ACS : 949
pp.1–29 (col.fig.: NIL)
PROD.TYPE: COM
3B2
ED: KAVYASHREE
PAGN: G.PRAKASH SCAN:

INTERNATIONAL JOURNAL OF ADAPTIVE CONTROL AND SIGNAL PROCESSING
Int. J. Adapt. Control Signal Process. 2007; **20**:000–000
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/acs.949

Learning for non-stationary Dirichlet processes

A. Quinn[1,*,†] and M. Kárný[2]

[1] *Department of Electronic and Electrical Engineering, Trinity College Dublin, Ireland*
[2] *Department of Adaptive Systems, Institute of Information and Automation, Academy of Sciences of the Czech Republic, Czech Republic*

SUMMARY

The Dirichlet process prior (DPP) is used to model an unknown probability distribution, F. This eliminates the need for parametric model assumptions, providing robustness in problems where there is significant model uncertainty. Two important parametric techniques for learning are extended to this non-parametric context for the first time. These are (i) *sequential stopping*, which proposes an optimal stopping time for on-line learning of F using i.i.d. sampling; and (ii) *stabilized forgetting*, which updates the DPP in response to changes in F, but without the need for a formal transition model. In each case, a practical and highly tractable algorithm is revealed, and simulation studies are reported. Copyright © 2007 John Wiley & Sons, Ltd.

Received 28 September 2005; Revised 4 November 2006; Accepted 9 November 2006

KEY WORDS: non-parametric process; non-parametric sequential stopping rule; non-stationary Dirichlet process; non-parametric stabilized forgetting

1. INTRODUCTION

All Bayesian methods for inference of an unknown quantity, x, require a probability distribution (i.e. model), $F(x)$, to be elicited. It can be difficult to propose such a model, and, if we do, resulting inferences and decisions may not be robust, in the sense that they may be affected greatly by modelling errors reflected in F. A parametric model, $F(x|\theta)$, involving a finite-dimensional unknown parameter, θ, is more flexible. Here, we elicit a prior on θ, and can therefore explore the set of distributions generated by the allowed values of θ. A countable

*Correspondence to: A. Quinn, Department of Electronic and Electrical Engineering, Trinity College Dublin, Ireland.
† E-mail: aquinn@tcd.ie

Contract/grant number: GA AVČR 1ET100750401
Contract/grant number: MŠ MTČR 1M6798555601

number of parametric models, $F_i(x|\theta_i)$, can also be compared [1], but, once again, there may be strong dependence on the choice of models, and their associated parameter priors, $F_i(\theta_i)$.

The classical non-parametric approach [2] is to build F only from the data, *via* the empirical distribution, $\sum_n \delta_{x_n}$, which places probability mass only at i.i.d. observations, x_n, of x. Kernel density estimation is concerned with finding smooth variants of this, *via* convolution for example. Other density estimation techniques—such as the maximum entropy (MaxEnt) method [3]—match selected moments of the empirical distribution, subject to some desirable regularization property (smoothness in the case of MaxEnt).

Bayesian non-parametrics [4–6] generalize the empirical approaches above by placing a probability distribution, $\mathscr{F}$, on the unknown distribution, F. Hence, a hierarchical modelling approach is taken

$$x \sim F \quad \text{and} \quad F \sim \mathscr{F}$$

Here, '$\sim$' denotes 'is distributed as'. F is the *non-parametric process* and $\mathscr{F}$ is the *non-parametric process prior*. Since, $F(\cdot|F) = F(\cdot)$, we may also view F as an infinite-dimensional 'parameter', generalizing the rôle of θ in parametric inference [7]. A special case is when x is finite-state (i.e. discrete) *a priori*, in which case F is expressible as a finite set of unknown probabilities, p. Then, $\mathscr{F} = F_p$ is a parametric prior on the 'parameters' p in the finite measurable simplex, Δ (see (4) in Section 2). The Dirichlet distribution, D, is the key example of F_p, occupying the important rôle of conjugate prior for independent, identically distributed (i.i.d.) sampling from multinomials [1]. Its generalization to a non-parametric measure on continuous probability distributions yields the Dirichlet process prior (DPP) [8], $\mathscr{D}$, which is the non-parametric model adopted in this paper. The Dirichlet process prior is favoured in the literature for its convenient property of being conjugate under i.i.d. sampling from the unknown measure, F.

Parametric modelling dominates in signal processing and control, and so it is not surprising that a number of important techniques for on-line learning have been framed only in the parametric context, relying, apparently, on prior elicitation of $F(x, \theta)$. The main purpose of this paper is to extend two important techniques for on-line learning to the non-parametric context. These are:

(1) to assess the convergence of $\mathscr{D} = \mathscr{D}_n$, under i.i.d. sampling, $\{x_1, \ldots, x_n\}$, from F, and to use this to design *sequential stopping rules* for i.i.d. sampling. This extends previous results on parametric stopping [9, 10];
(2) to track slowly non-stationary Dirichlet processes, F_t, *via* a tractable updating of our knowledge, expressed by $\mathscr{D}_t$. This is achieved by extending the *stabilized forgetting* procedure [11] to the non-parametric case.

In Section 2, we review those properties of the DPP which we will use later, notably its relationship to the Dirichlet distribution, as well as its moment properties. In Section 3, we address Problem 1 above, using Bayesian decision theory to design the stopping rules. Hence, the Kullback–Leibler Divergence (KLD) [12, 13] will be central to assessing convergence of $\mathscr{D}_n$. The Dirichlet process prior infers discrete distributions almost surely (a.s.), and so a technical difficulty arises in attempting to derive a partition-independent KLD. The problem is overcome *via* data-dependent partitioning (Section 3.3), and the appropriate algorithm (Algorithm 1) is presented. In Section 4, we address Problem 2 above, showing that (i) the stabilized forgetting operator may be applied to non-parametric priors, and, in particular, that (ii) the class of DPPs,

DOI: 10.1002/acs

$\mathscr{D}_t$, is closed under this operator. This leads to a very flexible and tractable algorithm for i.i.d. learning of a non-stationary Dirichlet process (Section 5). The stopping and tracking algorithms are explored in simulations in Section 6. Discussion of the potential scope of these non-parametric techniques follows in Section 7. Overall conclusions (Section 7) close the paper.

2. THE DIRICHLET PROCESS, $F \sim \mathscr{D}(\hat{F}_0, v_0)$

Let X^* be the space of elementary events, x, and let $\mathbb{A}$ be a σ-algebra of subsets of X^*, being measurable events for x. Let F be the *unknown* probability distribution (measure) on the measurable space, $(X^*, \mathbb{A})$, of x. Consider the special case when $(X^*, \mathbb{A}) = (\mathbb{R}^m, \mathbb{B})$, where $\mathbb{B}$ is the σ-algebra of Borel subsets of $\mathbb{R}^m$. Then the distribution of $x \in \mathbb{R}^m$ may be specified by $F \equiv F(x)$, which we use to denote either the unknown probability (density) function (p.(d.)f.) or cumulative distribution function of x [14], with the context making clear which is meant.

In general, F is a non-parametric stochastic process [8], whose measurable space is $(F^*, \mathbb{A}_F)$. Hence, F^* is a function space in the case $X^* = \mathbb{R}^m$ above. Details of the required σ-algebra, $\mathbb{A}_F$, of F^* may be found in [7] or [8]. Let $\mathscr{F}$ be a probability distribution on $(F^*, \mathbb{A}_F)$, being an appropriate prior for the non-parametric process, F. The distribution, $\mathscr{F}$, is defined by specifying the distribution of the finite set of unknown probabilities, $(F(A_1), \ldots, F(A_q))$, induced by F on *every* finite set of pairwise disjoint sets, $A_i \in \mathbb{A}$. The conditions under which $\mathscr{F}$ is *uniquely* defined are given, for example, in Theorem 1 of [6].

In this paper, we will employ the DPP as our non-parametric prior

$$F \sim \mathscr{D}(\hat{F}_0, v_0) \equiv \mathscr{D}_0 \tag{1}$$

Here, the unknown distribution, F, is the *Dirichlet process*, $\hat{F}_0$ is an arbitrary known probability measure on $(X^*, \mathbb{A})$, and $0 < v_0 < \infty$ is a known real scalar. The rôle of the subscripts, '0', will emerge in Section 3. Qualitatively, (i) $\mathscr{D}_0$ places mass on a space of distributions 'centred' on $\hat{F}_0$, with the mass concentrating on $\hat{F}_0$ as v_0 increases; and (ii) for every finite measurable partition of X^* (defined below), the unknown probabilities, p, induced by F have the (parametric) Dirichlet distribution, D. An important limitation of (1) is that it generates discrete distributions with probability one. The practical construction of these discrete realizations from $\mathscr{D}_0$ is given in [7]. The a.s. discreteness of the Dirichlet process will create difficulties for us when we attempt to refine the partition of X^* (Section 3). This limitation can be overcome by using extended DPPs, such as the mixture of Dirichlet process model. A thorough review of this and other non-parametric process priors is available in [6].

We now summarize more formally the consequences of the DPP (1) relevant to our work, noting that **I** below constitutes the formal definition.

2.1. Relationship to the Dirichlet distribution

Definition 1 (Quantization operator, $\mathsf{Q}_{\mathbb{P}_K}$, and induced measure)
Let

$$\mathbb{P}_K = \{X_1^*, \ldots, X_K^*\} \subset \mathbb{A}, \quad K < \infty$$

DOI: 10.1002/acs

be any finite measurable partition of X^*. Define the associated quantization operator, $\mathsf{Q}_{\mathbb{P}_K}[x]$, on X^*

$$\mathsf{Q}_{\mathbb{P}_K} : X^* \to \mathbb{N}^+$$

$$\mathsf{Q}_{\mathbb{P}_K}[x] = \{k : \chi_{X_k^*}(x) = 1\} \tag{2}$$

where $\chi_A(x) = 1$, if $x \in A$, zero otherwise, is the *indicator function on the set A*. If F is a probability measure on $(X^*, \mathbb{A})$, then the *induced measure* on $\mathbb{P}_K$—i.e. on the random variable (r.v.) k (2)—is the multinomial distribution, $p = [p_1, \ldots, p_K]'$, where $p_k = F(X_k^*)$, $k = 1, \ldots, K$. The following notation is used:

$$F \overset{\mathsf{Q}_{\mathbb{P}_K}}{\to} p$$

Note

If $\{\bar{x}_1, \ldots, \bar{x}_K\}$ is an *alphabet of symbols* representing the K partition cells, respectively, then $\bar{x}_k$ is called the *quantized value* of x, if $\mathsf{Q}_{\mathbb{P}_K}[x] = k$ (2). For convenience, we will assume that these symbols are chosen such that $\bar{x}_k \in X_k^*$, $\forall k$.

Consider the Dirichlet process, $F \sim \mathscr{D}(\hat{F}_0, v_0)$

$$\hat{F}_0 \overset{\mathsf{Q}_{\mathbb{P}_K}}{\to} \hat{p}_0$$

$$F \overset{\mathsf{Q}_{\mathbb{P}_K}}{\to} p$$

Then the unknown multinomial, p, has a *Dirichlet Distribution*, expressed *via* its p.d.f., with parameters $\hat{p}_0$ and v_0

$$\mathscr{D}(\hat{F}_0, v_0) \overset{\mathsf{Q}_{\mathbb{P}_K}}{\to} D(\hat{p}_0, v_0) = \alpha^{-1}(\hat{p}_0, v_0) \prod_{i=1}^{K} p_i^{v_0 \hat{p}_{0,i} - 1} \chi_{\Delta_K}(p)$$

$$p \sim D(\hat{p}_0, v_0) \tag{3}$$

Here

$$\Delta_K = \left\{ p \,\middle|\, p_k \geqslant 0,\ k = 1, \ldots, K,\ \sum_{k=1}^{K} p_k = 1 \right\} \tag{4}$$

is the standard simplex in $\mathbb{R}^K$ for the K-term multinomial, p, equipped with a σ-algebra of Borel sets induced by $\mathbb{A}_F$. Furthermore

$$\alpha(\hat{p}_0, v_0) = \frac{\prod_{k=1}^{K} \Gamma(v_0 \hat{p}_{0,k})}{\Gamma(v_0)}$$

is the normalizing constant, where $\Gamma(\cdot)$ is the Gamma function [15].

We recall that $D(\hat{p}_0, v_0)$ [8] has the following mean and variances, respectively:

$$\mathsf{E}_{D(\hat{p}_0, v_0)}[p] = \hat{p}_0$$

$$\mathsf{VAR}_{D(\hat{p}_0, v_0)}[p_k] = \frac{\hat{p}_{0,k}(1 - \hat{p}_{0,k})}{v_0 + 1}, \quad k = 1, \ldots, K \tag{5}$$

where the subscript of E specifies the distribution used in the expectation.

DOI: 10.1002/acs

Later, we will use the KLD [12], $\mathsf{KLD}[f\|\tilde{f}]$, which measures the proximity of a density, $f(x)$, to another density, $\tilde{f}(x)$:

$$\mathsf{KLD}[f\|\tilde{f}] = \int f(x) \ln\left[\frac{f(x)}{\tilde{f}(x)}\right] \mathrm{d}x \tag{6}$$

We note that (i) if $\tilde{f}(x) = 0$ implies that $f(x) = 0$ a.s., then $\mathsf{KLD}[f\|\tilde{f}] < \infty$; and (ii) $\mathsf{KLD}[f\|\tilde{f}] = 0$ iff $\tilde{f}(x) = f(x)$ a.s.

Lemma 1

Let $p, q \in \Delta_K$ be two multinomials, with $p \sim D(\hat{p}, v_p)$, $q \sim D(\hat{q}, v_q)$. Then

$$\mathsf{KLD}[D(\hat{p}, v_p)\|D(\hat{q}, v_q)] = \sum_{k=1}^{K} \left[(v_p\hat{p}_k - v_q\hat{q}_k)\psi(v_p\hat{p}_k) + \ln\left(\frac{\Gamma(v_q\hat{q}_k)}{\Gamma(v_p\hat{p}_k)}\right)\right] - (v_p - v_q)\psi(v_p) + \ln\left(\frac{\Gamma(v_p)}{\Gamma(v_q)}\right)$$

where $\psi(v) = (\mathrm{d}/\mathrm{d}v)\ln(\Gamma(v))$ is the digamma (psi) function [15].

2.2. Learning under i.i.d. sampling

Let

$$x_i \overset{\text{iid}}{\sim} F, \quad i = 1, \ldots, n$$

$$\{x\}_n \equiv \{x_1, \ldots, x_n\} \tag{7}$$

be n i.i.d. samples from the unknown distribution, F (1). Occasionally, we will refer to $\{x\}_n$ (7) as the *data*. The formal meaning of i.i.d. sampling from a non-parametric process is given as Definition 2 in [8]. Under i.i.d. learning, the *a posteriori* distribution of F is also Dirichlet

$$\begin{aligned} F|\{x\}_n &\sim \mathscr{D}(\hat{F}_n, v_n) \equiv \mathscr{D}_n \\ v_n &= v_0 + n \\ \hat{F}_n &= \frac{1}{v_n}[v_0\hat{F}_0 + n\tilde{F}_n] \end{aligned} \tag{8}$$

$\tilde{F}_n$ is the *empirical distribution* on $(X^*, \mathbb{A})$, given i.i.d. samples $\{x\}_n$ (7) [2, 6]

$$\tilde{F}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} \tag{9}$$

Here, δ_{x_i} is the probability measure with unit mass (i.e. degenerate) at x_i [8].

2.3. The mean distribution

$$\mathsf{E}_{\mathscr{D}(\hat{F}_0, v_0)}[F] = \hat{F}_0 \tag{10}$$

DOI: 10.1002/acs

2.4. The expectation of $g(x)$

Consider a real-valued (moment) transformation, g_x, defined on $(X^*, \mathbb{A})$. $\mathsf{E}_F[g_x]$ is therefore a random variable, and

$$\mathsf{E}_{\mathscr{D}(\hat{F}_0, v_0)}[\mathsf{E}_F[g_x]] = \mathsf{E}_{\hat{F}_0}[g_x] = \int g_x \, \mathrm{d}\hat{F}_0$$

assuming that $\mathsf{E}_{\hat{F}_0}[|g_x|] < \infty$ [8]. When $(X^*, \mathbb{A}) = (\mathbb{R}^m, \mathbb{B})$, then, from (8), the *posterior mean* of $g(x)$, given i.i.d. samples $\{x\}_n$ (7), is

$$\mathsf{E}_{\mathscr{D}(\hat{F}_n, v_n)}[\mathsf{E}_F[g(x)]] \equiv \hat{g}(x)_n = \frac{1}{v_n}\left[v_0 \hat{g}(x)_0 + \sum_{i=1}^{n} g(x_i)\right] \tag{11}$$

where $\hat{g}(x)_0 = \int g(x) \, \mathrm{d}\hat{F}_0(x)$ is the prior mean.

2.5. The predictive distribution

Lemma 2
If $F \sim \mathscr{D}(\hat{F}_0, v_0)$ (1), then F_n—the predictive distribution on $(X^*, \mathbb{A})$ given i.i.d. samples $\{x\}_n$ (7) from F—is $\hat{F}_n$ (8).

Proof
Since $F|\{x\}_n \sim \mathscr{D}(\hat{F}_n, v_n)$ (8), then, by definition

$$F_n \equiv \mathsf{E}_{\mathscr{D}(\hat{F}_n, v_n)}[F] = \hat{F}_n$$

using (10). □

From 2.1–2.5, we conclude the following:

- The Dirichlet process prior, $\mathscr{D}(\hat{F}_0, v_0)$ (1), which places probability on the (function) space, F^*, is a generalization of the Dirichlet Distribution, $D(\hat{p}_0, v_0)$ (3), which places probability on the multinomial simplex, Δ_K. $\mathscr{D}(\hat{F}_0, v_0)$ allows the unknown probabilities on *any* sets in $\mathbb{A}$ to be modelled. In particular, the unknown multinomial induced on any finite measurable partition, $\mathbb{P}_K$, $K \geqslant 1$, of X^* is modelled, whereas $D(\hat{p}_0, v_0)$ is specific to just one such partition. $\mathscr{D}(\hat{F}_0, v_0)$ may be understood as a non-parametric model at the *input* to a specific quantizer of x (2), and $D(\hat{p}_0, v_0)$ is the (parametric) model for the quantizer output, $y = Q_{\mathbb{P}_K}[x]$, consistent with the input model and this *specific* quantizer.
- $\mathscr{D}(\hat{F}_0, v_0)$ is the conjugate non-parametric prior for i.i.d. sampling from an unknown distribution F (see (1) and (8)). This generalizes the rôle fulfilled by $D(\hat{p}_0, v_0)$ as the conjugate prior for Bayesian learning of an unknown multinomial, p, under i.i.d. sampling.
- From (10), we recognize $\hat{F}_0$ as the *mean function* of the Dirichlet process, sometimes known as the 'centre' or 'base measure' [5] of $\mathscr{D}$. From (5), v_0 may be interpreted as the precision parameter of $\mathscr{D}(\hat{F}_0, v_0)$, controlling the degree to which probability mass is localized around $\hat{F}_0$. From (8), v_0 may also be interpreted as the unnormalized *weight* (in units of 'number of i.i.d. samples') of the prior.
- Known base measure, $\hat{F}_0$, may itself be parameterized. These parameters, and v_0, can be modelled hierarchically, as described in [6, 16].

DOI: 10.1002/acs

- From (8), we note that the sufficient statistics for identification of F are the *complete* i.i.d. sample set, $\{x\}_n$, itself. This generalizes—to uncountable spaces X^*—the case of the Dirichlet distribution for the unknown measure, p, on a fixed partition, $\mathbb{P}_K$. In this case, the sufficient data statistics are the counts (see (3))

$$\kappa_n \equiv (v_n \hat{F}_n(X_1^*), \ldots, v_n \hat{F}_n(X_K^*)) \tag{12}$$

2.5.1. Non-informative prior, $\mathscr{D}(0)$

If $v_0 = 0$ for any F_0, then, from (8)

$$v_n = n \quad \text{and} \quad \hat{F}_n = \tilde{F}_n$$

From (10), the minimum Bayes' risk estimate of F under quadratic loss is the classical choice in this case, namely the empirical distribution, $\tilde{F}_n$ [6, 17]. For this reason, we will regard

$$\mathscr{D}(\hat{F}_0, 0) \equiv \mathscr{D}(0)$$

as the *non-informative non-parametric prior* for F [6]. The induced Dirichlet distributions, $D(0)$, are improper.

3. A STOPPING RULE FOR ON-LINE I.I.D. LEARNING OF F

A fundamental problem in designing a learning algorithm is to propose an optimal number, N, of data for reliable inference of a quantity of interest. The Bayesian perspective views this as a decision task, minimizing the expected loss (i.e. maximizing the expected utility) [17] associated with a particular choice of N. Bayesian parametric stopping is reviewed in [18], while a non-parametric method, using the Bayesian bootstrap, has recently been proposed in [19]. These are *a priori* methods, in that the decision is taken *before* sampling begins. A more useful paradigm for on-line learning is *sequential stopping*, where a choice $N \geqslant n$ is made, based on the current data $\{x\}_n$ (7). Bayesian sequential stopping for particular parametric models was derived in [9] using a quadratic loss function. More recently, the KLD [12] between consecutive parametric densities was used for sequential stopping [10]. The parametric treatment has two shortcomings:

(1) the tractability of the computations is highly dependent on the choice of models;
(2) in the initial stages of sampling, when the number of samples, n, is small, there is need for robustness to the choice of model, since model checking is unreliable [6, 19].

In order to overcome both of these difficulties, we will relax the parametric assumption *via* an unknown distribution, F, and model our evolving knowledge of F using the non-parametric DPP (1). Firstly, we review the parametric case.

3.1. Parametric sequential stopping

We assume that the posterior p.d.f., $f(\theta|D_n) \equiv f_n$, on unknown parameters, $\theta \in \Theta^*$, is available, given sequential observations, $D_n = [d_1, \ldots, d_n]$. The notation, D_n, emphasizes the fact that observations may be dynamic (correlated) [10]. For stopping, we assess f_n as a functional approximation of the p.d.f. given more data. Thus, f_n can be accepted as an approximation of f_{n+k}, $k = 1, 2, \ldots$, if f_n is shown to converge and be close to its asymptotic value.

DOI: 10.1002/acs

The Bayesian decision framework [17] requires quantification of the *loss function*, $L(f_n, f_{n+k})$, associated with using f_n as the approximation of f_{n+k}. In [13], it was shown that the choice

$$L(f_n, f_{n+k}) = \ln\left(\frac{f_{n+k}}{f_n}\right) \tag{13}$$

is appropriate for density approximation under very general conditions. Its expected value—i.e. the Bayesian risk—is

$$R_{n,k} = \int \ln\left(\frac{f_{n+k}}{f_n}\right) \mathrm{d}F(\theta, d_{n+k}, \ldots, d_{n+1}|D_n) = \mathsf{E}_{n+k|n}[\mathsf{KLD}(f_{n+k}\|f_n)], \quad k, n \geqslant 1$$

using (6). $\mathsf{E}_{n+k|n}[\cdot]$ denotes expectation with respect to the k-step-ahead predictor, $f(d_{n+k}, \ldots, d_{n+1}|D_n)$. Expanding (13), then

$$R_{n,k} = \sum_{\kappa=1}^{k} \mathsf{E}_{n+\kappa|n}[\mathsf{KLD}(f_{n+\kappa}\|f_{n+\kappa-1})]$$

If f_n is a bounded martingale with respect to the σ-algebra generated by the observations, D_n, then f_n converges almost surely [20], and so $\mathsf{E}_{n+k|n}[\mathsf{KLD}(f_{n+k}\|f_n)] \to^{n\to\infty} 0$. Given these considerations, a minimum Bayes' risk criterion for stopping after N observations is

$$N = \min\{n : R_{n,k} < \varepsilon, \ \forall k \geqslant 1\} \tag{14}$$

for a chosen small stopping threshold, ε. This sequential stopping rule [10] is computationally expensive, since multi-step predictors, $f(d_{n+\kappa}, \ldots, d_{n+1}|D_n)$, $\kappa = 1, 2, \ldots$, must be computed at each sampling time, n. A simpler version of the stopping rule examines only the *realized* risk, given n observations, in accepting f_{n-1} as an approximation of f_n

$$N = \min\{n : \mathsf{KLD}_n < \varepsilon\} \quad \text{where } \mathsf{KLD}_n \equiv \mathsf{KLD}[f_n\|f_{n-1}] \tag{15}$$

and $\mathsf{KLD}[\cdot]$ is defined in (6). Note that a computationally tractable stopping rule is essential, if the cost of its implementation is not to outweigh the cost of the sampling it proposes to stop. The following result provides guidance in setting the value of ε.

Lemma 3
If

$$\left|\frac{f_n - f_{n-1}}{f_{n-1}}\right| < \varepsilon \quad \text{a.s.} \quad \forall \theta \in \Theta^*$$

where ε is a small positive constant, then $0 \leqslant \mathsf{KLD}_n < \varepsilon$.

Proof
Given the stated condition, then $\ln(f_n/f_{n-1}) \approx (f_n/f_{n-1}) - 1 \in (-\varepsilon, \varepsilon)$. Hence

$$\mathsf{E}_{f_n}[\ln(f_n/f_{n-1})] = \mathsf{KLD}[f_n\|f_{n-1}] \in (-\varepsilon, \varepsilon)$$

Since $\mathsf{KLD} \geqslant 0$, it follows that $0 \leqslant \mathsf{KLD} < \varepsilon$. □

DOI: 10.1002/acs

The necessity of the condition is not proved. Nevertheless, it encourages the setting of ε as the maximum relative change allowed in the update $f_{n-1} \to f_n$. A typical value for conservative stopping is 0.01 (i.e. 1%). A more detailed analysis can be based on the results in [21].

Remark 1 (Modelling of KLD_n*)*

The stopping rule (15) may be strongly dependent on the data realization, D_n. Outlier sensitivity can be reduced by modelling the sequence of realized KLDs, KLD_n. The following choice is appropriate:

$$\mathsf{KLD}_n = \left(\frac{1}{n^c}\right)\zeta_n, \quad \zeta_n \overset{\text{i.i.d.}}{\sim} \mathscr{LN}(1, r), \ \ n = 1, 2, \ldots \tag{16}$$

where $c > 0$, and $\mathscr{LN}(1, r)$ denotes the log-normal distribution [1] for positive, multiplicative, modelling error, ζ_n. In this case, $\ln(\zeta_n) \sim \mathscr{N}(0, r)$, the normal density with zero mean and variance r. (16) is an appropriate choice for modelling a simple monotonic decrease in positive quantity, KLD_n. The least squares (LS) estimate of c after n samples is found by standard methods [22] to be

$$\hat{c}_n = -\frac{\sum_{i=1}^{n} \ln(\mathsf{KLD}_i)\ln(i)}{\sum_{i=1}^{n} \ln^2(i)}, \quad n = 1, 2, \ldots \tag{17}$$

Note that (i) this may be estimated recursively using just one multiplication and one division, which is an acceptable overhead for stopping, and (ii) the posterior mean estimate of c—which, for this model, differs from the LS estimate above—requires estimation of r, and, correspondingly, more computations. Given (ii), and the fact that $-\hat{c}_n \ln(n)$ is the modelled value of $\ln(\mathsf{KLD}_n)$, it cannot be used directly in the predictive stopping rule (14). However, (15) may be replaced by the following criterion:

$$N = \min\{n : \mathsf{LKLD}_n < \ln(\varepsilon)\}$$
$$\mathsf{LKLD}_n = -\hat{c}_n \ln(n) \tag{18}$$

An appropriate choice of ε is the n-dependent value $\varepsilon_n = \sqrt{\hat{r}_n}$, though, once again, computations can be saved by employing a fixed value, such as in Lemma 3.

3.2. Convergence of $\mathscr{D}_n = \mathscr{D}(\hat{F}_n, v_n)$

We return to i.i.d. learning of the non-parametric Dirichlet process, F (1). After n i.i.d. observations of x, i.e. given the current i.i.d. set, $\{x\}_n$, our knowledge of F is expressed by $\mathscr{D}_n = \mathscr{D}(\hat{F}_n, v_n)$ (8). For stopping, the question arises as to when the sequence, $\mathscr{D}_n$, has converged in some manner, so that learning *via* i.i.d. sampling might be considered to be 'complete'. Our approach is to examine the KLD for the induced (parametric) Dirichlet distributions on increasingly refined partitions, $\mathbb{P}_K$, of X^*. For the present, we assume that F is stationary, an assumption we will relax in Section 4.

Theorem 1

Let $F \sim \mathscr{D}(\hat{F}_0, v_0) \equiv \mathscr{D}_0$, with $\hat{F}_0 > 0$ a.s., and $v_0 > 0$. Then, $F|\{x\}_n \sim \mathscr{D}(\hat{F}_n, v_n) \equiv \mathscr{D}_n$, $n = 1, 2, \ldots$, using (8), and the associated sequence of predictors is $F_n = \hat{F}_n$, using Lemma 2.

DOI: 10.1002/acs

Consider a K-cell finite measurable partition, $\mathbb{P}_K = \{X_1^*, \ldots, X_K^*\}$, $K \geqslant 1$, in $(X^*, \mathbb{A})$, such that $F \to^{\mathsf{Q}_{\mathbb{P}_K}} p \in \Delta_K$ (Definition 1). Then:

(i) The sequence of measures induced by $\mathscr{D}_n$ forms a non-zero, bounded martingale with respect to the σ-algebra generated by $\{x\}_n$. The convergent sequence of associated KLDs is

$$\mathsf{KLD}[\mathscr{D}_n || \mathscr{D}_{n-1}; \mathbb{P}_K] = \psi(v_n \hat{F}_n(X_{k_n}^*)) - \ln[v_n \hat{F}_n(X_{k_n}^*) - 1] - [\psi(v_n) - \ln(v_n - 1)] \quad (19)$$

where, using (2)

$$k_n = \mathsf{Q}_{\mathbb{P}_K}[x_n] \quad (20)$$

(ii) The sequence of predictors induced by $\mathscr{D}_n$ forms a non-zero, bounded martingale with respect to the σ-algebra generated by $\{x\}_n$. The convergent sequence of associated KLDs is

$$\mathsf{KLD}[F_n || F_{n-1}; \mathbb{P}_K] = \ln\left(\frac{v_n - 1}{v_n}\right) + \hat{F}_n(X_{k_n}^*) \ln\left(\frac{v_n \hat{F}_n(X_{k_n}^*)}{v_n \hat{F}_n(X_{k_n}^*) - 1}\right), \quad K \geqslant 2$$

$$= 0, \quad K = 1 \quad (21)$$

The convergent sequence of associated *reverse* KLDs is

$$\mathsf{KLD}[F_{n-1} || F_n; \mathbb{P}_K] = \ln\left(\frac{v_n}{v_n - 1}\right) + \frac{v_n \hat{F}_n(X_{k_n}^*) - 1}{v_n - 1} \ln\left(\frac{v_n \hat{F}_n(X_{k_n}^*) - 1}{v_n \hat{F}_n(X_{k_n}^*)}\right), \quad K \geqslant 2$$

$$= 0, \quad K = 1 \quad (22)$$

Proof

(i) The proof of the first statement follows trivially from Definition 1; i.e. from (3)

$$F \overset{\mathsf{Q}_{\mathbb{P}_K}}{\to} p \sim D(\hat{p}_n, v_n)$$

This sequence of induced Dirichlet distributions is known to be a bounded martingale with respect to σ-algebra generated by $\{x\}_n$ [10], positive given the stated condition. From (8)

$$v_n \hat{F}_n = v_{n-1} \hat{F}_{n-1} + \delta_{x_n} \quad (23)$$

x_n falls in the k_nth bin of the partition (20). Hence, from (23)

$$v_n \hat{F}_n(X_{k_n}^*) = v_{n-1} \hat{F}_{n-1}(X_{k_n}^*) + 1$$

and so

$$v_n \hat{p}_n = v_{n-1} \hat{p}_{n-1} + \mathbf{1}_{k_n} \quad (24)$$

Here, $\mathbf{1}_k$, $k = 1, 2, \ldots, K$, is the kth elementary vector in $\mathbb{R}^K$. Using (24) in Lemma 1, noting that $v_n = v_{n-1} + 1$ (8), and recalling that $\hat{p}_{n,k_n} \equiv \hat{F}_n(X_{k_n}^*)$, the result (19) follows.

(ii) From Lemma 2 and Definition 1

$$F_n = \hat{F}_n \overset{\mathsf{Q}_{\mathbb{P}_K}}{\to} \hat{p}_n$$

DOI: 10.1002/acs

Since $D(\hat{p}_n, v_n)$ is a bounded martingale with respect to the σ-algebra generated by $\{x\}_n$, and since $\hat{p}_n$ is its expectation, then $\hat{p}_n$ is itself a bounded martingale, positive given the stated condition. Hence

$$\begin{aligned}\mathsf{KLD}[F_n||F_{n-1};\mathbb{P}_K] &= \mathsf{KLD}[\hat{p}_n||\hat{p}_{n-1}]\\ &\overset{(6)}{=} \sum_{k=1}^{K} \hat{p}_{n,k} \ln\left(\frac{\hat{p}_{n,k}}{\hat{p}_{n-1,k}}\right)\\ &\overset{K\geqslant 2}{=} (1-\hat{p}_{n,k_n}) \ln\left(\frac{v_n-1}{v_n}\right) + \hat{p}_{n,k_n} \ln\left(\frac{\hat{p}_{n,k_n}}{\hat{p}_{n-1,k_n}}\right)\end{aligned}$$

where we have used (24) in the first term on the right-hand side. Using (24) once again in the final term above, result (21) follows. When $K = 1$, $\hat{p}_n = \hat{F}_n(X^*) = 1$, and, similarly, $\hat{p}_{n-1} = 0$. Hence, $\mathsf{KLD}[\cdot] = 0$ (6). The result (22) for the reverse KLD follows in the same way. Note, finally, that all the expressions in the Theorem are bounded since the term

$$v_n \hat{F}_n(X^*_{k_n}) - 1 \geqslant v_0 \hat{F}_0(X^*_{k_n}) > 0 \quad \text{a.s.}$$

under the conditions of the theorem. □

Remark 2 (Behaviour for large n)
Using standard expansions [15] of $\psi(x)$ and $\ln(x)$, then it may be shown that

$$\mathsf{KLD}[\mathscr{D}_n||\mathscr{D}_{n-1};\mathbb{P}_K] \to \frac{1}{v_n}\left[\frac{1-\hat{F}_n(X^*_{k_n})}{2\hat{F}_n(X^*_{k_n})}\right], \quad n \text{ large}$$

This confirms the martingale requirement that (19) converge to zero, and is in agreement with the model (16) for the choice $c = 1$. Similarly

$$\begin{aligned}\mathsf{KLD}[F_n||F_{n-1};\mathbb{P}_K] &\to \mathsf{KLD}[F_{n-1}||F_n;\mathbb{P}_K]\\ &\to \frac{1}{v_n^2}\left[\frac{1-\hat{F}_n(X^*_{k_n})}{2\hat{F}_n(X^*_{k_n})}\right], \quad n \text{ large}\end{aligned}$$

in agreement with (16) for the choice $c = 2$.

Remark 3 (Partition-dependent stopping)
Theorem 1, along with Remark 1, can provide an operational stopping rule for i.i.d. learning of a Dirichlet process. Essentially, a fixed partition, $\mathbb{P}_K$, is chosen *a priori*, and the counts (12), $\kappa_{n,k}$, $k = 1, \ldots, K$, are accumulated for each partition cell, X^*_k, according to update (24)

$$\kappa_n = \kappa_{n-1} + \mathbf{1}_{k_n}, \quad n = 1, 2, 3, \ldots$$

initialized by counts derived from the parameters of the DPP

$$\kappa_{0,k} = v_0 \hat{F}_0(X^*_k), \quad k = 1, \ldots, K$$

A stopping rule consistent with Bayes' risk minimization (15) would then employ either of the KLDs (19) or (21), modelled *via* (16), i.e. the stopping rule (18). The choice (19) will be more conservative, for the reason given in Remark 2.

This stopping rule is, *per se*, parametric, and so does not address the two limitations of parametric stopping listed at the beginning of Section 3. In particular, the dependence of the

DOI: 10.1002/acs

stopping on the choice and cardinality, K, of the chosen partition, $\mathbb{P}_K$, is of concern. For example, in the trivial case of one-symbol quantization ($K = 1$), then $\mathsf{KLD} = 0 \ \forall n \geqslant 1$, so $N = 1$ for all KLD choices in Theorem 1.

Note, finally, that

$$\mathscr{D}(\hat{F}_n(x_1, \ldots, x_n), v_n) \overset{\mathsf{Q}_{\mathbb{P}_K}}{\to} D(\hat{p}_n, v_n)$$

$$\mathscr{D}(\hat{F}_n(\bar{x}_{k_1}, \ldots, \bar{x}_{k_n}), v_n) \overset{\mathsf{Q}_{\mathbb{P}_K}}{\to} D(\hat{p}_n, v_n)$$

where the statistics of $\hat{F}_n$ (8) are now shown explicitly, and $\bar{x}_k$ are the *quantized values* of the i.i.d. samples (Definition 1). This expresses the fact that i.i.d. learning of induced (parametric) multinomial, p, *via* Dirichlet distribution, D_n, requires storage only of the counts, κ_n (12), for the *quantized* data, $\{\bar{x}\}_n$. In contrast, learning of F *via* $\mathscr{D}_n$ requires storage of the *exact* record, $\{x\}_n$, *via* the *sufficient function*, $\hat{F}_n$ (8).

3.3. A partition-independent KLD for the Dirichlet process

Consider any sequence of *increasingly refined*, finite partitions, $\mathbb{P}_K \subset \mathbb{A}$, with increasing cardinality K. The phrase in italics is to mean that

$$\lim_{K \to \infty} \hat{F}_n(X_k^*) = 0 \quad \text{a.s.} \quad \forall k, n \tag{25}$$

Then, from (3)

$$\mathscr{D}(\hat{F}_n, v_n) \overset{\mathsf{Q}_{\mathbb{P}_K}}{\to} D(\hat{p}_n^{(K)}, v_n) \equiv D_n^{(K)}$$

i.e. $D_n^{(K)}$ is the sequence of Dirichlet distributions induced on $\Delta_K \subset \mathbb{R}^K$ with respect to the *partition refinement schedule*, $\mathbb{P}_K$, $K = 1, 2, 3, \ldots$. The general concept of divergence between parametric measures was studied in [23], in the context of quantization. In particular, the properties of a sequence of divergences between the measures induced by a partition refinement schedule was studied. If this sequence converges for *any* refinement schedule, then it converges for *all* refinement schedules. This result provides the essential pathway to construction of a *partition-independent* KLD between non-parametric DPPs. We exploit the fact that the Dirichlet distribution, D_n, induced by $\mathbb{P}_K$ is indeed *parametric* for $K < \infty$. The non-parametric case is found in the limit as $K \to \infty$, but only if such a limit exists. The Lemma which follows will provide pointers to how we might construct such a partition-independent limit.

Lemma 4

Consider the sequence, $\mathscr{D}_n \equiv \mathscr{D}(\hat{F}_n, v_n)$, of posterior distributions of the Dirichlet process, F, under i.i.d. learning, with $1 \leqslant n < \infty$, $v_0 > 0$ and $\hat{F}_0 > 0$ a.s. The associated sequence of predictors of x is $F_n = \hat{F}_n$ (Lemma 2). Then the following properties hold for associated (partition-free) KLDs

(i) $$\mathsf{KLD}[\mathscr{D}_n \| \mathscr{D}_{n-1}] = +\infty \quad \text{a.s.}$$

(ii) $$\mathsf{KLD}[F_n \| F_{n-1}] = +\infty \quad \text{a.s.}$$

(iii) $$\mathsf{KLD}[F_{n-1} \| F_n] = \ln\left(\frac{v_n}{v_{n-1}}\right) \quad \text{a.s.} \tag{26}$$

DOI: 10.1002/acs

Proof

All results follow immediately from the identities in Theorem 1, by considering any partition refinement schedule, $\mathbb{P}_K$, as defined by (25). Under the stated conditions, From (8)

$$v_0 \hat{F}_0(X^*_{k_n}) \to 0^+ \quad \text{a.s.}$$

where the superscript '+' denotes 'from above'. Hence, from (8)

$$\hat{F}_n(X^*_{k_n}) \to \frac{1^+}{v_n} \quad \text{a.s.} \tag{27}$$

Qualitatively, the nth i.i.d. sample falls, a.s., into a cell devoid of point masses (arising either from degeneracies in $\hat{F}_0$ or from the previous i.i.d. samples, $\{x\}_{n-1}$), in the limit of *any* partition refinement schedule. Substituting (27) into (19) and (21), the first two results follow, respectively. Substituting (27) into (22), and using l'Hopîtal's rule, the result (26) follows. □

Notes

(a) From (26),

$$\mathsf{KLD}[F_{n-1} \| F_n] \to \frac{1}{v_n}, \quad n \text{ large}$$

in agreement with model (16) for the choice $c = 1$.

(b) Using (26), the following simple stopping rule may be used, without the need for model (16):

$$N = \min\left\{ n : \ln\left(\frac{v_n}{v_{n-1}}\right) < \varepsilon \right\} \tag{28}$$

with unique deterministic solution

$$N = \left\lceil \frac{\exp(\varepsilon)}{\exp(\varepsilon) - 1} - v_0 \right\rceil \tag{29}$$

Here, $\lceil \cdot \rceil$ denotes the smallest integer greater than or equal to the argument. An effective stopping rule should take account of the disposition of x_n with respect to previous samples, $\{x\}_{n-1}$, and with respect to the prior base measure, $\hat{F}_0$. The rule (28) fails in these respects. Furthermore, (26) is a *reverse* KLD, and so a Bayes' risk interpretation (Section 3.1) of stopping rule (28) cannot be advanced. Nevertheless, Lemma 3 shows that the stopping rule does bound the relative change in predictor F_n. This consideration, along with the intuitive appeal of the test statistic (28), recommends it as a stopping rule for i.i.d. learning with Dirichlet processes.

(c) For completeness, we note that

$$\mathsf{KLD}[\mathscr{D}_{n-1} \| \mathscr{D}_n] = +\infty \quad \text{a.s.}$$

under the conditions of Lemma 4.

(d) Result (i) above is a direct consequence of the fact that $\mathscr{D}_n$, $\forall n$, assigns probability zero to any continuous probability measure on $(X^*, \mathbb{A})$ [6, 8]. Partition refinement induces these zero-probability continuous distributions, causing divergence of the associated KLDs. Other non-parametric priors [6] might be considered as a means of overcoming this difficulty.

DOI: 10.1002/acs

3.4. Data-dependent stopping

Convergence of the KLD with increasing n is certain for any *finite* partition (Remark 2). We now consider a schedule where the partition is refined *in tandem* with the number of i.i.d. samples, so as to achieve a bounded partition-free KLD suitable for non-parametric stopping.

Lemma 5

Consider i.i.d. sampling, $x_n \sim F$, $n = 1, 2, 3, \ldots$, from the Dirichlet process F on $(X^*, \mathbb{A})$, such that $F|\{x\}_n \sim \mathscr{D}(\hat{F}_n, v_n)$ (8). Define the data-dependent sequence of quantizers (2) $\mathsf{Q}_{\mathbb{P}_{K_n}}$, such that $F \to^{\mathsf{Q}_{\mathbb{P}_{K_n}}} p_{K_n}$. Here, $K \equiv K_n$ is the number of cells in the data-dependent partition, $\mathbb{P}_{K_n} \subset \mathbb{A}$, after n i.i.d. samples, and p_{K_n} is the finite parameter synonymous with the induced multinomial in Δ_{K_n} (3). Let $K_n = O(g(n))$ when n is large. Then

$$\lim_{n\to\infty} \mathsf{KLD}[\mathscr{D}_n \| \mathscr{D}_{n-1}; \mathbb{P}_{K_n}] = \lim_{n\to\infty} \mathsf{KLD}[\mathscr{D}_n \| \mathscr{D}_{n-1}] = 0 \quad \text{a.s.} \tag{30}$$

if and only if

$$0 < g'(n) < 1 \tag{31}$$

where $g'(n)$ denotes the derivative of $g(n)$. In words, the KLD converges and is partition-independent iff *the partition is refined more slowly than the rate of accumulation of samples.*

Proof

(i) For partition independence in the limit, $\mathbb{P}_{K_n}$ must constitute a partition refinement schedule [23], such that (25) be satisfied. Hence, K_n must be monotonically increasing for *large* n, requiring that $g(n)$ be a monotonically increasing function. This proves the lower bound in (31).

(ii) If condition (30) is imposed on (19), then we require that

$$\lim_{n\to\infty} v_n \hat{F}_n(X^*_{k_n}) = +\infty \quad \text{a.s.} \tag{32}$$

This, in turn, requires that n increase more rapidly than the number of cells, K_n, when n is large. This proves the upper bound in (31). □

Notes

(a) (30) is a necessary condition for $\mathscr{D}_n$ to be a bounded martingale with respect to the σ-algebra generated by the i.i.d. samples $\{x\}_n$.

(b) (30) suggests the following data-dependent stopping rule:

$$N = \min\{n : \mathsf{KLD}[\mathscr{D}_n \| \mathscr{D}_{n-1}; \mathbb{P}_{K_n}] < \varepsilon\} \tag{33}$$

which—though again partition-dependent—is guaranteed to achieve partition independence as $n \to \infty$ (i.e. for $\varepsilon \to 0$). The rule is informal in the sense that this limit is not reached if $\varepsilon > 0$, meaning that $K_n < n \leqslant N < \infty$. Nevertheless, the stopping rule greatly improves on the fixed partition case, using (19) and (21).

(c) There are many partition refinement schedules that satisfy the conditions of the Lemma.

DOI: 10.1002/acs

(d) If (30) holds for the sequence of predictors, F_n (21), then the associated data-dependent stopping rule is

$$N = \min\{n : \mathsf{KLD}[F_n \| F_{n-1}; \mathbb{P}_{K_n}] < \varepsilon\} \tag{34}$$

In fact, inspection of (21) reveals that (30) is satisfied for the predictor sequence, F_n, under the weaker condition

$$0 < g'(n) \leqslant 1 \tag{35}$$

(e) The realized KLD sequences in (33) and (34) may be modelled as in Remark 1—i.e. using the modelled KLD in the stopping rule (18)—so as to improve robustness to outliers occurring in the realized i.i.d. data, $\{x\}_n$.

3.5. *A proposal for a data-dependent partition refinement schedule*

The final choice of schedule satisfying Lemma 5 must be made pragmatically

(1) the computational cost of evaluating the KLD at each n is strongly influenced by the way in which the repartitioning, $\mathbb{P}_{K_{n-1}} \to \mathbb{P}_{K_n}$, takes place;
(2) we must examine the influence of $\mathbb{P}_{K_n}$ on the sequence of KLDs at finite n.

Note, from (19) and (21), that the principal computational overhead is associated with re-evaluation of $\hat{F}_n(X^*_{k_n})$, i.e. the measure on the partition cell occupied by the new sample, x_n, in the re-defined partition, $\mathbb{P}_{K_n}$. This, in general, requires re-quantization (classification) of the *entire* i.i.d. sample set, $\forall n$

$$\{x\}_n \overset{\mathsf{Q}_{K_n}}{\to} \{\bar{x}\}_n$$

The effort can be significantly reduced if the repartitioning minimally disturbs the cells, but care is needed to ensure that the partition is actually being refined in this case.

Consider a partition refinement schedule where the partition vertices, v_k, are chosen coincident with the i.i.d. samples, $\{x\}_n$. In this case, the number of cells is $K_n = n + 1$. In order to quantize x_n in this case, the following *ordered vertex set*, must be maintained:

$$\mathbb{V}_{(n)} = \{v_{(1)}, \ldots, v_{(n+2)}\} = \{\underline{x}, \{x\}_{(n)}, \bar{x}\}, \quad n = 0, 1, 2, \ldots \tag{36}$$

where $\{x\}_{(n)} = \mathsf{sort}[\{x\}_n]$ denotes the ordered set of i.i.d. samples with kth element $x_{(k)}$, $\underline{x}$ and $\bar{x}$ are appropriate bounding elements of X^*, and $\{x\}_{(0)} = \{x\}_0 = \{\}$ by convention. Update (24) now becomes

$$v_n \hat{F}_n(X^*_{k_n}) = v_0 \hat{F}_0(X^*_{k_n}) + 1 \tag{37}$$

where $X^*_{k_n}$ is the new cell delimited by x_n and its neighbour, $v_{(k_n+1)}$, in the ordered set $\mathbb{V}_{(n)}$. (37) follows from (8) and the fact that $X^*_{k_n}$ is solely occupied by x_n under this partition refinement schedule. A straightforward and admissible ordering scheme for $X^* \subseteq \mathbb{R}^m$ is to sort $\{x\}_n$ in any one of the m co-ordinates, with $\mathbb{P}_{n+1}$ then defined by partitioning $\mathbb{R}^m$ along this co-ordinate. In general, this requires that the *marginal* prior base measure be available for this co-ordinate, so that $\hat{F}_0(X^*_{k_n})$ (37) can be evaluated.

Under this proposal, $g'(n) = 1$, failing condition (31) in Lemma 5. However, (35) is satisfied, and so this partition refinement schedule is appropriate for use with the stopping rule for predictors (34).

DOI: 10.1002/acs

By way of illustration, we now give the algorithm for data-dependent stopping using the proposed partition refinement schedule. For further clarity, we assume that $X^* = [\underline{x}, \bar{x}] \subset \mathbb{R}$, equipped with the usual σ-algebra of Borel subsets, $\mathbb{B}$. Thus, the Dirichlet process, F, is expressible as an unknown univariate, finitely supported distribution. The realized KLD sequence (34) is modelled recursively *via* (16)–(18).

Algorithm 1 (Realization-dependent stopping (scalar case))

$n = 0$
$\mathbb{V}_{(0)} = \{v_{(1)}, v_{(2)}\} = \{\underline{x}, \bar{x}\}$ % ordered partition vertices at n = 0
choose $\bar{N}$, ε, v_0, $\hat{F}_0$, $\hat{g}(x)_0$, a_0, b_0 % choose $\hat{g}(x)_0$ consistent with $\hat{F}_0$
initialize $\mathsf{LKLD}_0 = \ln(\varepsilon) + 1$ % ensure starting
for ($\mathsf{LKLD}_n > \ln(\varepsilon)$) AND ($n < \bar{N}$)
$n = n + 1$
$v_n = v_{n-1} + 1$
realize $x_n \sim F$
$\hat{g}(x)_n = \hat{g}(x)_{n-1} + \frac{1}{v_n}[g(x_n) - \hat{g}(x)_{n-1}]$ % recursive moment tracking via Eq. (11)
$\mathbb{V}_{(n)} = \{v_{(1)}, \ldots, v_{(n+2)}\} = \mathsf{sort}\{\mathbb{V}_{(n-1)}, x_n\}$ % insert x_n into ordered vertex set
$k_n = \{k : x_n = v_{(k)}\}$ % $\mathsf{Q}_{\mathbb{P}_{n+1}}[x_n]$ (2) is the position of x_n in $\mathbb{V}_{(n)}$
$\hat{p}_{0,k_n} = \hat{F}_0(v_{(k_n+1)}) - \hat{F}_0(x_n)$ % interpreting $\hat{F}_0$ as a c.d.f.
$\hat{p}_{0,k_n} = \frac{v_{(k_n+1)} - x_n}{\bar{x} - \underline{x}}$ % special case when $\hat{F}_0$ is the uniform on $[\underline{x}, \bar{x}]$
$\mathsf{KLD}_n = \ln[1 - \frac{1}{v_n}] + \frac{1}{v_n}(1 + v_0\hat{p}_{0,k_n}) \ln[1 + \frac{1}{v_0\hat{p}_{0,k_n}}]$ % Eq. (21), using Eq. (37)
$a_n = a_{n-1} + \ln(\mathsf{KLD}_n) \ln(n)$
$b_n = b_{n-1} + \ln^2(n)$ % recursive update of (17)
$\hat{c}_n = -\frac{a_n}{b_n}$
$\mathsf{LKLD}_n = -\hat{c}_n \ln n$
end
$N = n$
report $\hat{g}(x)_N$

4. THE NON-STATIONARY DIRICHLET PROCESS, $F_t \sim \mathscr{D}(\hat{F}_t, v_t)$

We now consider the case where the Dirichlet process, F (1), is non-stationary. We examine how our state of knowledge, expressed by the DPP, $\mathscr{D}_0$, can be updated in order to track this non-stationary behaviour. The update is supplemental to any data-based learning $\mathscr{D}_0 \to \mathscr{D}_n$ (8) which may be taking place as a result of i.i.d. sampling, $\{x\}_n$, from F.

Let F_τ be a non-stationary, non-parametric, unknown marginal distribution (i.e. random process) on $(X^*, \mathbb{A})$, $\forall\tau$, where $\tau \in \mathbb{R}$ is the independent index against which the process varies (it is referred to as 'time', but might equally denote frequency, space, etc.).

Definition 2 (Stationarity interval)
The *stationarity interval* is a prior knowledge object associated with the non-stationary process F_τ, and is defined as follows:

$$T = \max\{\zeta | F_\tau = F_{\tau+\zeta} \text{ a.s. } \forall\tau\}$$

DOI: 10.1002/acs

We will assume that $T > 0$. The stationarity interval has a 'natural' meaning in most applications, corresponding, for example, to (i) the sampling period of incoming data for a system, (ii) the minimum possible time between operator interventions in an industrial process, etc. It is used to calibrate the time axis.

Let $t = 0, 1, \ldots$ (a discrete time index) be the total number of complete intervals, T, observed since $\tau = 0$ (i.e. since the beginning of the observation window). Equivalently, t is the index into the *changepoint set*, $\tau_t = tT$. Let the associated sequence of unknown (marginal) distributions on $(X^*, \mathbb{A})$ be denoted by F_t. As before, our task is to model the non-parametric process F_t, $\forall t$. In the context of the DPP (1), we need to elicit, for example, the following marginals:

$$F_t \sim \mathscr{D}(\hat{F}_t, v_t) \equiv \mathscr{D}_t, \quad t = 0, 1, \ldots$$

For the time being, we suppress the subscript '$n = 0, 1, \ldots$' associated with i.i.d. learning (1), (8), returning to it in the next section. Under Definition 2, this prior measure is a.s. constant in the time interval $[\tau_t, \tau_{t+1})$. One way to model the transitions is to justify closure of the marginal Dirichlet measure under the update, i.e. $\mathscr{D}_t \to \mathscr{D}_{t+1}$, and deduce appropriate transition rules on the parameters, $\hat{F}_t \to \hat{F}_{t+1}$ and $v_t \to v_{t+1}$ (1). We will approach the problem in this way. This avoids full modelling of the time series to arbitrary order $q = 1, 2, \ldots$. The latter requires, for example, a measure, $\mathscr{F}^{(q)}$, on the unknown qth-order non-parametric distribution, $F^{(q)}$, on extended measure space, $((X^*)^q, (\mathbb{A})^q)$. We assume that this is unavailable.

4.1. Non-parametric stabilized forgetting

An optimized forgetting operator was proposed in [11] as a means of parameter tracking in non-stationary parametric time-series analysis. It copes with situations where no explicit transition model is available. In this section, we verify that this concept extends successfully and tractably in the non-parametric context. Following [11], but in the context of *non-parametric* measures, we make the following modelling assumptions about the unknown distributions, F_t, $t = 0, 1, 2, \ldots$:

$$\begin{aligned} F_0 &\sim \mathscr{F}_0 \\ F_{t+1} &\sim \mathscr{F}_t \quad \text{with probability } \lambda_t \\ F_{t+1} &\sim \mathscr{F}^a_{t+1} \quad \text{with probability } 1 - \lambda_t \end{aligned} \tag{38}$$

λ_t is a sequence of known *forgetting factors* and $\mathscr{F}^a_t$ is a sequence of known *alternative* distributions for non-parametric F_t. For convenience, we assume that all non-parametric process distributions, $\mathscr{F}$, are finite, non-null, and defined on identical measurable spaces, $(F^*, \mathbb{A}_F)$ (Section 2).

Denote by F_p the parametric measure induced on Δ_K by $\mathscr{F}$, *via* quantization operator $Q_{\mathbb{P}_K}$ (Definition 1). In the non-stationary case

$$\mathscr{F}_t \overset{Q_{\mathbb{P}_K}}{\to} F_{p,t} \quad \text{and} \quad F_t \overset{Q_{\mathbb{P}_K}}{\to} p_t \tag{39}$$

so that $p_t \sim F_{p,t}$. For *any* such partition, then, from (38), the minimum Bayes' risk decision (Section 3.1) in respect of the updated parametric measure, $F_{p,t+1}$, is

$$F_{p,t+1} = \arg\min_{F_p} \{\lambda_t \mathsf{E}_{F_{p,t}}[L(F_p, F_{p,t})] + (1 - \lambda_t)\mathsf{E}_{F^a_{p,t+1}}[L(F_p, F^a_{p,t+1})]\} \tag{40}$$

where $L(F_p, F_{p,t})$, for example, is the loss associated with the decision F_p, when $F_{p,t}$ is the true distribution on $(X^*, \mathbb{A})$. In common with [11], we use the *reverse* KLD to approximate the

DOI: 10.1002/acs

expected loss

$$\mathsf{E}_{F_{p,t}}[L(F_p, F_{p,t})] \to \mathsf{KLD}[F_p \| F_{p,t}] \tag{41}$$

Under the stated conditions for $\mathscr{F}$, the required KLDs exist and are finite. It has been shown in [11] that the unique solution of (40) under assignment (41) is

$$F_{p,t+1} \propto (F_{p,t})^{\lambda_t} (F^a_{p,t+1})^{(1-\lambda_t)} \tag{42}$$

We choose this minimizer for *every* finite measurable partition, $\mathbb{P}_K \subset \mathbb{A}$. More generally, it is this operator (42) that we use to construct the measure on the unknown probabilities, $(F_{t+1}(A_1), \ldots, F_{t+1}(A_q))$, for *every* finite set, $(A_1, \ldots, A_q)$, of pairwise disjoint sets, $A_i \in \mathbb{A}$ (Section 2). Then there exists a *unique* non-parametric process prior on $(F^*, \mathbb{A}_F)$ which induces them [8]. It is denoted by

$$F_{t+1} \sim \mathscr{F}_{t+1} \propto (\mathscr{F}_t)^{\lambda_t} (\mathscr{F}^a_{t+1})^{(1-\lambda_t)} \tag{43}$$

Notes

(1) (42), with (39), defines the *non-parametric stabilized forgetting operator*, (43), being true for any partition, $\mathbb{P}_K \subset \mathbb{A}$.
(2) (42) is not the minimum Bayes' risk decision, since the unreversed KLD, i.e. $\mathsf{KLD}[F_{p,t} \| F_p]$, is the Bayes' risk in approximating $F_{p,t}$ by F_p, as discussed in Section 3.1 [13]. (41) was chosen since any parametric distribution belonging to the exponential family [1] is closed under this operator. This is confirmed for the case of the Dirichlet distribution in the next Lemma. The minimum Bayes' risk decision chooses the *arithmetic* mean operator in place of the geometric mean (42). The resulting binary mixture requires a projection step back to the exponential family [24].

Lemma 6
The space of Dirichlet process distributions is closed under non-parametric stabilized forgetting (43).

Proof
Let F be a non-parametric process on $(X^*, \mathbb{A})$, with prior

$$F \sim [\mathscr{D}(\hat{F}, v)]^{\lambda} [\mathscr{D}(\hat{F}^a, v^a)]^{(1-\lambda)}$$

$0 \leqslant \lambda \leqslant 1$. Consider any $\mathbb{P}_K \subset \mathbb{A}$, such that $F \to^{\mathsf{Q}_{\mathbb{P}_K}} p$ (Definition 1). Using the definition (42) of the non-parametric operator (43), and recalling the definition of the Dirichlet distribution, D (3), then

$$p \sim [D(\hat{p}, v)]^{\lambda} [D(\hat{p}^a, v^a)]^{(1-\lambda)}$$

Here, $\hat{p}$ and $\hat{p}^a$ are the multinomials induced on $\mathbb{P}_K$ by $\hat{F}$ and $\hat{F}^a$, respectively. Using (3)

$$p \sim \beta^{-1}(v, v^a, \hat{p}, \hat{p}^a, \lambda) \prod_{k=1}^{K} p_k^{\lambda v \hat{p}_k + (1-\lambda) v^a \hat{p}^a_k - 1} \chi_{\Delta_K}(p) \tag{44}$$

Noting that $\sum_{k=1}^{K} \hat{p}_k = \sum_{k=1}^{K} \hat{p}^a_k = 1$, then, from (44)

$$p \sim D(\hat{p}_\lambda, v_\lambda) \tag{45}$$

DOI: 10.1002/acs

with

$$v_\lambda = \lambda v + (1-\lambda)v^a \tag{46}$$

$$\hat{p}_\lambda = \frac{1}{v_\lambda}[\lambda v\hat{p} + (1-\lambda)v^a\hat{p}^a] \tag{47}$$

with normalizing constant, $\beta(\cdot)$ in (44), given by $\alpha(\hat{p}_\lambda, v_\lambda)$ (3). Since (45) is true for all finite, measurable partitions, then, by Definition 1 of the DPP [8]

$$F \sim \mathscr{D}(\hat{F}_\lambda, v_\lambda)$$

with v_λ given by (46) and

$$\hat{F}_\lambda = \frac{1}{v_\lambda}[\lambda v\hat{F} + (1-\lambda)v^a\hat{F}^a] \qquad \square$$

4.2. Stabilized forgetting and flattening for the non-stationary Dirichlet process

We make the following modelling assumptions concerning the non-stationary process F_t:

$$F_0 \sim \mathscr{D}(\hat{F}_0, v_0)$$
$$\mathscr{F}_t^a = \mathscr{D}(\hat{F}_t^a, v_t^a), \quad t = 1, 2, \ldots \tag{48}$$

with an associated sequence of forgetting factors λ_t (38). Then, using non-parametric stabilized forgetting (Lemma 6), F_t is distributed approximately as a Dirichlet process $\forall t$

$$F_{t+1} \sim \mathscr{D}(\hat{F}_{t+1}, v_{t+1}), \quad t = 0, 1, 2, \ldots$$
$$\hat{F}_{t+1} = \frac{1}{v_{t+1}}[\lambda_t v_t \hat{F}_t + (1-\lambda_t)v_{t+1}^a \hat{F}_{t+1}^a]$$
$$v_{t+1} = \lambda_t v_t + (1-\lambda_t)v_{t+1}^a \tag{49}$$

In this sense, F_t is approximately a non-stationary Dirichlet process. As will be seen in the next section, this is important for ensuring a tractable learning schedule for F_t.

Consider two special cases:

I *Stabilization via the Prior*:

$$\mathscr{F}_t^a = \mathscr{D}(\hat{F}_0, v_0), \quad t = 1, 2, \ldots$$

Substituting into (49)

$$\mathscr{F}_t = \mathscr{D}(\hat{F}_0, v_0), \quad t = 0, 1, 2, \ldots$$

II *Exponential forgetting* (*flattening*): If $v_t^a = 0$ (48), then $\mathscr{F}_t^a = \mathscr{D}(0)$, $t = 1, 2, \ldots$, the non-informative DPP (Section 2.5.1). Then, from (46) and (47), and assuming for simplicity that $\lambda_t = \lambda$, we have

$$F_t \sim \mathscr{D}(\hat{F}_0, \lambda^t v_0), \quad t = 0, 1, 2, \ldots$$

This is a much weaker update than the stabilized update (49), since $\lim_{t\to\infty} \mathscr{D}(\hat{F}_t, v_t) = \mathscr{D}(0)$ in this case.

DOI: 10.1002/acs

5. LEARNING FOR THE NON-STATIONARY DIRICHLET PROCESS

We now consider the substantive task, namely, learning for F_t under i.i.d. sampling at all times $t = 0, 1, 2, \ldots$. We summarize the following key points:

- The stabilized forgetting operator commutes with the Bayes' rule operator (i.e. prior-to-posterior updating) in the parametric case [11]. By considering the measures induced for any partition, $\mathbb{P}_K \subset \mathbb{A}$ (42), we see that the non-parametric operator (43) possesses the same property.
- The space of Dirichlet process distributions is closed (i.e. conjugate) under i.i.d. sampling from F_t (8).
- The space of Dirichlet process distributions is closed under stabilized forgetting (Lemma 6).

Taken together, these reveal an important rôle for the DPP in ensuring a tractable algorithm for tracking of non-parametric processes.

Definition 3 ($\bar{N}$)
Consider the task of learning F_t *via* i.i.d. samples, $\{x_{t,1}, x_{t,2}, \ldots\}$, taken during the tth stationarity interval, whose duration is T (Definition 2). If δ is the time taken to generate one i.i.d. sample and complete the associated updates (8), then $\bar{N} = \lfloor T/\delta \rfloor$, is the maximum allowable number of i.i.d. samples per stationarity interval. Here, $\lfloor \cdot \rfloor$ denotes the greatest integer less than or equal to the argument.

We assume that there are $0 \leqslant N_t \leqslant \bar{N}$ i.i.d. samples available from F_t at each time $t \geqslant 0$

$$\{x\}_{t,N_t} \equiv \{x_{t,1}, \ldots, x_{t,N_t}\} \tag{50}$$

Our knowledge, modelled by the Dirichlet process distribution $\mathscr{D}_{t,n}$, evolves in response to two distinct and interleaved events:

(i) stabilized forgetting at each changepoint, indexed by the *first* subscript, t. We adopt stabilization *via* the prior (case I in Section 4.2), and a constant forgetting factor, $\lambda_t = \lambda$;
(ii) i.i.d. learning—indexed by the *second* subscript n—during each stationarity interval. The required updates are governed by (8).

From (8) and (49), the posterior distribution of the non-stationary Dirichlet process, F_t, is ($t \geqslant 0$, $n \geqslant 0$)

$$\begin{aligned} F_t | \{x\}_{t,n} &\sim \mathscr{D}(\hat{F}_{t,n}, v_{t,n}) \\ v_{t,n} &= v_{0,0} + \chi_{\mathbb{N}^+}(t) \sum_{j=0}^{t-1} \lambda^{t-j} N_j + n \\ \hat{F}_{t,n} &= \frac{1}{v_{t,n}} \left[v_{0,0} \hat{F}_{0,0} + \chi_{\mathbb{N}^+}(t) \sum_{j=0}^{t-1} \lambda^{t-j} N_j \tilde{F}_{j,N_j} + n \tilde{F}_{t,n} \right] \end{aligned} \tag{51}$$

where

$$\begin{aligned} F_0 &\sim \mathscr{D}(\hat{F}_{0,0}, v_{0,0}) \\ \tilde{F}_{j,N_j} &= \frac{1}{N_j} \sum_{i=1}^{N_j} \delta_{x_{j,i}} \end{aligned}$$

DOI: 10.1002/acs

are, respectively, the DPP and the empirical distribution (9) based on the i.i.d. sample set, $\{x\}_{j,N_j}$ (50), gathered during the jth stationarity interval. Hence, the sufficient statistics are the entire archive of i.i.d. samples, $j = 0, \ldots, t$.

5.1. The stopping rule for i.i.d. learning of F_t

Algorithm 1 may be used to determine the stopping number, N_t, in each stationarity interval. The data-dependent partition refinement schedule proposed in Section 3.4 implies that the required set of ordered partition vertices, $\mathbb{V}_{(t,n)}$, after n i.i.d. samples in stationarity interval, t, be comprised of the *entire* archive of i.i.d. samples. Two immediate difficulties arise:

(1) the cost of storing and sorting this set prohibits its use for stopping when t is large;
(2) $K_{t,n} \to +\infty$ for t large, where $K_{t,n}$ is the current number of cells in the partition (Lemma 5). In contrast, forgetting ensures that $v_{t,n}$ remains finite (51). Hence, the condition (31) is violated, as is the weaker condition (35), leading to failure of the associated stopping rules (33) and (34).

An appropriate adaptation of the partition refinement proposal (Section 3.4) is to decimate the partition vertex set, $\mathbb{V}_{(t-1,N_{t-1})}$ (36), at each changepoint, t; i.e. to transmit only a fraction of the i.i.d. samples for use in partitioning X^* during the *next* stationarity interval. If the fraction is chosen equal to λ (51), then the number of (interior) vertices is always equal to $v_{t,n} - v_{0,0}$. Of course, the active cell, $X^*_{k_{t,n}}$ (20), is now no longer solely occupied by $x_{t,n}$. Hence, evaluation of the required probability measure $\hat{F}_{t,n}(X^*_{k_{t,n}})$ (in (21)) necessitates quantization (and, therefore, storage and sorting) of the entire i.i.d. archive, obviating the benefits of the vertex decimation proposed above. Consider, therefore, an approximation which—at changepoint t—replaces the latest i.i.d. set, $\{x\}_{t-1,N_{t-1}}$, with a set quantized with respect to the newly decimated vertex set, $\mathbb{V}_{t,0}$. Then, the i.i.d. samples are *always* coincident with the partition vertices, sole occupancy is re-established for all partition cells, and so $\hat{F}_{t,n}(X^*_{k_{t,n}})$ is again evaluated simply, *via* (37). Note that this quantization of $\{x\}_{t-1,N_{t-1}}$ does not have to be implemented, and so the net computational requirement at each changepoint is merely *to decimate the vertex set*, $\mathbb{V}_{(t-1,N_{t-1})}$ (36), *keeping fraction* λ.

From (51), and using the decimated refinement schedule above, the following procedure is revealed for (on-line) learning of the non-stationary Dirichlet process, F_t. For convenience, we again assume that $(X^*, \mathbb{A}) = (\mathbb{R}^m, \mathbb{B})$, and track the posterior mean $\hat{g}(x)_{t,n}$ of $g(x)$ recursively, using (11). In the algorithm below, recall, from (36), that $\mathbb{V}_{(t,n)} = \{v_{(1)}, v_{(2)}, \ldots\}$ refers to the current set of *ordered* vertices, $v_{(k)}$.

Algorithm 2 (Learning for the non-stationary Dirichlet process (scalar case))

$\mathbb{V}_{(0,0)} = \{v_{(1)}, v_{(2)}\} = \{\underline{x}, \bar{x}\}$ % ordered partition vertices at n = 0
choose $\bar{N}$, ε, $v_{0,0}$, $\hat{F}_{0,0}$, $\hat{g}(x)_{0,0}$, $a_{0,0}$, $b_{0,0}$, λ
for $t = 0, 1, 2, \ldots$
  $n = 0$
  initialize $\mathsf{LKLD}_{t,n} = \ln(\varepsilon) + 1$ % ensure starting
  for ($\mathsf{LKLD}_{t,n} > \ln(\varepsilon)$) AND ($n < \bar{N}$)
    $n = n + 1$
    $v_{t,n} = v_{t,n-1} + 1$

DOI: 10.1002/acs

realize $x_{t,n} \sim F_t$
$\hat{g}(x)_{t,n} = \hat{g}(x)_{t,n-1} + \frac{1}{v_{t,n}}[g(x_{t,n}) - \hat{g}(x)_{t,n-1}]$
$\mathbb{V}_{(t,n)} = \mathsf{sort}\{\mathbb{V}_{(t,n-1)}, x_{t,n}\}$ % insert $x_{t,n}$ into ordered vertex set
$k_{t,n} = \{k : x_{t,n} = v_{(k)}\}$
$\hat{p}_{0,0,k_{t,n}} = \hat{F}_{0,0}(v_{(k_{t,n}+1)}) - \hat{F}_{0,0}(x_n)$ % interpreting $\hat{F}_{0,0}$ as a c.d.f.
$\mathsf{KLD}_{t,n} = \ln[1 - \frac{1}{v_{t,n}}] + \frac{1}{v_{t,n}}(1 + v_{0,0}\hat{p}_{0,0,k_{t,n}}) \ln[1 + \frac{1}{v_{0,0}\hat{p}_{0,0,k_{t,n}}}]$
$a_{t,n} = a_{t,n-1} + \ln(\mathsf{KLD}_{t,n}) \ln(n)$
$b_{t,n} = b_{t,n-1} + \ln^2(n)$ % recursive update of (17)
$\hat{c}_{t,n} = -\frac{a_{t,n}}{b_{t,n}}$
$\mathsf{LKLD}_{t,n} = -\hat{c}_{t,n} \ln n$

end

$N_t = n$
$v_{t+1,0} = \lambda v_{t,N_t} + (1 - \lambda)v_{0,0}$
$\hat{g}(x)_{t+1,0} = \frac{1}{v_{t+1,0}}[\lambda v_{t,N_t}\hat{g}(x)_{t,N_t} + (1 - \lambda)v_{0,0}\hat{g}(x)_{0,0}]$
$\hat{a}_{t+1,0} = \hat{a}_{t,N_t}$; $\hat{b}_{t+1,0} = \hat{b}_{t,N_t}$
$\mathbb{V}_{(t+1,0)} = \mathsf{decimate}\{\mathbb{V}_{(t,N_t)}\}$ % keep fraction λ of vertices, preserving $\underline{x}$ and $\bar{x}$

end

report $\hat{g}(x)_{t,N_t}$, $t = 0, 1, \ldots$

Notes

- the prior moment should be chosen consistently with the prior base measure: $\hat{g}(x)_{0,0} = \mathsf{E}_{\hat{F}_{0,0}}[g(x)]$ (11).
- The algorithm initializes the LS estimator of c_t (17) *via* $\hat{c}_{t+1,0} = \hat{c}_{t,N_t}$. This is important in ensuring a stable recursive update even in cases when N_t is small.
- When $\lambda \to 1$, most of the information in the i.i.d. sample, $\{x\}_{t,N_t}$ is propagated across the changepoint at $t + 1$. This corresponds to an assumption of very slowly non-stationary F_t. Conversely, when $\lambda \to 0$, then, from (51), $F_{t+1} \sim \mathscr{D}(\hat{F}_{0,0}, v_{0,0})$, i.e. the prior distribution. While this may be appropriate in coping with fast non-stationarities, all learning from previous i.i.d. sampling has been lost. Hence, the proposed algorithm is appropriate for slowly non-stationary processes.

5.2. The choice of λ

Consider the case of i.i.d. learning of F_t *via* non-parametric stabilized forgetting (case I in Section 4.2). Taking $N_t = \bar{N}$ (Definition 3) $\forall t$, and assuming that $\lambda < 1$, then, from (51)

$$\lim_{t\to\infty} v_{t,N_t} \equiv v_\infty = v_0 + \frac{\bar{N}}{1 - \lambda} = v_0 + s_\lambda \bar{N} \tag{52}$$

is the Dirichlet weight at a changepoint, in the long-run. However, $\bar{N}$ is the actual number of i.i.d. samples gathered in the stationarity interval, and so an immediate interpretation of $s_\lambda \geqslant 1$ is as the number of stationarity intervals since the last *actual* changepoint in the random process. This is modelled by the observer, *via* the prior setting of λ, to reflect the expected dynamics of F_t. For $s_\lambda = 1 + \delta_\lambda$, then a factor δ_λ of the i.i.d. record is 'remembered' from the previous stationarity interval (or intervals). For $\lambda = \delta_\lambda$ small, $s_\lambda = (1 - \delta_\lambda)^{-1} \approx 1 + \delta_\lambda$, in which case $100\lambda\%$ of the previous record, $\{x\}_{t-1,N_{t-1}}$,

DOI: 10.1002/acs

is remembered. For example, about 10% are remembered when $\lambda = 0.1$. Since stabilized forgetting is operating here in batch mode, with $N_t \gg 1$ typically, the considerations above encourage the setting of λ to small or moderate values. This is in contrast to the usual *on-line* context for forgetting [11], where, effectively, $N_t = 1$, $\forall t$.

6. SIMULATION STUDIES

Our purpose in this section is to provide illustrative examples of the application of the Dirichlet learning algorithms in two important contexts: (i) density estimation (Algorithm 1), and (ii) tracking of non-stationary random processes (Algorithm 2).

6.1. Comparison of stopping rules for i.i.d. learning

Stopping rules for i.i.d. learning of an unknown probability measure, $F(x)$, on $(\mathbb{R}, \mathbb{B})$ are compared. Student's t-distribution [1] is chosen as the true underlying measure. Its parametric probability density is

$$\mathscr{S}(x|m,r,\zeta) = \frac{\Gamma((\zeta+1)/2)}{\Gamma(\zeta/2)\Gamma(\frac{1}{2})}\left(\frac{1}{r\zeta}\right)^{1/2}\left[1+\frac{1}{r\zeta}(x-m)^2\right]^{-(\zeta+1)/2}$$

where $\mathsf{E}[x] = m \in \mathbb{R}$, $r \in \mathbb{R}^+$ is a scaling parameter, and $\zeta \in \mathbb{R}^+$ is known as the 'degrees-of-freedom' parameter. In the current simulations, we choose

$$x_n \overset{\text{iid}}{\sim} \mathscr{S}(x|20, \tfrac{1}{3}, 3)$$

whose small value of ζ induces a strongly non-Gaussian density with heavy tails (Figure 1). The non-parametric prior is, as always in this work, Dirichlet (1)

$$F \sim \mathscr{D}(\mathscr{U}_{(-100,+100]}, 3)$$

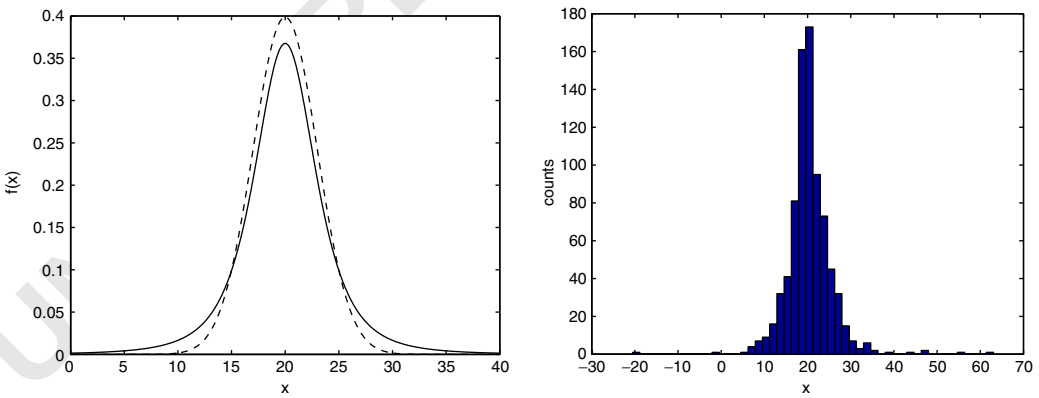

Figure 1. Left: Student's t-distribution, $f(x) = \mathscr{S}(x|m,r,\zeta)$, for $m = 20$ and $r = \frac{1}{3}$. The degrees-of-freedom parameter is $\zeta = 3$ (full) and $\zeta = 300$ (dashed), for which case $f(x) \approx \mathscr{N}(m,r)$. Right: Typical sample set, $\{x\}_N$, realized from $\mathscr{S}(x|20, 1/3, 3)$, with stopping at $N = 811$ (Algorithm 1).

DOI: 10.1002/acs

Here, $\hat{F}_0 = \mathscr{U}$ denotes the uniform measure on the stated interval. Three proposals for stopping are compared

(1) Algorithm 1;
(2) the data-independent criterion (28);
(3) a simple density estimation approach is considered, based on the principle of maximum entropy (MaxEnt) [3]. MaxEnt estimates a maximally smooth density consistent with any evaluated moments $\hat{g}_i(x)_n$ (11). When the mean, $\hat{x}_n$, and variance, $\hat{\sigma}_n^2$, are tracked, then the MaxEnt density estimate is Gaussian, $\mathscr{N}_n \equiv \mathscr{N}(\hat{x}_n, \hat{\sigma}_n^2)$. A heuristic stopping criterion examines the KLD between consecutive Gaussian density estimates

$$N = \min\{n : \mathsf{KLD}(\mathscr{N}_n \| \mathscr{N}_{n-1}) < \varepsilon\} \tag{53}$$

where [10]

$$\mathsf{KLD}(\mathscr{N}_n \| \mathscr{N}_{n-1}) = \frac{1}{2}\left[\ln\left(\frac{\hat{\sigma}_{n-1}^2}{\hat{\sigma}_n^2}\right) - 1 + \frac{\hat{\sigma}_n^2}{\hat{\sigma}_{n-1}^2} + \frac{(\hat{x}_n - \hat{x}_{n-1})^2}{\hat{\sigma}_{n-1}^2}\right]$$

In all three cases of stopping, the threshold is set at $\varepsilon = 0.01$.

A typical sample set at stopping, $\{x\}_N$, is illustrated in Figure 1 (right), with stopping induced at $N = 811$, using Algorithm 1. Several outliers are realized, being characteristic of this heavy-tailed distribution. The associated realized sequence of predictive KLDs (21) is plotted in Figure 2 (left), along with the recursively modelled KLD, $n^{-\hat{c}_n}$ (18), and the data-independent reverse KLD, $\ln(v_n/v_{n-1})$ (26). The latter is a far less conservative criterion, and induces stopping deterministically at $N = 98$ (29) for the chosen values of ε and v_0.

The data-dependence of the stopping criterion in Algorithm 1 is explored in a Monte Carlo (MC) simulation. The realized stopping numbers, N, for 200 repetitions is illustrated in Figure 2 (right). In each trial, the terminal posterior mean, $\hat{x}_N$ (11), is evaluated (Figure 3) (left). The realized terminal means for deterministic stopping with $\ln(v_n/v_{n-1})$ are shown in Figure 3 (centre). Their variability suggests that stopping has occurred prematurely. Finally, the realized

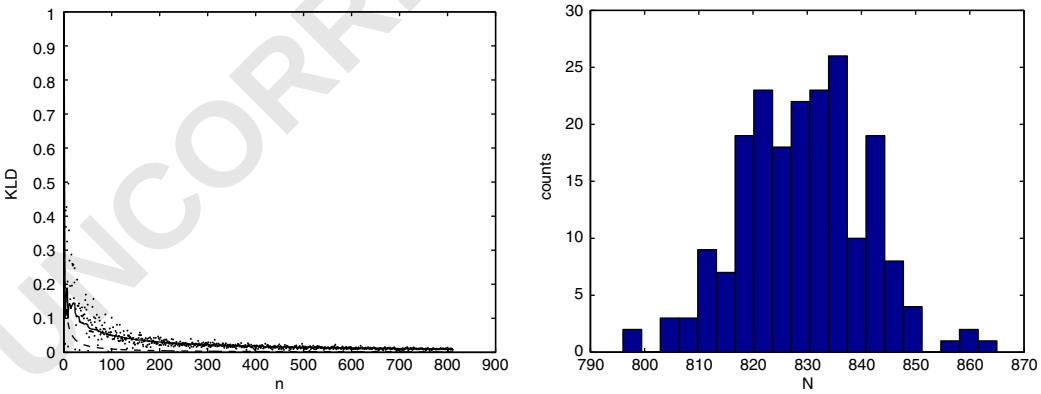

Figure 2. Left: realized KLD sequence, $\mathsf{KLD}[F_n \| F_{n-1}; \mathbb{P}_{n+1}]$ (21), up to stopping, using Algorithm 1 (dots); modelled KLD, $n^{-\hat{c}_n}$ (18) (full line); $\ln(v_n/v_{n-1})$ (26) (dashed line). Right: histogram of stopping numbers in 200 trials with Algorithm 1.

DOI: 10.1002/acs

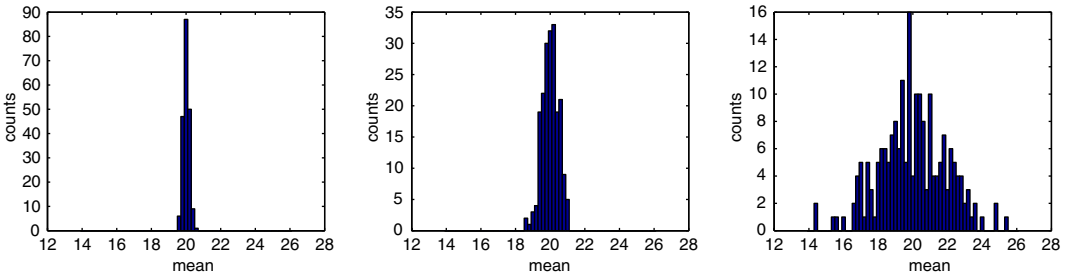

Figure 3. Terminal posterior means, $\hat{x}_N$ (11), realized in a Monte Carlo simulation with 200 repetitions. Stopping was implemented *via* three different KLDs: $\mathsf{KLD}[F_n \| F_{n-1}; \mathbb{P}_{n+1}]$ (21) (left), $\ln(v_n/v_{n-1})$ (26) (centre), and $\mathsf{KLD}(\mathcal{N}_n \| \mathcal{N}_{n-1})$ (53) (right).

means under MaxEnt-based stopping (53) are shown in Figure 3 (right). In this case, stopping occurs a.s. at $N = 5$, and is therefore unsuccessful.

Clearly, the KLD induced by the Dirichlet process (34) can track higher-order moments of $\mathscr{S}$ not assessed by the MaxEnt-based Gaussian density approximation. Of course, higher-order moment-matching techniques might be attempted, using MaxEnt or other parametric density estimators. However, a major advantage of the Dirichlet learning algorithms over any such parametric methods is revealed by these simulations; namely, these non-parametric techniques do not depend on any prior choice of moments.

6.2. *Tracking of a non-stationary random process*

A slowly non-stationary (i.e. broadband) random process is simulated with the following non-stationary marginal distribution on $(\mathbb{R}, \mathbb{B})$, $\forall t$

$$x_t \sim F_t = \mathcal{N}(m_t, r)$$

The time-variant mean is realized from the following AR(1) (AutoRegressive of order 1) process

$$m_t = -(1-\rho)m_{t-1} + \gamma e_t$$

ρ controls the bandwidth of m_t, and is set to $\rho = 10^{-4}$ in this simulation, giving a baseband process. Taking $r = 0.3$ and $\gamma^2 = 1.4 \times 10^{-4}$, then the signal-to-noise ratio (SNR) of x_t is $\mathsf{E}[m_t^2]/r = 3.7$ dB. Once again, the non-parametric prior is chosen as Dirichlet

$$F_0 \sim \mathscr{D}(\mathscr{U}_{(-100,+100]}, 5)$$

Algorithm 2 was used to track the posterior mean of x_t under i.i.d. sampling, i.e. $\hat{x}_{t,n}$ (11). Two choices of forgetting factor were considered, $\lambda = 0.1$ and 0.7, respectively, and $\varepsilon = 0.01$. The *terminal* posterior mean, $\hat{x}_{t,N_t}$, $t = 0, 1, \ldots$, is plotted along with the realized mean m_t, in Figure 4. We note the following:

(a) The mean squared error (MSE) in tracking m_t was found to be -36 dB ($\lambda = 0.1$) and -32 dB ($\lambda = 0.7$) in this simulation. A average saving of about 75% in the amount of i.i.d. sampling per stationarity interval T (Definition 2) has therefore been achieved, with only a small reduction in the quality of tracking. $\lambda > 0$ allows transmission of sampling statistics across changepoints, t, as discussed in Section 5.2, compensating successfully for

DOI: 10.1002/acs

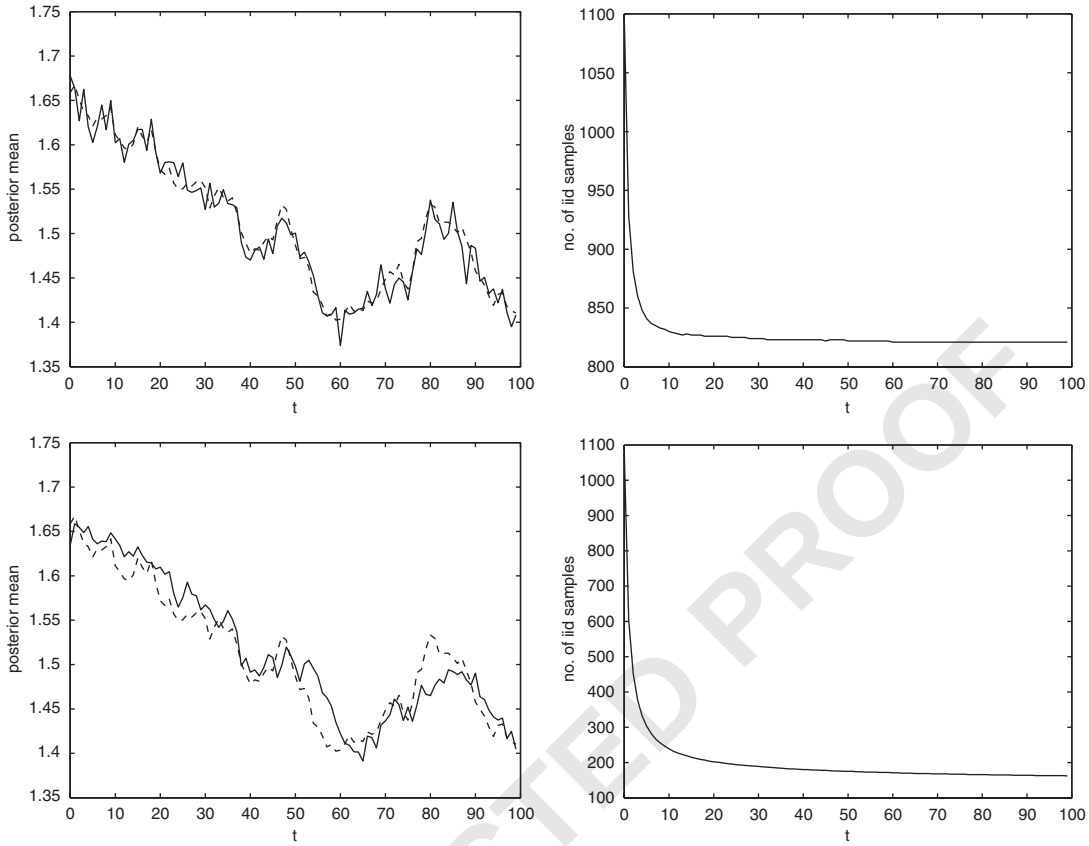

Figure 4. The posterior mean, $\hat{x}_{t,N_t}$ (11) (full line), of $x_t \sim F_t$, with $F_t|\{x\}_{t,N_t} \sim \mathscr{D}(\hat{F}_{t,N_t}, v_{t,N_t})$ (8), and forgetting factor $\lambda = 0.1$ (top), $\lambda = 0.7$ (bottom). The realized mean, m_t, is also shown (dashed line). The numbers of i.i.d. samples at stopping, $\forall t$, are plotted on the right-hand side in each case.

this reduction in sampling. To underline this point, we note that the MSE rises to -25 dB when the $\lambda = 0.7$ stopping numbers, N_t, are used in a naïve scheme that does not propagate statistics across the changepoints.

(b) The optimal choice of λ depends on the bandwidth of the random process, as explained in Section 2.5. This is controlled by the parameter ρ in these simulations. In Figure 4 (bottom) ($\lambda = 0.7$), there is some evidence of slow adaptation of the learning algorithm. Since $s_\lambda = 3.3$ (52) in this case, an assumption of stationarity over a period of $3.3T$, or, equivalently, over as many as $3.3\bar{N}$ i.i.d. samples, is being made. This underlines the need to keep λ small in most practical situations. The choice $\lambda = 0.1$ ($s_\lambda = 1.1$) provides a good compromise between the opposing goals of i.i.d. savings and effective tracking.

(c) In all cases, $N_t \to$ const. a.s. as $t \to \infty$. The rate of convergence and the terminal value are functions of the prior bounds, $\underline{x}$ and $\bar{x}$ (36), and of ε.

DOI: 10.1002/acs

7. DISCUSSION AND CONCLUSIONS

The Dirichlet process prior (DPP), $\mathscr{D}$, has proved to be a convenient non-parametric model in most respects. In particular, we have exploited the following three properties:

(1) it is conjugate with respect to i.i.d. sampling from the unknown distribution, F, leading to extremely tractable updates of moments (11);
(2) the set of $\mathscr{D}$ is closed under the non-parametric stabilized forgetting operator (43), yielding a very simple algorithm for adaptation of statistics when the Dirichlet process is slowly non-stationary;
(3) the martingale property of the induced Dirichlet distributions in the case of *finite* partitions of X^* allowed a stopping rule to be formulated.

However, a technical difficulty arose in the third task above, when attempting to derive a partition-independent divergence between successive DPPs under i.i.d. sampling. The DPP assigns zero probability to the resulting continuous distributions on $(X^*, \mathbb{A})$, and so the effect of the new i.i.d. sample on our state of knowledge cannot be assessed in such cases. The problem was circumvented by ensuring that the number of i.i.d. samples grew at least as fast as the number of partition cells. Some of the other non-parametric process priors available in the literature specifically overcome this limitation of the DPP, and therefore warrant consideration for the problem of designing partition-independent non-parametric stopping rules. The mixture of DPP assigns unit probability to the space of continuous distributions [6], and therefore warrants consideration in the current context.

In Section 3.4, we described one technique for refining the partition successfully, using the data themselves as the partition vertices. The advantages of the approach were (i) the a.s. convergence of the predictive KLD (Lemma 5) with an increasing number of i.i.d. samples; and (ii) the dependence of the stopping criterion only on the realized data and the prior, $\mathscr{D}(\hat{F}_0, v_0)$. Two disadvantages are also evident: (i) the partition refinement schedule is too fast to allow convergence of the KLD for the Dirichlet distributions themselves (Lemma 5); and (ii) the computational overhead in maintaining (i.e. storing and sorting) the set of *all* i.i.d. sample sets, $\{x\}_{t,N_t}$, $t = 0, 1, 2, \ldots$, can become prohibitive, possibly outweighing the cost of maximal i.i.d. sampling (i.e. up to $\bar{N}$), when t is large. The problem was overcome *via* a vertex decimation procedure (Section 5.1). The proposed non-parametric sequential stopping rule (Algorithm 1) performed well in simulation, leading to a reliable stopping schedule for both stationary and non-stationary non-parametric processes. The algorithms presented in this paper can be understood as a Bayesian generalization of simple histogram comparison criteria for stopping, in that it provides both prior-based regularization, and a schedule for data-based partitioning. The $\ln(v_n/v_{n-1})$ rule (28) performed reasonably, but is insensitive to (i) realized values, $\{x\}_n$, and (ii) the prior base measure, $\hat{F}_0$.

The simulation examples in Section 6 point to the relevance of the non-parametric learning algorithms in density estimation and tracking of non-stationary random processes. Work will be reported shortly on the use of these stopping rules in more ambitious practical contexts involving multivariate density estimation.

Many other data-dependent partition refinement schedules can be proposed to satisfy the requirements of Lemma 5, and merit further study.

The extension of the stabilized forgetting framework to the non-parametric case can be important in a wide variety of problems where adaptation is appropriate. An application in

DOI: 10.1002/acs

rejection sampling for difficult parametric distributions will be reported shortly. Its relevance to Markov chain Monte Carlo (MCMC) techniques, and, in particular, to particle filtering techniques, merits further study.

The paper addressed only the case of i.i.d. sampling from the Dirichlet process, F. It would be interesting to examine the possibilities for extending the reported results to dynamic data, using non-parametric modelling to 'relax' the parametric assumptions which are typically made in this case. The need then arises to model a non-parametric posterior process, F_n, which changes in response to the arrival of data. Hence, the non-stationary Dirichlet process (Section 4) may be an appropriate model in this context.

The use of non-parametric Bayesian techniques is synonymous with robustness. Their eschewing of a known parametric family in favour of a measurable space, $(F^*, \mathbb{A}_F)$, of distributions allows robust and flexible inference for problems with significant model uncertainty. Learning algorithms—such as the stopping and forgetting procedures developed in this paper—can potentially achieve far greater applicability by relaxing the parametric assumptions in favour of a non-parametric distribution modelled with a Bayesian non-parametric process prior.

To conclude, the DPP has been used to derive practical algorithms for learning of non-parametric processes *via* i.i.d. sampling. A tractable data-dependent sequential stopping rule was derived, using the KLD adapted to this non-parametric context. Likewise, a schedule for stabilized forgetting of i.i.d. samples was derived for non-stationary Dirichlet processes, by extending the appropriate parametric theory to the non-parametric case. The implied algorithm for on-line learning of a non-stationary Dirichlet process was reported. Effective tracking of the process was demonstrated in simulation.

ACKNOWLEDGEMENTS

The research was partially supported by grant GA AVČR 1ET100750401, and grant MŠ MTČR 1M6798555601.

REFERENCES

1. Bernardo JM, Smith AFM. *Bayesian Theory* (2nd edn). Wiley: Chichester, New York, Brisbane, Toronto, Singapore, 1997.
2. Bosq D. *Nonparametric Statistics for Stochastic Processes*: *Estimation and Prediction*. Springer: Berlin, 1998.
3. Rosenkrantz RD. In *Papers on Probability*, *Statistics and Statistical Physics*, Jaynes ET (ed.). D. Reidel: Dordrecht, Holland, 1983.
4. Dey D, Müller P, Sinha D. *Practical Nonparametric and Semi-Parametric Bayesian Statistics*. Springer: New York, 1998.
5. Müller P, Quintana FA. Nonparametric Bayesian data analysis. *Statistical Science* 2004; **19**(1):95–110.
6. Walker SG, Damien P, Laud PW, Smith AFM. Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society* 1999; **61**:485–527 (with Discussion).
7. Sethuraman J. A constructive definition of Dirichlet priors. *Statistica Sinica* 1994; **4**:639–650.
8. Ferguson TS. A Bayesian analysis of some nonparametric problems. *Annals of Statistics* 1973; **1**:209–230.
9. Rojíček J, Kárný M. A sequential stopping rule for extensive simulations. In *Preprints of the 3rd European IEEE Workshop on Computer-Intensive Methods in Control and Data Processing*, Rojíček J, Valečková M, Kárný M, Warwick K (eds), Praha, September 1998; 145–150 (ÚTIA AV ČR).
10. Kárný M, Kracík J, Nagy I, Nedoma P. When has estimation reached a steady state? The Bayesian sequential test. *International Journal of Adaptive Control and Signal Processing* 2005; **19**(1):41–60.
11. Kulhavý R, Zarrop MB. On a general concept of forgetting. *International Journal of Control* 1993; **58**(4):905–924.
12. Kullback S, Leibler R. On information and sufficiency. *Annals of Mathematical Statistics* 1951; **22**:79–87.
13. Bernardo JM. Expected information as expected utility. *Annals of Statistics* 1979; **7**(3):686–690.

DOI: 10.1002/acs

14. Ferguson TS. Prior distributions on spaces of probability measures. *Annals of Statistics* 1974; **2**(4):615–629.
15. Abramowitz M, Stegun IA. *Handbook of Mathematical Functions*. Dover Publications: New York, 1972.
16. Antoniak CE. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics* 1974; **2**:1152–1174.
17. Berger JO. *Statistical Decision Theory and Bayesian Analysis*. Springer: New York, 1985.
18. Lindley DV. The choice of sample size. *The Statistician* 1997; **46**(2):129–138.
19. Walker SG. How many samples: a Bayesian nonparametric approach. *The Statistician* 2003; **52**:475–492.
20. Loeve M. *Probability Theory II* (4th edn). Springer: Berlin, 1977.
21. Perez A. Information, ε-sufficiency and data reduction problem. *Kybernetika* 1965; **1**:299–310.
22. Ljung L. *System Identification*: *Theory for the User*. Prentice-Hall: London, 1987.
23. Vajda I. On convergence of information contained in quantized observations. *IEEE Transactions on Information Theory* 2002; **48**(8):2163–2172.
24. Andrýsek J. Approximate recursive Bayesian estimation of dynamic probabilistic mixtures. In *Multiple Participant Decision Making*, Andrýsek *et al*. (eds). Advanced Knowledge Int.: 2004; 39–54.

DOI: 10.1002/acs

John Wiley & Sons Ltd
The Atrium, Southern Gate, Chichester West, Sussex PO19 8SQ

Author Queries For ACS 949

While preparing this paper/manuscript for typesetting, the following queries have arisen

Query No.	Proof Page/line no	Details required	Authors response
1	References	Please provide the forename initial for the editor Andrýsek and publisher's location in Ref. 24.	

COPYRIGHT TRANSFER AGREEMENT

Wiley Production No. .....

Re: Manuscript entitled .....

(the "Contribution") written by .....

(the "Contributor") for publication in.....

(the "Journal) published by John Wiley & Sons Ltd ("Wiley").

In order to expedite the publishing process and enable Wiley to disseminate your work to the fullest extent, we need to have this Copyright Transfer Agreement signed and returned to us with the submission of your manuscript. If the Contribution is not accepted for publication this Agreement shall be null and void.

A. COPYRIGHT

1. The Contributor assigns to Wiley, during the full term of copyright and any extensions or renewals of that term, all copyright in and to the Contribution, including but not limited to the right to publish, republish, transmit, sell, distribute and otherwise use the Contribution and the material contained therein in electronic and print editions of the Journal and in derivative works throughout the world, in all languages and in all media of expression now known or later developed, and to license or permit others to do so.
2. Reproduction, posting, transmission or other distribution or use of the Contribution or any material contained therein, in any medium as permitted hereunder, requires a citation to the Journal and an appropriate credit to Wiley as Publisher, suitable in form and content as follows: (Title of Article, Author, Journal Title and Volume/Issue Copyright © [year] John Wiley & Sons Ltd or copyright owner as specified in the Journal.)

B. RETAINED RIGHTS

Notwithstanding the above, the Contributor or, if applicable, the Contributor's Employer, retains all proprietary rights other than copyright, such as patent rights, in any process, procedure or article of manufacture described in the Contribution, and the right to make oral presentations of material from the Contribution.

C. OTHER RIGHTS OF CONTRIBUTOR

Wiley grants back to the Contributor the following:

1. The right to share with colleagues print or electronic "preprints" of the unpublished Contribution, in form and content as accepted by Wiley for publication in the Journal. Such preprints may be posted as electronic files on the Contributor's own website for personal or professional use, or on the Contributor's internal university or corporate networks/intranet, or secure external website at the Contributor's institution, but not for commercial sale or for any systematic external distribution by a third party (eg: a listserver or database connected to a public access server). Prior to publication, the Contributor must include the following notice on the preprint: "This is a preprint of an article accepted for publication in [Journal title] Copyright © (year) (copyright owner as specified in the Journal)". After publication of the Contribution by Wiley, the preprint notice should be amended to read as follows: "This is a preprint of an article published in [include the complete citation information for the final version of the Contribution as published in the print edition of the Journal]" and should provide an electronic link to the Journal's WWW site, located at the following Wiley URL: http://www.interscience.wiley.com/. The Contributor agrees not to update the preprint or replace it with the published version of the Contribution.
2. The right, without charge, to photocopy or to transmit on-line or to download, print out and distribute to a colleague a copy of the published Contribution in whole or in part, for the Contributor's personal or professional use, for the advancement of scholarly or scientific research or study, or for corporate informational purposes in accordance with paragraph D2 below.
3. The right to republish, without charge, in print format, all or part of the material from the published Contribution in a book written or edited by the Contributor.
4. The right to use selected figures and tables, and selected text (up to 250 words) from the Contribution, for the Contributor's own teaching purposes, or for incorporation within another work by the Contributor that is made part of an edited work published (in print or electronic format) by a third party, or for presentation in electronic format on an internal computer network or external website of the Contributor or the Contributor's employer. The abstract shall not be included as part of such selected text.
5. The right to include the Contribution in a compilation for classroom use (course packs) to be distributed to students at the Contributor's institution free of charge or to be stored in electronic format in datarooms for access by students at the Contributor's institution as part of their course work (sometimes called "electronic reserve rooms") and for in-house training programmes at the Contributor's employer.

D. CONTRIBUTIONS OWNED BY EMPLOYER

1. If the Contribution was written by the Contributor in the course of the Contributor's employment (as a "work-made-for-hire" in the course of employment), the Contribution is owned by the company/employer which must sign this Agreement (in addition to the Contributor's signature), in the space provided below. In such case, the company/employer hereby assigns to Wiley, during the full term of copyright, all copyright in and to the Contribution for the full term of copyright throughout the world as specified in paragraph A above.
2. In addition to the rights specified as retained in paragraph B above and the rights granted back to the Contributor pursuant to paragraph C above, Wiley hereby grants back, without charge, to such company/employer, its subsidiaries and divisions, the right to make copies of and distribute the published Contribution internally in print format or electronically on the Company's internal network. Upon payment of the Publisher's reprint fee, the institution may distribute (but not re-sell) print copies of the published Contribution externally. Although copies so made shall not be available for individual re-sale, they may be included by the company/employer as part of an information package included with software or other products offered for sale or license. Posting of the published Contribution by the institution on a public access website may only be done with Wiley's written permission, and payment of any applicable fee(s).

E. GOVERNMENT CONTRACTS

In the case of a Contribution prepared under US Government contract or grant, the US Government may reproduce, without charge, all or portions of the Contribution and may authorise others to do so, for official US Government purposes only, if the US Government contract or grant so requires. (Government Employees: see note at end.)

F. COPYRIGHT NOTICE

The Contributor and the company/employer agree that any and all copies of the Contribution or any part thereof distributed or posted by them in print or electronic format as permitted herein will include the notice of copyright as stipulated in the Journal and a full citation to the Journal as published by Wiley.

G. CONTRIBUTOR'S REPRESENTATIONS

The Contributor represents that the Contribution is the Contributor's original work. If the Contribution was prepared jointly, the Contributor agrees to inform the co-Contributors of the terms of this Agreement and to obtain their signature(s) to this Agreement or their written permission to sign on their behalf. The Contribution is submitted only to this Journal and has not been published before, except for "preprints" as permitted above. (If excerpts from copyrighted works owned by third parties are included, the Contributor will obtain written permission from the copyright owners for all uses as set forth in Wiley's permissions form or in the Journal's Instructions for Contributors, and show credit to the sources in the Contribution.) The Contributor also warrants that the Contribution contains no libelous or unlawful statements, does not infringe on the right or privacy of others, or contain material or instructions that might cause harm or injury.

Tick one box and fill in the appropriate section before returning the original signed copy to the Publisher

☐ ***Contributor-owned work***

Contributor's signature ..... Date .....

Type or print name and title .....

Co-contributor's signature ..... Date .....

Type or print name and title .....

Attach additional signature page as necessary

☐ ***Company/Institution-owned work (made-for-hire in the course of employment)***

Contributor's signature ..... Date .....

Type or print name and title .....

Company or Institution
(Employer-for Hire) .....

Authorised signature of Employer ..... Date .....

Type or print name and title .....

☐ ***US Government work***

Note to US Government Employees

A Contribution prepared by a US federal government employee as part of the employee's official duties, or which is an official US Government publication is called a "US Government work", and is in the public domain in the United States. In such case, the employee may cross out paragraph A1 but must sign and return this Agreement. If the Contribution was not prepared as part of the employee's duties or is not an official US Government publication, it is not a US Government work.

☐ ***UK Government work (Crown Copyright)***

Note to UK Government Employees

The rights in a Contribution by an employee of a UK Government department, agency or other Crown body as part of his/her official duties, or which is an official government publication, belong to the Crown. In such case, the Publisher will forward the relevant form to the Employee for signature.

WILEY AUTHOR DISCOUNT CARD

As a highly valued contributor to Wiley's publications, we would like to show our appreciation to you by offering a **unique 25% discount** off the published price of any of our books*.

To take advantage of this offer, all you need to do is apply for the **Wiley Author Discount Card** by completing the attached form and returning it to us at the following address:

The Database Group
John Wiley & Sons Ltd
The Atrium
Southern Gate
Chichester
West Sussex PO19 8SQ
UK

In the meantime, whenever you order books direct from us, simply quote promotional code **S001W** to take advantage of the 25% discount.

The newest and quickest way to order your books from us is via our new European website at:

http://www.wileyeurope.com

Key benefits to using the site and ordering online include:

- Real-time SECURE on-line ordering
- The most up-to-date search functionality to make browsing the catalogue easier
- Dedicated Author resource centre
- E-mail a friend
- Easy to use navigation
- Regular special offers
- Sign up for subject orientated e-mail alerts

So take advantage of this great offer, return your completed form today to receive your discount card.

Yours sincerely,

Verity Leaver
E-marketing and Database Manager

*TERMS AND CONDITIONS

This offer is exclusive to Wiley Authors, Editors, Contributors and Editorial Board Members in acquiring books (excluding encyclopaedias and major reference works) for their personal use. There must be no resale through any channel. The offer is subject to stock availability and cannot be applied retrospectively. This entitlement cannot be used in conjunction with any other special offer. Wiley reserves the right to amend the terms of the offer at any time.

REGISTRATION FORM
FOR 25% BOOK DISCOUNT CARD

To enjoy your special discount, tell us your areas of interest and you will receive relevant catalogues or leaflets from which to select your books. Please indicate your specific subject areas below.

Accounting [ ]
- Public [ ]
- Corporate [ ]

Chemistry [ ]
- Analytical [ ]
- Industrial/Safety [ ]
- Organic [ ]
- Inorganic [ ]
- Polymer [ ]
- Spectroscopy [ ]

Encyclopedia/Reference [ ]
- Business/Finance [ ]
- Life Sciences [ ]
- Medical Sciences [ ]
- Physical Sciences [ ]
- Technology [ ]

Earth & Environmental Science [ ]

Hospitality [ ]

Genetics [ ]
- Bioinformatics/Computational [ ]
  Biology [ ]
- Proteomics [ ]
- Genomics [ ]
- Gene Mapping [ ]
- Clinical Genetics [ ]

Medical Science [ ]
- Cardiovascular [ ]
- Diabetes [ ]
- Endocrinology [ ]
- Imaging [ ]
- Obstetrics/Gynaecology [ ]
- Oncology [ ]
- Pharmacology [ ]
- Psychiatry [ ]

Non-Profit [ ]

Architecture [ ]

Business/Management [ ]

Computer Science [ ]
- Database/Data Warehouse [ ]
- Internet Business [ ]
- Networking [ ]
- Programming/Software [ ]
  Development [ ]
- Object Technology [ ]

Engineering [ ]
- Civil [ ]
- Communications Technology [ ]
- Electronic [ ]
- Environmental [ ]
- Industrial [ ]
- Mechanical [ ]

Finance/Investing [ ]
- Economics [ ]
- Institutional [ ]
- Personal Finance [ ]

Life Science [ ]

Landscape Architecture [ ]

Mathematics/Statistics [ ]

Manufacturing [ ]

Material Science [ ]

Psychology [ ]
- Clinical [ ]
- Forensic [ ]
- Social & Personality [ ]
- Health & Sport [ ]
- Cognitive [ ]
- Organizational [ ]
- Developmental and Special Ed [ ]
- Child Welfare [ ]
- Self-Help [ ]

Physics/Physical Science [ ]

[ ] I confirm that I am a Wiley Author/Editor/Contributor/Editorial Board Member of the following publications:

SIGNATURE: .....

PLEASE COMPLETE THE FOLLOWING DETAILS IN BLOCK CAPITALS:

TITLE AND NAME: (e.g. Mr, Mrs, Dr) .....

JOB TITLE: .....

DEPARTMENT: .....

COMPANY/INSTITUTION: .....

ADDRESS: .....

.....

.....

.....

TOWN/CITY: .....

COUNTY/STATE: .....

COUNTRY: .....

POSTCODE/ZIP CODE: .....

DAYTIME TEL: .....

FAX: .....

E-MAIL: .....

YOUR PERSONAL DATA
We, John Wiley & Sons Ltd, will use the information you have provided to fulfil your request. In addition, we would like to:

1. Use your information to keep you informed by post, e-mail or telephone of titles and offers of interest to you and available from us or other Wiley Group companies worldwide, and may supply your details to members of the Wiley Group for this purpose.

   [ ] Please tick the box if you do not wish to receive this information

2. Share your information with other carefully selected companies so that they may contact you by post, fax or e-mail with details of titles and offers that may be of interest to you.

   [ ] Please tick the box if you do not wish to receive this information.

If, at any time, you wish to stop receiving information, please contact the Database Group (databasegroup@wiley.co.uk) at John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, UK.

E-MAIL ALERTING SERVICE
We offer an information service on our product ranges via e-mail. If you do not wish to receive information and offers from John Wiley companies worldwide via e-mail, please tick the box [ ].

This offer is exclusive to Wiley Authors, Editors, Contributors and Editorial Board Members in acquiring books (excluding encyclopaedias and major reference works) for their personal use. There should be no resale through any channel. The offer is subject to stock availability and may not be applied retrospectively. This entitlement cannot be used in conjunction with any other special offer. Wiley reserves the right to vary the terms of the offer at any time.

Ref: S001W