An asymptotic analysis of a class of discrete nonparametric priors

Pierpaolo De Blasi
University of Torino and Collegio Carlo Alberto, Italy
Antonio Lijoi
University of Pavia and Collegio Carlo Alberto, Italy
Igor Prünster
University of Torino and Collegio Carlo Alberto, Italy

Abstract

In this paper we analyze the asymptotic behaviour of a large class of nonparametric priors, namely Gibbs-type priors, which represent a natural generalization of the Dirichlet process. After determining their topological support, we specifically investigate consistency of such priors according to the "what if", or frequentist, approach, which postulates the existence of a "true" distribution P_0. We provide a full taxonomy of their limiting behaviours: consistency holds essentially always for discrete P_0, whereas inconsistency may occur for diffuse P_0. Such findings are further illustrated by means of three specific priors admitting closed form expressions and exhibiting a wide range of asymptotic behaviours. For both Gibbs-type priors and discrete nonparametric priors in general, the possible inconsistency should not be interpreted as evidence against their use *tout court*. It rather represents an indication that they are designed for modeling discrete distributions, at which consistency holds true, and a neat evidence against their use in the case of diffuse P_0.

Key words and phrases: Asymptotics, Bayesian consistency, Bayesian nonparametrics, Gibbs–type priors, Foundations, Species sampling.

1 Introduction

In this paper we study the posterior consistency of Gibbs–type priors recently introduced in Gnedin and Pitman (2005). They identify a large class of *discrete* nonparametric priors, which means they select almost surely (a.s.) discrete distributions, and represent probably the most natural generalization of the Dirichlet process as will be argued in Section 2. Several members of this class of nonparametric priors are widely used in practice, for instance, in the contexts of mixture models (Ishwaran and James, 2001; Ishwaran and James, 2003; Lijoi, Mena and Prünster, 2007c), linguistics and information retrieval in document modeling (Teh, 2006; Teh and Jordan, 2010), species sampling (Lijoi, Mena and Prünster, 2007a,b; Navarrete, Quintana and Müller, 2008)

and, implicitly, in the context of exchangeable product partition models (Hartigan, 1990; Quintana and Iglesias, 2003).

A simple way to introduce Gibbs-type priors is through the system of predictive distributions they induce. To this end, we first lay out the basic framework. Let $(X_n)_{n\geq 1}$ be an (ideally) infinite sequence of observations, with each X_i taking values in a complete and separable metric space $\mathbb{X}$. Moreover, $\mathbf{P}_{\mathbb{X}}$ is the set of all probability measures on $\mathbb{X}$ endowed with the topology of weak convergence. In the most commonly employed Bayesian models $(X_n)_{n\geq 1}$ is assumed to be *exchangeable* which means there exists a probability distribution Q on $\mathbf{P}_{\mathbb{X}}$ such that

$$X_i|\tilde{p} \overset{\text{iid}}{\sim} \tilde{p}, \quad \tilde{p} \sim Q \tag{1}$$

Hence, $\tilde{p}$ is a random probability measure on $\mathbb{X}$ whose probability distribution Q is also termed *de Finetti measure* and acts as a prior for Bayesian inference. Whenever Q degenerates on a finite dimensional subspace of $\mathbf{P}_{\mathbb{X}}$, the inferential problem is usually called *parametric*. On the other hand, when the support of Q is infinite-dimensional then one typically speaks of a *nonparametric* inferential problem and it is generally agreed (Ferguson, 1974) that having a large topological support is a desirable property for a nonparametric prior: we will come back to this point later in Section 2. Given a sample $(X_1, \ldots, X_n)$, the predictive distribution coincides with the posterior expected value of $\tilde{p}$, that is

$$\mathrm{P}(X_{n+1} \in \cdot\,|X_1, \ldots, X_n) = \int_{\mathbf{P}_{\mathbb{X}}} p(\cdot)\, Q(\mathrm{d}p|X_1, \ldots, X_n). \tag{2}$$

As mentioned above, we will deal with discrete priors Q, which implies that a sample $(X_1, \ldots, X_n)$ will feature ties with positive probability: $X_1^*, \ldots, X_k^*$ denote the $k \leq n$ distinct observations and $n_1, \ldots, n_k$ their frequencies for which $\sum_{i=1}^k n_i = n$. Gibbs-type priors are characterized by predictive distributions (2) of the form

$$\mathrm{P}(X_{n+1} \in \cdot\,|X_1, \ldots, X_n) = \frac{V_{n+1,k+1}}{V_{n,k}} P^*(\cdot) + \frac{V_{n+1,k}}{V_{n,k}} \sum_{i=1}^{k} (n_i - \sigma)\delta_{X_i^*}(\cdot), \tag{3}$$

where $\sigma \in (-\infty, 1)$, $P^*(\mathrm{d}x) := \mathrm{E}[\tilde{p}(\mathrm{d}x)]$ is a diffuse probability measure representing the prior guess at the shape of $\tilde{p}$ and $\{V_{n,k} : k = 1, \ldots, n;\, n \geq 1\}$ is a set of non-negative weights satisfying the recursion

$$V_{n,k} = (n - \sigma k)V_{n+1,k} + V_{n+1,k+1}. \tag{4}$$

Therefore, Gibbs-type priors are characterized by predictive distributions, which are a linear combination of the prior guess and a weighted version of the empirical measure. The most widely known prior within this class is the Dirichlet process (Ferguson, 1973).

In this paper we will focus on the asymptotic behaviour of Gibbs-type priors and, in particular, investigate posterior consistency according to the "what if" approach of Diaconis and Freedman (1986). Such an approach consists in assuming that the data $(X_n)_{n\geq 1}$ are independent and identically distributed from some "true" $P_0 \in \mathbf{P}_X$ and in verifying whether the posterior distribution $Q(\cdot\,|X_1, \ldots, X_n)$ accumulates in any

neighborhood of P_0, under a suitable topology. Since Gibbs-type priors are defined on $\mathbf{P}_{\mathbb{X}}$ and are discrete, the appropriate notion of convergence is convergence in the weak topology. Therefore, we aim at establishing whether $Q(A_\epsilon|X_1,\ldots,X_n) \to 1$, a.s.-$P_0^\infty$, as $n \to \infty$ and for any $\epsilon > 0$, where A_ϵ denotes a weak neighborhood of P_0 of radius ε and P_0^∞ is the infinite product measure $P_0 \times P_0 \times \cdots$. In pursuing this plan we first show that "genuinely nonparametric" Gibbs-type priors (a notion that will be clarified in Section 2) have full weak support. We then prove a general structural result on Gibbs-type priors showing that the posterior distribution converges to a point mass at the limiting predictive distribution

$$\alpha\, P^* + (1-\alpha)P_0 \qquad \alpha \in [0,1] \tag{5}$$

which is a linear combination of the prior guess P^* and the "true" distribution P_0. This points out that Gibbs-type priors are well-behaved in the limit in the sense of convergence taking place rather than implying consistency. As for the latter to happen, one needs $\alpha = 0$ in (5), a feature clearly satisfied in the Dirichlet case. Since a few particular cases of Gibbs-type priors with $\sigma \in (0,1)$ have already been considered in Jang, Lee and Lee (2010) and James (2008), attention is focused on the case of $\sigma \in (-\infty, 0)$ for which nothing is known to date and which yield competitive estimators for species estimation in Ecology (Favaro, Lijoi, Mena and Prünster, 2012). A full taxonomy of the asymptotic behaviours is provided. In fact, in deriving the results it is fundamental to distinguish the cases of P_0 discrete and diffuse: in the former case one essentially always has consistency, whereas in the latter we provide a sufficient condition for consistency, which has the merit of being close to necessary. This is shown by exhibiting specific priors, which, by a minimal violation of the sufficient condition, already lead to inconsistency. Moreover, we are able to provide two new and completely explicit priors exhibiting the two extreme limit behaviours, $\alpha = 0$ and $\alpha = 1$. In particular, the latter corresponds to the worst case scenario where the posterior tends to concentrate around the prior guess P^* and no learning at all takes place: we will refer to such a pathological situation as "total" inconsistency. A third specific prior yields the whole spectrum of $\alpha \in (0,1)$ and serves as interpretation of the two extreme cases.

The results of the present paper briefly outlined above are to be read at two distinct levels. The first immediate one is that we provide a comprehensive analysis of consistency properties of a large and intuitive class of nonparametric priors. This fills in a gap in the current rapidly growing literature on asymptotic properties of Bayesian nonparametric procedures. See Ghosal (2010) for a recent review. The relevance of the results we achieve is further witnessed by the increasing use of these priors in statistical practice. The second level at which the results of the present paper should be read concerns general foundational and methodological questions. In particular, by providing neat asymptotic results we highlight that discrete nonparametric priors are actually designed to model discrete distributions and that they should under no circumstance be used to model data coming from diffuse distributions. This is an important point since most Bayesian nonparametric approaches to survival analysis rely on discrete priors such as neutral to the right processes or cumulative hazards modeled as independent increments processes. The fact that discrete nonparametric priors typically have full

weak support, as the ones considered in the present paper, has often led to think that they could represent suitable models also for diffuse distributions. Consequently, the famous example of inconsistency due to Diaconis and Freedman (1986), involving the use of a Dirichlet process in a semiparametric location problem, was interpreted as an indication of the fact that one needs to be careful with Bayesian nonparametric models in general and more specifically with modeling diffuse data with the Dirichlet process. In our opinion, this essentially represented a misunderstanding: its reason probably lies in the fact that the Dirichlet process combines full weak support with consistency for independent and identically distributed data generated from a diffuse P_0, which is more of a coincidence than a structural property nonparametric priors should possess. It is simply wrong to use discrete priors in such contexts and we hope to be able to demonstrate such a claim by means of our explicit illustrations. In particular, we also exhibit a specific nonparametric prior which in the case of diffuse P_0 may produce either consistency or "total" inconsistency by simply tuning a scalar parameter. On the other hand, and this is not a coincidence, consistency is the rule for discrete data generating distributions P_0 or even for diffuse P_0 provided the Gibbs-type prior is used as mixing measure in a hierarchical model. The latter claim immediately follows from Ghosal, Ghosh and Ramamoorthi (1999); Lijoi, Prünster and Walker (2005).

The outline of the paper is as follows. In Section 2 Gibbs–type priors are concisely reviewed and their topological support is investigated. Section 3 contains the general results on the asymptotic behaviour whereas Section 4 illustrates the above mentioned specific priors which highlight the various possible asymptotic regimes. Some concluding remarks, concerning mainly the foundational implications, are provided in Section 5.

2 Gibbs-type priors and their topological support

As mentioned in the Introduction, modeling data according to a discrete prior Q implies that a sample $(X_1, \ldots, X_n)$ will feature ties with positive probability. Recall that $X_1^*, \ldots, X_k^*$ denotes the $k \leq n$ distinct observations and $n_1, \ldots, n_k$ their frequencies for which $\sum_{i=1}^k n_i = n$. In choosing a specific predictive structure the key quantity to consider is the probability of obtaining a new distinct observation, that is

$$\mathrm{P}(X_{n+1} = \text{"new"} \mid X_1, \ldots, X_n). \tag{6}$$

If $\boldsymbol{\Theta}$ is a finite-dimensional parameter entering the specification of $\tilde{p}$, there are three possibilities for modeling (6): (i) $\mathrm{P}(X_{n+1} = \text{"new"}|X_1, \ldots, X_n) = f(n, \boldsymbol{\Theta})$, which means that the probability of obtaining a new observation depends on the sample size n but not on the number of distinct observations k and on their frequencies $n_1, \ldots, n_k$; (ii) $\mathrm{P}(X_{n+1} = \text{"new"}|X_1, \ldots, X_n) = f(n, k, \boldsymbol{\Theta})$, which means that dependence is now on both n and k but not on the frequencies $n_1, \ldots, n_k$; (iii) $\mathrm{P}(X_{n+1} = \text{"new"}|X_1, \ldots, X_n) = f(n, k, n_1, \ldots, n_k, \boldsymbol{\Theta})$ which depends on all the sample information. As shown in Zabell (1982), (i) holds if and only if the prior is a Dirichlet process with parameter measure θP^*, which corresponds to $\mathrm{P}(X_{n+1} = \text{"new"}|X_1, \ldots, X_n) =$

$f(n, \mathbf{\Theta}) = \theta/(\theta + n)$. Case (ii) corresponds to Gibbs-type priors for which

$$\mathrm{P}(X_{n+1} = \text{“new”} \mid X_1, \ldots, X_n) = \frac{V_{n+1,k+1}}{V_{n,k}} \tag{7}$$

with the $V_{n,k}$'s satisfying (4). In the general situation (iii), which in principle would be the most desirable, serious tractability issues arise: priors have to be studied on a case-by-case basis and typically lead to quite complicated expressions (Favaro, Prünster and Walker, 2011). In light of the above considerations, the simplifying assumption underlying Gibbs–type priors seems to represent the right compromise between flexibility and tractability. In fact, it is only the probability of obtaining a new observation that does not depend on the frequencies and not the complete prediction rule (3). To clarify this point it is useful to interpret (3) by means of a two step procedure. The first step concerns the probability of X_{n+1} being “new” or “old”: X_{n+1} is new with probability $V_{n+1,k+1}/V_{n,k}$, whereas it coincides with one of the previously observed $X_1^*, \ldots, X_k^*$ with probability $1 - V_{n+1,k+1}/V_{n,k} = (n - k\sigma)V_{n+1,k}/V_{n,k}$; as mentioned above these probabilities do not depend on the frequencies $n_1, \ldots, n_k$. The second step is as follows: given X_{n+1} is new, it is sampled independently from P^*; given X_{n+1} is “old”, it coincides with X_i^* with probability $(n_i - \sigma)/(n - k\sigma)$ for $i = 1, \ldots, k$, which depends explicitly on $n_1, \ldots, n_k$. Moreover, when compared to the Dirichlet process, the Gibbs–type framework leads to apparent advantages in species sampling problems (Lijoi, Mena and Prünster, 2007a,b) and also to more robust estimates of the number of components in mixture models (Lijoi, Mena and Prünster, 2007c). As for species sampling, it is enough to think of having two samples of size $n = 10$ featuring, respectively, $k^{'} = 1$ and $k^{''} = n$ distinct species: with the Dirichlet process the probability of observing a new species is $\theta(\theta + n)^{-1}$ in both situations, whereas it explicitly, and meaningfully, depends on k for other Gibbs-type priors. For instance, if one uses the two-parameter Poisson-Dirichlet process (Pitman, 1996), a notable member of the family of Gibbs-type priors, one has a wide spectrum of modeling possibilities for the specific application at hand: indeed, one has that

$$\mathrm{P}(X_{n+1} = \text{“new”} \mid X_1, \ldots, X_n) = \frac{\theta + k\sigma}{\theta + n}, \tag{8}$$

where the possible value of the parameters (σ, θ) are $\sigma \in [0, 1)$ and $\theta > -\sigma$ or $\sigma \in (-\infty, 0)$ and $\theta = x|\sigma|$ for some $x \in \mathbb{N}$; therefore, (8) is monotonically increasing in k for $\sigma \in (0, 1)$ and monotonically decreasing in k for $\sigma < 0$.

In addition, to the predictive structure which completely characterizes Gibbs-type priors, it is also worth to recall some features of the underlying de Finetti measure Q whose posterior expected value yields the predictive distributions (2). Gibbs-type priors are species sampling models (Pitman, 1996) and therefore they can be seen as laws of random probability measures representable as

$$\tilde{p}(\,\cdot\,) = \sum_{i \geq 1} \tilde{p}_i \delta_{Y_i}(\,\cdot\,), \tag{9}$$

where the weights $(\tilde{p}_n)_{n \geq 1}$ take value on the infinite probability simplex, while the (Y_i)'s are independent and identically distributed from a diffuse P^* and are independent from

the p_i's. Clearly, $\mathrm{E}[\tilde{p}(\,\cdot\,)] = P^*(\,\cdot\,)$ which explains the terminology *prior guess* adopted for P^*. Such a framework allows to give an alternative definition of Gibbs-type priors, which coincides with the original one in Gnedin and Pitman (2005): Gibbs-type priors are species sampling models (9) for which the probability of obtaining in a n-size sample k distinct observations with frequencies $n_1, \ldots, n_k$ has, for any $n \geq 1$, product form

$$V_{n,k} \prod_{i=1}^{k} (1-\sigma)_{n_i-1}, \tag{10}$$

with $\sigma \in (-\infty, 1)$, the $V_{n,k}$'s satisfying (4) and $(a)_m$ denoting the rising factorial $(a)_m = a(a+1)\cdots(a+m-1)$. Such a distribution, which in fact corresponds to a distribution of the partition of the positive integers $\mathbb{N}$ induced by an exchangeable sequence, is known as exchangeable partition probability function. This concept was introduced by J. Pitman and plays a major role in modern probability theory. See Pitman (2006) and references therein. The above mentioned special case of the two-parameter Poisson-Dirichlet process corresponds to

$$V_{n,k} = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta+1)_{n-1}}, \tag{11}$$

with, as before, $(\sigma \in [0,1); \theta > -\sigma)$ or $(\sigma \in (-\infty, 0); \theta = x|\sigma|; x \in \mathbb{N})$. From (11) one immediately obtains (8) via (7). For our purposes it is to be noted that the two-parameter Poisson-Dirichlet model with $(\sigma \in (-\infty, 0); \theta = x|\sigma|; x \in \mathbb{N})$ corresponds to an x-variate symmetric Dirichlet distribution with parameter vector $(|\sigma|, \ldots, |\sigma|)$.

In Gnedin and Pitman (2005) a complete characterization of the underlying de Finetti measure Q is also provided and distinguishes three cases according to the value of σ: (i) if $\sigma = 0$, $\tilde{p}$ is either a Dirichlet process or a mixture of Dirichlet processes w.r.t. the total mass parameter θ; (ii) if $\sigma \in (0,1)$, then Q is essentially a Poisson-Kingman model based on the stable random measure, whose description goes beyond the scope of the present paper and we refer the reader to Pitman (2006) and references therein; (iii) if $\sigma < 0$, Q is a mixture of the corresponding two-parameter model (11) with $(\sigma \in (-\infty, 0); \theta = |\sigma| x; x \in \mathbb{N})$, that is

$$V_{n,k} = \sum_{x \geq k} \frac{\prod_{i=1}^{k-1}(x|\sigma| + i\sigma)}{(x|\sigma|+1)_{n-1}} \pi(x), \tag{12}$$

where π is a probability measure on $\mathbb{N}$ and the sum obviously runs over $x \geq k$ since the numerator in the summands corresponding to $x < k$ is 0. Therefore, since in the case of negative σ the two-parameter model coincides with a x-variate symmetric Dirichlet distribution, one can also describe such Gibbs-type priors in terms of a mixture model

$$\begin{array}{rcl} (\tilde{p}_1, \ldots, \tilde{p}_k) & \sim & \text{Dirichlet}(|\sigma|, \ldots, |\sigma|) \\ k & \sim & \pi(\cdot) \end{array} \tag{13}$$

Using the species metaphor, one can describe (12) or equivalently (13) as putting a prior π on the number of species k and, conditionally on the number of species being

x, these are distributed as a x-variate symmetric Dirichlet distribution. In contrast to the case of $\sigma \geq 0$ where the model assumes the existence of an infinite number of species, the case of $\sigma < 0$ assumes a possibly random but finite number of species. Therefore, in light of the previous considerations, one deduces that if the probability of observing a new species is assumed to depend on n and k but not on $n_1, \ldots, n_k$ and moreover the a priori number of species is assumed to be finite (either random or not random), then the model is necessarily (13).

Before proceeding to studying the support properties of Gibbs-type priors, let us restrict attention to "genuinely nonparametric" Gibbs-type priors: specifically, we will henceforth consider Gibbs-type priors whose realizations are discrete distributions whose support contains a finite number of points that can be equal to any positive integer. This is the same as saying that we concentrate on Gibbs-type priors with either $\sigma \in [0, 1)$ or $\sigma < 0$ such that the support of π in (12) is the whole set of positive integers $\mathbb{N}$. Note that for the "parametric" case of $\sigma < 0$ and π supported by a finite subset of $\mathbb{N}$ one immediately has consistency for any P_0 in its support by the results of Freedman (1963).

Now we move on to considering the topological support of Gibbs-type priors. It is widely accepted (Ferguson, 1974) that nonparametric priors should have a large topological support. Since we are dealing with a class of discrete nonparametric priors this requirement translates in asking Q to have large support in the weak topology. In fact, the next result shows that Gibbs-type priors have full weak support, that is their topological support coincides with the space of probability measures whose support is included in the support of the prior guess P^*. In particular, if the support of P^* coincides with $\mathbb{X}$, the support of Q is the whole space $\mathbf{P}_{\mathbb{X}}$. Such a property is already known in the Dirichlet process case (Ferguson, 1973; Majumdar, 1992) and has been recently extended to a class of predictor-dependent nonparametric priors, known as dependent Dirichlet processes in Barrientos, Jara and Quintana (2011).

Proposition 1. *Let Q be a Gibbs-type prior with prior guess P^* and, in the case $\sigma < 0$, mixing measure π such that $\pi(x) > 0$ for any $x \in \mathbb{N}$. Then the topological support of Q coincides with*

$$\{p \in \mathbf{P}_{\mathbb{X}} : \ \text{supp}(p) \subset \text{supp}(P^*)\}.$$

The proof is deferred to the Appendix. The property stated in Proposition 1 makes Gibbs-type priors even more appealing for modeling purposes. When used to model directly the data in species sampling contexts, it ensures that weak neighborhoods of any given distribution (whose support is included in the support of the prior guess P^*) have a priori positive probability. It is also a desirable property in the context of mixture models where $\tilde{p}$ acts as a mixing distribution: indeed, it ensures a high degree of flexibility of the model for any given kernel and has relevant implications in terms of consistency since one can extend results known for Dirichlet mixtures (Ghosal, Ghosh and Ramamoorthi, 1999; Lijoi, Prünster and Walker, 2005).

3 Posterior consistency of Gibbs–type priors

Having provided a concise account of Gibbs-type priors and a result concerning their support, we now move on to studying their asymptotic behaviour. For brevity, in the sequel we use the notation Q_n for denoting the posterior distribution $Q(\cdot \,|\, X_1, \ldots, X_n)$ of the random probability measure $\tilde{p}$ in (1), conditional on the sample $X_1, \ldots, X_n$. Assuming the data are independent and identically distributed from some "true" distribution P_0 in $\mathbf{P}_{\mathbb{X}}$, we are interested in checking whether Q_n tends to concentrate, as the sample size n increases, in a weak neighbourhood of some element, say P', in $\mathbf{P}_{\mathbb{X}}$, almost surely with respect to the infinite product measure P_0^∞. If A'_ε is a weak neighbourhood of P' with radius $\varepsilon > 0$, we intend to establish conditions under which

$$Q_n(A'_\varepsilon) \;\to\; 1 \qquad\qquad \text{a.s.-}P_0^\infty \tag{14}$$

as $n \to \infty$ and for any $\varepsilon > 0$. More importantly, we would like to identify cases when $P' = P_0$, which corresponds to Q being *weakly consistent* in the frequentist sense.

Weak consistency of the Dirichlet process prior is quite straightforward to prove by investigating the asymptotic behaviour of the corresponding posterior expected value (i.e. the predictive distributions (2)) and the posterior expected variance. Given the Dirichlet process prior is a special case of Gibbs-type prior, we adopt a similar strategy in this more general framework. Since the predictive distributions in (3) characterize Gibbs-type priors, it is apparent that the validity of (14) will depend on the limiting behaviour of the weights V_{n,κ_n}. We shall use the notation κ_n, in which the dependence on n is made explicit, to denote the number of blocks in the partition of the first n observations, that is $\kappa_n := 1 + \sum_{j=2}^n \mathbb{1}_{D_{j-1}}(X_j)$ with $D_{j-1} = \{X_1, \ldots, X_{j-1}\}^c$. Given the asymptotics of κ_n with respect P_0^∞ is considered, different choices of P_0 yield different limiting behaviours for κ_n. On the one hand, if P_0 is discrete with N point masses, for any $N \in \mathbb{N} \cup \{\infty\}$, then $P_0^\infty(\lim_n \kappa_n = N) = 1$ and $P_0^\infty(\lim_n n^{-1}\kappa_n = 0) = 1$ even if $N = \infty$. On the other hand, if P_0 is diffuse, $P_0^\infty(\kappa_n = n) = 1$ for any $n \geq 1$. Henceforth we shall focus on these two cases and adopt the shorter notation $\kappa_n \ll_{a.s.} n$ and $\kappa_n \sim_{a.s.} n$, which stand for $\kappa_n/n \to 0$ and $\kappa_n/n \to 1$ a.s.-P_0^∞, respectively. See Remark 2 below for a discussion of the case where P_0 is a combination of a discrete and a diffuse component.

Based on the above intuition, in order to establish the validity of (14), for some P', one needs to investigate the asymptotics for $V_{n+1,\kappa_n+1}/V_{n,\kappa_n}$ under P_0^∞. Indeed, in what follows we shall assume that the probability of recording a new distinct observation at step $n+1$

$$\frac{V_{n+1,\kappa_n+1}}{V_{n,\kappa_n}} \qquad \text{converges} \qquad \text{a.s.-}P_0^\infty \tag{H}$$

as $n \to \infty$, and the limit is identified by some constant $\alpha \in [0, 1]$. For all Gibbs-type priors for which an explicit expression of the V_{n,κ_n}'s is known, (H) holds true regardless as to whether P_0 is discrete or diffuse. Hence, assuming (H) does not significantly restrict the generality of our results given it serves only to exclude pathological behaviours. The role of condition (H) is also transparent: since it determines the asymptotics of the predictive distribution, it also identifies the possible element P'

in $\mathbf{P}_{\mathbb{X}}$ for which (14) holds true. The following theorem shows that (H) is actually sufficient to establish weak convergence at such P'.

Theorem 1. *Let $\tilde{p}$ be a Gibbs-type prior with prior guess $P^* = \mathrm{E}[\tilde{p}]$, whose support coincides with $\mathbb{X}$, and assume condition (H) holds true. Moreover $(X_i)_{i\geq 1}$ is a sequence of independent and identically distributed random elements from some probability distribution P_0 which is either discrete or diffuse. Then the posterior converges weakly, a.s.-P_0^∞, to a point mass at $\alpha P^*(\cdot) + (1-\alpha)P_0(\cdot)$.*

According to Theorem 1, achievement of weak consistency is guaranteed in the trivial case of $P^* = P_0$, which will be excluded henceforth, and when $\alpha = 0$: therefore, it is sufficient to check whether the probability of obtaining a new observation, given previously recorded data, converges to 0, a.s.-P_0^∞. One might also wonder whether there are circumstances leading to $\alpha = 1$, which corresponds to the posterior concentrating around the prior guess P^*, a situation we refer to as "total" inconsistency. A specific prior exhibiting such a pathological behaviour is provided in Section 4. Note also that Theorem 1 includes as a special case Proposition 1 of James (2008) which is confined to $\tilde{p}$ being a two-parameter Poisson-Dirichlet process with parameters (σ, θ) such that $\sigma \in [0, 1)$ and $\theta > -\sigma$. In fact, in the two-parameter Poisson-Dirichlet case, it is immediate to see from (8) that when P_0 is discrete ($\kappa_n \ll_{a.s.} n$) we have $\alpha = 0$, implying consistency. When P_0 is diffuse ($\kappa_n \sim_{a.s.} n$), we have $\alpha = \sigma$, hence inconsistency, unless $\sigma = 0$, which corresponds to the Dirichlet case. See also Jang, Lee and Lee (2010, Theorem 1). Let us now provide a proof of the stated result. The key ingredient is represented by an upper bound on the posterior variance $\mathrm{Var}[\tilde{p}(A) \,|\, \boldsymbol{X}_{\kappa_n}^{(n)}]$, which is of independent interest and is discussed in some detail in Remark 1 below.

Proof of Theorem 1. The strategy we are going to use in the proof amounts to showing that, under (H), the posterior variance of $\tilde{p}(A)$, given a sample $\boldsymbol{X}_{\kappa_n}^{(n)} = (X_1, \ldots, X_n)$ featuring $\kappa_n \leq n$ distinct values, converges to 0, a.s.-P_0^∞. To this end, we shall deduce an upper bound for $\mathrm{Var}[\tilde{p}(A) \,|\, \boldsymbol{X}_{\kappa_n}^{(n)}]$.

As in Freedman and Diaconis (1983), we consider the class of semi-norms on $\mathbf{P}_{\mathbb{X}}$ defined by $\|P_1 - P_2\|_{\mathcal{A}}^2 = \sum_{i=1}^\infty [P_1(A_i) - P_2(A_i)]^2$ for a generating sequence of measurable partitions $\mathcal{A} = \{A_i\}_{i=1}^\infty$ of $\mathbb{X}$. Indeed, convergence under such semi-norms implies weak convergence. Note that $\mathrm{E}\big[\| \tilde{p} - \mathrm{E}[\, \tilde{p} | \boldsymbol{X}_{\kappa_n}^{(n)}] \|_{\mathcal{A}}^2 \, | \boldsymbol{X}_{\kappa_n}^{(n)}\big] = \sum_{i=1}^\infty \mathrm{Var}[\tilde{p}(A_i) | \boldsymbol{X}_{\kappa_n}^{(n)}]$. Hence, we are going to show that

$$\sum_{i=1}^{\infty} \mathrm{Var}[\tilde{p}(A_i) | \boldsymbol{X}_{\kappa_n}^{(n)}] \to 0 \qquad \text{a.s.-}P_0^\infty$$

as $n \to \infty$ for any partition $\mathcal{A}$.

This would imply that the posterior concentrates in a weak-neighbourhood of the predictive distribution. See also James (2008) for a similar approach in the specific case of the two-parameter Poisson-Dirichlet process. To this end, let us first simplify the notation and set $g_{c,d}^{a,b}(n) = V_{n+a,\kappa_n+b}/V_{n+c,\kappa_n+d}$ with a, b, c and d non-negative integers such that $a \geq c$ and $b \geq d$. Exchangeability implies

$$E[\tilde{p}(A)^2 | \boldsymbol{X}_{\kappa_n}^{(n)}] = \int_A \mathrm{P}(X_{n+2} \in A \,|\, \boldsymbol{X}_{\kappa_n}^{(n)}, X_{n+1} = x) \mathrm{P}(X_{n+1} \in \mathrm{d}x | \boldsymbol{X}_{\kappa_n}^{(n)})$$

$$= g^{1,1}_{0,0}(n) \int_A \mathrm{P}(X_{n+2} \in A \,|\, \boldsymbol{X}^{(n)}_{\kappa_n}, X_{n+1} = x)\, P^*(\mathrm{d}x)$$

$$+ g^{1,0}_{0,0}(n) \sum_{j=1}^{\kappa_n} \delta_{X^*_j}(A)\,(n_j - \sigma)\,\mathrm{P}(X_{n+2} \in A \,|\, \boldsymbol{X}^{(n)}_{\kappa_n}, X_{n+1} = X^*_j)$$

for any $A \in \mathscr{X}$, where $X^*_1, \ldots, X^*_{\kappa_n}$ are the κ_n distinct values that partition $\boldsymbol{X}^{(n)}_{\kappa_n}$. After some tedious and lengthy algebra, one gets to

$$\mathrm{E}[\tilde{p}(A)^2 | \boldsymbol{X}^{(n)}_{\kappa_n}] = g^{2,0}_{0,0}(n) \sum_{i,j=1}^{k} (n_i - \sigma)(n_j + \delta_{i,j} - \sigma)\delta_{X^*_i}(A)\delta_{X^*_j}(A)$$

$$+ 2g^{2,1}_{0,0}(n) \sum_{i=1}^{\kappa_n} (n_i - \sigma)\delta_{X^*_i}(A)P^*(A) + g^{2,2}_{0,0}(n)\, P^*(A)^2 + g^{2,1}_{0,0}(n)(1-\sigma)P^*(A),$$

where $\delta_{i,j}$ is the Kronecker δ function and note that we have also relied on the diffuseness of P^*. Letting

$$\tilde{P}_{n,k} = \frac{1}{n - \kappa_n \sigma} \sum_{j=1}^{\kappa_n} (n_j - \sigma)\delta_{X^*_j} \tag{15}$$

denote a weighted empirical distribution at the distinct observations, one can use the above expression for the posterior second moment of $\tilde{p}(A)$ and obtain

$$\mathrm{Var}[\tilde{p}(A) | \boldsymbol{X}^{(n)}_{\kappa_n}] = \Big(g^{2,0}_{0,0}(n) - (g^{1,0}_{0,0}(n))^2 \Big)(n - \sigma\kappa_n)^2 \tilde{P}_{n,\kappa_n}(A)^2$$

$$+ g^{2,0}_{0,0}(n)(n - \sigma\kappa_n)\tilde{P}_{n,\kappa_n}(A)$$

$$+ 2\Big(g^{2,1}_{0,0}(n) - g^{1,0}_{0,0}(n) g^{1,1}_{0,0}(n) \Big)(n - \sigma\kappa_n)\tilde{P}_{n,\kappa_n}(A)P^*(A)$$

$$+ \Big(g^{2,2}_{0,0}(n) - (g^{1,1}_{0,0}(n))^2 \Big) P^*(A)^2 + g^{2,1}_{0,0}(n)(1-\sigma)P^*(A).$$

This can be re-expressed in a more convenient form in terms of the quantity

$$I(n, \kappa_n) := 1 - \frac{V_{n+2,\kappa_n+1}}{V_{n+1,\kappa_n+1}} \frac{V_{n,\kappa_n}}{V_{n+1,\kappa_n}}. \tag{16}$$

Indeed, one has

$$\mathrm{Var}[\tilde{p}(A) | \boldsymbol{X}^{(n)}_{\kappa_n}] = -I(n, \kappa_n) \Big(\mathrm{E}[\tilde{p}(A) | \boldsymbol{X}^{(n)}_{\kappa_n}] \Big)^2 + W_{n,\kappa_n}(A),$$

where, using the identities (A1) and (A2) of Lemma 1 in Appendix,

$$W_{n,\kappa_n}(A) = g^{2,1}_{0,0}(n)(n - \sigma\kappa_n)\tilde{P}_{n,\kappa_n}(A)$$

$$\times \left[(g_{1,0}^{2,0}(n) - g_{1,1}^{2,1}(n))(n - \sigma\kappa_n)\tilde{P}_{n,\kappa_n}(A) + g_{1,0}^{2,0}(n)\right]$$

$$+ g_{0,0}^{1,1}(n)P^*(A)\left[(g_{1,1}^{2,2}(n) - g_{1,0}^{2,1}(n))P^*(A) + g_{1,1}^{2,1}(n)(1-\sigma)\right]$$

$$= I(n,\kappa_n)\mathrm{E}[\tilde{p}(A)|\boldsymbol{X}_{\kappa_n}^{(n)}] + g_{0,0}^{1,1}(n)(g_{1,1}^{2,2}(n) - g_{1,0}^{2,1}(n))P^*(A)\Big[P^*(A) - 1\Big]$$

$$+ g_{0,0}^{1,0}(n)\left(g_{1,0}^{2,0}(n) - g_{1,0}^{2,1}(n)\right)(n - \sigma\kappa_n)^2\tilde{P}_{n,k}(A)\Big[\tilde{P}_{n,k}(A) - 1\Big].$$

Since $(\tilde{P}_{n,k}(A) \vee P^*(A)) \leq 1$, the following upper bound for $\mathrm{Var}[\tilde{p}(A)|\boldsymbol{X}_{\kappa_n}k^{(n)}]$ holds true

$$\mathrm{Var}[\tilde{p}(A)|\boldsymbol{X}_k^{(n)}] \leq I(n,\kappa_n)\mathrm{E}[\tilde{p}(A)|\boldsymbol{X}_{\kappa_n}^{(n)}]\Big(1 - \mathrm{E}[\tilde{p}(A)|\boldsymbol{X}_{\kappa_n}^{(n)}]\Big) + Z_{n,\kappa_n}(A),$$

where

$$Z_{n,\kappa_n}(A) = g_{0,0}^{1,0}(n)(n - \sigma\kappa_n)^2\tilde{P}_{n,k}(A)\,(g_{1,1}^{2,1}(n) - g_{1,0}^{2,0}(n))_+$$

$$+ g_{0,0}^{1,1}(n)\,P^*(A)\,(g_{1,0}^{2,1}(n) - g_{1,1}^{2,2}(n))_+ \quad (17)$$

where, for any a in $\mathbb{R}$, $a_+ := \max\{a, 0\}$. Use again identities (A1) and (A2) of Lemma 1 to get

$$Z_{n,\kappa_n}(A) = g_{0,0}^{1,0}(n)(n - \sigma\kappa_n)\tilde{P}_{n,\kappa_n}(A)(g_{1,0}^{2,0}(n) - I(n,\kappa_n))_+$$

$$+ g_{0,0}^{1,1}(n)P^*(A)(g_{1,1}^{2,1}(n)(1-\sigma) - I(n,\kappa_n))_+$$

Set now, for any $a \in \mathbb{R}$, $a_- := a - a_+$ and define

$$J(n,\kappa_n) := \left(\frac{V_{n+2,\kappa_n+1}}{V_{n+1,\kappa_n+1}}(1 - \sigma_-) - I(n,\kappa_n)\right)_+ \quad (18)$$

One, thus, notes that $(g_{1,1}^{2,1}(n)(1-\sigma) - I(n,\kappa_n))_+ \leq J(n,\kappa_n)$, and

$$(g_{1,0}^{2,0}(n) - I(n,\kappa_n))_+ \leq (g_{1,1}^{2,1}(n) - I(n,\kappa_n))_+ \leq J(n,\kappa_n).$$

This implies that $Z_{n,\kappa_n}(A) \leq J(n,\kappa_n)\mathrm{E}[\tilde{p}(A)|\boldsymbol{X}_{\kappa_n}^{(n)}]$, which in turn yields

$$\mathrm{Var}[\tilde{p}(A)|\boldsymbol{X}_{\kappa_n}^{(n)}] \leq I(n,\kappa_n)\mathrm{E}[\tilde{p}(A)|\boldsymbol{X}_{\kappa_n}^{(n)}]\Big(1 - \mathrm{E}[\tilde{p}(A)|\boldsymbol{X}_{\kappa_n}^{(n)}]\Big)$$

$$+ J(n,\kappa_n)\mathrm{E}[\tilde{p}(A)|\boldsymbol{X}_{\kappa_n}^{(n)}] \quad (19)$$

for any A in $\mathscr{X}$. The upper bound (19), combined with $x(1-x) \leq 1$ for any $x \in [0,1]$, leads to

$$\sum_{i=1}^{\infty}\mathrm{Var}[\tilde{p}(A_i)|\boldsymbol{X}_{\kappa_n}^{(n)}] \leq I(n,\kappa_n) + J(n,\kappa_n).$$

Therefore, we need to show that $J(n,\kappa_n)+I(n,\kappa_n)\to 0$ a.s.-P_0^∞ as $n\to\infty$. In the sequel we shall omit the a.s.-P_0^∞ specification and explicitly use it when possible confusion may arise. By virtue of condition (H), with the limit identified by a value α in $[0,1]$, one has $(V_{n+1,\kappa_n}/V_{n,\kappa_n})(n-\sigma\kappa_n)\to(1-\alpha)$. Hence

$$1-I(n,\kappa_n)=\frac{V_{n+2,\kappa_n+1}}{V_{n+1,\kappa_n+1}}\Big/\frac{V_{n+1,\kappa_n}}{V_{n,\kappa_n}}\sim\frac{n-\kappa_n\sigma}{n+1-(\kappa_n+1)\sigma}$$

and one can conclude that $I(n,\kappa_n)\to 0$, as $n\to\infty$. It follows also that $J(n,k)\to 0$ as long as $(1-\sigma_-)V_{n+2,\kappa_{n+1}}/V_{n+1,\kappa_{n+1}}\to 0$, but the latter is also implied by condition (H) since $V_{n+2,\kappa_n+1}/V_{n+1,\kappa_n+1}\sim(1-\alpha)/(n+1-\sigma(\kappa_n+1))$. The proof is completed after noting that, if P_0 is either discrete or diffuse, the weighted empirical distribution $\tilde{P}_{n,k}$ in (15) converges uniformly to P_0 as $n\to\infty$, a.s.-P_0^∞, as it can be shown by a suitable adaptation of Glivenko-Cantelli's theorem. □

Remark 1. The upper bound for the posterior variance (19) derived within the proof of Theorem 1 is crucial for the determination of the asymptotic behaviour of the posterior distribution and it sheds some light on a distributional property of $\tilde{p}$ that is of independent interest. Its usefulness is also motivated by the fact that the exact expression of posterior variances is typically involved. See, e.g., Jang, Lee and Lee(2010) for species sampling models and James, Lijoi and Prünster (2006) for normalized random measures with independent increments. It should be noted that the bound can be simplified under some further assumptions. Indeed, a close inspection of the arguments used in the proof of Theorem 1 suggests that

$$\mathrm{Var}[\tilde{p}(A)|\boldsymbol{X}_{\kappa_n}^{(n)}]\leq I(n,\kappa_n)\mathrm{E}[\tilde{p}(A)|\boldsymbol{X}_{\kappa_n}^{(n)}]\Big(1-\mathrm{E}[\tilde{p}(A)|\boldsymbol{X}_{\kappa_n}^{(n)}]\Big),\tag{20}$$

namely $J(n,\kappa_n)=0$, whenever one of the following two inequalities is satisfied

$$\frac{V_{n+2,\kappa_n}}{V_{n+1,\kappa_n}}-\frac{V_{n+2,\kappa_n+1}}{V_{n+1,\kappa_n+1}}\geq 0,\tag{21}$$

$$\frac{V_{n+2,\kappa_n+2}}{V_{n+1,\kappa_n+1}}-\frac{V_{n+2,\kappa_n+1}}{V_{n+1,\kappa_n}}\geq 0,\tag{22}$$

see (17). Specifically, (21) implies (22) when $\sigma\in[0,1)$, and (22) implies (21) when $\sigma<0$ as implied by inequality (A3) of Lemma 1 in the Appendix. Since $\mathrm{Var}[\tilde{p}(A)|\boldsymbol{X}_{\kappa_n}^{(n)}]\leq\mathrm{E}[\tilde{p}(A)|\boldsymbol{X}_{\kappa_n}^{(n)}]\big(1-\mathrm{E}[\tilde{p}(A)|\boldsymbol{X}_{\kappa_n}^{(n)}]\big)$, for any $n\geq 1$ and A in $\mathscr{X}$, the validity of one of (21)-(22) implies that a sharper bound is obtained with the addition of the multiplicative factor $I(n,\kappa_n)$. Such a simplification indeed occurs for the two most widely used instances of Gibbs-type priors. For example, when $\tilde{p}$ is a Dirichlet process with baseline measure θP^*, then $I(n,\kappa_n)=1/(\theta+n+1)$ and

$$\mathrm{Var}[\tilde{p}(A)|\boldsymbol{X}_{\kappa_n}^{(n)}]=\frac{1}{\theta+n+1}\mathrm{E}[\tilde{p}(A)|\boldsymbol{X}_{\kappa_n}^{(n)}]\Big(1-\mathrm{E}[\tilde{p}(A)|\boldsymbol{X}_{\kappa_n}^{(n)}]\Big)\tag{23}$$

For the two-parameter Poisson-Dirichlet process model with $\theta>0$ and $\sigma\in(0,1)$, we recover the bound given in James (2008) as a special case of our general result. Indeed,

one can easily check that (21) is valid, $I(n, \kappa_n) = 1/(\theta + n + 1)$ and (23) holds true with equality replaced by strict inequality. □

In order to complete the picture one needs to identify those situations in which $\alpha = 0$ so that $P' = P_0$ and weak consistency is achieved. As already mentioned, for the case of $\sigma \in (0, 1)$, some results for the special instances of Gibbs-type priors admitting closed form predictive structure have been derived in James (2008) and Jang, Lee and Lee (2010). In contrast, for the case $\sigma < 0$, to date no results are available in the literature and therefore we focus attention on this subclass of Gibbs-type priors. Theorem 2 gives neat sufficient conditions for consistency in terms of the tail behaviour of the mixing distribution π on the positive integers $\mathbb{N}$ in (13).

Theorem 2. *Let $\tilde{p}$ be a Gibbs-type prior with parameter $\sigma < 0$, mixing measure π and prior guess P^* whose support coincides with $\mathbb{X}$. Then the posterior is consistent*

(i) *at any discrete P_0 if for sufficiently large x*

$$\frac{\pi(x+1)}{\pi(x)} \leq 1; \tag{T1}$$

(ii) *at any diffuse P_0 if for sufficiently large x and for some $M < \infty$*

$$\frac{\pi(x+1)}{\pi(x)} \leq \frac{M}{x}. \tag{T2}$$

Before proving the result, it is worth remarking a few implications arising from (T1) and (T2) above. Note that condition (T1) is an extremely mild assumption on the regularity of the tail of the mixing π: it requires $x \mapsto \pi(x)$ to be ultimately decreasing, a condition met by the commonly used probability measures on $\mathbb{N}$. Nonetheless, one could construct *ad hoc* examples where such a condition fails to be true. For instance, a mixture of geometric distributions of the type

$$\pi(x) = a(1-p_1)p_1^{x-1}\,\mathbb{1}_{\cup_k\{2k\}}(x) + (1-a)(1-p_2)p_2^{x-1}\,\mathbb{1}_{\cup_k\{2k+1\}}(x)$$

for some a, p_1 and p_2 in $(0, 1)$, does not satisfy (T1). However, this would not be a sign of inconsistency but rather of presumably consistent cases not covered by the sufficient condition (T1) which stands only for technical reasons needed for pinning down the proof. On the other hand, condition (T2) requires the tail of π to be sufficiently light. This is indeed a binding condition and it is particularly interesting to note that such a condition is also close to being necessary. This will become clear when we deal with some specific priors in Section 4. As a matter of fact, we will describe situations ranging from weak consistency, where (T2) holds true, to inconsistency and "total" inconsistency according as to the heaviness of the tails of the mixing distribution π that is chosen. The heavier the tails and the further apart from P_0 the limiting P' in (14) will be.

Proof of Theorem 2. The proof amounts to showing that, under the stated hypothesis, (H) holds true with $\alpha = 0$ so that consistency follows by Theorem 1. Let $V_{n,\kappa_n} =$

$\sum_{x\geq\kappa_n} V^{\sigma,x}_{n,\kappa_n}\,\pi(x)$ where

$$V^{\sigma,x}_{n,\kappa_n} = \frac{|\sigma|^{k-1}\,\prod_{i=1}^{k-1}(x-i)}{(x|\sigma|-1)_{n+1}}.$$

One, then, has $V_{n,\kappa_n} = \sum_{y\geq 0} v_{n,\kappa_n}(y)$ where

$$v_{n,\kappa_n}(y) = \frac{|\sigma|^{\kappa_n-n}\,(y+1)_{\kappa_n-1}}{(\kappa_n+y+\frac{1}{|\sigma|})\cdots(\kappa_n+y+\frac{n-1}{|\sigma|})}\,\pi(y+\kappa_n).$$

After some algebra,

$$V_{n+1,\kappa_n+1} = \sum_{y\geq 0} x_{n,\kappa_n}(y)\frac{\pi(\kappa_n+y+1)}{\pi(\kappa_n+y)}v_{n,\kappa_n}(y), \tag{24}$$

where $x_{n,\kappa_n}(y) = (\kappa_n+y)a_{n,\kappa_n}(y)/(n/|\sigma|+\kappa_n+y+1)$ and

$$a_{n,\kappa_n}(y) = \prod_{i=1}^{n-1}\frac{\left(\kappa_n+y+\frac{i}{|\sigma|}\right)}{\left(\kappa_n+y+1+\frac{i}{|\sigma|}\right)}.$$

We start by considering the case of P_0 discrete. This yields $\kappa_n \ll_{a.s.} n$ and we shall assume $P_0^\infty[\lim_n \kappa_n = \infty] = 1$: indeed when $P_0^\infty[\lim_n \kappa_n < \infty] = 1$ the proof of Theorem 2(i) is straightforward. For n large enough

$$\begin{aligned}\frac{V_{n+1,\kappa_n+1}}{V_{n,\kappa_n}} &\leq \frac{1}{V_{n,\kappa_n}}\sum_{y\geq 0} x_{n,\kappa_n}(y)v_{n,\kappa_n}(y)\\ &\leq x_{n,\kappa_n}(\kappa_n)\frac{\sum_{y=0}^{\kappa_n} v_{n,\kappa_n}(y)}{V_{n,\kappa_n}} + \sum_{y\geq\kappa_n+1} x_{n,\kappa_n}(y)\frac{v_{n,\kappa_n}(y)}{V_{n,\kappa_n}}\\ &\leq x_{n,\kappa_n}(\kappa_n) + \frac{1}{V_{n,\kappa_n}}\sum_{y\geq\kappa_n+1} v_{n,\kappa_n}(y),\end{aligned} \tag{25}$$

where we used (T1) in the first inequality, the monotonicity of $y\mapsto x_{n,\kappa_n}(y)$ in the second inequality and $x_{n,\kappa_n}(y)\leq 1$ in the last inequality. Note that, as $n\to\infty$,

$$x_{n,\kappa_n}(\kappa_n) = \frac{2\kappa_n}{\frac{n}{|\sigma|}+2\kappa_n+1}a_{n,\kappa_n}(\kappa_n)\to 0 \tag{26}$$

a.s.-P_0^∞, since $a_{n,\kappa_n}(y)\leq 1$ for any y and n. As for the second summand in (25), note that

$$\begin{aligned}\frac{1}{V_{n,\kappa_n}}\sum_{y\geq\kappa_n+1} v_{n,\kappa_n}(y) &= \frac{v_{n,\kappa_n}(\kappa_n)}{V_{n,\kappa_n}}\sum_{y\geq 0}\frac{v_{n,\kappa_n}(\kappa_n+y+1)}{v_{n,\kappa_n}(\kappa_n)}\\ &\leq \frac{v_{n,\kappa_n}(\kappa_n)}{v_{n,\kappa_n}(\kappa_n-1)}\sum_{y\geq 0}\prod_{j=0}^{y}\frac{v_{n,\kappa_n}(\kappa_n+j+1)}{v_{n,\kappa_n}(\kappa_n+j)}.\end{aligned}$$

By virtue of

$$\frac{v_{n,\kappa_n}(y+1)}{v_{n,\kappa_n}(y)} = \frac{\kappa_n + y}{y+1}\, a_{n,\kappa_n}(y)\, \frac{\pi(\kappa_n + y + 1)}{\pi(\kappa_n + y)}$$

and (T1), for n large enough one has

$$\frac{1}{V_{n,\kappa_n}} \sum_{y \geq \kappa_n + 1} v_{n,\kappa_n}(y) \leq 2a_{n,\kappa_n}(\kappa_n - 1) \sum_{y \geq 0} \prod_{j=0}^{y} \frac{2\kappa_n + j}{\kappa_n + j + 1} a_{n,\kappa_n}(\kappa_n + j).$$

In view of (25) and (26) one just needs to prove that

$$a_{n,\kappa_n}(\kappa_n - 1) \to 0 \tag{27}$$

as $n \to \infty$, and

$$\sum_{y \geq 0} \prod_{j=0}^{y} \frac{2\kappa_n + j}{\kappa_n + j + 1} a_{n,\kappa_n}(\kappa_n + j) < \infty \tag{28}$$

for sufficiently large n. To this aim, note that

$$a_{n,\kappa_n}(\kappa_n - 1) = \prod_{j=1}^{n-1} \left(1 - \frac{|\sigma|}{2\kappa_n |\sigma| + j}\right).$$

If $S_{n,k} = |\sigma| \sum_{j=1}^{n-1} (k|\sigma| + j)^{-1}$, basing on the inequalities $(1-R)^{x/R} \leq 1 - x \leq \mathrm{e}^{-x}$ for any $0 \leq x \leq R \leq 1$, it easily follows that

$$\left(1 - \frac{1}{2\kappa_n}\right)^{2\kappa_n S_{n,2\kappa_n}} \leq a_{n,\kappa_n}(\kappa_n - 1) \leq \mathrm{e}^{-S_{n,2\kappa_n}}. \tag{29}$$

Moreover, $(1 - 1/(2\kappa_n))^{2\kappa_n} \to \mathrm{e}^{-1}$ and, as $n \to \infty$,

$$S_{n,2\kappa_n} \sim \log \left(\frac{n + 2\kappa_n |\sigma| - 1}{2\kappa_n |\sigma|}\right)^{|\sigma|}. \tag{30}$$

These, combined with (29), lead to the following asymptotic evaluation

$$a_{n,\kappa_n}(\kappa_n - 1) \sim \left(\frac{|\sigma| 2\kappa_n}{n + |\sigma| 2\kappa_n - 1}\right)^{|\sigma|} \tag{31}$$

as $n \to \infty$. As for (28), the y-th term of the series can be written as $\varphi_n(0)\varphi_n(1) \cdots \varphi_n(y)$ where

$$\varphi_n(y) = \frac{2\kappa_n + y}{\kappa_n + y + 1} a_{n,\kappa_n}(\kappa_n + y) = \frac{2k_n + y}{\kappa_n + y + 1} \prod_{j=1}^{n-1} \left(1 - \frac{|\sigma|}{(2\kappa_n + y + 1)|\sigma| + j}\right).$$

Adapting the arguments used in (29) and (30), it can be shown that

$$\varphi_n(y) \sim \frac{2\kappa_n + y}{\kappa_n + y + 1} \left(\frac{|\sigma|(2\kappa_n + y + 1)}{|\sigma|(2\kappa_n + y + 1) + n - 1}\right)^{|\sigma|}$$

as $n \to \infty$, cfr. (31). Next, for $y \to \infty$, use a first order Taylor expansion that yields

$$\begin{aligned}
\varphi_n(y) &\sim \left(1 + \frac{\kappa_n - 1}{\kappa_n + y + 1}\right)\left(1 - \frac{n-1}{|\sigma|(2\kappa_n + y + 1) + n - 1}\right)^{|\sigma|} \\
&= \left(1 + \frac{\kappa_n - 1}{\kappa_n + y + 1}\right)\left(1 - \frac{|\sigma|(n-1)}{|\sigma|(2\kappa_n + y + 1) + n - 1}\right) + O(y^{-2}) \\
&= 1 + \left(\frac{\kappa_n - 1}{\kappa_n + y + 1} - \frac{|\sigma|(n-1)}{|\sigma|(2\kappa_n + y + 1) + n - 1}\right) + O(y^{-2}) \\
&= 1 - \frac{n - \kappa_n}{y} + O(y^{-2}).
\end{aligned}$$

Finally, the series in (28) is convergent since $n - \kappa_n > 0$ (Pólya, and Szegö, 1978). This completes the proof of (i).

Let us now deal with the case where P_0 is diffuse. Hence $\kappa_n = n$, a.s.-P_0^∞ and

$$\begin{aligned}
\frac{V_{n+1,n+1}}{V_{n,n}} &\leq \frac{1}{V_{n,n}} \sum_{y \geq 0} x_{n,n}(y) \frac{M}{n+y} v_{n,n}(y) \\
&\leq \frac{1}{V_{n,n}} \frac{M}{n/|\sigma| + n + 1} \sum_{y \geq 0} a_{n,n}(y) v_{n,n}(y) \\
&\leq \frac{1}{V_{n,n}} \frac{M}{n/|\sigma| + n + 1} \sum_{y \geq 0} v_{n,n}(y) = \frac{M}{n/|\sigma| + n + 1},
\end{aligned}$$

where we used (T2) for $x = n + y$ in the first inequality, $n/|\sigma| + n + y + 1 > n/|\sigma| + n + 1$ in the second inequality and $a_{n,n}(y) \leq 1$ in the last inequality. Since the last term goes to 0 for $n \to \infty$, the proof is complete. □

4 Illustrations

We have learned from Theorem 2 that Gibbs-type priors are consistent when P_0 is discrete, condition (T1) being valid for most commonly used mixing measures π. On the other hand, when P_0 is diffuse one needs to closely investigate the tail behaviour of π and check whether (T2) holds true. One is then naturally led to wondering what happens when (T2) is not satisfied: may in such a case consistency fail to occur even if the "true" P_0 is in the weak support of $\tilde{p}$?

In this section we consider three different Gibbs-type priors with $\sigma = -1$: each prior is characterized by a specific choice of the mixing distribution π. For all such elicitations we immediately have consistency at a discrete discrete P_0 by Theorem 2(i) and therefore we focus on the case of P_0 diffuse, for which different conclusions are reached. As we shall see, according as to heaviness of the tails of π one can move from a situation where the weight α of the convex linear combination of P^* and P_0 in Theorem 1 is equal to 0, thus yielding consistency, to a situation where α increases up to its largest value $\alpha = 1$. We shall note that the heavier the tail of π and the larger α, i.e. the lighter the weight assigned to the "true" P_0 in the limiting distribution identified in Theorem 1.

The first prior is characterized by a heavy-tailed mixing distribution π, which does not admit a finite expected value: condition (T2) is not met and it turns out that $\alpha = 1$ so that the posterior concentrates around the prior guess P^*, referred to as "total" inconsistency. The second specific prior, where the mixing π has light tails that satisfy (T2) in Theorem 2, results in a consistent asymptotic behaviour. In the third case α takes values over the whole unit interval $[0, 1]$ according to a parameter that determines the heaviness of the tail of π.

The illustration we provide is also useful to appreciate the role of condition (T2) in Theorem 2. As we shall see, if the upper bound in (T2) on the ratio $\pi(x+1)/\pi(x)$ does not hold true, consistency is not achieved and, therefore, (T2) cannot be replaced by a milder condition intermediate between (T1) and (T2). This leads to infer that (T2) is also close to being necessary.

4.1. Gnedin's Gibbs–type prior. We consider a family of Gybbs-type priors with $\sigma = -1$ recently introduced by A.V. Gnedin in Gnedin (2010). It is characterized by the mixing distribution

$$\pi(x) = \frac{\gamma(1-\gamma)_{x-1}}{x!}\, \mathbb{1}_{\{1,2,\dots\}}(x)$$

for some $\gamma \in (0, 1)$. This distribution arises in discrete renewal theory (Feller, 1971, Chapter XII) and in connection with the two-parameter Poisson-Dirichlet process (Pitman, 2006). It is characterized by a heavy tail admitting moments of order less than γ. In order to establish consistency one would like to apply Theorem 2. For a discrete P_0, condition (T1) clearly holds true and weak consistency is achieved. In contrast, for a diffuse P_0, the corresponding sufficient condition (T2) is not satisfied: $\pi(x+1)/\pi(x) = (x-\gamma)/(x+1)$ for any positive integer x and cannot be eventually bounded by M/x for some constant M. Therefore one has to establish by direct calculation whether consistency or inconsistency occurs.

In terms of the corresponding Gibbs–type prior, in Gnedin (2010) it is shown that the V_{n,κ_n}'s admit a simple closed form expression given by

$$V_{n,\kappa_n} = \frac{(\kappa_n - 1)!(1-\gamma)_{\kappa_n - 1}(\gamma)_{n-\kappa_n}}{(n-1)!(1+\gamma)_{n-1}}$$

and, consequently, the weights of the prediction rule simplify to

$$\frac{V_{n+1,\kappa_n+1}}{V_{n,\kappa_n}} = \frac{\kappa_n(\kappa_n - \gamma)}{n(\gamma + n)}. \tag{32}$$

From (32) it is easy to to see that, if P_0 is diffuse implying $\kappa_n = n$, condition (H) holds true with $\alpha = 1$. Therefore, by Theorem 1 it follows that the weak limit coincides with the prior guess P^*, whatever the "true" distribution of the data P_0. This means we are in the case of "total" inconsistency apart of the trivial case of $P^* = P_0$ we have already excluded. In this completely explicit setup, it is also interesting to have a closer look at the structure of the bound on the posterior variance discussed in Remark 1: it is easy to check that (22) holds true and, thus, the bound (23) applies with $I(n, \kappa_n) = (2n + \gamma + 1)/[(n+1)(\gamma + n + 1)]$, which does not depend on κ_n. Now, since

$I(n, \kappa_n) \to 0$, as $n \to \infty$, the posterior concentrates, as n increases, at some P' in $\mathbf{P}_{\mathbb{X}}$, in accordance with the general result stated in Theorem 1. Finally note that consistency for the case of discrete P_0, already established by means of Theorem 2(i), can also be deduced from (32) combined with Theorem 1: if P_0 is discrete then $\kappa_n \ll_{a.s.} n$ and (32) converges to $\alpha = 0$ implying convergence to P_0 in Theorem 1.

4.2. Gibbs–type prior with Poisson mixing. The second Gibbs-type prior we consider is characterized by $\sigma = -1$ and a Poisson mixing distribution π with parameter $\lambda > 0$ restricted to the positive integers, i.e.

$$\pi(x) = \frac{\mathrm{e}^{-\lambda}}{1 - \mathrm{e}^{-\lambda}} \frac{\lambda^x}{x!} \mathbb{1}_{\{1,2,\ldots\}}(x).$$

Such a π has light tails and condition (T2) is satisfied since $\pi(x+1)/\pi(x) = \lambda/(x+1)$. Therefore, by Theorem 2(ii), the posterior is consistent when P_0 is diffuse and, *a fortiori*, when P_0 is discrete. Given the Gibbs-type prior at issue admits closed form expressions, the same conclusion can be drawn by direct calculation. The V_{n,κ_n}'s can be expressed as

$$V_{n,\kappa_n} = \pi(\kappa_n) V_{n,\kappa_n}^{-1,\kappa_n} \, {}_1F_1(\kappa_n; \kappa_n + n; \lambda),$$

where ${}_1F_1(a; b; z) = \sum_{j \geq 0} \frac{(a)_j}{j!\,(b)_j} z^j$ is, for any a, b and z in $\mathbb{R}$, the confluent hypergeometric function. Therefore, one has that

$$\frac{V_{n+1,\kappa_n+1}}{V_{n,\kappa_n}} = \frac{\lambda \kappa_n}{(n + \kappa_n + 1)(n + \kappa_n)} \frac{{}_1F_1(\kappa_n; \kappa_n + n; \lambda)}{{}_1F_1(\kappa_n + 1; \kappa_n + n + 2; \lambda)}. \tag{33}$$

With P_0 diffuse, $\kappa_n = n$ for any n a.s.–P_0^∞. Now, by virtue of Eq. (17) of Erdélyi, Magnus, Oberhettinger and Tricomi (1953, Section 6.13.2), the functions ${}_1F_1(n; 2n; \lambda)$ and ${}_1F_1(n+1; 2n+2; \lambda)$ have the same asymptotic expansion as $n \to \infty$, namely

$$\frac{\sqrt{2\pi}\Gamma(2n)}{\sqrt{n/2}\Gamma(n)\Gamma(n)} \, \mathrm{e}^{\lambda/2} \left(\frac{1}{2}\right)^{2n} \big[1 + O(1/n)\big].$$

This means that

$$\frac{{}_1F_1(n; 2n; \lambda)}{{}_1F_1(n+1; 2n+2; \lambda)} \to 1$$

as $n \to \infty$, and

$$\frac{V_{n+1,n+1}}{V_{n,n}} \sim \frac{\lambda}{2(2n+1)} \to 0 \tag{34}$$

as $n \to \infty$. Hence, for the case of diffuse P_0, we have shown by direct calculation that the probability of observing a new species converges to $\alpha = 0$, which by Theorem 1 implies consistency. This is clearly in agreement with the conclusion drawn from Theorem 2(ii) by looking at the tails of the mixing distribution π.

4.3. Gibbs–type prior with geometric mixing. The last sub-family of Gibbs-type priors with $\sigma = -1$ is identified by a geometric mixing distribution

$$\pi(x) = (1 - \eta)\eta^{x-1} \mathbb{1}_{\{1,2,\ldots\}}(x)$$

for some $\eta \in (0,1)$. Note that $\pi(x+1)/\pi(x) = \eta$ so that condition (T2) does not hold true. Therefore, in this case, one can only apply Part (i) of Theorem 2 leading to state consistency solely for the case of discrete P_0. It is therefore interesting to investigate what happens for the case of P_0 diffuse which is not covered by Theorem 2. By direct calculation it turns out that

$$V_{n,\kappa_n} = \pi(\kappa_n) V_{n,\kappa_n}^{-1,\kappa_n} \, {}_2F_1(\kappa_n, \kappa_n + 1; \kappa_n + n; \eta),$$

where ${}_2F_1(a,b;c;z) = \sum_{j\geq 0} \frac{(a)_j\,(b)_j}{j!\,(c)_j}\, z^j$ for any a, b, and c in $\mathbb{R}$ and for any z such that $|z| < 1$, is the Gauss hypergeometric function. Moreover, one has

$$\frac{V_{n+1,\kappa_n+1}}{V_{n,\kappa_n}} = \frac{\eta\kappa_n(\kappa_n+1)}{(n+\kappa_n+1)(n+\kappa_n)} \frac{{}_2F_1(\kappa_n, \kappa_n+1; \kappa_n+n; \eta)}{{}_2F_1(\kappa_n+1, \kappa_n+2; \kappa_n+n+2; \eta)}. \tag{35}$$

With P_0 diffuse, one can replace κ_n with n in the ratio above. Then, by Eq. (16) in Erdélyi, Magnus, Oberhettinger and Tricomi (1953, Section 2.3.2), one obtains the following asymptotic expansions, as $n \to \infty$,

$${}_2F_1(n+1, n+2; 2n+2; \eta) \sim \left(\frac{2}{\eta}\right)^{4+2n} \left(2 - \eta - 2\sqrt{1-\eta}\right)^{n+2} C(\eta)$$

$${}_2F_1(n, n+1; 2n; \eta), \sim \left(\frac{2}{\eta}\right)^{2+2n} \left(2 - \eta - 2\sqrt{1-\eta}\right)^{n+1} C(\eta),$$

where $C(\eta) = \left[(1 + 2\sqrt{1-\eta}/\eta)^2 - ((2-\eta)/\eta)^2\right]^{-\frac{3}{2}}$. On the basis of these asymptotic equivalences one has

$$\frac{V_{n+1,n+1}}{V_{n,n}} \to \alpha = \frac{2 - \eta - 2\sqrt{1-\eta}}{\eta} \in [0,1]. \tag{36}$$

The limit α in (36) can be any point in $[0,1]$ according to the value of η: by Theorem 1 it follows that we can obtain the whole spectrum of weak limits $\alpha P^*(\cdot) + (1-\alpha)P_0(\cdot)$ ranging from consistency ($\alpha = 0$) to "total" inconsistency ($\alpha = 1$). In particular, α is increasing in η, so the larger η, the heavier the limiting mass assigned to the prior guess. Small values of η identify a situation similar to the one discussed in Section 4.2 since they yield a light-tailed π. Conversely, large values of η are more in line with what happens with in Section 4.1 giving rise to heavy-tailed π. Finally, it is worth remarking that a minimal deviation from condition (T2) already produces inconsistent behaviours, even extreme ones, showing that (T2) is close to being necessary.

5 Concluding remarks

Among various criteria one can use for the validation of a statistical model, and of the corresponding inferences, consistency plays a major role. Even in a Bayesian framework, an important prerequisite to any inferential procedure is the specification of a prior that, among others, is consistent according to the frequentist approach.

If $\mathbb{X}$ is finite, P_0 in the weak support of a discrete nonparametric prior $\tilde{p}$ guarantees consistency (Freedman, 1963). When $\mathbb{X}$ is infinite, inconsistent behaviours may appear. To approach such a problem one can essentially undertake two paths: (i) try to identify classes of priors which are consistent whatever the choice of P_0 and dismiss the others; (ii) try to identify the data generating mechanisms the various classes of nonparametric priors are designed for and study consistency w.r.t. choices of P_0 that are compatible with such mechanisms. The seminal contribution to (i) is due to Freedman (1963) (see also Fabius, 1964), where the author identifies a class of nonparametric priors, the family of "tail-free" priors, which are consistent for any P_0, either discrete or diffuse, in its weak support. Notably, the Dirichlet process and Pólya-tree priors (Ferguson, 1974; Lavine, 1992) belong to this class. However, ensuring consistency for any P_0 is not for free. On the one hand, all tail-free priors (with the exception of the Dirichlet process), and the inferential results they yield, heavily depend on the sequence of nested partitions defining them. On the other hand "tail-freeness" appears to be a quite fragile property: as shown in Freedman and Diaconis (1983) and Diaconis and Freedman (1986), inconsistency can already appear when one considers mixtures of the Dirichlet process. Therefore, one might wonder whether it is worth pursuing such a path. Or, alternatively, whether it is not better to establish what kind of inferential issues a prior can address and study consistency for compatible P_0's. In this paper we adhered to this second option: Gibbs-type priors are discrete nonparametric priors and therefore consistency has to be investigated w.r.t. discrete P_0's. The answer we have been able to provide is completely positive in the sense that they are (essentially) always consistent w.r.t. discrete P_0's. It is also worth noting that, given the nature of the phenomenon to be studied, one can establish in advance whether the "true" distribution of the data is discrete or not. When one considers a diffuse data generating P_0, which does not fit a framework within which Gibbs-type priors are used, not surprisingly inconsistency may arise and one can even face completely erratic behaviours such as "total" inconsistency. However, this should not be interpreted as indication to dismiss Gibbs-type priors, thus dropping very natural prediction rules as pointed out in Section 2. Such inconsistent behaviours, combined with consistency in the case of discrete P_0, should rather be seen as strong general methodological evidence against the use of discrete nonparametric priors for modeling data generated from diffuse distributions, a common practice, for instance, in survival analysis applications.

Appendix

A.1 Proof of Proposition 1. Without loss of generality we assume that the support of the prior guess $\mathrm{E}[\tilde{p}(\,\cdot\,)] = P^*$ coincides with $\mathbb{X}$. Let us start by considering the case of $\sigma < 0$. Let $d_{\mathbb{X}}$ be the distance on $\mathbb{X}$ and let d_w denote the Prokhorov distance on $\mathbf{P}_{\mathbb{X}}$. We wish to show that any weak-neighborhood of G_0 has positive Q mass for any probability measure $G_0 \in \mathbf{P}_X$. Since $\mathbb{X}$ is separable, it is well-known that the set of discrete distributions with a finite number of point masses is dense in $\mathbf{P}_{\mathbb{X}}$, with respect to d_w. Hence, for any $\epsilon > 0$ there exists a positive integer k_0, vector of weights $(p_1^0, \ldots, p_{k_0}^0)$ in the k_0-dimensional simplex Δ_{k_0} and points $x_1^0, \ldots, x_{k_0}^0$ in $\mathbb{X}$ such that

$d_w(G_{\mathbf{p}^0,\mathbf{x}^0}, G_0) < \epsilon/2$, where $G_{\mathbf{p}^0,\mathbf{x}^0} = \sum_{i=1}^{k_0} p_i^0 \delta_{x_i^0}$. For any $\eta, \delta > 0$ introduce the sets

$$U_0(\eta) = \{\mathbf{p} = (p_1, \ldots, p_{k_0}) \in \Delta_{k_0} : \ |p_i - p_i^0| < \eta \text{ for any } i = 1, \ldots, k_0\}$$
$$V_0(\eta) = \{\mathbf{x} = (x_1, \ldots, x_{k_0}) \in \mathbb{X}^{k_0} : \ d_{\mathbb{X}}(x_i, x_i^0) < \delta \text{ for any } i = 1, \ldots, k_0\}$$

and $W_0(\eta, \delta)$ stands for the set of discrete probability distributions $G_{\mathbf{p},\mathbf{x}} = \sum_{i=1}^{k_0} p_i \delta_{x_i}$ for $\mathbf{p} \in U_0(\eta)$ and $\mathbf{x} \in V_0(\delta)$. Recall that, conditionally on $K = k_0$, the vector $(p_1, \ldots, p_{k^0})$ has symmetric Dirichlet distribution with parameter $|\sigma|$. This fact, combined with the assumptions on π and on P^*, entails $Q(W_0(\eta, \delta)) > 0$. The proof is completed by showing that, for appropriate choices of η and δ, any $G_{\mathbf{p},\mathbf{x}}$ in $W_0(\eta, \delta)$ is such that $d_w(G_{\mathbf{p},\mathbf{x}}, G_0) < \epsilon$. But this follows by standard arguments. Since

$$d_w(G_{\mathbf{p},\mathbf{x}}, G_0) \leq d_w(G_{\mathbf{p},\mathbf{x}}, G_{\mathbf{p}^0,\mathbf{x}^0}) + \frac{\epsilon}{2},$$

we next show that $\eta = \delta/k_0$ implies that $d_w(G_{\mathbf{p},\mathbf{x}}, G_{\mathbf{p}^0,\mathbf{x}^0}) < \delta$ so that $\delta = \epsilon/2$ would work. For $A \in \mathscr{X}$, the set A^ρ stands for A enlarged by its $d_{\mathbb{X}}$-neighbourhood with radius ρ, $A^\rho = \{x : d_{\mathbb{X}}(x, A) < \rho\}$. When $\rho > \delta$, it is obvious that $x_i^0 \in A$ implies that $x_i \in A^\rho$ whenever $\mathbf{x} = (x_1, \ldots, x_{k_0})$ is in $V_0(\delta)$. One can equivalently say that if $I_0 = \{i : x_i^0 \in A\}$ and $I = \{i : x_i \in A^\rho\}$, then $I \supset I_0$ and

$$\begin{aligned} G_{\mathbf{p}^0,\mathbf{x}^0}(A) - G_{\mathbf{p},\mathbf{x}}(A^\rho) &= \sum_{i \in I_0 \cap I} (p_i^0 - p_i) - \sum_{i \in I \setminus I_0} p_i \\ &\leq \sum_{i \in I} (p_i^0 - p_i) \leq \eta \, \mathrm{card}(I) = \eta k_0 = \delta < \rho. \end{aligned}$$

On the other hand, if $\rho < \delta$, then there exists some set A in $\mathscr{X}$ such that $x_i^0 \in A$ and $x_i \notin A^\rho$ so that is not possible to bound $G_{\mathbf{p}^0,\mathbf{x}^0}(A) - G_{\mathbf{p},\mathbf{x}}(A^\rho)$ by ρ. This completes the proof of the case $\sigma < 0$. The Dirichlet case is well known (Ferguson, 1973; Majumdar, 1992) and the general $\sigma = 0$ case follows by direct extension of the results concerning the Dirichlet process. The case of $\sigma > 0$ follows immediately from the representation of Gibbs-type partitions with $\sigma > 0$ in terms of stable completely random measures (Gnedin and Pitman, 2005, Theorem 12 (iii)). □

A.2 Auxiliary results. Here we collect some useful results on various quantities related to the weights $V_{n,k}$ defining the partition distribution induced by Gibbs–type prior. The main ingredient is the application of the backward recursion (4) for various combination of n and k. Recall that $I(n, k)$ is the factor (16) appearing in the bound (19) of Theorem 1.

Lemma 1. *Let $I(n, k)$ be defined as in (3.3). Then*

$$I(n,k) = \left(\frac{V_{n+2,k}}{V_{n+1,k}} - \frac{V_{n+2,k+1}}{V_{n+1,k+1}} \right)(n - \sigma k) + \frac{V_{n+2,k}}{V_{n+1,k}} \tag{A1}$$

$$I(n,k) = \frac{V_{n+2,k+2}}{V_{n+1,k+1}} - \frac{V_{n+2,k+1}}{V_{n+1,k}} + \frac{V_{n+2,k+1}}{V_{n+1,k+1}}(1 - \sigma) \tag{A2}$$

$$\frac{V_{n+2,k}}{V_{n+1,k}} - \frac{V_{n+2,k+1}}{V_{n+1,k+1}} \tag{A3}$$

$$\begin{matrix}\leq\\(>)\end{matrix} [n+1-\sigma(k+1)]\left(\frac{V_{n+2,k+2}}{V_{n+1,k+1}}-\frac{V_{n+2,k+1}}{V_{n+1,k}}\right) \quad \textit{for} \quad \begin{matrix}0\leq\sigma<1\\(\sigma<0)\end{matrix}$$

Proof. The proof relies on the backward recursion defining the weights of Gibbs-type priors, which is stated in (1.4). As for equation (A1),

$$\begin{aligned}\left(\frac{V_{n+2,k}}{V_{n+1,k}}-\frac{V_{n+2,k+1}}{V_{n+1,k+1}}\right)(n-\sigma k)+\frac{V_{n+2,k}}{V_{n+1,k}}&\\ =\frac{V_{n+2,k}}{V_{n+1,k}}(n+1-\sigma k)-\frac{V_{n+2,k+1}}{V_{n+1,k+1}}(n-\sigma k)&\\ =1-\frac{V_{n+2,k+1}}{V_{n+1,k}}-\frac{V_{n+2,k+1}}{V_{n+1,k+1}}(n-\sigma k)&\\ =1-\frac{V_{n+2,k+1}}{V_{n+1,k}}\left(1+\frac{V_{n+1,k}}{V_{n+1,k+1}}(n-\sigma k)\right)&\\ =1-\frac{V_{n+2,k+1}}{V_{n+1,k}}\frac{V_{n,k}}{V_{n+1,k+1}}=I(n,k)&\end{aligned}$$

where we used the backward recursion (1.4) for $(n+1,k)$ in the second equality and for (n,k) in the last equality.

As for equation (A2), we use the backward recursion (1.4) for $(n+1,k+1)$ to get $V_{n+2,k+2}+V_{n+2,k+1}(1-\sigma)=V_{n+1,k+1}-(n-\sigma k)V_{n+2,k+1}$. Then

$$\begin{aligned}\frac{V_{n+2,k+2}}{V_{n+1,k+1}}-\frac{V_{n+2,k+1}}{V_{n+1,k}}+\frac{V_{n+2,k+1}}{V_{n+1,k+1}}(1-\sigma)&\\ =\frac{V_{n+1,k+1}-V_{n+2,k+1}(n-\sigma k)}{V_{n+1,k+1}}-\frac{V_{n+2,k+1}}{V_{n+1,k}}&\\ =1-\frac{V_{n+2,k+1}}{V_{n+1,k+1}}(n-\sigma k)-\frac{V_{n+2,k+1}}{V_{n+1,k}}&\\ =1-\frac{V_{n+2,k+1}}{V_{n+1,k+1}}\left((n-\sigma k)+\frac{V_{n+1,k+1}}{V_{n+1,k}}\right)&\\ =1-\frac{V_{n+2,k+1}}{V_{n+1,k}}\frac{V_{n,k}}{V_{n+1,k}}=I(n,k)&\end{aligned}$$

where we used again the backward recursion for (n,k) in the last equality.

As for equation (A3), use the backward recursion (1.4) for $(n+1,k+1)$ and $(n+1,k)$ on the right hand side, respectively, to get

$$\begin{aligned}\frac{V_{n+2,k+2}}{V_{n+1,k+1}}-\frac{V_{n+2,k+1}}{V_{n+1,k}}&\\ =\left(1-\frac{V_{n+2,k+1}}{V_{n+1,k+1}}(n+1-\sigma(k+1))\right)-\left(1-\frac{V_{n+2,k}}{V_{n+1,k}}(n+1-\sigma k)\right)&\\ =(n+1-\sigma(k+1))\left(\frac{V_{n+2,k}}{V_{n+1,k}}\frac{n+1-\sigma k}{n+1-\sigma(k+1)}-\frac{V_{n+2,k+1}}{V_{n+1,k+1}}\right).&\end{aligned}$$

Finally, consider that $\frac{n+1-\sigma k}{n+1-\sigma(k+1)}\geq 1$ for $0\leq\sigma\leq 1$ implies the first inequality and that $\frac{n+1-\sigma k}{n+1-\sigma(k+1)}<1$ for $\sigma<0$ implies the second inequality. □

Acknowledgments
The authors are grateful to Subhashis Ghosal, Sasha Gnedin and Judith Rousseau for several insightful discussions. This work was supported by the European Research Council (ERC) through StG "N-BNP" 306406.

References

Barrientos, A.F., Jara, A. and Quintana, F.A. (2011). On the support of MacEachern's dependent Dirichlet processes and extensions. *Bayesian Anal.* **7**, 277–310.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates *Ann. Statist.* **14**, 1–26.

Erdélyi, A., Magnus, W., Oberhettinger, F. and Tricomi, F.G. (1953). *Higher Trascendental Functions*, Vol. 1,McGraw-Hill, New York.

Fabius, J. (1964). Asymptotic behavior of Bayes' estimates. *Ann. Math. Statist.* **35**, 846–856.

Favaro, S., Lijoi, A., Mena, R. H. and Prünster, I. (2012). Bayesian nonparametric estimators derived from Gibbs-type priors with finitely many types. *In preparation.*

Favaro, S., Prünster, I. and Walker, S.G. (2011). On a class of random probability measures with general predictive structure. *Scand. J. Stat.* **38**, 359–376.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, Vol. 2, 2nd edn. Wiley, New York.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230.

Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2**, 615–629.

Freedman, D. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.* **34**, 1386–1403.

Freedman, D. and Diaconis, P. (1983). On inconsistent Bayes estimates in the discrete case. *Ann. Statist.* **11**, 1109–1118.

Ghosal, S., Ghosh, J.K. and Ramamoorthi, R.V. (1999). Posterior consistency of Dirichlet mixtures in density estimation *Ann. Statist.* **27**, 143–158.

Ghosal, S. (2010). The Dirichlet process, related priors, and posterior asymptotics. In *Bayesian Nonparametrics* (Eds. Hjort, N., Holmes, C., Müller, P., Walker, S.), pp. 35-79. Cambridge Univ. Press, Cambridge.

Gnedin, A. (2010). A species sampling model with finitely many types. *Elect. Comm. Probab.* **15**, 79–88.

Gnedin, A. and Pitman, J. (2005). Exchangeable Gibbs partitions and Stirling triangles. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)* **325**, 83-102.

Hartigan, J.A. (1990). Partition models. *Comm. Statist. Theory Methods* **19**, 2745–2756.

Ishwaran, H. and James, L.F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Stat. Ass.* **96**, 161–173.

Ishwaran, H. and James, L.F. (2003). Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statist. Sinica* **13**, 1211–1235.

James, L.F. (2008). Large sample asymptotics for the two parameter Poisson Dirichlet process. In *Pushing the Limits of Contemporary Statistics.* (B. Clarke, S. Ghosal Eds.), 187–199, IMS, Hayward.

James, L.F., Lijoi, A. and Prünster, I. (2006). Conjugacy as a distinctive feature of the Dirichlet process. *Scand. J. Statist.* **33**, 105–120.

Jang, G.H., Lee, J. and Lee, S. (2010). Posterior consistency of species sampling priors. *Statistica Sinica* **20**, 581–593.

Lavine, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.* **20**, 1222–1235.

Lijoi, A., Mena, R.H. and Prünster, I. (2007a). Bayesian nonparametric estimation of the probability of discovering a new species *Biometrika*. **94**, 769–786.

Lijoi, A., Mena, R.H. and Prünster, I. (2007b). A Bayesian nonparametric method for prediction in EST analysis. *BMC Bioinformatics*, **8**: 339.

Lijoi, A., Mena, R.H. and Prünster, I. (2007c). Controlling the reinforcement in Bayesian nonparametric mixture models. *J. Roy. Statist. Soc. Ser. B* **69**, 715-740.

Lijoi, A., Prünster, I. and Walker, S.G. (2005). On consistency of nonparametric normal mixtures for Bayesian density estimation. *J. Amer. Statist. Assoc.* **100**, 1292–1296.

Majumdar, S. (1992). On topological support of Dirichlet prior. *Statist. Probab. Lett.* **15**, 385–388.

Navarrete, C., Quintana, F. and Müller, P. (2008). Some issues on nonparametric Bayesian modeling using species sampling models. *Stat. Modell.* **41**, 3–21.

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory* (T.S. Ferguson, L.S. Shapley, J.B. MacQueen Eds.), 245–267. IMS Lecture Notes Monogr. Ser., Vol. **30**, Hayward.

Pitman, J. (2006). *Combinatorial Stochastic Processes.* Ecole d'Eté de Probabilités de Saint-Flour XXXII. Lecture Notes in Math., vol. **1875**. Springer, Berlin.

Pólya, G. and Szegö, G. (1978). *Problems and theorems in analysis. I. Series, integral calculus, theory of functions.* Springer-Verlag, Berlin-New York.

Quintana, F. A. and Iglesias, P.L. (2003). Bayesian clustering and product partition models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65**, 557–574.

Teh, Y.W. (2006). A Hierarchical Bayesian Language Model based on Pitman-Yor Processes". In *Proceedings of Coling/ACL 2006*, 985-992.

Teh, Y.W. and Jordan, M.I. (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian Nonparametrics* (Hjort, N.L., Holmes, C.C. Müller, P., Walker, S.G. Eds.), pp. 80–136. Cambridge University Press, Cambridge.

Zabell, S.L. (1982). W. E. Johnson's "sufficientness" postulate. *Ann. Statist.* **10**, 1090–1099.