{d}\theta \mid -\kappa_t) = \frac{1}{na_0}\left[\frac{1}{4}\mathsf{v}''(t) + \frac{1}{2}\mathsf{v}'(t)\frac{p'(t)}{p(t)} - \frac{3}{4}\frac{a_1}{a_0}\mathsf{v}'(t)\right] + o\left(\frac{1}{n}\right)$$

and

$$\int_{\Theta} \Delta^2(\theta,\tau)\pi_n^*(\mathrm{d}\theta \mid -\kappa_t) = \Delta^2(t,\tau) + \frac{1}{na_0}\left[\frac{1}{2}\frac{\partial}{\partial\theta}\Delta^2(\theta,\tau)\Big|_{\theta=t}\frac{p'(t)}{p(t)} + \frac{1}{2}\frac{\partial^2}{\partial\theta^2}\Delta^2(\theta,\tau)\Big|_{\theta=t} - \frac{3}{4}\frac{a_1}{a_0}\frac{\partial}{\partial\theta}\Delta^2(\theta,\tau)\Big|_{\theta=t}\right] + o\left(\frac{1}{n}\right)$$

completing the proof just by mere substitutions.

5.6. Theorem 2

The core of the proof hinges on the identity (45). Now, the asymptotic expansion of its left-hand side is provided by (34), while the analogous expansion for right-hand side follows from a combination of (40) with (41). It is now straightforward to notice that the validity of (43) and (44) entails (45). At this stage, the validity of (46) and (47) for $\hat{R}_{0,m}$ and $\hat{R}_{1,m}$ follows directly by substitution. As to the same identities for $R^*_{0,m}$ and $R^*_{1,m}$, the argument rests on the combination of (3) with (32), exploiting the fact that the additional term (33) is of $o(1/n)$-type. Thus, the asymptotic expansion of the left-hand side of (3) is given in terms of integrals with respect to ν of the sum of the two left-hand sides of (40) and (41), respectively. Resorting once again to (45), one gets the desired identities for $R^*_{0,m}$ and $R^*_{1,m}$ by substitution.

5.7. Theorem 3

The core is the proof of (10), with the same expressions (46) and (47) also for $\tilde{R}_{0,m}$ and $\tilde{R}_{1,m}$, respectively. As in the proof of Theorem 2, the left-hand side of (10) is analyzed by resorting to inequality (32), exploiting the fact that the ensuing additional term, similar to that in (33), is of $o(1/n)$-type. Now, the argument is very similar to that of the preceding proof, with the variant that now the expansion (35) is not optimized in τ, but it is just evaluated at $\tau = \hat{\tau}_n$. The conclusion reduces once again to a matter of substituting the expressions (48)–(50) into the two expansions (35) and (41).

6. Conclusions and Future Developments

This paper should be seen as a pioneering work in the field of predictive problems, whose main aim is to show how the practical construction of efficient estimators of random quantities (that depend on future and/or past observations) entails non-standard metrizations of the parameter space Θ. This is the essence of the compatibility Equations (43) and (44). Of course, all the lines of research proposed in this paper deserve much more attention, in order to produce new results of wider validity.

The first issue deals with the extension of the compatibility equations to higher dimensions, including the infinite dimension. For finite dimensions, this is only a technical fact. Indeed, the question relies on extending the asymptotic expansion given in Proposition 3 from dimension 1 to dimension $d > 1$. This is done in [39] as far as the Bayesian setting,

and in [53,54] for a general mathematical setting. See also [55] [Section 2.2]. For the infinite dimension, the mathematical literature is rather scant. Some interesting results on asymptotic expansions of Laplace type for separable Hilbert spaces with Gaussian measure are contained in [56]. Finally, the topic is still in its early stage as far as metric measure spaces (i.e., the full nonparametric setting) are concerned. See [57,58].

Another mathematical tool that proves to be critical to our study is the Wasserstein distance. As explained in specific monographs like [27,59], the Wasserstein distance has several connections with other fields of mathematical analysis, such as optimal transportation and the theory of PDEs. Actually, the achievement of some estimators within our theory (like the one in (55)) is tightly connected with some optimization issues in transport theory. In this respect, an interesting mathematical area to explore is represented by the theory of Wasserstein barycenters and the ensuing numerical algorithms. See [60]. Research on this is ongoing.

Then, all the extensions of de Finetti's Law of Large Numbers for the log-likelihood process, stated in Theorem 1, Proposition 1 and Lemma 1 in Section 2.2, are worth being reconsidered, independently of their use for the purposes of this paper. As to possible extensions, the first hint is concerned with the analysis of dominated, parametric non-regular models, as those considered in [61–63]. Here, in fact, we never used the properties of the MLE as the root of the gradient of the log-likelihood, so that the asymptotic results contained in the quoted works should be enough to extend our statements. Subsequently, it would be also very interesting to consider dominated models which are parametrized by infinite-dimensional objects, where typically the MLE does not exist. See, e.g., the recent book [64] for plenty of examples.

As to more statistical objectives, it would be interesting to further deepen the connection between our approach and some relevant achievements obtained within the empirical Bayes theory, such as those contained in [22,23,65–68]. See also the book [69] for plenty of applications. In particular, the discussion contained in Section 4.4 about the original Poisson-mixture setting considered by Herbert Robbins deserves more attention.

A very fertile area of the application of predictive inference is that of *species sampling problems*. The pioneering works on this topic can be identified with the works [66,67,70]. Nowadays, the Bayesian approach (especially of nonparametric type) has received much attention, and has produced noteworthy new results in this field. See [17,71–73] and also [55,74,75] for novel asymptotic results. Indeed, it would be interesting to investigate whether it is possible to derive, within the approach of this paper, both asymptotic results and new estimators, hopefully more competitive than the existing ones.

Another prolific field of application is that of *density estimation*, aimed at solving clustering and/or classification problems. See [76] for a Bayesian perspective. Here, there is an additional technical difficulty due to the fact that the parameter is an element of some infinite-dimensional manifold, so that the characterization of any metric on Θ will prove mathematically more complex.

A last mention is devoted to predictive problems with "compressed data". This kind of research comes directly from computer science, where the complexity of the observed data make the available sample essentially useless for statistical inference purposes. For this reason, many algorithms have been conceived to compress the information in order to make it useful in some sense. See, e.g., [77]. Here, the Bayesian approach is in its early stage (see [78]), and the results of this paper can provide a valuable contribution.

Funding: This research received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 817257.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: I wish to express my enormous gratitude and admiration to Eugenio Regazzini. He has represented for me a constant source of inspiration, transmitting enthusiasm and method for the development of my own research. This paper represents a small present for his 75-th birthday.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Cifarelli, D.M.; Dolera, E.; Regazzini, E. Note on "Frequentist Approximations to Bayesian prevision of exchangeable random elements" [Int. J. Approx. Reason. 78 (2016) 138–152]. *Int. J. Approx. Reason.* **2017**, *86*, 26–27. [CrossRef]
2. Cifarelli, D.M.; Dolera, E.; Regazzini, E. frequentist approximations to Bayesian prevision of exchangeable random elements. *Int. J. Approx. Reason.* **2016**, *78*, 138–152. [CrossRef]
3. Dolera, E. On an asymptotic property of posterior distributions. *Boll. Dell'Unione Mat. Ital.* **2013**, *6*, 741–748. (In Italian)
4. Dolera, E.; Regazzini, E. Uniform rates of the Glivenko–Cantelli convergence and their use in approximating Bayesian inferences. *Bernoulli* **2019**, *25*, 2982–3015. [CrossRef]
5. de Finetti, B. Bayesianism: Its unifying role for both the foundations and applications of statistics. *Int. Stat. Rev.* **1974**, *42*, 117–130. [CrossRef]
6. de Finetti, B. La prévision: Ses lois logiques, ses sources subjectives. *Ann. L'Inst. Henri Poincaré* **1937**, *7*, 1–68.
7. Ferguson, T.S. *Mathematical Statistics: A Decision Theoretic Approach*; Academic Press: Cambridge, MA, USA, 1967.
8. Lehmann, E.L.; Casella, G. *Theory of Point Estimation*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 1998.
9. Aldous, D.J. *Exchangeability and Related Topics*; Ecole d'Eté de Probabilités de Saint-Flour XIII, Lecture Notes in Mathematics; Springer: Berlin/Heidelberg, Germany, 1985; pp. 1–198.
10. Berti, P.; Pratelli, L.; Rigo, P. Exchangeable sequences driven by an absolutely continuous random measure. *Ann. Probab.* **2013**, *78*, 138–152. [CrossRef]
11. Fortini, S.; Ladelli, L.; Regazzini, E. Exchangeability, predictive distributions and parametric models. *Sankhya* **2000**, *62*, 86–109.
12. Rubin, D.B. Bayesianly justifiable and relevant frequency calculations for the applied statisticians. *Ann. Stat.* **1984**, *12*, 1151–1172. [CrossRef]
13. Lijoi, A.; Prünster, I. Models beyond the Dirichlet process. In *Bayesian Nonparametrics*; Hjort, N.L., Holmes, C.C., Müller, P., Walker, S.G., Eds.; Cambridge University Press: Cambridge, UK, 2010; pp. 80–136.
14. Robbins, H. The empirical Bayes approach to statistical decision problems. *Ann. Math. Stat.* **1964**, *35*, 1–20. [CrossRef]
15. Ghosh, J.K.; Sinha, B.K.; Joshi, S.N. Expansions for posterior probability and integrated Bayes risk. In *Statistical Decision Theory and Related Topics III*; Gupta, S., Berger, J., Eds.; Academic Press: Cambridge, MA, USA, 1982; pp. 403–456.
16. Favaro, S.; Nipoti, B.; Teh, Y.W. Rediscovery of Good-Turing estimators via Bayesian nonparametrics. *Biometrics* **2016**, *72*, 136–45. [CrossRef] [PubMed]
17. Lijoi, A.; Mena, R.H.; Prünster, I. Bayesian Nonparametric Estimation of the Probability of Discovering New Species. *Biometrika* **2009**, *94*, 769–786. [CrossRef]
18. de Finetti, B. Probabilità di una teoria e probabilità dei fatti. In *Studi di Probabilità, Statistica e Ricerca Operativa in onore di Giuseppe Pompilj*; Oderisi: Gubbio, Italy, 1971; pp. 86–101. (In Italian)
19. Rao, R.C. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **1945**, *37*, 81–91.
20. Amari, S.-I. *Information Geometry and Its Applications*; Applied Mathematical Sciences; Springer: Berlin/Heidelberg, Germany, 2016; Volume 194.
21. Oller, J.M.; Corcuera, J.M. Intrinsic analysis of statistical estimation. *Ann. Stat.* **1995**, *23*, 1562–1581. [CrossRef]
22. Zhang, C.-H. Estimation of sums of random variables: Example and information bounds. *Ann. Stat.* **2005**, *33*, 2022–2041. [CrossRef]
23. Robbins, H. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*; Statistical Laboratory of the University of California: Davis Davis, CA, USA, 1956; Volume I, pp. 157–163.
24. Berezin, S.; Miftakhov, A. On barycenters of probability measures. *Bull. Pol. Acad. Sci. Math.* **2020**, *68*, 11–20. [CrossRef]
25. Karcher, H. Riemannian center of mass and mollifier smoothing. *Commun. Pure Appl. Math.* **1977**, *30*, 509–541. [CrossRef]
26. Kim, Y.-H.; Pass, B. Nonpositive curvature, the variance functional, and the Wasserstein barycenter. *Proc. Am. Math. Soc.* **2000**, *148*, 1745–1756. [CrossRef]
27. Ambrosio, L.; Gigli, N.; Savaré, G. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, 2nd ed.; Birkhäuser: Basel, Switzerland, 2008.
28. Billingsley, P. *Probability and Measure*, 3rd ed.; John Wiley & Sons: Hoboken, NJ, USA, 1995.
29. do Carmo, M.P. *Riemannian Geomerty*; Birkhäuser: Basel, Switzerland, 2013.
30. Heinonen, J.; Kilpeläinen, T.; Martio, O. *Nonlinear Potential Theory of Degenerate Elliptic Equations*; Oxford Science Publications: Oxford, UK, 2008.
31. Kufner, A. *Weighted Sobolev Spaces*; John Wiley & Sons: Hoboken, NJ, USA, 1985.
32. de Finetti, B. La legge dei grandi numeri nel caso dei numeri aleatori equivalenti. *Rend. Della R. Accad. Naz. Lincei* **1933**, *18*, 203–207. (In Italian)

33. Bauschke, H.H.; Combettes, P.L. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2017.
34. Borwein, J.M.; Noll, D. Second order differentiability of convex functions in Banach spaces. *Trans. Am. Math. Soc.* **1994**, *132*, 43–81. [CrossRef]
35. Dolera, E.; Favaro, S. Rates of convergence in de Finetti's representation theorem, and Hausdorff moment problem. *Bernoulli* **2020**, *26*, 1294–1322. [CrossRef]
36. Mijoule, G.; Peccati, G.; Swan, Y. On the rate of convergence in de Finetti's representation theorem. *Lat. Am. J. Probab. Math. Stat.* **2016**, *13*, 1–23. [CrossRef]
37. Dolera, E. Estimates of the approximation of weighted sums of conditionally independent random variables by the normal law. *J. Inequal. Appl.* **2013**, *2013*, 320. [CrossRef]
38. Götze, F. On the rate of convergence in the central limit theorem in Banach Spaces. *Ann. Probab.* **1986**, *14*, 922–942. [CrossRef]
39. Tierney, L.; Kadane, J.B. Accurate approximations for posterior moments and marginal densities. *J. Am. Stat. Assoc.* **1986**, *81*, 82–86. [CrossRef]
40. DasGupta, A. *Asymptotic Theory of Statistics and Probability*; Springer: Berlin/Heidelberg, Germany, 2008.
41. Dall'Aglio, G. Sugli estremi dei momenti delle funzioni di ripartizione doppia. *Ann. Della Sc. Norm. Super. Pisa Cl. Sci.* **1956**, *10*, 35–74. (In Italian)
42. Dolera, E.; Mainini, E. On Uniform Continuity of Posterior Distributions. *Stat. Probab. Lett.* **2020**, *157*, 108627. [CrossRef]
43. Dolera, E.; Mainini, E. Lipschitz continuity of probability kernels in the optimal transport framework. *arXiv* **2020**, arXiv:2010.08380.
44. Dowson, D.C.; Landau, B.V. The Fréchet distance between multivariate normal distributions. *J. Multivar. Anal.* **1982**, *12*, 450–455. [CrossRef]
45. Olkin, I.; Pukelsheim, F. The distance between two random vectors with given dispersion matrices. *Linear Algebra Its Appl.* **1982**, *48*, 257–263. [CrossRef]
46. Malagó, L.; Montrucchio, L.; Pistone, G. Wasserstein Riemannian geometry of positive definite matrices. *Inf. Geom.* **2018**, *1*, 137–179. [CrossRef]
47. Efron, B.; Hastie, T. *Computer Age Statistical Inference. Algorithms, Evidence, and Data Science*; Cambridge University Press: Cambridge, UK, 2016.
48. Thyrion, P. Contribution à l'étude du bonus pour non sinistre en assurance automobile. *ASTIN Bull. J. IAA* **1960**, *1*, 142–162. (In French) [CrossRef]
49. van Houwelingen, J.C. Monotonizing empirical Bayes estimators for a class of discrete distributions with monotone likelihood ratio. *Stat. Neerl.* **1977**, *31*, 95–104. [CrossRef]
50. Carlin, B.P.; Louis, T.A. *Bayesian Methods for Data Analysis*, 3rd ed.; Chapman and Hall: Boca Raton, FL, USA, 2009.
51. Ledoux, M.; Talagr, M. *Probability in Banach Spaces*; Springer: Berlin/Heidelberg, Germany, 1991.
52. Wong, R. *Asymptotic Approximations of Integrals*; SIAM: Philadelphia, PA, USA, 2001.
53. McClure, J.P.; Wong, R. Error bounds for multidimensional Laplace approximation. *J. Approx. Theory* **1983**, *37*, 372–390. [CrossRef]
54. Olver, F.W.J. Error bounds for the Laplace approximation for definite integrals. *J. Approx. Theory* **1968**, *1*, 293–313. [CrossRef]
55. Dolera, E.; Favaro, S. A Berry–Esseen theorem for Pitman's α–diversity. *Ann. Appl. Probab.* **2020**, *30*, 847–869. [CrossRef]
56. Albeverio, S.; Steblovskaya, V. Asymptotics of infinite-dimensional integrals with respect to smooth measures. (I). *Infin. Dimens. Anal. Quantum Probab. Relat. Top.* **1999**, *2*, 529–556. [CrossRef]
57. Gigli, N. Second order analysis on $(\mathscr{P}_2(M), W_2)$. *Mem. Am. Math. Soc.* **2012**, *216*, xii+154.
58. Gigli, N.; Ohta, S.I. First variation formula in Wasserstein spaces over compact Alexandrov spaces. *Can. Math. Bull.* **2010**, *55*, 723–735. [CrossRef]
59. Villani, C. *Optimal Transport. Old and New*; Springer: Berlin/Heidelberg, Germany, 2009.
60. Cuturi, M.; Doucet, A. Fast Computation of Wasserstein Barycenters. In Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 21–26 June 2014; Volume 32, pp. 685–693.
61. Smith, R.L. Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **1985**, *72*, 67–90. [CrossRef]
62. Woodroofe, M. Maximum likelihood estimation of a translation parameter of a truncated distribution. *Ann. Math. Stat.* **1972**, *43*, 113–122. [CrossRef]
63. Woodroofe, M. Maximum likelihood estimation of a translation parameter of a truncated distribution (II). *Ann. Stat.* 1974, *2*, 474–488. [CrossRef]
64. Giné, E.; Nickl, R. *Mathematical Foundations of Infinite-Dimensional Statistical Models*; Cambridge Series in Statistical and Probabilistic Mathematics: Cambridge, UK, 2016.
65. Efron, B.; Thisted, R. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **1976**, *63*, 435–447. [CrossRef]
66. Good, I.J. The population frequencies of species and the estimation of population parameters. *Biometrika* **1953**, *40*, 237–264. [CrossRef]
67. Good, I.J.; Toulmin, G.H. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **1956**, *43*, 45–63. [CrossRef]
68. Orlitsky, A.; Suresh, A.T.; Wu, Y. Optimal prediction of the number of unseen species. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 13283–13288. [CrossRef]

69. Maritz, J.S.; Lwin, T. *Empirical Bayes Methods with Applications*; Chapman and Hall: Boca Raton, FL, USA, 1989
70. Fisher, R.A.; Corbet, A.S.; Williams, C.B. The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **1943**, *12*, 42–58. [CrossRef]
71. Favaro, S.; Lijoi, A.; Mena, R.H.; Prünster, I. Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. *J. Roy. Statist. Soc. Ser. B* 2009, *71*, 993–1008. [CrossRef]
72. Favaro, S.; Lijoi, A.; Prünster, I. A new estimator of the discovery probability. *Biometrics* **2012**, *68*, 1188–1196. [CrossRef]
73. Arbel, J.; Favaro, S.; Nipoti, B.; Teh, Y.W. Bayesian nonparametric inference for discovery probabilities: Credible intervals and large sample asymptotic. *Stat. Sin.* **2017**, *27*, 839–858. [CrossRef]
74. Dolera, E.; Favaro, S. A compound Poisson perspective of Ewens–Pitman sampling model. *Mathematics* **2021**, *9*, 2820. [CrossRef]
75. Pitman, J. *Combinatorial Stochastic Processes*; Ecole d'Eté de Probabilités de Saint-Flour XXXII, Lecture Notes in Mathematics; Springer: Berlin/Heidelberg, Germany, 2006.
76. Sambasivan, R.; Das, S.; Sahu, S.K. A Bayesian perspective of statistical machine learning for big data. *Comput. Stat.* **2020**, *35*, 893–930. [CrossRef]
77. Cormode, G.; Yi, K. *Small Summaries for Big Data*; Cambridge University Press: Cambridge, UK, 2020
78. Dolera, E.; Favaro, S.; Peluchetti, S. Learning-augmented count-min sketches via Bayesian nonparametrics. *arXiv* **2021** arXiv:2102.04462.