Approximating predictive probabilities of Gibbs-type priors

Julyan Arbel, Stefano Favaro

Approximating predictive probabilities of Gibbs-type priors

Julyan Arbel

Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK
38000 Grenoble, France
e-mail: julyan.arbel@inria.fr

Stefano Favaro*

Department of Economics and Statistics, University of Torino
Corso Unione Sovietica 218/bis, 10134, Torino, Italy
e-mail: stefano.favaro@unito.it

Abstract: Gibbs-type random probability measures, or Gibbs-type priors, are arguably the most "natural" generalization of the celebrated Dirichlet prior. Among them the two parameter Poisson–Dirichlet prior certainly stands out for the mathematical tractability and interpretability of its predictive probabilities, which made it the natural candidate in several applications. Given a sample of size n, in this paper we show that the predictive probabilities of any Gibbs-type prior admit a large n approximation, with an error term vanishing as $o(1/n)$, which maintains the same desirable features as the predictive probabilities of the two parameter Poisson–Dirichlet prior.

AMS 2000 subject classifications: Primary 62F15, 60G57.
Keywords and phrases: Bayesian nonparametrics, first and second order asymptotic approximations, Gibbs-type prior, predictive probabilities, normalized generalized Gamma prior, two parameter Poisson–Dirichlet prior.

1. Introduction

Gibbs-type random probability measures, or Gibbs-type priors, are arguably the most "natural" generalization of the Dirichlet process by Ferguson [17]. They have been first introduced in the seminal works of Pitman [36] and Gnedin and Pitman [18], and their importance in Bayesian nonparametrics have been extensively discussed in Lijoi and Prünster [30], De Blasi et al. [12] and Bacallado et al. [4]. Gibbs-type priors have been widely used in the context of Bayesian nonparametric inference for species sampling problems, where their mathematical tractably allowed to obtain explicit expressions for the posterior distributions of various population's features, and to predict features of additional unobservable samples. See, e.g., Lijoi et al. [27], Lijoi et al. [29], Favaro et al. [14], Favaro et al. [15], Bacallado et al. [3] and Arbel et al. [1]. The class of Gibbs-type priors has been also applied in the context of nonparametric mixture modeling, thus

*Also affiliated to Collegio Carlo Alberto, Moncalieri, Italy.

generalizing the celebrated Dirichlet process mixture model of Lo [31]. Nonparametric mixture models based in Gibbs-type priors are characterized by a more flexible parameterization than Dirichlet process mixture model, thus allowing for a better control of the clustering behaviour. See, e.g., Ishwaran and James [22], Lijoi et al. [26], Lijoi et al. [28], Favaro and Walker [16] and Lomeli et al. [32]. Most recently, Gibbs-type priors have been used in Bayesian nonparametric inference for ranked data (Caron et al. [9]), sparse exchangeable random graphs and networks (Caron and Fox [8] and Herlau [20]), exchangeable feature allocations (e.g., Teh and Görür [41], Broderick et al. [5], Heaukulani and Roy [19], Roy [38] and Battiston et al. [6]), reversible Markov chains (Bacallado et al. [2]), dynamic textual data (Chen et al. [10] and Chen et al. [11]), and bipartite graphs (Caron [7]).

The definition of Gibbs-type random probability measures relies on the notion of α-stable Poisson–Kingman model introduced by Pitman [36]. Specifically, let $(J_i)_{i\geq 1}$ be the decreasing ordered jumps of an α-stable subordinator, i.e. subordinator with Lévy measure $\rho(\mathrm{d}x) = C_\alpha x^{-\alpha-1}\mathrm{d}x$ for some constant C_α, and let $P_i = J_i/T_\alpha$ with $T_\alpha = \sum_{i\geq 1} J_i < +\infty$ almost surely; in particular T_α is a positive α-stable random variable, and we denote its density function by f_α. If $\mathrm{PK}(\alpha; t)$ denotes the conditional distribution of $(P_i)_{i\geq 1}$ given $T_\alpha = t$, and if $T_{\alpha,h}$ is a random variable with density function $f_{T_{\alpha,h}}(t) = h(t)f_\alpha(t)$, for any nonnegative function h, then an α-stable Poisson–Kingman model is defined as the discrete random probability measure $P_{\alpha,h} = \sum_{i\geq 1} P_{i,h}\delta_{X_i^*}$, where $(P_{i,h})_{i\geq 1}$ is distributed as $\int_{(0,+\infty)} \mathrm{PK}(\alpha; t) f_{T_{\alpha,h}}(t)\mathrm{d}t$ and $(X_i^*)_{i\geq 1}$ are random variables, independent of $(P_{i,h})_{i\geq 1}$, and independent and identically distributed according to a nonatomic probability measure ν_0. An α-stable Poisson–Kingman model thus provides with a generalization of the normalized α-stable process in Kingman [25], which is recovered by setting $h = 1$. According to Gnedin and Pitman [18], Gibbs-type random probability measures are defined as a class of discrete random probability measures indexed by a parameter $\alpha < 1$ such that: i) for any $\alpha < 0$ they are M-dimensional symmetric Dirichlet distribution, with M being a nonnegative random variable on $\mathbb{N}$; ii) for $\alpha = 0$ they coincide with the Dirichlet process; iii) for any $\alpha \in (0, 1)$ they are α-stable Poisson–Kingman models.

In this paper we focus on the predictive probabilities of Gibbs-type priors with $\alpha \in (0, 1)$, i.e. the posterior expectation $\mathbb{E}[P_{\alpha,h}(\cdot) \,|\, \boldsymbol{X}_n]$, with $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ being a random sample from $P_{\alpha,h}$. Due to the (almost sure) discreteness of the Gibbs-type random probability measure $P_{\alpha,h}$, we expect ties in a sample $\boldsymbol{X}_n$ from $P_{\alpha,h}$, that is $\boldsymbol{X}_n$ features $K_n = k_n \leq n$ distinct types, labelled by $X_1^*, \ldots, X_{K_n}^*$, with corresponding frequencies $(N_1, \ldots, N_{K_n}) = (n_1, \ldots, n_{k_n})$ such that $\sum_{1\leq i\leq k_n} n_i = n$. In other terms the sample $\boldsymbol{X}_n$ induces a random partition of the set $\{1, \ldots, n\}$; see Pitman [37] for details on Gibbs-type random partitions. According to Pitman [36], the predictive probabilities of $P_{\alpha,h}$ are

$$\Pr[X_{n+1} \in \cdot \,|\, \boldsymbol{X}_n] = \frac{V_{n+1,k_n+1}}{V_{n,k_n}}\nu_0(\cdot) + \frac{V_{n+1,k_n}}{V_{n,k_n}}\sum_{i=1}^{k_n}(n_i - \alpha)\delta_{X_i^*}(\cdot) \qquad (1)$$

for $n \geq 1$, where

$$V_{n,k_n} = \frac{\alpha^{k_n}}{\Gamma(n-k_n\alpha)} \int_0^{+\infty} \int_0^1 t^{-k_n\alpha} p^{n-k_n\alpha-1} h(t) f_\alpha((1-p)t) \mathrm{d}t\mathrm{d}p, \quad (2)$$

with $\Gamma(\cdot)$ being the Gamma function. See, e.g., Pitman [36] and Gnedin and Pitman [18] for a detailed account on (1) and (2). Hereafter we briefly recall two noteworthy examples of Gibbs-type random probability measures: the two parameter Poisson–Dirichlet process and the normalized generalized Gamma process.

Example 1. *Let $(a)_n$ be the rising factorial of a of order n, i.e. $(a)_n = \prod_{0\leq i\leq n-1}(a+i)$, for $a > 0$. For any $\alpha \in (0,1)$ and $\theta > -\alpha$ the two parameter Poisson–Dirichlet process, say $P_{\alpha,\theta}$, is a Gibbs-type random probability measure with*

$$h(t) = \frac{\alpha\Gamma(\theta)}{\Gamma(\theta/\alpha)} t^{-\theta} \quad (3)$$

such that

$$V_{n,k_n} = \frac{\prod_{i=0}^{k_n-1}(\theta+i\alpha)}{(\theta)_n}. \quad (4)$$

The normalized α-stable process is $P_{\alpha,0}$, whereas the Dirichlet process may be recovered as a limiting special case for $\alpha \to 0$. See, e.g., Perman et al. [34], Pitman and Yor [35], James [23], Pitman [36] and James [24] for detailed accounts on $P_{\alpha,\theta}$.

Example 2. *Let $\Gamma(\cdot,\cdot)$ be the incomplete Gamma function, i.e., $\Gamma(a,b) = \int_b^\infty x^{a-1}\exp\{-x\}\mathrm{d}x$ for $(a,b) \in \mathbb{R}\times\mathbb{R}^+$. For any $\alpha \in (0,1)$ and $\tau \geq 0$ the normalized generalized Gamma process, say $G_{\alpha,\tau}$, is a Gibbs-type random probability measure with*

$$h(t) = \mathrm{e}^{\tau^\alpha - \tau t} \quad (5)$$

such that

$$V_{n,k_n} = \frac{\alpha^{k_n}\mathrm{e}^\tau}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-\tau^{1/\alpha})^i \Gamma\left(k_n - \frac{i}{\alpha}, \tau\right). \quad (6)$$

The normalized α-stable process coincides with $G_{\alpha,0}$, whereas $G_{1/2,\tau}$ is the normalized inverse Gaussian process. See James [23], Pitman [36], Lijoi et al. [26], Lijoi [27], Lijoi et al. [29] and James [24] for detailed accounts on $G_{\alpha,\tau}$ and applications.

Within the large class of predictive probabilities of the form (1), those of the two parameter Poisson–Dirichlet process $P_{\alpha,\theta}$ certainly stand out for their mathematical tractability, and for having an intuitive interpretability with respect to the parameter $\alpha \in (0,1)$ and $\theta > -\alpha$. See De Blasi et al. [12] for details, and Zabell [42] and Bacallado et al. [4] for a description of the predictive probabilities of $P_{\alpha,\theta}$ in terms of a simple Pólya like urn scheme. These desirable features of $P_{\alpha,\theta}$ arise from the product form of the V_{n,k_n}'s in (4), which makes the ratio

$V_{n+1,k_n+1}/V_{n,k_n}$ a simple linear function of k_n, and the ratio $V_{n+1,k_n}/V_{n,k_n}$ independent of k_n. Specifically, the predictive probabilities of $P_{\alpha,\theta}$ reduce to the following

$$\Pr[X_{n+1} \in \cdot \mid \boldsymbol{X}_n] = \frac{\theta + k_n\alpha}{\theta + n}\nu_0(\cdot) + \frac{1}{\theta + n}\sum_{i=1}^{k_n}(n_i - \alpha)\delta_{X_i^*}(\cdot), \tag{7}$$

for $n \geq 1$. The weight attached to ν_0 in (7) can be read as a sum of two terms with distinct asymptotic orders of magnitude: i) αk_n, referred to as the first order term, and θ, referred to as the second order term. An analogous two-term decomposition holds for the weight attached to the empirical part of (7). Our distinction and phrasing is formally captured by writing the two weights as follows

$$\frac{\theta + k_n\alpha}{\theta + n} = \frac{k_n\alpha}{n} + \frac{\theta}{n} + o\left(\frac{1}{n}\right) \tag{8}$$

and

$$\frac{1}{\theta + n} = \frac{1}{n} - \frac{\theta}{n^2} + o\left(\frac{1}{n^2}\right), \tag{9}$$

where o is almost sure, recovering both contributions in a two-term asymptotic decomposition. Equations (8) and (9) lead to two large n approximations of the predictive distribution displayed in (7). In particular: i) a first order approximation of (7), denoted by $\sim$, is obtained by combining (7) with the first term on the right-hand side of (8) and (9); ii) a second order approximation of (7), denoted by $\approx$, is obtained by combining (7) with the first two terms on the right-hand side of (8) and (9).

Ruggiero et al. [39] and Arbel et al. [1] first extended the decompositions (8) and (9) to the case of the normalized inverse Gaussian process and the normalized generalized Gamma process, respectively, thus covering the setting of Example 2. In the next theorem we generalize (8) and (9) to the entire class of Gibbs-type priors, that is, for any continuously differentiable function h and any $\alpha \in (0, 1)$ we provide a two-term asymptotic decomposition for the weights $V_{n+1,k_n+1}/V_{n,k_n}$ and $V_{n+1,k_n}/V_{n,k_n}$ of the predictive probabilities (1).

Theorem 1. *Let $\boldsymbol{X}_n$ be a sample from $P_{\alpha,h}$ featuring $K_n = k_n \leq n$ distinct types, labelled by $X_1^*, \ldots, X_{K_n}^*$, with frequencies $(N_1, \ldots, N_{K_n}) = (n_1, \ldots, n_{k_n})$. Assume that function h is continuously differentiable and denote its derivative by h'. Then*

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = \frac{k_n\alpha}{n} + \frac{\beta_n}{n} + o\left(\frac{1}{n}\right) \tag{10}$$

and

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = \frac{1}{n} - \frac{\beta_n}{n^2} + o\left(\frac{1}{n^2}\right) \tag{11}$$

for any $n \geq 1$, where $\beta_n = \varphi_h(nk_n^{-1/\alpha})$ with φ_h being defined as $\varphi_h(t) = -th'(t)/h(t)$.

Theorem 1 may be applied to obtain a first and a second order approximations of the predictive probabilities of any Gibbs-type prior $P_{\alpha,h}$, thus remarkably simplifying the evaluation (1) for any choice of the function h. Besides that, Theorem 1 highlights, for large n, the role of the function h from a purely predictive perspective. According to Theorem 1, h does not affect the first order term in the asymptotic decompositions (10) and (11), and it is sufficient to consider a second order term in order to take into account h. This leads to two meaningful approximations of the predictive probabilities (1). In particular, by considering the sole first order term in (10) and (11), one obtains the first order approximation

$$\Pr[X_{n+1} \in \cdot \mid \boldsymbol{X}_n] \sim \frac{k_n \alpha}{n} \nu_0(\cdot) + \frac{1}{n} \sum_{i=1}^{k_n} (n_i - \alpha) \delta_{X_i^*}(\cdot), \tag{12}$$

which is the predictive probability of the normalized α-stable process, i.e. $h = 1$. By including the second order term in (10) and (11), one obtains the second order approximation

$$\Pr[X_{n+1} \in \cdot \mid \boldsymbol{X}_n] \approx \frac{\beta_n + k_n \alpha}{\beta_n + n} \nu_0(\cdot) + \frac{1}{\beta_n + n} \sum_{i=1}^{k_n} (n_i - \alpha) \delta_{X_i^*}(\cdot), \tag{13}$$

which resembles the predictive probabilities (7) of the two parameter Poisson–Dirichlet process $P_{\alpha,\theta}$, with the parameter θ replaced by a suitable function of h, α and the number k_n of distinct types in the sample $\mathbf{X}_n$. The predictive probabilities of any Gibbs-type prior thus admit a second order approximation, for large n, with an error term vanishing as $o(1/n)$. More importantly, such a second order approximation maintains the same mathematical tractability and interpretability as the predictive probability of the two parameter Poisson–Dirichlet prior.

The paper is structured as follows. In Section 2 we prove Theorem 1 and the approximate predictive probabilities displayed in Equation (12) and Equation (13). In Section 3 we present a numerical illustration of our approximate predictive probabilities, thus showing their usefulness from a practical point of view; the R code for generating the plots presented in Section 3, including the functions for obtaining predictive weights approximations, is available at http://www.julyanarbel.com/software. Section 4 contains a brief discussion of our results.

2. Proof of Theorem 1, Equation (12) and Equation (13)

Throughout this section, we will use the notation $a_n \asymp b_n$ when $a_n/b_n \to 1$ as $n \to \infty$, almost surely. The main argument of the proof consists in a Laplace approximation of the integral form for V_{n,k_n} in (2) as $n \to \infty$. This approximation basically replaces an exponentially large term in an integrand by a Gaussian kernel which matches both mean and variance of the integrand. From evaluating

the Gibbs-type predictive probabilities (1) on the whole space it is clear that we have

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = 1 - (n - \alpha k_n)\frac{V_{n+1,k_n}}{V_{n,k_n}}. \tag{14}$$

Denote the integrand function of (2) by $f_n(p,t) = t^{-\alpha k_n} p^{n-1-k_n\alpha} h(t) f_\alpha((1-p)t)$, and denote integration over its domain $(0,1)\times\mathbb{R}_+^*$ by $\iint$. Then we can write

$$\frac{V_{n+1,k_n}}{V_{n,k_n}} = \frac{1}{n-\alpha k_n}\frac{\iint p f_n}{\iint f_n}. \tag{15}$$

Note that this ratio of integrals coincides with $\mathbb{E}_n(P)$, that is the expectation under the probability distribution with density proportional to f_n. This, combined with (14) provides $V_{n+1,k_n+1}/V_{n,k_n} = \mathbb{E}_n(1-P)$. In order to apply the Laplace approximation method, write the nonnegative integrand f_n in exponential form $f_n = \mathrm{e}^{nl_n}$, and further define functions $g(p,t) = 1-p$ and $\tilde{g}(p,t) = 1$. Then

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = \frac{\iint g\mathrm{e}^{nl_n}}{\iint \tilde{g}\mathrm{e}^{nl_n}}. \tag{16}$$

The mode (t_n, p_n) of f_n (or equivalently of l_n) is determined by the root of the partial derivatives

$$n\frac{\partial l_n(p,t)}{\partial p} = \frac{n-\alpha k_n - 1}{p} - t\frac{f'_\alpha(t(1-p))}{f_\alpha(t(1-p))} \tag{17}$$

and

$$n\frac{\partial l_n(p,t)}{\partial t} = \frac{-\alpha k_n}{t} + \frac{h'(t)}{h(t)} + (1-p)\frac{f'_\alpha(t(1-p))}{f_\alpha(t(1-p))}, \tag{18}$$

where f'_α and h' denote respectively the derivatives of the α-stable density f_α and of the function h. Now consider the Laplace approximations to the numerator and the denominator of the ratio (16) with the notations set forth in Section 6.9 of Small [40]. The exponential term is identical in both integrands of the ratio (16), hence the term involving $\det f_n$, the Hessian of f_n, is also identical and equal to

$$C_n = (2\pi/n)^{2/2}(-\det f_n)^{-1/2}\mathrm{e}^{nl_n(t_n,p_n)}.$$

Thus it simplifies in the ratio. One needs only to consider the asymptotic series expansions, where we require a second order term $a(t_n,p_n)$ for the numerator, that is

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = \frac{C_n \times \left(g(t_n,p_n) + \frac{1}{n}a(t_n,p_n) + \mathcal{O}\left(\frac{1}{n^2}\right)\right)}{C_n \times \left(\tilde{g}(t_n,p_n) + \mathcal{O}\left(\frac{1}{n}\right)\right)}.$$

The expression of $a(t_n, p_n)$ is provided in Equation (6.14) of Small [40]. In our case, $a(t_n, p_n) = o(1/n)$, hence with $\tilde{g} = 1$, the previous display simplifies to the following

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = g(t_n, p_n) + o\left(\frac{1}{n}\right). \tag{19}$$

Let $\varphi_h(t) = -th'(t)/h(t)$. Note that, adding $(1-p_n)\times$(17) and $t_n\times$(18) we can write

$$g(t_n, p_n) = 1 - p_n = \frac{\alpha k_n + \varphi_h(t_n)}{n + \varphi_h(t_n) - 1} \tag{20}$$

so, in view of (19),

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = \frac{\alpha k_n + \varphi_h(t_n)}{n + \varphi_h(t_n) - 1} + o\left(\frac{1}{n}\right). \tag{21}$$

Let $\psi(x) = (xf'_\alpha(x))/(\alpha f_\alpha(x))$. By (17), $\psi((1-p_n)t_n) = (1-p_n)(n - \alpha k_n - 1)/\alpha p_n$. By Theorem 2 in Arbel et al. [1], $1 - p_n \asymp \alpha k_n/n$. Hence, $\psi((1-p_n)t_n)$ grows to infinity when $n \to \infty$ at the same rate as k_n. But studying the variations of the α-stable density f_α, Nolan [33] shows that the only infinite limit of ψ is in 0^+ according to

$$\psi(x) \underset{0^+}{\asymp} (\alpha/x)^{\frac{\alpha}{1-\alpha}}.$$

In order that $\psi((1-p_n)t_n)$ matches with its infinite limit when $n \to \infty$, its argument $(1-p_n)t_n$ needs go to 0^+, which yields to the following asymptotic equivalence

$$k_n \asymp \psi((1-p_n)t_n) \asymp \left(\frac{\alpha}{(1-p_n)t_n}\right)^{\frac{\alpha}{1-\alpha}},$$

which in turn gives

$$t_n \asymp \alpha \frac{k_n^{1-1/\alpha}}{1-p_n} \asymp \alpha \frac{k_n^{1-1/\alpha}}{\alpha k_n/n} \asymp \frac{n}{k_n^{1/\alpha}} \asymp T_{\alpha,h},$$

where the last equivalence is from [36]. Since function h is assumed to be positive and continuous differentiable, $\varphi_h(T_{\alpha,h})$ is a.s. well defined (and finite) and $\varphi_h(t_n) \asymp \varphi_h(nk_n^{-1/\alpha}) \asymp \varphi_h(T_{\alpha,h})$ a.s., so (21) can be rewritten

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = \frac{\alpha k_n}{n} + \frac{\beta_n}{n} + o\left(\frac{1}{n}\right),$$

where we set $\beta_n = \varphi_h(nk_n^{-1/\alpha})$. In other terms, to match the expression of the second order approximate predictive probability displayed in Equation (13), we have

$$\frac{V_{n+1,k_n+1}}{V_{n,k_n}} = \frac{\beta_n + k_n\alpha}{\beta_n + n} + o\left(\frac{1}{n}\right).$$

The expression of the second weight in the predictive of the theorem follows from (14), i.e.,

$$\begin{aligned}\frac{V_{n+1,k_n}}{V_{n,k_n}} &= \frac{1 - V_{n+1,k_n+1}/V_{n,k_n}}{n - \alpha k_n} \\ &= \left(1 - \frac{\alpha k_n}{n} + \frac{\beta_n}{n} + o\left(\frac{1}{n}\right)\right)\left(\frac{1}{n} + \frac{\alpha k_n}{n^2} + o\left(\frac{k_n}{n^2}\right)\right), \\ &= \frac{1}{n} - \frac{\alpha k_n}{n^2} - \frac{\beta_n}{n^2} + \frac{\alpha k_n}{n^2} + o\left(\frac{1}{n^2}\right) = \frac{1}{n} - \frac{\beta_n}{n^2} + o\left(\frac{1}{n^2}\right),\end{aligned}$$

or, to match the expression of the second order approximate predictive of equation (13),

$$\frac{V_{n+1,k_n}}{V_{n,k_n}} = \frac{1}{\beta_n + n} + o\left(\frac{1}{n^2}\right).$$

3. Numerical illustrations

As we recalled in Example 1, the two parameter Poisson–Dirichlet process $P_{\alpha,\theta}$ is a Gibbs-type random probability measure with $\alpha \in (0,1)$ and $h(t) = t^{-\theta}\Gamma(\theta + 1)/\Gamma(\theta/\alpha + 1)$, for any $\theta > -\alpha$. Then, by an application of Theorem 1, the predictive probabilities of $P_{\alpha,\theta}$ admit a first order approximation of the form (12) and a second order approximation of the form (13) with $\varphi_h(t) = \theta$, and such that

$$\beta_n = \theta.$$

Among Gibbs-type random probability measures with $\alpha \in (0,1)$, the two parameter Poisson–Dirichlet process stands out for a predictive structure which admits a simple numerical evaluation. This is certainly one of the reasons that made the two parameter Poisson–Dirichlet prior a natural candidate in several applications within the large class of Gibbs-type priors. Hereafter we present a brief numerical illustration to compare the predictive probabilities of $P_{\alpha,\theta}$ with their first and second order approximations given in terms of Equation (8) and Equation (9). While there is no practical reason to make use our approximate predictive probabilities, because of the simple expression of (7), the illustration is useful to show the accuracy of our approximations. We then present the same numerical illustration for the normalized generalized Gamma process $G_{\alpha,\tau}$ of Example 2. We will see that, differently from the two parameter Poisson–Dirichlet process, the predictive probabilities of the normalized generalized Gamma process do not admits a simple numerical evaluation. This motivates the use of Theorem 1.

We consider 500 data points sampled independently and identically distributed from the ubiquitous Zeta distribution. For any $\sigma > 1$ this is a distribution with probability mass function $\Pr(Z = z) \propto z^{-\sigma}$, for $z \in \mathbb{N}$. Here we choose $\sigma = 1.5$. For each $n = 1, \ldots, 500$ we record the number k_n of distinct types at the n-th draw, and we evaluate the predictive weight $V_{n+1,k_n+1}/V_{n,k_n}$

for the two parameter Poisson–Dirichlet prior, i.e. the left-hand side of (8). We consider the following pairs of parameters (α, θ): $(0.25, 1)$, $(0.25, 3)$, $(0.25, 10)$, $(0.5, 1)$, $(0.5, 3)$, $(0.5, 10)$, $(0.75, 1)$, $(0.75, 3)$ and $(0.75, 10)$. For each of these pairs we compare the left-hand side of Equation (8) with the first term of the right-hand side of Equation (8) (first oder approximation) and with the first two terms of the right-hand side of Equation (8) (second order approximation), that are

$$\frac{\theta + k_n \alpha}{\theta + n}, \tag{22}$$

$$\frac{k_n \alpha}{n} \tag{23}$$

and

$$\frac{k_n \alpha}{n} + \frac{\theta}{n}, \tag{24}$$

respectively. Figure 1 shows the curve, as functions of n, of the "exact" predictive weight (22) and its first order approximation (23) and second order approximation (24). The first order approximation consistently underestimates the "exact" predictive weight, while the second order approximation consistently overestimates it. This is due to the fact that the parameter θ is positive. The discrepancy between the first order approximation and (22) stays substantial even for large values of n, all the more for large θ. On the contrary, the second order approximation consistently outperforms the first order approximation, closely following (22). Importantly, for $n = 500$, the "exact" predictive weight and its second order approximation are barely distinguishable in all the considered pairs of parameters.

Figure 1 about here

3.1. The normalized generalized Gamma process

As we recalled in Example 2, the normalized generalized Gamma process is a Gibbs-type random probability measure with $\alpha \in (0, 1)$ and $h(t) = \exp\{\tau^\alpha - \tau t\}$, for any $\tau \geq 0$. Then, according to Theorem 1, the predictive probabilities of the normalized generalized Gamma process admit a first order approximation of the form (12) and a second order approximation of the form (13) with $\varphi_h(t) = \tau t$, and such that

$$\beta_n = \frac{\tau n}{k_n^{1/\alpha}}.$$

The predictive probabilities of the normalized generalized Gamma process are of the form (1), with the predictive weights $V_{n+1,k_n+1}/V_{n,k_n}$ and $V_{n+1,k_n}/V_{n,k_n}$ admitting an explicit (closed-form) expression in terms of (6). However, differently from the two parameter Poisson–Dirichlet process, the evaluation of the predictive weights is cumbersome, thus preventing their practical implementation. In particular, as pointed out in Lijoi et al. [28] in the context of mixture models with a normalized generalized Gamma prior, the evaluation of (6) gives

rise to severe numerical issues, even for not too large values of n. These issues are mainly due to the evaluation of the incomplete gamma function, as well as with handling very small terms and very large terms within the summation (6). Here we give an example of the numerical issues encountered in the evaluation of (6). We consider 500 data points sampled independently and identically distributed from the Zeta distribution with parameter $\sigma = 1.5$. Figure 2 shows the predictive weight $V_{n+1,k_n+1}/V_{n,k_n}$ computed for the following pairs of parameters (α, τ): $(0.5, 1)$, $(0.5, 3)$ and $(0.5, 10)$. The V_{n,k_n}'s in (6) are evaluated using the software Mathematica, which allows to set the desired numerical precision. Despite that, the resulting values of $V_{n+1,k_n+1}/V_{n,k_n}$ appear very unstable, notably for large values of the τ parameter, leading to evaluate $V_{n+1,k_n+1}/V_{n,k_n}$ outside of $(0, 1)$ even for not too large values of n. In contrast, the first order approximation and the second order approximation of $V_{n+1,k_n+1}/V_{n,k_n}$ behave well.

Figure 2 about here

Due to the aforementioned numerical issues in the evaluation of (6), hereafter we propose an alternative approach to evaluate the V_{n,k_n}'s of the normalized generalized Gamma process. This is a Monte Carlo approach, and it relies on the fact that V_{n,k_n} in (6) can be written as the expectation of a suitable ratio of independent random variables. Recall that f_α denotes the density function of a positive α-stable random variable. Then, using (2) with $h(t) = \exp\{\tau^\alpha - \tau t\}$, we can write

$$\begin{aligned} V_{n,k_n} &= \frac{\alpha^{k_n}}{\Gamma(n - k_n\alpha)} \int_0^{+\infty} \int_0^1 p^{n-1-k_n\alpha} t^{-k_n\alpha} \exp\{\tau^\alpha - \tau t\} f_\alpha(t(1-p)) \mathrm{d}p\mathrm{d}t \\ &= \frac{\alpha^{k_n-1}\Gamma(k_n)}{\Gamma(n)} \int_0^{+\infty} \exp\{\tau^\alpha - \tau t\} \frac{\alpha\Gamma(n)}{\Gamma(k_n)\Gamma(n - k_n\alpha)} t^{-k_n\alpha} \\ &\qquad \int_0^1 (1-p)^{n-k_n\alpha-1} f_\alpha(tp) \mathrm{d}p\mathrm{d}t \\ &= \frac{\alpha^{k_n-1}\Gamma(k_n)}{\Gamma(n)} \mathbb{E}\left[\exp\left\{\tau^\alpha - \frac{\tau X}{Y}\right\}\right], \end{aligned} \tag{25}$$

where X and Y are two independent random variables such that Y is distributed according to a Beta distribution with parameter $(k_n\alpha, n - k_n\alpha)$, and X is distributed according to a polynomially tilted positive α-stable random variable, i.e.,

$$\Pr[X \in \mathrm{d}x] = \frac{\Gamma(k_n\alpha + 1)}{\Gamma(k_n + 1)} x^{-k_n\alpha} f_\alpha(x) \mathrm{d}x. \tag{26}$$

We refer to Pitman [36], Pitman [37] and Devroye [13] for a detailed account on the polynomially tilted α-stable random variable X. Given the representation (25) we can perform a Monte Carlo evaluation of V_{n,k_n} by simply sampling from the Beta random variable Y and from the random variable X with distribution (26).

Sampling form the Beta random variable Y is straightforward. The random variable X can be sampled by using an augmentation argument that reduces the problem of sampling X to the problem of sampling a Gamma random variable and, given that, an exponentially tilted α-stable random variable, i.e. a random variable with density function $\exp\{c^\alpha - cx\} f_\alpha(x)$, for some constant $c > 0$. The problem of sampling exponentially tilted α-stable random variables has been considered in Devroye [13] and Hofert [21]. Specifically, we can write (26) as follows

$$\begin{aligned}\frac{\Gamma(k_n\alpha+1)}{\Gamma(k_n+1)} x^{-k_n\alpha} f_\alpha(x) &= \frac{\alpha}{\Gamma(k_n)} \int_0^{+\infty} c^{k_n\alpha-1} \exp\{-c^\alpha\} \frac{\exp\{-cx\} f_\alpha(x)}{\exp\{-c^\alpha\}} \mathrm{d}c \\ &= \int_0^{+\infty} f_C(c) f_{X|C=c}(x) \mathrm{d}c,\end{aligned}$$

where $f_{X|C=c}$ is the density function of an exponentially tilted positive α-stable random variable, and f_C is the density function of the random variable $C = G^{1/\alpha}$, where G being a Gamma random variable with parameter $(k_n, 1)$. We use the rejection sampler of Hofert [21] for sampling the exponentially tilted positive α-stable random variable with density function $f_{X|C=c}$. Note that, as k_n grows, the tilting parameter $C = G^{1/\alpha}$ gets larger in distribution. As a result, the acceptance probability decreases and the Monte Carlo algorithm slows down. Let Be, Ga and tSt respectively denote Beta, Gamma and exponentially tilted positive α-stable distributions, and let Γ_l represents the logarithm of the Γ function. Hereafter is the step-by-step pseudocode for the Monte Carlo evaluation of the V_{n,k_n}'s:

1. Set $M = 10^4$, n, k_n, α, τ;
2. Sample $Y \sim \mathrm{Be}(\alpha k_n, n - \alpha k_n)$ of size M;
3. Sample $G \sim \mathrm{Ga}(k_n, 1)$ of size M;
4. Sample $X \sim \mathrm{tSt}(\alpha, G^{1/\alpha})$ of size M;
5. Set $v = (k_n - 1)\log\alpha + \Gamma_l(k_n) - \Gamma_l(n) + \tau^\alpha - \tau X/Y$;
6. Set $V = \exp(v)$.

(27)

We perform a numerical study in the same setting described for the two parameter Poisson–Dirichlet process. That is, 500 data points are sampled independently and identically distributed from the Zeta distribution with parameter $\sigma = 1.5$. We consider the following pairs of parameters (α, τ): $(0.25, 1)$, $(0.25, 3)$, $(0.25, 10)$, $(0.5, 1)$, $(0.5, 3)$, $(0.5, 10)$, $(0.75, 1)$, $(0.75, 3)$ and $(0.75, 10)$. For these pairs of parameters the predictive weight $V_{n+1,k_n+1}/V_{n,k_n}$ is evaluated by means of the above steps 1-6, and this evaluation is compared with the first order approximation and with the second order approximation of $V_{n+1,k_n+1}/V_{n,k_n}$ given by Theorem 1, i.e.

$$\frac{k_n\alpha}{n} \tag{28}$$

and

$$\frac{k_n\alpha}{n} + \frac{\tau}{k_n^{1/\alpha}}, \tag{29}$$

respectively. Figure 3 shows that the Monte Carlo evaluation of $V_{n+1,k_n+1}/V_{n,k_n}$ lays between the first order approximation and the second order approximation of $V_{n+1,k_n+1}/V_{n,k_n}$. As n moves, the difference between the resulting Monte Carlo curve and the approximate curves is imperceptible for $\alpha = 0.25$; such a difference is also very small for $\tau = 1$. Larger values of α and/or τ lead to larger discrepancies between the Monte Carlo curve and the approximate curves. The second order approximation is consistently closer to the Monte Carlo value than the first order approximation. In particular we observe that for $n = 500$ the second order approximation and the Monte Carlo value are indistinguishable, whereas the first order approximation may still be far from the Monte Carlo value for several choices of the parameters, e.g. $(\alpha, \tau) = (0.75, 3)$ and $(\alpha, \tau) = (0.75, 10)$.

Figure 3 about here

We conclude by motivating the use of the second order approximation instead of the Monte Carlo evaluation. First of all, for pairs of parameters with large α and large τ, e.g. $(\alpha, \tau) = (0.75, 10)$ in our numerical study, the Monte Carlo evaluation is extremely noisy, although we have used a large number of iterations, i.e 10^4. In particular, as shown in Figure 3, the noise does not vanish as n grows. On the contrary, the second order approximation has a more stable behavior, and for $(\alpha, \tau) = (0.75, 10)$ it converges to the bulk of the Monte Carlo curve, which makes it more reliable than the latter for large values of n. Furthermore, evaluating the second order approximation is fast. On the other hand, the computational burden of the Monte Carlo evaluation is very heavy, e.g. 35 hours were required for the nine configurations of Figure 3, with 10^4 iterations for each weight. This is because of the sampling of the exponentially tilted α stable random variable. Indeed the rejection sampler originally proposed by Hofert [21] has an acceptance probability that decreases as n grows, making this approach prohibitive for large sample sizes. Although our Monte Carlo code could certainly be fastened, our empirical study suggests that the computing time increases exponentially with the sample size n. See the average Monte Carlo running time in Figure 4, as well as the running time and cumulated running time for each of the nine parameter configurations in Figure 5 and Figure 6.

Figure 4 about here

Figure 5 about here

Figure 6 about here

4. Discussion

Gibbs-type priors form a flexible class of nonparametric priors parameterized by $\alpha \in (0, 1)$ and a function h. According to the definition of Gibbs-type random probability measures in terms of α-stable Poisson-Kingman models, the function h has the primary role of enriching the parameterization of the normalized

α-stable process by introducing additional parameters other than α. See, e.g., Example 1 and Example 2. In this paper we introduced a first order approximation (12) and a second order approximation (13) for the predictive probabilities of Gibbs-type priors, for any $\alpha \in (0,1)$ and any function h. In particular we showed that at the level of the first order approximation the function h has no impact on the predictive probabilities. Indeed Equation (12) coincides with the predictive probability of the normalized α-stable process, i.e. a Gibbs-type random probability measure with $\alpha \in (0,1)$ and $h(t) = 1$. However, it is sufficient to consider a second order approximation in order to take into account h. Indeed, Equation (13) coincides with the predictive probability of the two parameter Poisson–Dirichlet process in which the parameter θ is replaced by a suitable function of h. The proposed approximations thus highlight the role of the function h from a purely predictive perspective, and at the same time provide practitioners with a way to easily handle the predictive probabilities of any Gibbs-type prior.

Acknowledgements

This work was partly conducted during a scholar visit of Julyan Arbel at the Department of Statistics & Data Science of the University of Texas at Austin, whose hospitality was greatly appreciated. Stefano Favaro is supported by the European Research Council through StG N-BNP 306406.

References

[1] Arbel, J., Favaro, S., Nipoti, B. and Teh, Y.W. (2017). Bayesian nonparametric inference for discovery probabilities: credible intervals and large sample asymptotics. *Statist. Sinica*, **27**, 839–858.

[2] Bacallado, S., Favaro, S. and Trippa, L. (2013). Bayesian nonparametric analysis of reversible Markov chains. *Ann. Statist.*, **41**, 870–896.

[3] Bacallado, S., Favaro, S. and Trippa, L. (2015). Looking-backward probabilities for Gibbs-type exchangeable random partitions, *Bernoulli*, **21**, 1–37

[4] Bacallado, S., Battiston, M., Favaro, S. and Trippa, L. (2017). Sufficientness postulates for Gibbs-type priors and hierarchical generalizations. *Statist. Sci.*, to appear.

[5] Broderick, T., Pitman, J. and Jordan, M. (2013). Feature allocations, probability functions, and paintboxes. *Bayesian Anal.*, **8**, 1–22.

[6] Battiston, M, Favaro, S. Roy, D.M. and Teh, Y.W. (2016). A characterization of product-form exchangeable feature probability functions. *Preprint arXiv:1607.02066*.

[7] Caron, F. (2012). Bayesian nonparametric models for bipartite graphs. *Adv. Neur. Inf. Proc. Sys.*

[8] Caron, F. and Fox, E.B. (2017). Sparse graphs using exchangeable random measures. *J. Roy. Statist. Soc. Ser. B*, to appear

[9] Caron, F., Teh, Y.W. and Murphy, T.B. (2014). Bayesian nonparametric Plackett-Luce models for the analysis of preferences for college degree programmes. *Ann. Appl. Statist.* **8**, 1145–1181.

[10] Chen, C., Ding, N. and Buntine, W (2012). Dependent hierarchical normalized random measures for dynamic topic modeling. *Int. Conf. Mach. Learn.*

[11] Chen, C., Rao, V.A., Buntine, W. and Teh, Y.W. (2013). Dependent normalized random measures. *Int. Conf. Mach. Learn.*

[12] De Blasi, P., Favaro, S., Lijoi, A., Mena, R.H., Prünster, I. and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *EEE Trans. Pattern Anal. Mach. Intell.*, **37**, 212–229.

[13] Devroye, L. (2009). Random variate generation for exponentially and polynomially tilted stable distributions. *ACM Trans. Model. Comput. Simul.*, **4**, 18.

[14] Favaro, S., Lijoi, A., Mena, R.H. and Prünster, I. (2009). Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. *J. Roy. Statist. Soc. Ser. B*, **71**, 993–1008.

[15] Favaro, S., Lijoi, A. and Prünster, I. (2012). A new estimator of the discovery probability. *Biometrics*, **68**, 1188–1196.

[16] Favaro, S. and Walker, S.G. (2013). Slice sampling σ-stable Poisson-Kingman mixture models. *J. Comput. Graph. Statist.*, **22**, 830–847.

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

[18] Gnedin, A. and Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *J. Math. Sci.*, **138**, 5674–5685.

[19] Heaukulani, C. and Roy, D.M. (2015). Gibbs-type Indian buffet processes. *Preprint arXiv:1512.02543*.

[20] Herlau, T., Schmidt, M. N., and Mørup, M. (2016). Completely random measures for modeling block-structured sparse networks. *Adv. Neur. Inf. Proc. Sys.*

[21] Hofert, M. (2011). Efficiently sampling nested archimedean copulas. *Comput. Statist. Data Anal.*, **55**, 57–70.

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

[23] James, L.F. (2002). Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics. *Preprint arXiv:math/0205093*.

[24] James, L.F. (2013). Stick-breaking PG(α, ζ)-generalized Gamma processes. *Preprint arXiv:1308.6570*.

[25] Kingman, J.F.C (1975). Random discrete distributions. *J. Roy. Statist. Soc., Series B*, **37**, 1–22.

[26] Lijoi, A., Mena, R.H. and Prünster, I. (2005). Hierarchical mixture modelling with normalized inverse-Gaussian priors. *J. Amer. Stat. Assoc.* **100** 1278–1291.

[27] Lijoi, A., Mena, R.H. and Prünster, I. (2007). Bayesian nonparamet-

ric estimation of the probability of discovering new species. *Biometrika*, **94**, 769–786.

[28] Lijoi, A., Mena, R.H. and Prünster, I. (2007a). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. Roy. Statist. Soc. Ser. B*, **69**, 769–786.

[29] Lijoi, A., Prünster, I. and Walker, S.G. (2008). Investigating nonparametric priors with Gibbs structure. *Statist. Sinica*, **18**, 1653–1668.

[30] Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics*, Hjort, N.L., Holmes, C.C. Müller, P. and Walker, S.G. Eds. Cambridge University Press.

[31] Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates. *Ann. Statist.*, **12**, 351–357.

[32] Lomeli, M., Favaro, S and Teh, Y.W. (2017). A marginal sampler for σ-stable Poisson-Kingman mixture models. *J. Comput. Graph. Statist.*, **26**, 44–53.

[33] Nolan, J. (2003). *Stable distributions: models for heavy-tailed data.* Birkhauser Boston.

[34] Perman, M., Pitman, J. and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Related Fields.* **92**, 21–39.

[35] Pitman, J. and Yor, M. (1997). The two parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**, 855–900.

[36] Pitman, J. (2003). Poisson-Kingman partitions. In *Science and Statistics: A Festschrift for Terry Speed*, Goldstein, D.R. Eds. Institute of Mathematical Statistics.

[37] Pitman, J. (2006). *Combinatorial Stochastic Processes.* Ecole d'Eté de Probabilités de Saint-Flour XXXII. Lecture notes in mathematics, Springer.

[38] Roy, D.M. (2014). The continuum-of-urns scheme, generalized beta and Indian buffet processes, and hierarchies thereof. *Preprint arXiv:1501.00208.*

[39] Ruggiero, M., Walker, S. G. and Favaro, S. (2013). Alpha-diversity processes and normalized inverse-Gaussian diffusions. *Ann. Appl. Probab.*, 23(1):386–425.

[40] Small, C.G. (2010). *Expansions and asymptotics for statistics.* CRC Press.

[41] Teh, Y.W. and Görür (2010). Indian buffet processes with power law behavior. *Adv. Neur. Inf. Proc. Sys.*

[42] Zabell, S.L. (2005). The continuum of inductive methods revisited. In *Symmetry and its discontents: essays on the history of inductive probability.* Cambridge Univ. Press, New York.

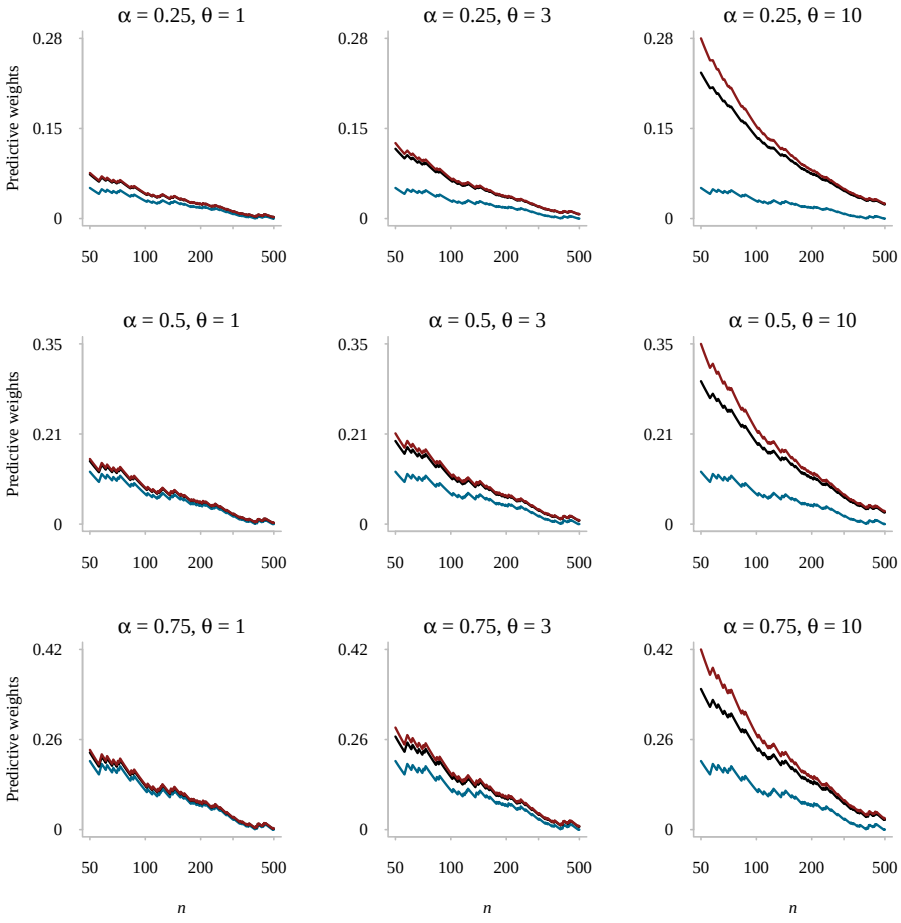

FIG 1. *Predictive weights $V_{n+1,k_n+1}/V_{n,k_n}$ in the two parameter Poisson–Dirichlet process case with data generated from a Zeta(1.5) distribution. In black: the "exact" value (22). In blue: the first order approximation (23). In red: the second order approximation (24). The following values for the parameters are considered: $\alpha = 0.25, 0.5$ and 0.75 in the top, middle and bottom rows respectively; $\theta = 1, 3$ and 10 for the left, middle and right columns respectively. The sample size on the x-axis in log scale runs from $n = 50$ to $n = 500$. The points are connected by straight lines only for visual simplification.*

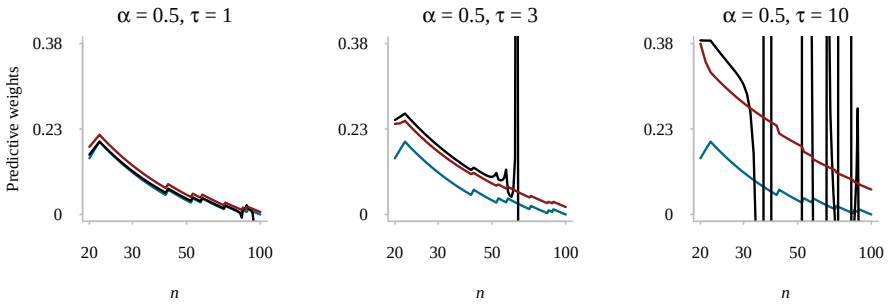

FIG 2. *Predictive weights $V_{n+1,k_n+1}/V_{n,k_n}$ in the normalized generalized Gamma process case with data generated from a Zeta(1.5) distribution. In black: the "exact" value (6) evaluated using Mathematica. In blue: the first order approximation (28). In red: the second order approximation (29). The following values for the parameters are considered: $(\alpha,\tau) = (0.5,1),(0.5,3)$ and $(0.5,10)$ for the left, middle and right panels respectively. The sample size on the x-axis in log scale runs from $n = 20$ to $n = 100$. The points are connected by straight lines only for visual simplification.*

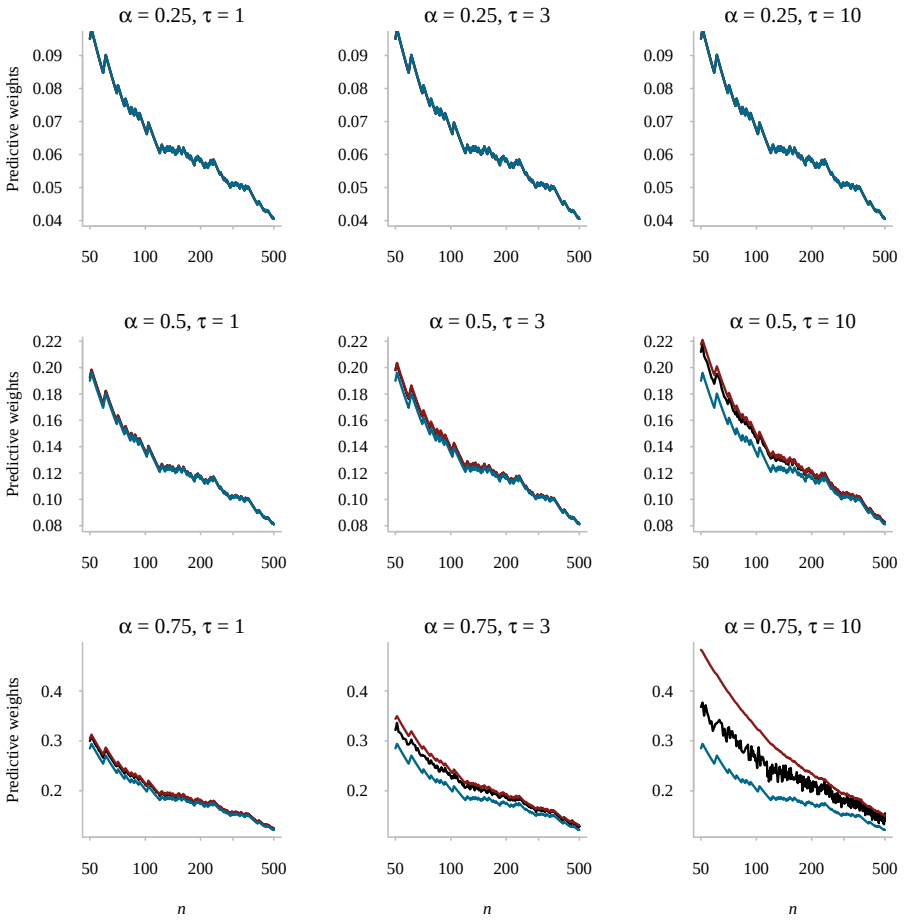

FIG 3. *Predictive weights $V_{n+1,k_n+1}/V_{n,k_n}$ in the normalized generalized Gamma process case with data generated from a Zeta(1.5) distribution. In black: the "exact" value evaluated by the Monte Carlo procedure (27). In blue: the first order approximation (28). In red: the second order approximation (29). The following values for the parameters are considered: $\alpha = 0.25, 0.5$ and 0.75 in the top, middle and bottom rows respectively; $\tau = 1, 3$ and 10 for the left, middle and right columns respectively. The sample size on the x-axis in log scale runs from $n = 50$ to $n = 500$. The points are connected by straight lines only for visual simplification.*

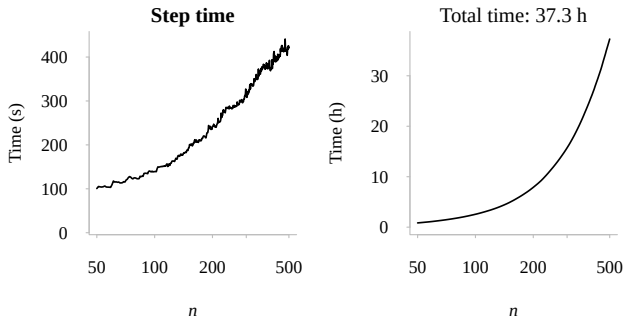

FIG 4. *Left panel: running time (in seconds) averaged over all nine parameter configurations, and right panel: cumulated running time (in hours) averaged over all nine parameter configurations, for the Monte Carlo evaluation* (27) *of the predictive weights* $V_{n+1,k_n+1}/V_{n,k_n}$ *in the normalized generalized Gamma process case. The sample size on the* x*-axis in log scale runs from* $n = 50$ *to* $n = 500$*. The points are connected by straight lines only for visual simplification.*

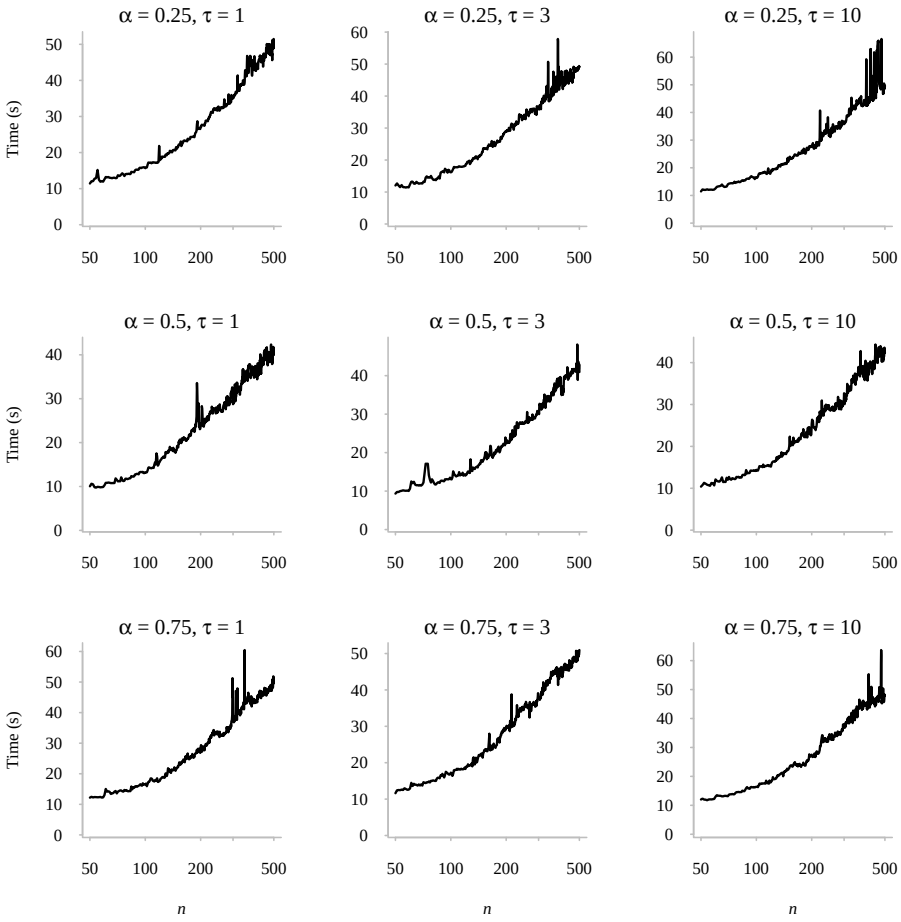

FIG 5. *Running time (in seconds) for the Monte Carlo evaluation (27) of the predictive weights $V_{n+1,k_n+1}/V_{n,k_n}$ in the normalized generalized Gamma process case. The following values for the parameters are considered: $\alpha = 0.25, 0.5$ and 0.75 in the top, middle and bottom rows respectively; $\tau = 1, 3$ and 10 for the left, middle and right columns respectively. The sample size on the x-axis in log scale runs from $n = 50$ to $n = 500$. The points are connected by straight lines only for visual simplification.*

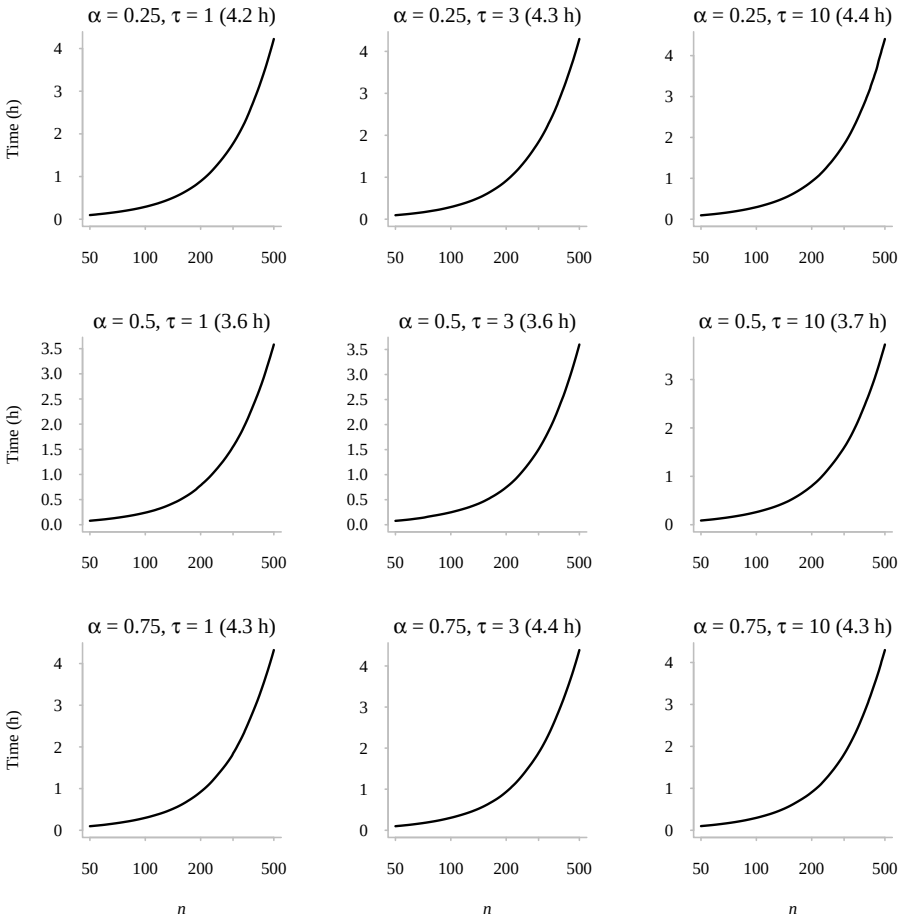

FIG 6. *Cumulated running time (in hours) for the Monte Carlo evaluation (27) of the predictive weights $V_{n+1,k_n+1}/V_{n,k_n}$ in the normalized generalized Gamma process case. The following values for the parameters are considered: $\alpha = 0.25, 0.5$ and 0.75 in the top, middle and bottom rows respectively; $\tau = 1, 3$ and 10 for the left, middle and right columns respectively. The sample size on the x-axis in log scale runs from $n = 50$ to $n = 500$. The points are connected by straight lines only for visual simplification.*