

To appear in *Statistics: A Journal of Theoretical and Applied Statistics*
Vol. 00, No. 00, Month 20XX, 1–17

Characterising variation of nonparametric random probability measures using the Kullback-Leibler divergence

J. Watson^{a,*}, L. Nieto-Barajas^{a,b} and C. Holmes^a

^a*Department of Statistics, University of Oxford, UK;* ^b*Department of Statistics, ITAM, Mexico*

(Received 00 Month 20XX; final version received 00 Month 20XX)

This work characterises the dispersion of some popular random probability measures, including the Bootstrap, the Bayesian Bootstrap, and the Pólya tree prior. This dispersion is measured in terms of the variation of the Kullback-Leibler divergence of a random draw from the process to that of its baseline centering measure. By providing a quantitative expression of this dispersion around the baseline distribution, our work provides insight for comparing different parameterisations of the models and for the setting of prior parameters in applied Bayesian settings. This highlights some limitations of the existing canonical choice of parameter settings in the Pólya tree process.

Keywords: Bayesian nonparametrics, Kullback-Leibler divergence, bootstrap methods, Pólya trees.

1. Introduction

Random probability models (RPMs) are key elements of Bayesian nonparametrics [1–3] used to express prior beliefs with wide support. Moreover random measures such as the Bootstrap are important tools in many non-Bayesian settings. The field of Bayesian nonparametrics has become increasingly popular in recent years due to the flexible modelling structures it supports, alleviating concerns over the “closed hypothesis space” of parametric Bayesian inference. The most well known models in Bayesian nonparametrics are the Dirichlet process (DP) and its generalizations, which include Pólya tree (PT) processes [for the main references see 1, 4, 5]. Priors based on the Dirichlet and the Pólya tree processes are of particular interest because of their analytical tractability and their conjugacy properties for inference problems.

RPMs are typically characterised via a baseline centering distribution F_0 and a dispersion parameter vector that quantifies how concentrated the nonparametric prior is around F_0 . For example, in the Pólya tree prior for continuous variables the user specifies a recursive partition (tree) structure Π on the support of F_0 and then a precision function, typically of the form $\alpha\rho(\cdot)$, where α is a concentration parameter and $\rho(\cdot)$ is a univariate monotone function of the level of the tree partition; Lavine [5] recommends $\rho(m) = m^2$ as a “sensible canonical choice” for level m , which has been adopted as the standard choice in many applications [see for example 6–10]. However little formal guidance exists on the setting and sensitivity of the models with respect to this choice. Moreover, in practical applications when using Pólya trees it is also necessary to truncate

*Corresponding author. Email: jameswatsonwork@gmail.com

the tree at a certain level M . This adds an additional parameter to the model that the user must set.

This paper considers the general question of how far random draws F are from a specific centring distribution F_0 . One aim is to parametrise an RPM such as the Pólya tree process in such a way as to sample distributions at a specific KL divergence from F_0 . We believe that addressing this question is important for the use of such nonparametric distributions in practice. For example in Bayesian inference when they are used as a nonparametric prior, our results can help avoid selecting a prior overly concentrated around the centring distribution. In this note we provide some guidance on how to answer these questions using the Kullback-Leibler (KL) divergence [11]. We concentrate on this divergence for its fundamental role played in information theory and Bayesian statistics [e.g. 12–14]. One consequence of our work is a better insight for both choosing the truncation level M and for choosing the parameters α and $\rho(\cdot)$. We also give comparable results for the more widely used bootstrap methods, namely the Bayesian and the frequentist bootstrap which could possibly help guide their use as nonparametric model extensions, for example see [15].

Section 2 introduces some notation and defines the Pólya tree as the principal Bayesian model considered. Section 3 presents several properties and results for the variation in KL divergence, considering random draws from RPMs including the Bayesian and frequentist Bootstrap. Section 4 concludes with a discussion on the implications of these findings for practitioners.

2. Notation

The Pólya tree will be our main object of interest, particularly as the Dirichlet process can be seen as a particular case of a Pólya tree, see [16]. We define it as follows.

The Pólya tree relies on a binary partition tree of the sample space. For simplicity of exposition we consider $(\mathbb{R}, \mathcal{B})$ as our measurable space with \mathbb{R} the real line and \mathcal{B} the Borel sigma algebra of subsets of \mathbb{R} . Using the notation in [17], the binary partition tree is denoted by $\Pi = \{B_{mj} : m \in \mathbb{N}, j = 1, \dots, 2^m\}$, where the index m denotes the level in the tree and j the location of the partitioning subset within the level. The sets at level 1 are denoted by (B_{11}, B_{12}) ; the partitioning subsets of B_{11} are (B_{21}, B_{22}) , and $B_{12} = B_{23} \cup B_{24}$, such that $(B_{21}, B_{22}, B_{23}, B_{24})$ denote the sets at level 2. In general, at level m , the set B_{mj} splits into two disjoint sets $(B_{m+1,2j-1}, B_{m+1,2j})$, where $B_{m+1,2j-1} \cap B_{m+1,2j} = \emptyset$ and $B_{m+1,2j-1} \cup B_{m+1,2j} = B_{mj}$.

We associate random branching probabilities Y_{mj} with every set B_{mj} . We will use F to denote a cdf or a probability measure in-distinctively, and f to denote a density. We define $Y_{m+1,2j-1} = F(B_{m+1,2j-1} \mid B_{mj})$, and $Y_{m+1,2j} = 1 - Y_{m+1,2j-1} = F(B_{m+1,2j} \mid B_{mj})$. We denote by $\mathcal{Y} = \{Y_{mj}\}$ the set of random branching probabilities associated with the elements of Π .

Definition 2.1 [5]. Let $\mathcal{A}_m = \{\alpha_{mj}, j = 1, \dots, 2^m\}$ be non-negative real numbers, $m = 1, 2, \dots$, and let $\mathcal{A} = \bigcup_m \mathcal{A}_m$. A random probability measure F on $(\mathbb{R}, \mathcal{B})$ is said to have a Pólya tree prior with parameters (Π, \mathcal{A}) if for $m = 1, 2, \dots$ there exist random variables $\mathcal{Y}_m = \{Y_{m,2j-1}, j = 1, \dots, 2^{m-1}\}$ such that the following hold:

- (i) All the random variables in $\mathcal{Y} = \bigcup_m \mathcal{Y}_m$ are independent.
- (ii) For every $m = 1, 2, \dots$ and every $j = 1, \dots, 2^{m-1}$, $Y_{m,2j-1} \sim \text{Be}(\alpha_{m,2j-1}, \alpha_{m,2j})$.

(iii) For every $m = 1, 2, \dots$ and every $j = 1, \dots, 2^m$

$$F(B_{mj}) = \prod_{k=1}^m Y_{m-k+1, j_{m-k+1}^{(m,j)}},$$

where $j_{k-1}^{(m,j)} = \lceil j_k^{(m,j)} / 2 \rceil$ is a recursive decreasing formula, whose initial value is $j_m^{(m,j)} = j$, that locates the set B_{mj} with its ancestors upwards in the tree. $\lceil \cdot \rceil$ denotes the ceiling function, and $Y_{m,2j} = 1 - Y_{m,2j-1}$ for $j = 1, \dots, 2^{m-1}$.

There are several ways of centring the process around a parametric probability measure F_0 . The simplest and most used method [9] consists of matching the partition with the dyadic quantiles of the desired centring measure and keeping α_{mj} constant within each level m . More explicitly, at each level m we take

$$B_{mj} = \left(F_0^{-1} \left(\frac{j-1}{2^m} \right), F_0^{-1} \left(\frac{j}{2^m} \right) \right], \quad (1)$$

for $j = 1, \dots, 2^m$, with $F_0^{-1}(0) = -\infty$ and $F_0^{-1}(1) = \infty$. If we further take $\alpha_{mj} = \alpha_m$ for $j = 1, \dots, 2^m$ we get $E\{F(B_{mj})\} = F_0(B_{mj})$.

In particular we take $\alpha_{mj} = \alpha \rho(m)$, which specifies the beta distributions of the branching probabilities (see part (ii) in Definition 2.1). Decreasing the value of α increases the variance of the beta random variables, therefore the parameter α is interpreted as a precision parameter of the Pólya tree [10]. In the same way, the function ρ controls the speed at which the variance of the branching probabilities changes as one moves down the tree. According to Ferguson [16], $\rho(m) = 1/2^m$ defines an a.s. discrete measure that coincides with the Dirichlet process [4], and $\rho(m) = 1$ defines a continuous singular measure. Moreover, if ρ is such that $\sum_{m=1}^{\infty} \rho(m)^{-1} < \infty$ it guarantees that F is absolutely continuous [18], e.g., $\rho(m) = m^2, m^3, 2^m, 4^m$.

In practice we need to stop partitioning the space at a finite level M to define a finite tree process. At the lowest level M , we can spread the probability within each set B_{Mj} according to f_0 . In this case the random probability measure defined will have a density of the form

$$f(x) = \prod_{m=1}^M Y_{m, j_m^{(x)}} 2^M f_0(x), \quad (2)$$

for $X \in \mathbb{R}$, and with $j_m^{(X)}$ identifying the set at level m that contains X . This maintains the condition $E(f) = f_0$. We denote a finite Pólya tree process as $\mathcal{PT}_M(\alpha, \rho, F_0)$. Taking $M \rightarrow \infty$ defines a draw from a Pólya tree.

Let us consider a set of functions $\rho(m)$ of the following types:

$$\rho_1(m) = 1/2^m, \quad \rho_2(m) = 1, \quad \rho_3(m) = m^\delta, \quad \text{and} \quad \rho_4(m) = \delta^m, \quad (3)$$

with $\delta > 1$, to define discrete, singular and two absolutely continuous measures, respectively. These four families of functions are broad enough to illustrate the results in this work, in particular ρ_3 and ρ_4 which are widely used in practice ($\rho_3(m) = m^2$ being the so-called ‘canonical choice’, see [5]).

To measure “distance” between probability distributions, we concentrate on the Kullback-Leibler divergence, which for densities f and g is defined as

$$\text{KL}(f||g) = E_f \left[\log \left\{ \frac{f(x)}{g(x)} \right\} \right] = \int \log \left\{ \frac{f(x)}{g(x)} \right\} f(x) dx. \quad (4)$$

3. Properties

3.1. Pólya Trees

If $F \sim \mathcal{PT}_M(\alpha, \rho, F_0)$ then it is not difficult to show that the KL between the centring distribution F_0 and a random draw F is a random variable that does not depend on F_0 , and is given by:

$$\text{KL}(f_0||f) = - \sum_{m=1}^M \sum_{j=1}^{2^m} (\log Y_{mj}) \frac{1}{2^m} - M \log 2, \quad (5)$$

where $Y_{m,2j-1} \sim \text{Be}(\alpha_{m,2j-1}, \alpha_{m,2j})$. Since the KL divergence measure is asymmetric, we can reverse the role of f and f_0 . In this case the reverse KL divergence becomes:

$$\text{KL}(f||f_0) = \sum_{m=1}^M \sum_{j=1}^{2^m} (\log Y_{mj}) \prod_{k=1}^m Y_{m-k+1, j_{m-k+1}^{(m,j)}} + M \log 2. \quad (6)$$

We now present some results that characterize the first two moments of these divergences.

PROPOSITION 3.1 *Let $F \sim \mathcal{PT}_M(\alpha, \rho, F_0)$. Then the Kullback-Leibler divergence between f_0 and f , defined in (5), has mean and variance given by*

$$E\{\text{KL}(f_0||f)\} = \sum_{m=1}^M \{\psi_0(2\alpha\rho(m)) - \psi_0(\alpha\rho(m)) - \log 2\}$$

and

$$\text{Var}\{\text{KL}(f_0||f)\} = \sum_{m=1}^M \frac{1}{2^m} \{\psi_1(\alpha\rho(m)) - 2\psi_1(2\alpha\rho(m))\},$$

where $\psi_0(\cdot)$ and $\psi_1(\cdot)$ denote the digamma and trigamma functions respectively ¹.

Proof. The expected value follows by noting that the geometric mean of a beta random variable is $E(\log Y_{mj}) = \psi_0(2\alpha\rho(m)) - \psi_0(\alpha\rho(m))$. For the variance, we use the fact that the random variables Y_{mj} are independent across m , and for the same m , Y_{mj} and Y_{mk} are independent for $|k-j| > 1$. Noting that $\text{Var}(\log Y_{mj}) = \psi_1(\alpha\rho(m)) - \psi_1(2\alpha\rho(m))$ and since $Y_{m,2j} = 1 - Y_{m,2j-1}$, for $j = 1, \dots, 2^{m-1}$, with $\text{Cov}\{\log Y_{m,2j-1}, \log(1 - Y_{m,2j})\} = -\psi_1(2\alpha\rho(m))$, the result follows. ■

¹The digamma function is defined as the logarithmic derivative of the gamma function, i.e. $\psi_0(x) = \frac{d}{dx} \log \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$. In similar fashion, the trigamma function is defined as the second derivative.

We now concentrate on the limiting behaviour of the expected KL value as a function of the finite tree level M . For some cases of the function $\rho(\cdot)$ this limit is finite. This is given in the following corollary.

COROLLARY 3.2 *Let $\mathcal{E}_M := E\{\text{KL}(f_0||f)\}$ be the expectation defined in Proposition 3.1 with explicit dependence on the truncation level M . The limit of the expected KL divergence, as $M \rightarrow \infty$, is bounded by:*

$$\lim_{M \rightarrow \infty} \mathcal{E}_M \leq \frac{1}{4\alpha} \zeta(\delta) + \frac{1}{\alpha^2} \zeta(\delta^2)$$

for $\rho_3(m)$ with $\zeta(\delta) = \sum_{n=1}^{\infty} n^{-\delta}$ the Riemann zeta function, and

$$\lim_{M \rightarrow \infty} \mathcal{E}_M \leq \frac{\alpha(\delta + 1) + 4}{4\alpha^2(\delta^2 - 1)}$$

for $\rho_4(m)$, where the ρ functions and δ are defined in (3).

Proof. The digamma function can be expanded as: $\psi_0(x) = \log x - (1/2)x^{-1} - \mathcal{O}(x^{-2})$, from which these inequalities follow. ■

Taking instead the reverse KL, we have the following properties.

PROPOSITION 3.3 *Let $F \sim \mathcal{PT}_M(\alpha, \rho, F_0)$. Then the Kullback-Leibler divergence between f and f_0 , defined in (6), has mean and variance given by*

$$E\{\text{KL}(f||f_0)\} = \sum_{m=1}^M \{\psi_0(\alpha\rho(m) + 1) - \psi_0(2\alpha\rho(m) + 1) + \log 2\}$$

and

$$\text{Var}\{\text{KL}(f||f_0)\} = A + B,$$

where

$$A = \sum_{m=1}^M \left[\left\{ \prod_{k=1}^m \left(\frac{\alpha\rho(k) + 1}{2\alpha\rho(k) + 1} \right) \right\} \lambda_5(m) - \left(\frac{1}{2} \right)^m \lambda_2^2(m) \right],$$

$$B = \sum_{m=1}^M \left(\left(\frac{\alpha\rho(m)}{2\alpha\rho(m) + 1} \right) \left\{ \prod_{k=1}^{m-1} \left(\frac{\alpha\rho(k) + 1}{2\alpha\rho(k) + 1} \right) \right\} \lambda_6(m) - \left(\frac{1}{2} \right)^m \lambda_2^2(m) \right.$$

$$\left. + \sum_{j=1}^{m-1} \left[\left(\frac{\alpha\rho(j)}{2\alpha\rho(j) + 1} \right) \left\{ \prod_{k=1}^{j-1} \left(\frac{\alpha\rho(k) + 1}{2\alpha\rho(k) + 1} \right) \right\} \lambda_2^2(m) - \left(\frac{1}{2} \right)^j \lambda_2^2(m) \right] \right)$$

$$+2 \left\{ \prod_{k=1}^{m-1} \left(\frac{\alpha\rho(k)+1}{2\alpha\rho(k)+1} \right) \right\} \sum_{j=m+1}^M \left\{ \left(\frac{\alpha\rho(m)+1}{2\alpha\rho(m)+1} \right) \lambda_3(m) \lambda_2(j) \right. \\ \left. + \left(\frac{\alpha\rho(m)}{2\alpha\rho(m)+1} \right) \lambda_4(m) \lambda_2(j) - \lambda_2(m) \lambda_2(j) \right\}$$

with

$$\lambda_2(m) = \psi_0(\alpha\rho(m)+1) - \psi_0(2\alpha\rho(m)+1),$$

$$\lambda_3(m) = \psi_0(\alpha\rho(m)+2) - \psi_0(2\alpha\rho(m)+2),$$

$$\lambda_4(m) = \psi_0(\alpha\rho(m)+1) - \psi_0(2\alpha\rho(m)+2),$$

$$\lambda_5(m) = \psi_1(\alpha\rho(m)+2) - \psi_1(2\alpha\rho(m)+2) + \{\psi_0(\alpha\rho(m)+2) - \psi_0(2\alpha\rho(m)+2)\}^2,$$

$$\lambda_6(m) = \{\psi_0(\alpha\rho(m)+1) - \psi_0(2\alpha\rho(m)+2)\}^2 - \psi_1(2\alpha\rho(m)+2).$$

Proof. The expected value follows by using independence properties and by noting that $E\{(\log Y_{mj})Y_{mj}\} = \lambda_2(m)/2$. For the variance, we first bring the variance operator within the sum by splitting it into the sum of variances of each element plus the sum of covariances². ■

Figures 1 and 2 respectively illustrate the behaviour of the mean and standard deviation, as a function of the truncation level M for the two KL measures (5) (empty dots) and (6) (solid dots) with random densities sampled from a Pólya tree. The four panels in each figure correspond to choices of $\rho(m) = 1/2^m, 1, m^\delta, \delta^m$, as given in (3). In all cases we use $\alpha = 1$, and $\delta = 2$ (the so-called canonical choice). The plots show that $E\{\text{KL}(f_0||f)\} \geq E\{KL(f||f_0)\}$ for all M and for all functions ρ . Apart from the singular continuous case, $\rho_2(m) = 1$, the variances of $\text{KL}(f_0||f)$ are also larger than those of $\text{KL}(f||f_0)$.

We see that for the case of $\rho_1(m) = 1/2^m$, which corresponds to the Dirichlet process, the mean value of the KL and the reverse KL diverge to infinity as $M \rightarrow \infty$ ³. The KL (5) increases at an exponential rate whereas for the reverse KL (6) the growth rate is constant. As for the standard deviations, that of the KL also diverges as $M \rightarrow \infty$, however, that of the reverse KL converges.

The precision function $\rho_2(m) = 1$, which defines a singular continuous random distribution [16], has asymptotic constant expected values for both KL and reverse KL in

²The variance of each element is defined in terms of first and second moments and rely on independence properties to compute them. Working out the algebra with patience and noting that $E\{(\log Y_{mj})Y_{mj}\} = \lambda_2(m)/2$, $E\{(\log Y_{mj})Y_{mj}^2\} = \frac{1}{2} \left(\frac{\alpha\rho(m)+1}{2\alpha\rho(m)+1} \right) \lambda_3(m)$, $E\{(\log Y_{mj})Y_{mj}(1 - Y_{mj})\} = \frac{1}{2} \left(\frac{\alpha\rho(m)}{2\alpha\rho(m)+1} \right) \lambda_4(m)$, $E\{(\log Y_{mj})^2 Y_{mj}^2\} = \frac{1}{2} \left(\frac{\alpha\rho(m)+1}{2\alpha\rho(m)+1} \right) \lambda_5(m)$, and $E\{(\log Y_{mj}) \log(1 - Y_{mj}) Y_{mj}(1 - Y_{mj})\} = \frac{1}{2} \left(\frac{\alpha\rho(m)}{2\alpha\rho(m)+1} \right) \lambda_6(m)$, the result is obtained.

³Figure 1 appears to show that $E\{\text{KL}(f||f_0)\}$ remains constant, but this is an artefact due to the scale.

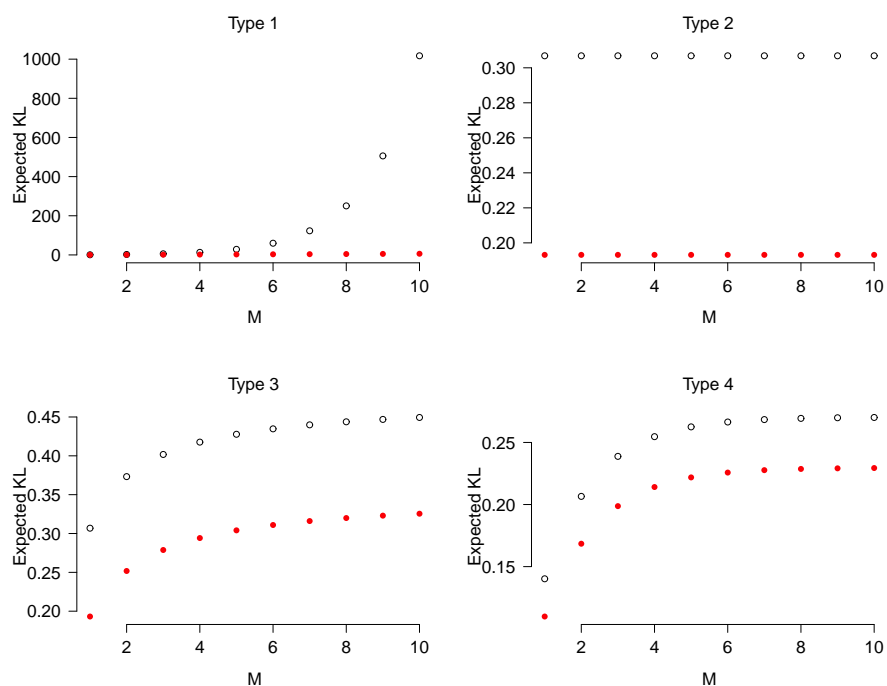


Figure 1. Expected values of KL divergence when f is sampled from a Pólya tree centred at f_0 for different truncation values M . $E\{KL(f_0||f)\}$ (empty dots) and $E\{KL(f||f_0)\}$ (solid dots). Types 1 to 4 denote the different ρ functions as in (3), with $\alpha = 1$ and $\delta = 2$.

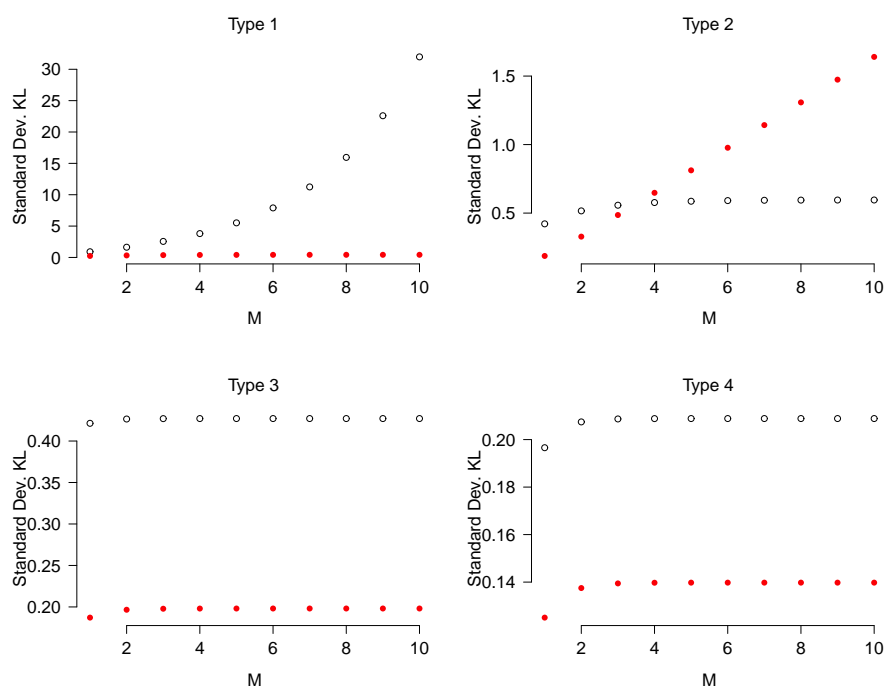


Figure 2. Standard deviations (SD) of KL divergence when f is sampled from a Pólya tree centred at f_0 for different truncation values M . $SD\{KL(f_0||f)\}$ (empty dots) and $SD\{KL(f||f_0)\}$ (solid dots). Types 1 to 4 denote the different ρ functions as in (3), with $\alpha = 1$ and $\delta = 2$.

the limit of M . The variance of the KL converges to a finite value as $M \rightarrow \infty$, but for the reverse KL the variance increases at a constant rate as a function of M . In the case of the two continuous processes, obtained with precision functions ρ_3 and ρ_4 , the expected values for KL and the reverse KL converge in the limit as shown from the upper bounds given in Corollary 3.2. Interestingly, the variances for the two KL divergences are asymptotically constant, with practically the same value for $M \geq 2$.

These results can be used to select the appropriate parametrisation of a Pólya tree prior when the precision function is $\rho_3(m) = m^\delta$. For this choice of precision function, the accepted method for choosing the values of δ and α is to first fix $\delta = 2$, and then vary α to completely control the variety of draws from the process. However, these two parameters are confounded and should not be chosen independently. For a given level of truncation M , a choice of δ and α will determine how concentrated the Pólya tree process is around the baseline distribution. Higher dispersion can be achieved by lower values of both parameters, although the variance of the draws in terms of KL is mainly governed by α . This is shown in Figure 3 which plots both the log-expected KL (black lines) and the log-variance of the expected KL (dashed lines) as a function of both parameters δ and α . This plot shows the dependence between the two parameters and we see that the δ parameter has a greater effect on the expected KL value of the random draws as α is increased.

Calibrating the KL divergence is not easy. A simple approach [e.g. 6, 19] consists in taking f_0 and f to be Bernoulli densities with probability of success $1/2$ and π respectively. If $\pi = 0.99$ (considerably furthest away from $1/2$) $\text{KL}(f_0||f) = 1.614$ which in log scale becomes approximately 0.5. Looking at Figure 3, appropriate choice of (δ, α) values in the range $\delta \in (1, 2)$ and $\alpha \in (0.30, 0.55)$ produce a log expected KL of 0.5, however a δ close to 1 and an α close to 0.55 would cause draws with lower variance. This suggests the pair $\delta = 1.01$ and $\alpha = 0.55$, for $M = 10$, to be a reasonable prior choice. In general, we hope Figure 3 provides a graphical tool which helps to choose a parametrisation (α, δ) of a Pólya tree process.

3.2. Frequentist and Bayesian Bootstrap

Let us now consider the setting where f_0 is a discrete density with n atoms $\{\xi_1, \dots, \xi_n\}$, i.e., $f_0(x) = \sum_{i=1}^n p_i \delta_{\xi_i}(x)$, with $p_i > 0$ for all $i = 1, \dots, n$ and $\sum_{i=1}^n p_i = 1$. Let $\mathbf{w} = (w_1, \dots, w_n)$ be random weights such that $w_i \geq 0$ and $\sum_{i=1}^n w_i = 1$ almost surely. Let f be a random density defined as a reweighing of the atoms of f_0 with the random weights \mathbf{w} . In notation, $f(x) = \sum_{i=1}^n w_i \delta_{\xi_i}(x)$.

The Kullback-Leibler divergence between f_0 and f does not depend on the atom locations and is given by:

$$\text{KL}(f_0||f) = \sum_{i=1}^n p_i \log \left(\frac{p_i}{w_i} \right), \quad (7)$$

and the reverse Kullback-Leibler has the form

$$\text{KL}(f||f_0) = \sum_{i=1}^n w_i \log \left(\frac{w_i}{p_i} \right). \quad (8)$$

Of particular interest is the setting where we have a random sample X drawn from a population (for example in the context of Bayesian inference, where X is a Monte Carlo

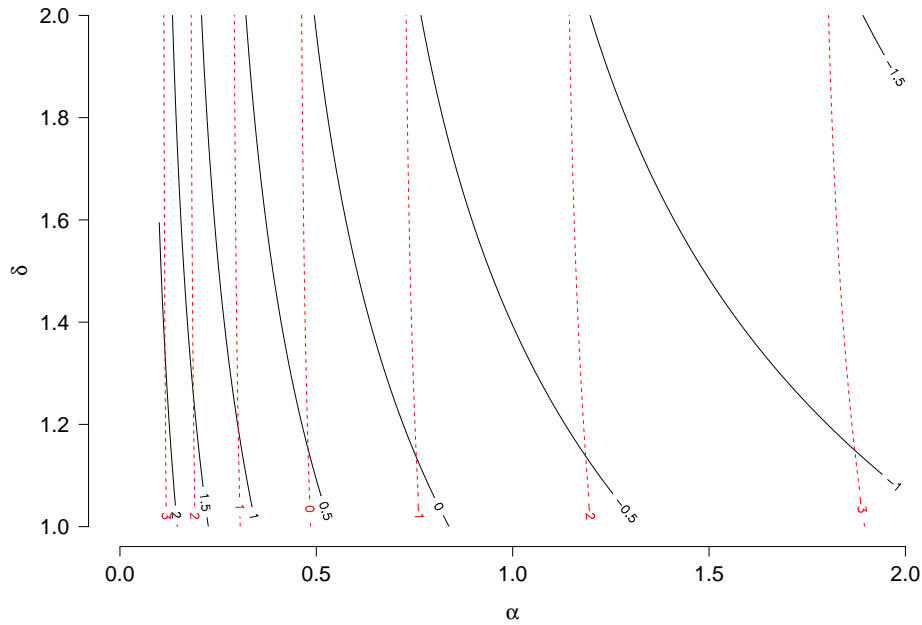


Figure 3. Overlaid contour plots of the log-expected KL (black lines) and the log-variance of KL (dashed red lines), given in Proposition 3.1, as functions of the two parameters (α, δ) , at regular intervals of $1/2$ and 1 respectively, of draws from a Pólya tree process with truncation level $M = 10$.

sample drawn from some posterior distribution) and where each atom has equal weight. Here f_0 would represent the ‘best guess’ of the true underlying posterior. Embedding f_0 into a nonparametric distribution allows for a formal assessment of the uncertainty of the sampling procedure in X . This is the motivation for the bootstrap methods. As noted before, the KL divergence can be used to calibrate this nonparametric embedding.

In this setting we first highlight an important property in the relationship between the divergences (7) and (8).

PROPOSITION 3.4 *Consider the KL divergences (7) and (8). If $p_i = 1/n$ for $i = 1, \dots, n$, then for any given re-weighting vector \mathbf{w} taken from the simplex $\mathcal{Q}_n := \{w : w_i \geq 0, \sum_{i=1}^n w_i = 1\}$ we have that*

$$\text{KL}(f_0||f) \geq \text{KL}(f||f_0).$$

Proof. Let $h(\mathbf{w}) := \text{KL}(f_0||f) - \text{KL}(f||f_0)$. Using expressions (7) and (8) we have $h(\mathbf{w}) = -\sum (1/n + w_i) \log(nw_i)$. Note that $h(\mathbf{w}^*) = 0$ for $\mathbf{w}^* = (1/n, \dots, 1/n)$, h is a convex function with $h''(\mathbf{w}^*)$ is positive, such that $h(\mathbf{w})$ is a strictly increasing function away from \mathbf{w}^* . Therefore the local minimum at $\mathbf{w}^* = (1/n, \dots, 1/n)$ is the global minimum and the result follows. ■

This result is consistent with the results from previous section. However, in this particular discrete setting $\text{KL}(f_0||f)$ dominates $\text{KL}(f||f_0)$.

Taking for instance $n\mathbf{w} \sim \text{Mult}(n, \mathbf{p})$ ⁴, a multinomial distribution with n trials and n categories with probability of success $\mathbf{p} = (p_1, \dots, p_n)$, means that the random f ’s will

⁴We use this notation to emphasise the fact that \mathbf{w} represents a random probability mass function, but taking

be centred at f_0 . It is not difficult to show that $E(f) = f_0$. Note that if $p_i = 1/n$ for $i = 1, \dots, n$ this choice of distribution for the weights \mathbf{w} coincides with the frequentist bootstrap [20] for which the atoms $\{\xi_i\}$ are replaced by i.i.d. random variables $\{X_i\}$.

We note that the KL divergence (7) will not in general be defined, as w_i can be zero. In fact, for large n and for $p_i = 1/n$ in the previous multinomial choice, approximately one third of the weights will be zero. However, $0 \log 0$ is defined by convention as 0, so the reverse KL (8) is well defined.

PROPOSITION 3.5 *The expected value of the Kullback-Leibler between a “bootstrap” draw f , with $n\mathbf{w} \sim \text{Mult}(n, \mathbf{p})$, and its centring distribution f_0 , defined in (8), has the following upper bound:*

$$E\{\text{KL}(f||f_0)\} \leq \sum_{i=1}^n p_i \log \left(p_i + \frac{1-p_i}{n} \right) - H(\mathbf{p})$$

where $H(\mathbf{p}) = \sum_{i=1}^n p_i \log p_i$, the entropy of the vector \mathbf{p} . For the special case when $p_i = 1/n$, we have $E\{\text{KL}(f||f_0)\} \leq \log(2 - 1/n) \leq \log 2$

Proof.

$$E\{\text{KL}(f||f_0)\} = \sum_{i=1}^n E\{w_i \log w_i\} - \sum_{i=1}^n E\{w_i\} \log p_i.$$

Working on the individual expected values,

$$E(w_i \log w_i) = \sum_{k=1}^n \binom{n}{k} p_i^k (1-p_i)^{n-k} \left(\frac{k}{n} \right) \log \left(\frac{k}{n} \right).$$

From which we get $E(w_i \log w_i) = p_i E\{\log((v_i + 1)/n)\}$, with $v_i \sim \text{Bin}(n-1, p_i)$. Using Jensen's inequality we get $E\{w_i \log w_i\} \leq p_i \log(p_i + (1-p_i)/n)$. Substituting this into the original sum and using $E\{w_i\} = p_i$ gives the result. ■

This upper bound provides insight into the relationship between Kullback-Leibler divergence and the bootstrap as a function of sample size n and may assist in the calibration and interpretation of the KL divergence between discrete distributions. KL is used in many applications of statistics [see for example 6, 21–23], however its calibration remains an open problem [15].

An alternative way of making the random f 's to be centered around f_0 is by sampling weights \mathbf{w} from a Dirichlet distribution with parameter vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n)$ such that $\beta_i = \alpha_n p_i$, $i = 1, \dots, n$, with $\alpha_n > 0$ a parameter changing as a function of the number of atoms. This is denoted $\mathbf{w} \sim \text{Dir}(\alpha_n \mathbf{p})$. It is straightforward to prove that $E(f) = f_0$, and that the form of α_n parametrises the precision, analogous to the Pólya tree case. If we take $\alpha_n = n$, $p_i = 1/n$ and replace the atoms $\{\xi_i\}$ by i.i.d. random variables $\{X_i\}$, we obtain the original Bayesian bootstrap proposed by Rubin [24]⁵. Sampling from a Dirichlet with parameter vector $\alpha_n \mathbf{p}$ gives a generalised version of this bootstrap procedure. Ishwaran & Zarepour considered this model albeit in a different context [25]. In this new setting,

values on the set $\{0, 1/n, 2/n, \dots, 1\}$. A factor of n is needed for the vector to be distributed according to a multinomial distribution.

⁵It is interesting to note that in the original work they only consider this special case.

both the $\text{KL}(f_0||f)$ and the reverse $\text{KL}(f||f_0)$, given in (7) and (8) respectively, are well defined since $w_i \neq 0$ almost surely. Their expected values and variances can be obtained in closed form as functions of α_n and \mathbf{p} .

PROPOSITION 3.6 *Let f be a “generalised Bayesian bootstrap” draw around f_0 with weights $\mathbf{w} \sim \text{Dir}(\alpha_n \mathbf{p})$. Then the Kullback-Leibler divergence given in (7) has mean and variance:*

$$E\{\text{KL}(f_0||f)\} = H(\mathbf{p}) - \sum_{i=1}^n p_i \{\psi_0(\alpha_n p_i) - \psi_0(\alpha_n)\}$$

$$\text{Var}\{\text{KL}(f_0||f)\} = \sum_{i=1}^n p_i^2 \psi_1(\alpha_n p_i) - \psi_1(\alpha_n)$$

where ψ_0 and ψ_1 are the digamma and trigamma functions.

Proof. This result follows from $E(\log w_i) = \psi_0(\alpha_n p_i) - \psi_0(\alpha_n)$ and linearity of expectation. The variance follows from $\text{Var}(\log w_i) = \psi_1(\alpha_n p_i) - \psi_1(\alpha_n)$, and $\text{Cov}(\log w_i, \log w_j) = \psi_1(\alpha_n p_i) \delta_{ij} - \psi_1(\alpha_n)$, where δ_{ij} is the Kronecker delta function taking value 1 when $i = j$ and 0 otherwise. ■

The limiting behaviour of the expected KL and its variance of Proposition 3.6, as n tends to infinity, can more easily be studied for the special case of $p_i = 1/n$, $i = 1, \dots, n$. When $\alpha_n = \alpha$, i.e. constant, they both diverge to infinity. In the limit, this is a well known construction of a Dirichlet process, when the atoms are sampled i.i.d. from a baseline measure G . However, if we make α_n grow linearly with n , say $\alpha_n = \alpha n$, then $\lim_{n \rightarrow \infty} E\{\text{KL}(f_0||f)\} = \log(\alpha) - \psi_0(\alpha)$ and $\lim_{n \rightarrow \infty} \text{Var}\{\text{KL}(f_0||f)\} = 0$. These values are obtained by noting that $\psi_0(n)$ behaves like $\log(n)$ for large n . Finally, if we increase the rate at which α_n grows with n , say $\alpha_n = \alpha n^2$, both mean and variance of the KL converge to zero as $n \rightarrow \infty$.

PROPOSITION 3.7 *Let f be a “generalised Bayesian bootstrap” draw around f_0 with weights $\mathbf{w} \sim \text{Dir}(\alpha_n \mathbf{p})$. Then the Kullback-Leibler divergence given in (8) has mean:*

$$E\{\text{KL}(f||f_0)\} = \sum_{i=1}^n p_i \{\psi_0(\alpha_n p_i + 1) - \psi_0(\alpha_n + 1)\} - H(\mathbf{p})$$

where $H(\mathbf{p}) := \sum_{i=1}^n p_i \log p_i$ the entropy of the vector \mathbf{p} , and the variance given by

$$\begin{aligned} \text{Var}\{\text{KL}(f||f_0)\} = & \sum_{i=1}^n \{ \text{Var}(w_i \log w_i) + (\log p_i)^2 \text{Var}(w_i) - 2(\log p_i) \text{Cov}(w_i \log w_i, w_i) \} \\ & + 2 \sum_{i < j} \{ \text{Cov}(w_i \log w_i, w_j \log w_j) + (\log p_i)(\log p_j) \text{Cov}(w_i, w_j) \\ & - 2(\log p_j) \text{Cov}(w_i \log w_i, w_j) \} \end{aligned}$$

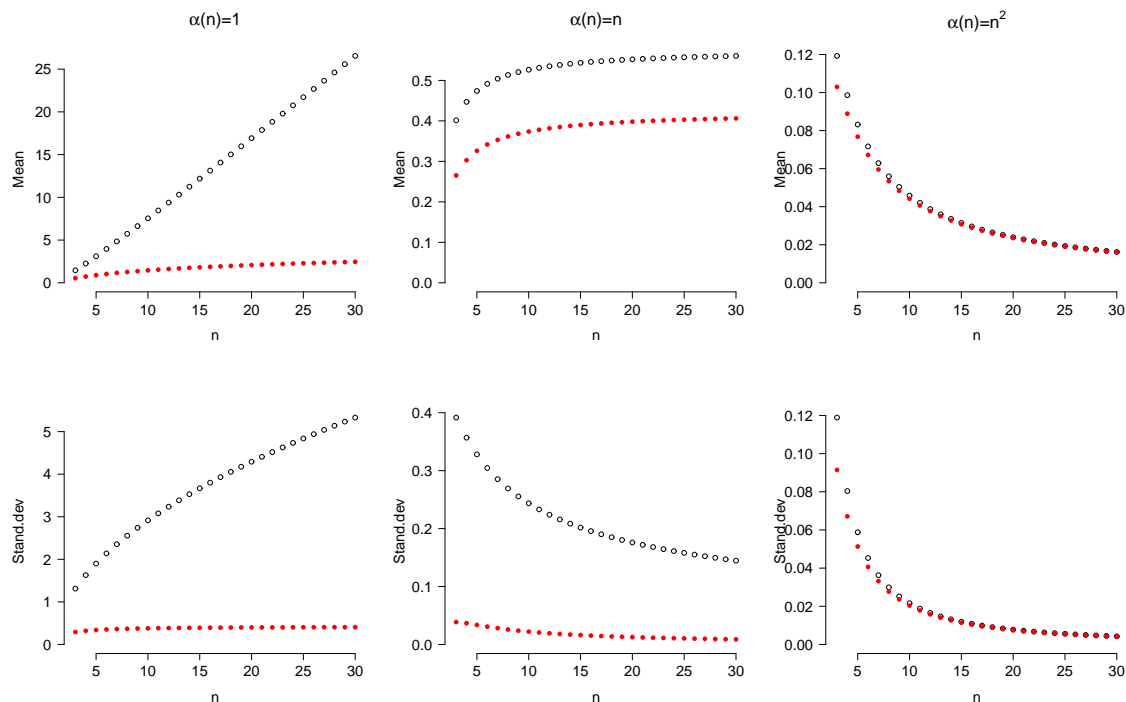


Figure 4. Expected values (top row) and standard deviations (bottom row) of $\text{KL}(f_0||f)$ (black empty dots) and $\text{KL}(f||f_0)$ (red solid dots), when f is sampled from a generalized Bayesian bootstrap with $p_i = 1/n$. In columns from left to right: $\alpha_n = 1$, $\alpha_n = n$ and $\alpha_n = n^2$.

where each of the elements are given in the footnote ⁶.

Proof. Note that each $w_i \sim \text{Be}\{\alpha_n p_i, \alpha_n(1 - p_i)\}$ and thus we have that $E(w_i \log w_i) = p_i\{\psi_0(\alpha_n p_i + 1) - \psi_0(\alpha_n + 1)\}$. Using linearity of expectation and substituting this expression we obtain the mean. Using properties of the variance and covariance of sums we get the second part of the result. ■

Similarly to the previous case, if we take $p_i = 1/n$ and $\alpha_n = \alpha n$ in Proposition 3.7, when $n \rightarrow \infty$ then $E\{\text{KL}(f||f_0)\} \rightarrow \psi_0(\alpha + 1) - \log(\alpha)$. It is also possible to show analytically that each term in the variance goes to zero as $n \rightarrow \infty$, but this can also be seen using the relation between the two KLs given in Proposition 3.4, and noting that the variance involves a monotonic transformation, hence we have that $\text{Var}\{\text{KL}(f_0||f)\} \geq \text{Var}\{\text{KL}(f||f_0)\}$. From the previous result it follows that $\lim_{n \rightarrow \infty} \text{Var}\{\text{KL}(f||f_0)\} = 0$ for these choices of p_i and α_n .

In Figure 4 we compare the expected value (top row) and standard deviation (bottom row) of both KL and reverse KL, in the generalized Bayesian bootstrap, for $p_i = 1/n$ and different values of α_n as a function of n . The first column corresponds to $\alpha_n = 1$, the second column to $\alpha_n = n$ and the third to $\alpha_n = n^2$, which induce high, moderate and small variance in the \mathbf{w} respectively. In accordance to what we have proved, the expected value and variance of $\text{KL}(f_0||f)$ are larger than those of $\text{KL}(f||f_0)$, and their

⁶ $\text{Var}(w_i) = p_i(1 - p_i)/(\alpha_n + 1)$, $\text{Cov}(w_i, w_j) = -p_i p_j / (\alpha_n + 1)$, $\text{Var}(w_i \log w_i) = p_i(\alpha_n p_i + 1)/(\alpha_n + 1)\{\psi_1(\alpha_n p_i + 2) - \psi_1(\alpha_n + 2) + [\psi_0(\alpha_n p_i + 2) - \psi_0(\alpha_n + 2)]^2\} - p_i^2\{\psi_0(\alpha_n p_i + 1) - \psi_0(\alpha_n + 1)\}^2$, $\text{Cov}(w_i \log w_i, w_i) = p_i(\alpha_n p_i + 1)/(\alpha_n + 1)\{\psi_0(\alpha_n p_i + 2) - \psi_0(\alpha_n + 2)\} - p_i^2\{\psi_0(\alpha_n p_i + 1) - \psi_0(\alpha_n + 1)\}$, $\text{Cov}(w_i \log w_i, w_j) = p_i p_j \{-\psi_0(\alpha_n p_i + 1)/(\alpha_n + 1) + \psi_0(\alpha_n + 1) - \alpha_n \psi_0(\alpha_n + 2)/(\alpha_n + 1)\}$, $\text{Cov}(w_i \log w_i, w_j \log w_j) = \alpha_n p_i p_j / (\alpha_n + 1)[\{\psi_0(\alpha_n p_i + 1) - \psi_0(\alpha_n + 2)\}\{\psi_0(\alpha_n p_j + 1) - \psi_0(\alpha_n + 2)\} - \psi_1(\alpha_n + 2) - p_i p_j \{\psi_0(\alpha_n p_i + 1) - \psi_0(\alpha_n + 1)\}\{\psi_0(\alpha_n p_j + 1) - \psi_0(\alpha_n + 1)\}]$.

limiting behaviours can also be assessed from the graphs. In particular, for $\alpha_n = n$ (middle column) the expected KL and reverse KL are asymptotically finite, whereas the corresponding variances go both to zero as $n \rightarrow \infty$. This plot provides a graphical visualisation of the effect of the parametrisation of the generalised Bayesian bootstrap procedure in terms of the KL divergence and its variance. In practice the value n would be fixed (for example the number of samples from a Monte Carlo simulation) and therefore the KL of bootstrapped draws would depend only on α and the exponent given to n .

If we replace the atoms $\{\xi_i\}$ by i.i.d. random variables $\{X_i\}$ from a distribution G and take $p_i = 1/n$ for $i = 1, \dots, n$, then f_0 represents the empirical density for the random variables X_i 's and f represents a random process centered around the empirical. [25] considered exactly this random probability process and derived results for the limiting behaviour for a variety of choices of α_n (see Theorem 3, page 948). Let F be the cdf associated to f . When $\alpha_n = \alpha$, then F is distributed according to a Dirichlet process $\mathcal{DP}(\alpha, G)$, in the limit as $n \rightarrow \infty$. If $\alpha_n = \alpha n$, then we have almost sure weak convergence of F to G , as $n \rightarrow \infty$. For the third case considered here, $\alpha_n = \alpha n^2$, F converges in probability to G , as $n \rightarrow \infty$.

The case where $\alpha_n = \alpha n$ is of particular interest. Although we have weak convergence of $F \rightarrow G$ as $n \rightarrow \infty$, the random distribution does not converge in KL divergence (see top middle panel in Figure 4). In other words, although functionals of f tend to the functionals of f_0 , the KL divergence between the two densities remains non zero. This becomes apparent when considering the random quantity nw_i , which comes into the equation (7), whose variance becomes asymptotically $1/\alpha$, as $n \rightarrow \infty$. Convergence in Kullback-Leibler is a strong statement, stronger than convergence of functionals and L_1 convergence. A more intuitive illustration is the posterior convergence of two Dirichlet processes with different baseline measures (that have the same support). By posterior consistency, both will weakly converge to the same measure, but their L_1 divergence will remain finite and their KL divergence will remain infinite.

4. Discussion

This note explores properties of the KL and reverse KL of draws F from some random probability models with respect to their centring distribution F_0 . These properties become relevant when applying a particular process as a modelling tool. For example, draws from the Dirichlet process prior have divergent expected KL (obtained in our Pólya tree setting with ρ_1 in (3) and $M \rightarrow \infty$, and also obtained in the Bayesian bootstrap setting with $\alpha_n = \alpha$ and $n \rightarrow \infty$). This is somewhat of a surprising result but in accordance with the full support (in the weak topology) property of the Dirichlet process⁷.

Our key result concerns the Pólya tree prior. In the majority of applications, it is usually constructed in its continuous version, i.e. the precision function ρ satisfies the continuity property, for example ρ_3 and ρ_4 as given in (3). In these cases, the first two moments of the distance (in KL units) of draws from their centring measure is given as an explicit function of the truncation level M and the precision function $\rho(m)$ and precision parameter α (see Propositions 3.1 and 3.3). Therefore the specification of M , α , and ρ are all highly important, where a careless choice may lead to a prior overly concentrated around f_0 . The vast majority of applications with Pólya tree priors use the family $\rho_3(m)$ with choice of exponent $\delta = 2$.

⁷Proposition 3 of Ferguson [4] states that any fixed density (measure) g absolutely continuous with respect to f_0 can be arbitrarily approximated pointwise with a draw f from a Dirichlet process.

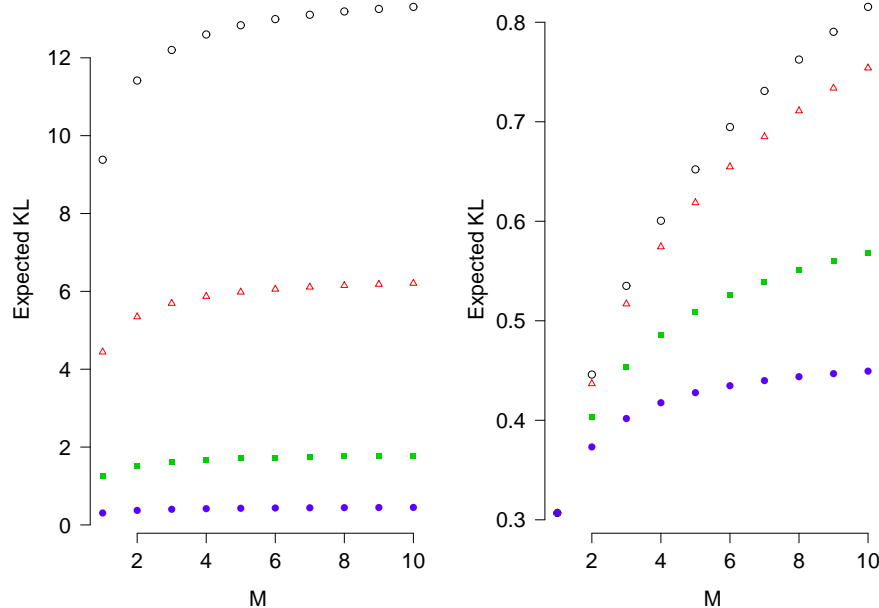


Figure 5. Expected $KL(f_0||f)$ for f sampled from a Pólya tree centred at f_0 with parameters α and $\rho(m) = m^\delta$. Varying α for fixed $\delta = 2$ (left panel), and varying δ for fixed $\alpha = 1$ (right panel). Left: $\alpha = 0.05$ (empty dots), $\alpha = 0.1$ (triangles), $\alpha = 0.3$ (squares), $\alpha = 1$ (solid dots). Right: $\delta = 1.01$ (empty dots), $\delta = 1.1$ (triangles), $\delta = 1.5$ (squares), $\delta = 2$ (solid dots). The solid dots in both panels correspond to the same values.

In Figure 5 we show that as M is increased, using $\rho_3(m)$ with a choice of δ closer to 1 (right panel) gives greater gains in expected KL than those obtained for the standard choice of $\delta = 2$ and only decreasing the parameter α (left panel). The concentration around the baseline measure is highly sensitive to this choice of exponent, thus questioning the “sensible canonical choice” of $\delta = 2$ given by Lavine [5].

Moreover, in practice Pólya trees are used in their finite versions, that is, finite M . In such cases M is usually chosen with a rule of thumb [e.g. 26], say $M = \log_2(n)$ with n being the data sample size. The authors note ‘a law of diminishing returns’ when the truncation level is increased from $M \rightarrow M + 1$. Our study confirms this by plotting the diversity of draws as measured in KL against M , and these findings suggest that a Pólya tree prior with as a low as $M = 4$ and $\rho(m) = 2^m$ can produce random draws that are equally far from the centring distribution as with a larger M (see two bottom panels in Figures 1 and 2). If it desired to make proper use of finite nature of the tree, the various possibilities in specification of the precision function ρ within families that satisfy the continuity property should be used.

In the discrete setting, we can always see f_0 as the empirical density obtained from a sample of size n taken from a continuous density. This is often the case when characterising a posterior distribution in Bayesian analysis, for example via MCMC sampling [e.g. 27]. One lesson from this work, is that by increasing n , the variance of the reverse KL in the frequentist bootstrap, and the variance of the KL and reverse KL for the Bayesian bootstrap with $\alpha_n = \alpha n$, converge to zero. This implies that for large n a frequentist or Bayesian bootstrap draw lies below $\log(2)$ and exactly at $\log(\alpha) - \psi_0(\alpha)$ or $\psi_0(\alpha + 1) - \log(\alpha)$ in KL units, respectively (see also comments following Propositions

3.6 and 3.7).

Our work provides the first analytical results on the KL divergence of draws from RPMs around their centering distribution. In future work it would be interesting to characterise other divergence criteria such as the total-variation norm as well as the KL variation for other random probability measures such as the class of Dirichlet process mixture models [e.g. 28]. Extending our results to infinite mixture models would appear to present difficult technical challenges that potentially might only be solved approximately using computational methods.

Acknowledgements

We are grateful to Judith Rousseau for helpful comments. Watson is supported by the Industrial Doctoral Training Centre (SABS-IDC) at Oxford University and Hoffman-La Roche. This work was done whilst Nieto-Barajas was visiting the Department of Statistics at the University of Oxford. He is supported by CONACYT grant 244459 and *Asociación Mexicana de Cultura, A.C.*–Mexico. Holmes gratefully acknowledges support for this research from the Oxford-Man Institute, the EPSRC i-Like programme grant EP/K014463/1 and the Medical Research Council program leader’s award MC_UP_A390_1107.

References

- [1] Hjort N, Holmes C, Müller P, Walker S. Bayesian nonparametrics. Cambridge University Press; 2010.
- [2] Ghosh J, Ramamoorthi R. Bayesian nonparametrics. Vol. 1. Springer; 2003.
- [3] Müller P, Quintana F. Nonparametric Bayesian data analysis. *Statistical science*. 2004; 19(1):95–110.
- [4] Ferguson T. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*. 1973;:209–230.
- [5] Lavine M. Some aspects of Pólya tree distributions for statistical modelling. *The Annals of Statistics*. 1992;20(3):1222–1235.
- [6] Karabatsos G. Bayesian nonparametric model selection and model testing. *Journal of Mathematical Psychology*. 2006;50(2):123–148.
- [7] Muliere P, Walker S. A Bayesian non-parametric approach to survival analysis using Pólya trees. *Scandinavian Journal of Statistics*. 1997;24(3):331–340.
- [8] Walker SG, Damien P, Laud PW, Smith A. Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*. 1999;61(3):485–527.
- [9] Hanson T, Johnson W. Modeling regression error with a mixture of Pólya trees. *Journal of the American Statistical Association*. 2002;97(460):1020–1033.
- [10] Walker S, Mallick B. Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 1997;59(4):845–860.
- [11] Kullback S, Leibler R. On information and sufficiency. *The Annals of Mathematical Statistics*. 1951;22:79–86.
- [12] Kullback S. *Information theory and statistics*. Courier Dover Publications; 1997.
- [13] Bernardo J, Smith A. *Bayesian theory*. John Wiley & Sons; 1994.
- [14] Cover T, Thomas J. *Elements of information theory*. Wiley, New York; 1991.
- [15] Watson J, Holmes C. Approximate models and robust decisions. *Statistical Science*, to appear. 2016;.

- [16] Ferguson T. Prior distributions on spaces of probability measures. *The Annals of Statistics*. 1974;2(4):615–629.
- [17] Nieto-Barajas LE, Müller P. Rubbery Pólya tree. *Scandinavian Journal of Statistics*. 2012; 39(1):166–184.
- [18] Kraft C. A class of distribution function processes which have derivatives. *Journal of Applied Probability*. 1964;1(2):385–388.
- [19] McCulloch RE. Local model influence. *Journal of the American Statistical Association*. 1989; 84(406):473–478.
- [20] Efron B. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*. 1979; 7(1):1–26.
- [21] Hansen LP, Sargent TJ. Robust control and model uncertainty. *The American Economic Review*. 2001;91(2):60–66.
- [22] Breuer T, Csiszr I. Measuring distribution model risk. *Mathematical Finance*. 2016; 26(2):395–411.
- [23] Simpson DP, Martins TG, Riebler A, Fuglstad GA, Rue H, Sørbye SH. Penalising model component complexity: A principled, practical approach to constructing priors. *arXiv preprint arXiv:14034630*. 2014;.
- [24] Rubin D. The Bayesian bootstrap. *The Annals of Statistics*. 1981;9(1):130–134.
- [25] Ishwaran H, Zarepour M. Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*. 2002;12(3):941–963.
- [26] Hanson T. Inference for mixtures of finite Pólya tree models. *Journal of the American Statistical Association*. 2006;101(476):1548–1564.
- [27] Gelman A, Carlin J, Stern H, Dunson D, Vehtari A, Rubin D. *Bayesian data analysis*. Chapman and Hall; 2013.
- [28] Lo AY. On a class of Bayesian nonparametric estimates: I. Density estimates. *Annals of Statistics*. 1984;12:351–357.