Bernoulli **27**(1), 2021, 702–726
https://doi.org/10.3150/20-BEJ1255

A class of models for Bayesian predictive inference

PATRIZIA BERTI[1], EMANUELA DREASSI[2], LUCA PRATELLI[3] and PIETRO RIGO[4]

[1]*Dipartimento di Matematica Pura ed Applicata "G. Vitali", Università di Modena e Reggio-Emilia, via Campi 213/B, 41100 Modena, Italy. E-mail: patrizia.berti@unimore.it*
[2]*Dipartimento di Statistica, Informatica, Applicazioni Università di Firenze, viale Morgagni 59, 50134 Firenze, Italy. E-mail: emanuela.dreassi@unifi.it*
[3]*Accademia Navale di Livorno, viale Italia 72, 57100 Livorno, Italy. E-mail: pratel@mail.dm.unipi.it*
[4]*Dipartimento di Scienze Statistiche "P. Fortunati", Università di Bologna, via delle Belle Arti 41, 40126 Bologna, Italy. E-mail: pietro.rigo@unibo.it*

In a Bayesian framework, to make predictions on a sequence $X_1, X_2, \dots$ of random observations, the inferrer needs to assign the predictive distributions $\sigma_n(\cdot) = P(X_{n+1} \in \cdot \mid X_1, \dots, X_n)$. In this paper, we propose to assign σ_n directly, without passing through the usual prior/posterior scheme. One main advantage is that no prior probability has to be assessed. The data sequence (X_n) is assumed to be conditionally identically distributed (c.i.d.) in the sense of (*Ann. Probab.* **32** (2004) 2029–2052). To realize this programme, a class Σ of predictive distributions is introduced and investigated. Such a Σ is rich enough to model various real situations and (X_n) is actually c.i.d. if σ_n belongs to Σ. Furthermore, when a new observation X_{n+1} becomes available, σ_{n+1} can be obtained by a simple recursive update of σ_n. If μ is the a.s. weak limit of σ_n, conditions for μ to be a.s. discrete are provided as well.

Keywords: Bayesian nonparametrics; conditional identity in distribution; exchangeability; predictive distribution; random probability measure; sequential predictions; strategy

1. Introduction

The object of this paper is Bayesian predictive inference for a sequence of random observations. Let $(X_n : n \geq 1)$ be a sequence of random variables with values in a measurable space $(S, \mathcal{B})$. Assuming that $(X_1, \dots, X_n) = x$, for some $n \geq 1$ and $x \in S^n$, the problem consists of predicting X_{n+1} based on the observed data x. In a Bayesian framework, this means to assess the *predictive distribution*, say

$$\sigma_n(x)(B) = P\big(X_{n+1} \in B \mid (X_1, \dots, X_n) = x\big) \quad \text{for all } B \in \mathcal{B}.$$

To address this problem, the X_n can be taken to be the coordinate random variables on S^∞. Accordingly, in the sequel, we let

$$X_n(s_1, \dots, s_n, \dots) = s_n$$

for each $n \geq 1$ and each $(s_1, \dots, s_n, \dots) \in S^\infty$. Also, to avoid needless technicalities, S is assumed to be a Borel subset of a Polish space and $\mathcal{B}$ the Borel σ-field on S.

Let $\mathcal{P}$ denote the collection of all probability measures on $\mathcal{B}$. Following Dubins and Savage [15], a *strategy* is a sequence

$$\sigma = (\sigma_0, \sigma_1, \dots)$$

such that

1350-7265 © 2021 ISI/BS

- $\sigma_0 \in \mathcal{P}$ and $\sigma_n = \{\sigma_n(x) : x \in S^n\}$ is a collection of elements of $\mathcal{P}$;
- The map $x \mapsto \sigma_n(x)(B)$ is $\mathcal{B}^n$-measurable for fixed $n \geq 1$ and $B \in \mathcal{B}$.

Here, σ_0 should be regarded as the marginal distribution of X_1 and $\sigma_n(x)$ as the conditional distribution of X_{n+1} given that $(X_1, \dots, X_n) = x$.

According to the Ionescu-Tulcea theorem, for any strategy σ, there is a unique probability measure P on $(S^\infty, \mathcal{B}^\infty)$ satisfying

$$P(X_1 \in \cdot) = \sigma_0 \quad \text{and} \quad P\big(X_{n+1} \in \cdot \mid (X_1, \dots, X_n) = x\big) = \sigma_n(x)$$
$$\text{for all } n \geq 1 \text{ and } P\text{-almost all } x \in S^n.$$

Such a P is denoted as P_σ in the sequel.

To make predictions on the sequence (X_n), a Bayesian inferrer needs precisely a strategy σ. The Ionescu-Tulcea theorem establishes that, for *any* strategy σ, the predictions based on σ are consistent with a unique probability distribution for (X_n).

1.1. The standard and non-standard approach for exchangeable data

The data sequence (X_n) is usually assumed to be exchangeable. In that case, there are essentially two procedures for selecting a strategy σ. For definiteness, we call them the *standard approach* (SA) and the *non-standard approach* (NSA). The only reason for using these terms is that the first approach is much more popular than the second. Both approaches can be adopted to make Bayesian predictive inference and both lead to a full specification of the probability distribution of (X_n).

According to SA, to obtain σ, the inferrer should:

- Select a prior π, namely, a probability measure on $\mathcal{P}$;
- Calculate the posterior of π given that $(X_1, \dots, X_n) = x$, say $\pi_n(x)$;
- Evaluate σ as

$$\sigma_0(B) = \int_{\mathcal{P}} p(B)\pi(dp) \quad \text{and} \quad \sigma_n(x)(B) = \int_{\mathcal{P}} p(B)\pi_n(x)(dp) \quad \text{for all } B \in \mathcal{B}.$$

To assess a prior π is not an easy task. In addition, once π is selected, it is also quite difficult to evaluate the posterior $\pi_n(x)$. Frequently, it happens that $\pi_n(x)$ cannot be written in closed form but only approximated numerically.

On the other hand, SA is not motivated by prediction alone. Another motivation, possibly the main one, is to make inference on other features of the data distribution, such as a mean, a quantile, a correlation, or more generally some random parameter (possibly, infinite dimensional). In all these cases, the posterior $\pi_n(x)$ is fundamental. In short, SA is a cornerstone of Bayesian inference, but, when prediction is the main target, is possibly quite involved.

Instead, NSA entails assigning σ_n directly, without passing through π and π_n. Merely, rather than choosing π and then evaluating π_n and σ_n, the inferrer just selects his/her predictive distribution σ_n. This procedure makes sense because of the Ionescu-Tulcea theorem. See [3,4,8,11,12,16,19,20,24]; see also [17,25,26,29] and references therein.

NSA is in line with de Finetti, Dubins and Savage, among others. Pitman's work is fundamental as well; see, for example, [27] and [28]. In fact, NSA is usually adopted (or at least implicit) in species sampling models; see [24].

Similarly, NSA is used in [19] to obtain a fast online Bayesian prediction. Suppose that $S = \mathbb{R}$ and $\sigma_n(x)$ admits a density, with respect to some fixed measure λ on $\mathcal{B}$, for all $n \geq 0$ and $x \in S^n$. In

[19], the update of predictive distributions is given a nice characterization in terms of copulas. Such a characterization, in turn, allows for making Bayesian predictions through an useful recursive procedure. In a sense, the present paper fits into the framework of [19].

From our point of view, NSA has essentially two merits. Firstly, it requires the assignment of probabilities on *observable facts* only. The value of the next observation X_{n+1} is actually observable, while π and π_n (being probabilities on $\mathcal{P}$) do not deal with observable facts. Secondly, as noted in [19], Section 6, NSA is much more efficient than SA when prediction is the main goal. In this case, why select the prior π explicitly? Rather than wondering about π, it seems reasonable to reflect on how X_{n+1} is affected by $(X_1, \dots, X_n)$.

Finally, NSA is even more appealing in a *nonparametric* framework, where selecting a prior with large support is usually difficult.

We discuss an example to make the above remarks clearer.

Example 1 (SA versus NSA). If (X_n) is exchangeable, de Finetti's theorem yields

$$P(X_1 \in B_1, \dots, X_n \in B_n) = \int_\Theta \prod_{i=1}^n P_\theta(B_i)\pi(d\theta)$$

for some parameter space Θ, some prior π on Θ, and some statistical model

$$M = \{P_\theta : \theta \in \Theta\} \quad \text{where } P_\theta \in \mathcal{P} \text{ for each } \theta.$$

In the parametric case, Θ is a Borel subset of $\mathbb{R}^k$ and M is dominated and smooth. In the nonparametric case, Θ is infinite-dimensional, typically $\Theta = \mathcal{P}$. In both cases, SA entails selecting π, evaluating the posterior π_n and calculating σ as

$$\sigma_n(x)(B) = \int_\Theta P_\theta(B)\pi_n(x)(d\theta).$$

In turn, NSA entails selecting σ directly, without passing through π and π_n.

In our opinion, SA may be unsuitable for prediction even in the parametric framework. Not only it is hard to choose π, but to evaluate π_n may be difficult as well. On the contrary, NSA usually takes the available information on the data into account more effectively. In fact, in various practical situations, arguing in terms of strategies is simpler than arguing in terms of priors. An obvious example are Polya urns, where the strategy σ is naturally determined by the sampling scheme, while the prior π is not. The merits of NSA increase further in the nonparametric framework. In that case, if prediction is the main goal, to assess a prior π and evaluate the posterior π_n is really too expensive.

One more remark is in order. Because of exchangeability, the probability distribution P of (X_n) can be written as above for some M and π. However, by assigning a strategy σ, the inferrer identifies $P = P_\sigma$, not the pair (M, π). In a sense, when applying NSA, the "model uncertainty" about θ is integrated out by the choice of σ. This appears reasonable after all, as when making predictions, the relevant object is σ not (M, π). An intriguing problem, pioneered by Diaconis, Ylvisaker and Freedman, is to give conditions on σ implying that the statistical model M underlying P_σ has a given form, for instance M is an exponential family. Such a problem, however, is not investigated in this paper. See [13,14,16,30] and references therein.

1.2. Conditionally identically distributed data

If (X_n) is assumed to be exchangeable, however, NSA has a gap. Given an arbitrary strategy σ, the Ionescu-Tulcea theorem does not grant exchangeability of (X_n) under P_σ. Therefore, for NSA to apply, one should first characterize those strategies σ which make (X_n) exchangeable under P_σ. A nice characterization is [16], Theorem 3.1. However, the conditions on σ for making (X_n) exchangeable are quite hard to check in real problems. This is possibly one of the reasons why NSA has not yet been developed. Another reason is the lack of constructive procedures for determining σ. It is precisely this lack which makes SA necessary for prediction, even if analytically more involved.

An obvious way to bypass the gap mentioned in the above paragraph is to weaken the exchangeability assumption. One option is to assume (X_n) to be *conditionally identically distributed* (c.i.d.), namely

$$P(X_k \in \cdot \mid \mathcal{F}_n) = P(X_{n+1} \in \cdot \mid \mathcal{F}_n) \quad \text{a.s. for all } k > n \geq 0$$

where $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$ and $\mathcal{F}_0$ is the trivial σ-field.

Roughly speaking, the above condition means that, at each time $n \geq 0$, the future observations $(X_k : k > n)$ are identically distributed given the past $\mathcal{F}_n$. Such a condition is actually weaker than exchangeability. Indeed, (X_n) is exchangeable if and only if it is stationary and c.i.d.

We refer to Section 2.1 for more information on c.i.d. sequences. Here, we just mention three reasons for taking c.i.d. data into account.

- It is not hard to characterize the strategies σ which make (X_n) c.i.d. under P_σ; see Theorem 3. Therefore, unlike the exchangeable case, NSA can be easily implemented.
- The asymptotic behavior of c.i.d. sequences is analogous to that of exchangeable ones.
- A number of meaningful strategies cannot be used if (X_n) is assumed to be exchangeable, but are available if (X_n) is only required to be c.i.d. See the examples in Sections 4–6.

To support the latter claim, we also note that conditional identity in distribution is a more appropriate assumption than exchangeability in some real problems. Examples occur in various fields, including clinical trials, generalized Polya urns, species sampling models and disease surveillance; see [1,2,5] and [10].

1.3. Further notation and conditions (a)–(b)

A *kernel* (or a *random probability measure*) on $(S, \mathcal{B})$ is a collection

$$\alpha = \{\alpha(x) : x \in S\}$$

such that $\alpha(x) \in \mathcal{P}$ for each $x \in S$ and the map $x \mapsto \alpha(x)(B)$ is measurable for fixed $B \in \mathcal{B}$. Here, $\alpha(x)(B)$ denotes the value taken at B by the probability measure $\alpha(x)$.

Let $\sigma_0 \in \mathcal{P}$ and α a kernel on $(S, \mathcal{B})$. In the sequel, σ_0 and α are such that:

(a) σ_0 is a stationary distribution for α, namely,

$$\sigma_0(B) = \int \alpha(x)(B)\sigma_0(dx) \quad \text{for all } B \in \mathcal{B};$$

(b) There is a set $A \in \mathcal{B}$ such that $\sigma_0(A) = 1$ and

$$\alpha(x)(B) = \int \alpha(z)(B)\alpha(x)(dz) \quad \text{for all } x \in A \text{ and } B \in \mathcal{B}.$$

Conditions (a)–(b) are not so unusual. For instance, they are satisfied whenever α is a regular conditional distribution for σ_0 given any sub-σ-field of $\mathcal{B}$; see Lemma 6. In particular, conditions (a)–(b) trivially hold if

$$A = S \quad \text{and} \quad \alpha(x) = \delta_x \quad \text{for all } x \in S$$

where δ_x denotes the point mass at x.

Finally, if $x = (x_1, \dots, x_n) \in S^n$ and $y \in S$, we write (x, y) to denote

$$(x, y) = (x_1, \dots, x_n, y).$$

In addition, for any strategy σ, we let

$$S^0 = \{\emptyset\}, \qquad \sigma_0(\emptyset) = \sigma_0, \qquad \sigma_1(\emptyset, y) = \sigma_1(y).$$

1.4. Content of this paper

We aim to develop NSA for c.i.d. data. To this end, we introduce and investigate a class Σ of strategies. Such a Σ is rich enough to model various real situations and (X_n) is c.i.d. under P_σ for each $\sigma \in \Sigma$. Furthermore, when a new observation X_{n+1} becomes available, σ_{n+1} can be obtained from a simple recursive update of σ_n.

Each $\sigma \in \Sigma$ can be described as follows. Fix $\sigma_0 \in \mathcal{P}$ and a kernel α on $(S, \mathcal{B})$ satisfying conditions (a)–(b). Also, for every $n \geq 0$, fix a measurable function $f_n : S^{n+2} \to [0, 1]$ such that

$$f_n(x, y, z) = f_n(x, z, y) \quad \text{for all } x \in S^n \text{ and } (y, z) \in S^2.$$

Given σ_0, α and $(f_n : n \geq 0)$, a strategy σ can be obtained via the recursive equation

$$\sigma_{n+1}(x, y)(B) = \int \alpha(z)(B) f_n(x, y, z)\sigma_n(x)(dz) + \alpha(y)(B)\left\{1 - \int f_n(x, y, z)\sigma_n(x)(dz)\right\}$$

for all $n \geq 0$, $B \in \mathcal{B}$, $x \in S^n$ and $y \in S$.

We define Σ as the collection of all the strategies σ obtained as above.

The simplest example corresponds to

$$f_n(x, y, z) = q_n(x),$$

where $q_n : S^n \to [0, 1]$ is any measurable map (with q_0 constant). In that case, the recursive equation reduces to

$$\sigma_{n+1}(x, y) = q_n(x)\sigma_n(x) + \{1 - q_n(x)\}\alpha(y) \tag{1}$$

for all $n \geq 0$, $x \in S^n$ and $y \in S$. (Here, for the sake of simplicity, we are assuming $A = S$ where A is the set involved in condition (b).)

In this specific case, the updating rule is quite transparent: $\sigma_{n+1}(x, y)$ is just a convex combination of the previous predictive distribution $\sigma_n(x)$ and the kernel α evaluated in the last observation y. In addition, the weight $q_n(x)$ does not depend on y. Note also that $\sigma_{n+1}(x, y)$ can be written explicitly (and not only in recursive form) as

$$\sigma_{n+1}(x, y) = \sigma_0 \prod_{i=0}^{n} q_i + \alpha(y)(1 - q_n) + \sum_{i=1}^{n} \alpha(x_i)(1 - q_{i-1}) \prod_{j=i}^{n} q_j,$$

where $x=(x_1,\dots,x_n)\in S^n$, $y\in S$ and q_i is a shorthand notation to denote

$$q_i=q_i(x_1,\dots,x_i).$$

In general, specifying f_n and α suitably, various meaningful strategies can be shown to be members of Σ. Some of these strategies are well known and some are new (in the sense that, to our knowledge, they have not been proposed to date). Examples of the former are the predictive distributions of Dirichlet sequences, species sampling sequences and generalized Polya urns. Examples of the latter are the strategies of Sections 5–6.

Another nice feature of Σ is that it also includes diffuse strategies, and this fact may be useful to model real situations. We recall that a probability measure is *diffuse* if it vanishes on singletons, and a strategy σ is diffuse if $\sigma_n(x)$ is diffuse for all $n\geq 0$ and $x\in S^n$.

In addition to introducing Σ, our main contributions are Theorems 4–5 and Theorems 18–20. The former state that (X_n) is c.i.d. under P_σ for each $\sigma\in\Sigma$, while the latter deal with the asymptotics of σ_n. A few words should be spent on Theorem 18.

Let $X_1^*, X_2^*,\dots$ denote the (finite or infinite) sequence of distinct values corresponding to the observations $X_1, X_2,\dots$ If (X_n) is c.i.d. under P_σ, where σ is *any* strategy (possibly not belonging to Σ), there is a random probability measure μ on $(S,\mathcal{B})$ such that

$$\sigma_n(B)\overset{\text{a.s.}}{\to}\mu(B)\quad\text{for every fixed } B\in\mathcal{B}$$

where "a.s." stands for "P_σ-a.s."; see Section 2.1. Theorem 18 states that

$$\mu\overset{\text{a.s.}}{=}\sum_k W_k\delta_{X_k^*},$$

for some random weights $W_k\geq 0$ such that $\sum_k W_k=1$, if and only if

$$\lim_n P_\sigma(X_n\neq X_i\ \text{for each } i<n)=0.$$

Furthermore, W_k admits the representation

$$W_k\overset{\text{a.s.}}{=}\lim_n\frac{1}{n}\sum_{i=1}^n 1_{\{X_i=X_k^*\}}.$$

By applying Theorem 18 to $\sigma\in\Sigma$, it is not hard to give conditions on f_n and α implying that μ is a.s. discrete. Conditions for $X_1^*, X_2^*,\dots$ to be i.i.d. and independent of the weights $W_1, W_2,\dots$ are given as well.

It is worth noting that Theorem 18 holds true for any strategy σ which makes (X_n) c.i.d. Hence, Theorem 18 extends a known fact concerning exchangeability to all c.i.d. sequences; see, for example, [23].

In addition to the results quoted above, other main contributions of this paper are the examples included in Sections 4–6. In our opinion, these examples should support the fact that Σ is rich enough to cover a wide range of problems.

2. Preliminaries

2.1. Conditional identity in distribution

C.i.d. sequences have been introduced in [5] and [21] and then investigated in various papers; see, for example, [1–4,6,7,10,18]. Here, we just recall a few basic facts.

Let $(\mathcal{G}_n : n \geq 0)$ be a filtration and $(Y_n : n \geq 1)$ a sequence of S-valued random variables. Then, (Y_n) is c.i.d. with respect to $(\mathcal{G}_n)$ if it is adapted to $(\mathcal{G}_n)$ and

$$P(Y_k \in \cdot \mid \mathcal{G}_n) = P(Y_{n+1} \in \cdot \mid \mathcal{G}_n) \quad \text{a.s. for all } k > n \geq 0.$$

When $(\mathcal{G}_n)$ is the canonical filtration of (Y_n), the filtration is not mentioned at all and (Y_n) is just called c.i.d. From a result in [21], (Y_n) is exchangeable if and only if it is stationary and c.i.d.

Let (Y_n) be c.i.d., $\mathcal{G}_n = \sigma(Y_1, \dots, Y_n)$, and

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{Y_i}$$

the empirical measure. In a sense, the asymptotic behavior of (Y_n) is similar to that of an exchangeable sequence. This claim can be supported by two facts.

First, there is a random probability measure μ on $(S, \mathcal{B})$ satisfying

$$\mu_n(B) \xrightarrow{\text{a.s.}} \mu(B) \quad \text{for every fixed } B \in \mathcal{B}.$$

As a consequence, for fixed $n \geq 0$ and $B \in \mathcal{B}$, one obtains

$$\begin{aligned} E\{\mu(B) \mid \mathcal{G}_n\} &= \lim_m E\{\mu_m(B) \mid \mathcal{G}_n\} \\ &= \lim_m \frac{1}{m} \sum_{i=n+1}^{m} P(Y_i \in B \mid \mathcal{G}_n) = P(Y_{n+1} \in B \mid \mathcal{G}_n) \quad \text{a.s.} \end{aligned}$$

Thus, as in the exchangeable case, the predictive distribution $P(Y_{n+1} \in \cdot \mid \mathcal{G}_n)$ can be written as $E\{\mu(\cdot) \mid \mathcal{G}_n\}$, where μ is the a.s. weak limit of the empirical measures μ_n. In particular, for each $B \in \mathcal{B}$, the martingale convergence theorem implies

$$P(Y_{n+1} \in B \mid \mathcal{G}_n) = E\{\mu(B) \mid \mathcal{G}_n\} \xrightarrow{\text{a.s.}} \mu(B). \tag{2}$$

Second, (Y_n) is asymptotically exchangeable, in the sense that

$$(Y_n, Y_{n+1}, \dots) \to (Z_1, Z_2, \dots) \quad \text{in distribution, as } n \to \infty,$$

where (Z_n) is an exchangeable sequence. Moreover, (Z_n) is directed by μ, namely

$$P(Z_1 \in B_1, \dots, Z_k \in B_k) = E\left\{\prod_{i=1}^{k} \mu(B_i)\right\}$$

for all $k \geq 1$ and $B_1, \dots, B_k \in \mathcal{B}$.

The role played by μ is not as crucial as in the exchangeable case, since the probability distribution of (Y_n) is not completely determined by μ; see Example 17. Nevertheless, μ is a meaningful random parameter for (Y_n). In fact, $\mu(B)$ is the long run frequency of the events $\{Y_n \in B\}$. Similarly, because of (2), $\mu(B)$ can be regarded as the asymptotically optimal predictor of the event {the next observation belongs to B}. And finally, μ is the directing measure of the exchangeable limit sequence (Z_n).

2.2. Stationarity, reversibility and characterizations

We first recall some definitions. Let $\tau \in \mathcal{P}$ and $\alpha = \{\alpha(x) : x \in S\}$ a kernel on $(S, \mathcal{B})$. Then:

- τ is a *stationary distribution* for α if
$$\int \alpha(x)(B)\tau(dx) = \tau(B) \quad \text{for all } B \in \mathcal{B};$$
- α is *reversible* with respect to τ if
$$\int_A \alpha(x)(B)\tau(dx) = \int_B \alpha(x)(A)\tau(dx) \quad \text{for all } A, B \in \mathcal{B};$$
- α is a *regular conditional distribution* for τ given $\mathcal{G}$, where $\mathcal{G} \subset \mathcal{B}$ is a sub-σ-field, if $x \mapsto \alpha(x)(B)$ is $\mathcal{G}$-measurable and
$$\int_A \alpha(x)(B)\tau(dx) = \tau(A \cap B) \quad \text{for all } A \in \mathcal{G} \text{ and } B \in \mathcal{B}.$$

Since S is nice (it is in fact a Borel subset of a Polish space), for any sub-σ-field $\mathcal{G} \subset \mathcal{B}$, there exists a τ-a.s. unique regular conditional distribution for τ given $\mathcal{G}$; see e.g. [22], page 107. Note also that reversibility implies stationarity (just take $A = S$) but not conversely. In addition, τ is a stationary distribution for α provided α is a regular conditional distribution for τ (take $A = S$ again).

We next characterize exchangeable and c.i.d. sequences in terms of strategies.

Theorem 2 ([16], Theorem 3.1). *For any strategy σ, (X_n) is exchangeable under P_σ if and only if*

(i) *The kernel $\{\sigma_{n+1}(x, y) : y \in S\}$ is reversible with respect to $\sigma_n(x)$ for all $n \geq 0$ and P_σ-almost all $x \in S^n$;*
(ii) *$\sigma_n(x) = \sigma_n(f(x))$ for all $n \geq 2$, all permutations f on S^n and P_σ-almost all $x \in S^n$.*

To deal with the c.i.d. case, it suffices to drop condition (ii) and to replace "reversible" with "stationary" in condition (i).

Theorem 3 ([6], Theorem 3.1). *For any strategy σ, (X_n) is c.i.d. under P_σ if and only if*

(i*) *The kernel $\{\sigma_{n+1}(x, y) : y \in S\}$ has stationary distribution $\sigma_n(x)$ for all $n \geq 0$ and P_σ-almost all $x \in S^n$.*

An obvious consequence of Theorem 3 is that (X_n) is c.i.d. under P_σ whenever $\{\sigma_{n+1}(x, y) : y \in S\}$ has stationary distribution $\sigma_n(x)$ for all $n \geq 0$ and all $x \in C^n$, where $C \in \mathcal{B}$ is any set with $\sigma_0(C) = 1$.

Theorem 3 also suggests how to assess a c.i.d. sequence stepwise. First, select $\sigma_0 \in \mathcal{P}$, the marginal distribution of X_1. Then, choose a kernel $\{\sigma_1(y) : y \in S\}$ with stationary distribution σ_0, where $\sigma_1(y)$ is the conditional distribution of X_2 given $X_1 = y$. Next, for each $x \in S$, select a kernel $\{\sigma_2(x, y) : y \in S\}$ with stationary distribution $\sigma_1(x)$, where $\sigma_2(x, y)$ is the conditional distribution of X_3 given $X_1 = x$ and $X_2 = y$. And so on. In other terms, for getting a c.i.d. sequence, it is sufficient to assign a kernel with a given stationary distribution at each step.

3. A sequential updating rule

Our starting point is the following simple fact.

Theorem 4. *Let $\tau \in \mathcal{P}$ and $f : S^2 \to [0,1]$ a measurable symmetric function. Fix a kernel $\alpha = \{\alpha(x) : x \in S\}$ on $(S, \mathcal{B})$ and define*

$$\beta(x)(B) = \int \alpha(z)(B) f(x,z)\tau(dz) + \alpha(x)(B) \int \big(1 - f(x,z)\big)\tau(dz)$$

for all $x \in S$ and $B \in \mathcal{B}$. Then, $\beta = \{\beta(x) : x \in S\}$ is a kernel on $(S, \mathcal{B})$. Moreover:

- *If τ is stationary for α, then τ is stationary for β;*
- *If $\alpha(x) = \delta_x$ for all $x \in S$, then β is reversible with respect to τ.*

Proof. Let $\phi(x) = \int f(x,z)\tau(dz)$. If $\phi(x) = 0$, then $\beta(x)$ is clearly a probability measure on $\mathcal{B}$. If $\phi(x) \in (0,1]$,

$$\beta(x)(B) = \phi(x) \frac{\int \alpha(z)(B) f(x,z)\tau(dz)}{\phi(x)} + \big(1 - \phi(x)\big)\alpha(x)(B).$$

Hence, $\beta(x) \in \mathcal{P}$ for all $x \in S$. Further, for fixed $B \in \mathcal{B}$, the map $x \mapsto \beta(x)(B)$ is measurable because of Fubini's theorem. Thus, β is a kernel on $(S, \mathcal{B})$.

Next, suppose τ is stationary for α. Since $f(x,z) = f(z,x)$, one obtains

$$\begin{aligned}\int \beta(x)(B)\tau(dx) &= \int\int \alpha(z)(B) f(x,z)\tau(dz)\tau(dx) \\ &\quad + \int \alpha(x)(B)\tau(dx) - \int \alpha(x)(B)\phi(x)\tau(dx) \\ &= \int \alpha(z)(B) \int f(z,x)\tau(dx)\tau(dz) + \tau(B) - \int \alpha(x)(B)\phi(x)\tau(dx) \\ &= \int \alpha(z)(B)\phi(z)\tau(dz) + \tau(B) - \int \alpha(x)(B)\phi(x)\tau(dx) = \tau(B)\end{aligned}$$

for all $B \in \mathcal{B}$. Thus, τ is stationary for β.

Finally, if $\alpha(x) = \delta_x$, then

$$\begin{aligned}\int_A \beta(x)(B)\tau(dx) &= \int\int 1_A(x)1_B(z) f(x,z)\tau(dz)\tau(dx) \\ &\quad + \int 1_A(x)1_B(x)\tau(dx) - \int 1_A(x)1_B(x)\phi(x)\tau(dx)\end{aligned}$$

for all $A, B \in \mathcal{B}$. It follows that

$$\begin{aligned}&\int_A \beta(x)(B)\tau(dx) - \int_B \beta(x)(A)\tau(dx) \\ &= \int\int 1_A(x)1_B(z) f(x,z)\tau(dz)\tau(dx) - \int\int 1_B(x)1_A(z) f(x,z)\tau(dz)\tau(dx) \\ &= \int 1_B(z) \int 1_A(x) f(z,x)\tau(dx)\tau(dz) - \int 1_B(x) \int 1_A(z) f(x,z)\tau(dz)\tau(dx) = 0.\end{aligned}$$

Thus, β is reversible with respect to τ. □

Heuristically, in the special case $\alpha(x)=\delta_x$, the idea underlying β reminds of Metropolis' algorithm. Starting from a state x, one first selects a new state z according to τ, and then goes to z or remains in x with probabilities $f(x,z)$ and $1-f(x,z)$, respectively. This naive idea can be adapted to an arbitrary kernel α as follows. First, select z according to τ. Then, the new state y is drawn from $\alpha(z)$ with probability $f(x,z)$, or from $\alpha(x)$ with probability $1-f(x,z)$. From our point of view, however, what is meaningful is that this idea provides a simple updating procedure.

Next, as in Section 1.4, fix $\sigma_0\in\mathcal{P}$, a kernel α on $(S,\mathcal{B})$ and a sequence of measurable functions $f_n:S^{n+2}\to[0,1]$ such that

$$f_n(x,y,z)=f_n(x,z,y)\quad\text{for all } n\geq 0, x\in S^n \text{ and } (y,z)\in S^2.$$

The kernel α is assumed to satisfy conditions (a)–(b) of Section 1.3. We recall that (a)–(b) are automatically true if α is a regular conditional distribution for σ_0 given any sub-σ-field $\mathcal{G}\subset\mathcal{B}$; see Lemma 6 below. In particular, conditions (a)–(b) hold if $\alpha(x)=\delta_x$ for all $x\in S$.

Given σ_0, α and $(f_n:n\geq 0)$, define a strategy σ according to

$$\sigma_{n+1}(x,y)(B)=\int\alpha(z)(B)f_n(x,y,z)\sigma_n(x)(dz)+\alpha(y)(B)\left\{1-\int f_n(x,y,z)\sigma_n(x)(dz)\right\}$$

for all $n\geq 0$, $x\in S^n$, $y\in S$ and $B\in\mathcal{B}$.

Note that, when a new observation y becomes available, $\sigma_{n+1}(x,y)$ is just a recursive update of $\sigma_n(x)$.

Let Σ denote the collection of all the strategies σ obtained in this way, for σ_0, α and $(f_n:n\geq 0)$ varying. Each $\sigma\in\Sigma$ makes (X_n) c.i.d.

Theorem 5. *Let $\sigma\in\Sigma$. Then, (X_n) is c.i.d. under P_σ. Moreover, if $\alpha(x)=\delta_x$ for all $x\in S$, then*

$$P_\sigma\left[(X_{n+1},X_{n+2})\in\cdot\mid\mathcal{F}_n\right]=P_\sigma\left[(X_{n+2},X_{n+1})\in\cdot\mid\mathcal{F}_n\right]\quad a.s. \tag{3}$$

for all $n\geq 0$, where $\mathcal{F}_0$ is the trivial σ-field and $\mathcal{F}_n=\sigma(X_1,\dots,X_n)$.

Proof. We show that there is $C\in\mathcal{B}$ such that $\sigma_0(C)=1$ and $\{\sigma_{n+1}(x,y):y\in S\}$ has stationary distribution $\sigma_n(x)$ for all $n\geq 0$ and all $x\in C^n$. By the remark after Theorem 3, this implies that (X_n) is c.i.d. under P_σ.

Let $A\in\mathcal{B}$ be the set involved in condition (b). Define

$$A_0=A\quad\text{and}\quad A_{n+1}=\left\{x\in A_n:\alpha(x)(A_n)=1\right\}\quad\text{for all } n\geq 0.$$

If $\sigma_0(A_n)=1$ for some $n\geq 0$, condition (a) yields

$$\int\alpha(x)(A_n)\sigma_0(dx)=\sigma_0(A_n)=1,$$

which in turn implies $\sigma_0(A_{n+1})=1$. Since $\sigma_0(A_0)=\sigma_0(A)=1$, by induction, one obtains $\sigma_0(A_n)=1$ for each $n\geq 0$. Let

$$C=\bigcap_{n=0}^{\infty}A_n.$$

If $x \in C$, then $\alpha(x)(A_n) = 1$ for all n, so that $\alpha(x)(C) = 1$. Also, $C \subset A$ and $\sigma_0(C) = 1$. To summarize, C satisfies

$$\sigma_0(C) = 1, \qquad \alpha(x)(C) = 1 \quad \text{and} \quad \int \alpha(z)(B)\alpha(x)(dz) = \alpha(x)(B) \quad \text{for all } x \in C \text{ and } B \in \mathcal{B}.$$

Next, if $\sigma_n(x)(C) = 1$ for some $n \geq 0$ and all $x \in C^n$, then

$$\begin{aligned}\sigma_{n+1}(x, y)(C) &= \int_C \alpha(z)(C) f_n(x, y, z)\sigma_n(x)(dz) + \alpha(y)(C)\left\{1 - \int f_n(x, y, z)\sigma_n(x)(dz)\right\} \\ &= \int_C f_n(x, y, z)\sigma_n(x)(dz) + 1 - \int f_n(x, y, z)\sigma_n(x)(dz) = 1 \quad \text{for all } (x, y) \in C^{n+1}.\end{aligned}$$

Arguing by induction again, $\sigma_0(C) = 1$ implies

$$\sigma_n(x)(C) = 1 \quad \text{for all } n \geq 0 \text{ and all } x \in C^n.$$

Finally, fix $(x, y) \in C^{n+1}$. Since $\sigma_n(x)(C) = 1$,

$$\begin{aligned}\int \alpha(v)(B)\sigma_{n+1}(x, y)(dv) &= \int_C \int \alpha(v)(B)\alpha(z)(dv) f_n(x, y, z)\sigma_n(x)(dz) \\ &\quad + \left\{1 - \int f_n(x, y, z)\sigma_n(x)(dz)\right\} \int \alpha(v)(B)\alpha(y)(dv) \\ &= \int_C \alpha(z)(B) f_n(x, y, z)\sigma_n(x)(dz) \\ &\quad + \left\{1 - \int f_n(x, y, z)\sigma_n(x)(dz)\right\}\alpha(y)(B) \\ &= \sigma_{n+1}(x, y)(B) \quad \text{for all } B \in \mathcal{B}.\end{aligned}$$

Therefore, $\sigma_{n+1}(x, y)$ is a stationary distribution for the kernel α. By Theorem 4, $\sigma_{n+1}(x, y)$ is still stationary for the kernel $\{\sigma_{n+2}(x, y, z) : z \in S\}$.

This concludes the proof that (X_n) is c.i.d. under P_σ. To conclude the proof of the whole theorem, suppose $\alpha(x) = \delta_x$ for all $x \in S$. Then, condition (3) is a direct consequence of Theorem 4 and the following well-known fact. Let X and Z be S-valued random variables, τ the probability distribution of X, and $\gamma = \{\gamma(x) : x \in S\}$ a regular version of the conditional distribution of Z given X. Then,

$$(X, Z) \sim (Z, X) \quad \Leftrightarrow \quad \gamma \text{ is reversible with respect to } \tau. \qquad \square$$

Condition (3) is stronger than the c.i.d. condition. As an example, (3) implies

$$(X_i, X_j) \sim (X_j, X_i) \quad \text{for all } i \neq j$$

and this may fail for an arbitrary c.i.d. sequence; see e.g. [6], Example 3. Therefore, when $\alpha(x) = \delta_x$, the updating procedure of this section yields a special type of c.i.d. sequences. On the other hand, just because of condition (3), Theorem 2 implies

$$(X_n) \text{ is exchangeable under } P_\sigma \quad \Leftrightarrow \quad \sigma_n(x) = \sigma_n\big(f(x)\big)$$

for all $n \geq 2$, all permutations f on S^n and P_σ-almost all $x \in S^n$. In other terms, if $\sigma \in \Sigma$ and $\alpha(x) = \delta_x$, the exchangeability of (X_n) can be easily characterized.

Finally, we turn to conditions (a)–(b). The next result is helpful for finding a kernel α satisfying (a)–(b).

Lemma 6. *If $\alpha = \{\alpha(x) : x \in S\}$ is a regular conditional distribution for σ_0 given a sub-σ-field $\mathcal{G} \subset \mathcal{B}$, then α satisfies conditions* (*a*)–(*b*).

Proof. Condition (a) (that is, σ_0 stationary for α) has been already noted in Section 2.2. In turn, the proof of (b) essentially agrees with that of [9], Lemma 10, but we report it for completeness. Let $\mathcal{G}_0$ be the σ-field over S generated by the maps $z \mapsto \alpha(z)(B)$ for all $B \in \mathcal{B}$. Then, α is also a regular conditional distribution for σ_0 given $\mathcal{G}_0$. In addition, since $\mathcal{B}$ is countably generated, $\mathcal{G}_0$ is countably generated as well. Hence, there is $A \in \mathcal{B}$ such that $\sigma_0(A) = 1$ and

$$\alpha(x)(B) = \delta_x(B) \quad \text{for all } x \in A \text{ and } B \in \mathcal{G}_0.$$

Fix $x \in A$ and $B \in \mathcal{B}$. Since the map $z \mapsto \alpha(z)(B)$ is $\mathcal{G}_0$-measurable, one obtains

$$\int \alpha(z)(B)\alpha(x)(dz) = \int \alpha(z)(B)\delta_x(dz) = \alpha(x)(B). \qquad \square$$

4. Examples: Discrete strategies

From now on, we fix $\sigma_0 \in \mathcal{P}$ and a sequence

$$q_n : S^n \to [0,1], \quad n \geq 0,$$

of measurable functions (with q_0 constant).

Moreover, in this section, we let

$$\alpha(x) = \delta_x \quad \text{for all } x \in S \quad \text{and}$$

$$f_n(x,y,z) = q_n(x) \quad \text{for all } x \in S^n \text{ and } (y,z) \in S^2.$$

With this choice of f_n, the calculation of $\sigma_n(x)$ is straightforward. Writing

$$x = (x_1, \dots, x_n) \quad \text{and} \quad q_i = q_i(x_1, \dots, x_i),$$

one obtains

$$\sigma_n(x) = \sigma_0 \prod_{i=0}^{n-1} q_i + \delta_{x_n}(1 - q_{n-1}) + \sum_{i=1}^{n-1} \delta_{x_i}(1 - q_{i-1}) \prod_{j=i}^{n-1} q_j. \tag{4}$$

The strategy (4) is connected to Beta-GOS processes, as meant in [1]. If σ_0 is diffuse, the q_i have the following interpretation. Let $x = (x_1, \dots, x_n)$. Since $\sigma_0(\{x_1, \dots, x_n\}) = 0$ and $\delta_{x_i}(\{x_1, \dots, x_n\}) = 1$ for $i \leq n$, it follows that

$$\begin{aligned} P_\sigma\bigl(X_{n+1} = X_i \text{ for some } i \leq n \mid (X_1, \dots, X_n) = x\bigr) &= \sigma_n(x)\bigl(\{x_1, \dots, x_n\}\bigr) \\ &= (1 - q_{n-1}) + \sum_{i=1}^{n-1}(1 - q_{i-1}) \prod_{j=i}^{n-1} q_j = 1 - \prod_{i=0}^{n-1} q_i. \end{aligned}$$

More importantly, by specifying the q_i suitably, a lot of meaningful predictive distributions can be obtained from (4).

Example 7 (Exponential smoothing). If $q_i = q$ for all $i \geq 0$, where $q \in [0, 1]$ is any constant, formula (4) reduces to

$$\sigma_n(x) = q^n \sigma_0 + (1-q) \sum_{i=1}^{n} q^{n-i} \delta_{x_i};$$

see also [2]. Roughly speaking, this choice of σ makes sense when the inferrer has only vague opinions on the dependence structure of the data, and yet he/she feels that the weight of the i-th observation x_i should be a decreasing function of $n - i$. Note that $\sigma_n(x)$ is not invariant under permutations of x, so that (X_n) fails to be exchangeable under P_σ. Yet, (X_n) is c.i.d. under P_σ because of Theorem 5.

Example 8 (Dirichlet sequences). If $q_i = \frac{i+c}{i+1+c}$ for some constant $c > 0$, formula (4) yields

$$\sigma_n(x) = \frac{c\sigma_0 + \sum_{i=1}^{n} \delta_{x_i}}{n+c}.$$

These are the predictive distributions of a Dirichlet sequence. In this case, (X_n) is exchangeable under P_σ.

Example 9 (Latent variables). Suppose q_i of the form

$$q_i = q_i(x_1, \ldots, x_i; \lambda_1, \ldots, \lambda_i)$$

where $\lambda_1, \ldots, \lambda_i$ take values in a Borel set T of some Polish space.

To cover this situation, fix a Borel probability measure σ_0^* on $S \times T$ such that

$$\sigma_0^*(B \times T) = \sigma_0(B) \quad \text{for all } B \in \mathcal{B},$$

and define

$$\sigma_n^*\big[(x_1, \lambda_1), \ldots, (x_n, \lambda_n)\big] = \sigma_0^* \prod_{i=0}^{n-1} q_i + \delta_{(x_n, \lambda_n)}(1 - q_{n-1}) + \sum_{i=1}^{n-1} \delta_{(x_i, \lambda_i)}(1 - q_{i-1}) \prod_{j=i}^{n-1} q_j.$$

Marginalizing σ_n^*, one obtains

$$\sigma_n^*\big[(x_1, \lambda_1), \ldots, (x_n, \lambda_n)\big](B \times T) = \sigma_n(x)(B) \quad \text{for all } B \in \mathcal{B}$$

where $\sigma_n(x)$ is given by (4). Also, up to replacing S with $S \times T$, Theorem 5 applies to the strategy σ^*. More precisely, let P_{σ^*} be the probability measure on the Borel sets of $(S \times T)^\infty$ induced by σ^* and let Λ_n be the n-th coordinate random variable on T^∞. Then, the sequence (X_n, Λ_n) is c.i.d. under P_{σ^*}. In other terms, (X_n) is c.i.d. (under P_{σ^*}) even if q_i depends on the latent variables $\lambda_1, \ldots, \lambda_i$.

A last remark, motivated by the next Example 10, is as follows. The above argument still applies if λ_1 is a known constant and

$$q_i = q_i(x_1, \ldots, x_i; \lambda_1, \ldots, \lambda_i, \lambda_{i+1}).$$

In fact, since λ_1 is constant, $q_0 = q_0(\lambda_1)$ is constant as well. Thus, it suffices to replace (x_n, λ_n) with (x_n, λ_{n+1}), namely, to define σ_n^* as

$$\sigma_n^*\big[(x_1, \lambda_2), \dots, (x_n, \lambda_{n+1})\big] = \sigma_0^* \prod_{i=0}^{n-1} q_i + \delta_{(x_n, \lambda_{n+1})}(1 - q_{n-1}) + \sum_{i=1}^{n-1} \delta_{(x_i, \lambda_{i+1})}(1 - q_{i-1}) \prod_{j=i}^{n-1} q_j.$$

Arguing as above, the sequence (X_n, Λ_{n+1}) is c.i.d. under P_{σ^*} and

$$\sigma_n^*\big[(x_1, \lambda_2), \dots, (x_n, \lambda_{n+1})\big](B \times T) = \sigma_n(x)(B) \quad \text{for all } B \in \mathcal{B}$$

where $\sigma_n(x)$ is given by (4).

Example 10 (Generalized Polya urns). An urn contains $a > 0$ white balls and $b > 0$ black balls. At each time $n \geq 1$, one ball is taken out and then replaced together with D_n more balls of the same color. In the classical scheme, $D_n = d$ for all n where $d \geq 0$ is a fixed constant. Here, instead, (D_n) is any sequence of non-negative random variables.

Let Y_n be the indicator of the event {white ball at time n}. Following [5], Example 1.3, it is natural to let

$$P(Y_{n+1} = 1 \mid Y_1, \dots, Y_n, D_1, \dots, D_n) = \frac{a + \sum_{i=1}^n D_i Y_i}{a + b + \sum_{i=1}^n D_i} \quad \text{a.s.}$$

Assuming D_1 constant, this is a special case of Example 9. Take in fact $S = \{0, 1\}$, $T = [0, \infty)$, and σ_0^* a Borel probability on $S \times T$ such that

$$\sigma_0^*\big(\{1\} \times [0, \infty)\big) = \frac{a}{a+b}.$$

Then, it suffices to let

$$q_i(x_1, \dots, x_i; \lambda_1, \dots, \lambda_i, \lambda_{i+1}) = \frac{a + b + \sum_{j=1}^{i} \lambda_j}{a + b + \sum_{j=1}^{i+1} \lambda_j} \quad \text{for all } i \geq 0.$$

5. Examples: Diffuse strategies

In this section, we still let $f_n(x, y, z) = q_n(x)$ but $\alpha = \{\alpha(x) : x \in S\}$ is any kernel on $(S, \mathcal{B})$ satisfying conditions (a)–(b). For the sake of simplicity, we suppose that (b) holds with $A = S$ (this can actually be assumed without loss of generality). We denote by $\sigma \in \Sigma$ the strategy induced by σ_0, α and $(q_n : n \geq 0)$. Such a σ can be written as

$$\sigma_n(x) = \sigma_0 \prod_{i=0}^{n-1} q_i + \alpha(x_n)(1 - q_{n-1}) + \sum_{i=1}^{n-1} \alpha(x_i)(1 - q_{i-1}) \prod_{j=i}^{n-1} q_j$$

for all $n \geq 1$ and $x \in S^n$, where $q_i = q_i(x_1, \dots, x_i)$.

To our knowledge, none of the strategies exhibited in this section have ever been proposed before. Two more remarks are in order.

First, σ is diffuse whenever σ_0 and α are diffuse. Instead, the strategy (4), as well as many popular strategies, has a discrete part in correspondence with the observed data.

Second, let $\mathcal{H} \subset \mathcal{B}$ be a countable partition of S and let $H(x)$ denote the unique $H \in \mathcal{H}$ which includes the point $x \in S$. For definiteness, suppose $\sigma_0(H) > 0$ for all $H \in \mathcal{H}$. Then, a simple kernel satisfying conditions (a)–(b) is

$$\alpha(x) = \sum_{H \in \mathcal{H}} 1_H(x)\sigma_0(\cdot \mid H) = \sigma_0\big(\cdot \mid H(x)\big) \quad \text{for all } x \in S. \tag{5}$$

Let us turn to specific examples.

Example 11 (Examples 7 and 8 continued). For $x = (x_1, \dots, x_n) \in S^n$, the strategies of Examples 7 and 8 turn into

$$\sigma_n(x) = q^n \sigma_0 + (1-q)\sum_{i=1}^{n} q^{n-i}\alpha(x_i) \quad \text{and} \quad \sigma_n(x) = \frac{c\sigma_0 + \sum_{i=1}^{n} \alpha(x_i)}{n+c},$$

respectively. For definiteness, we focus on the second one, which can be viewed as a version of the predictive distributions of Dirichlet sequences.

As a first example, fix a countable partition $\mathcal{H}$ of S and define α according to (5). Then,

$$\sigma_n(x) = \frac{c\sigma_0 + \sum_{i=1}^{n} \sigma_0(\cdot \mid H(x_i))}{n+c}.$$

Note that $\sigma_n(x)$ is absolutely continuous with respect to σ_0 for all $n \geq 0$ and $x \in S^n$. Such a strategy σ could be reasonable when $H(x_i)$ is the basic information coming from the observation x_i. Roughly speaking, due to the precision of the measuring tool, one is actually observing an element of $\mathcal{H}$ rather than a point of S.

For a more elaborate example, take $S = \mathbb{R}^2$ and denote by $\mathcal{R}$ the Borel σ-field on $\mathbb{R}$ (so that $\mathcal{B} = \mathcal{R}^2$). Fix a probability measure r on $\mathcal{R}$ and define

$$\sigma_0(A \times B) = \int_A N(u)(B)r(du) \quad \text{for all } A, B \in \mathcal{R},$$

where $N(u) = N(u, 1)$ is the Gaussian law on $\mathcal{R}$ with mean u and variance 1. Then, a kernel satisfying conditions (a)–(b) is

$$\alpha(u, v) = \delta_u \times N(u) \quad \text{for all } (u, v) \in \mathbb{R}^2.$$

Thus, if a point in the plane is selected through $\alpha(u, v)$, the abscissa agrees with u a.s. while the ordinate is distributed according to $N(u)$. Using this α, one obtains

$$\sigma_n(x)(A \times B) = \frac{c\sigma_0(A \times B) + \sum_{i=1}^{n} 1_A(u_i)N(u_i)(B)}{n+c} \quad \text{for all } A, B \in \mathcal{R},$$

where the i-th observation x_i is written as $x_i = (u_i, v_i)$. This strategy σ comes into play when the basic information which stems from $x_i = (u_i, v_i)$ is the abscissa u_i. Note that $\sigma_n(x)(A \times B)$ is small if n is large but $u_i \notin A$ for all i. In a sense, the classical Dirichlet prediction scheme is preserved as regards the abscissas $u_1, \dots, u_n$ of the observed data. Note also that $N(u)$ could be replaced by $Q(u)$, where $\{Q(u) : u \in \mathbb{R}\}$ is any measurable collection of probabilities on $\mathcal{R}$.

Next, given a countable class G of measurable maps $g : S \to S$, say that σ_0 is G-invariant if

$$\sigma_0\bigl(g^{-1}B\bigr) = \sigma_0(B) \quad \text{for all } g \in G \text{ and } B \in \mathcal{B}.$$

In this case, the inferrer may wish his/her predictions were G-invariant as well.

Example 12 (Invariant strategies). Suppose σ_0 is G-invariant and

$$\mathcal{G} = \bigl\{B \in \mathcal{B} : g^{-1}B = B \text{ for all } g \in G\bigr\}.$$

As noted in Section 2.2, since S is nice, there is a regular conditional distribution $\alpha = \{\alpha(x) : x \in S\}$ for σ_0 given $\mathcal{G}$. Because of Lemma 6, α satisfies conditions (a)–(b). By standard arguments, since G is countable and $\mathcal{B}$ countably generated, α can be taken in such a way that $\alpha(x)$ is G-invariant for all $x \in S$. In turn, this implies that $\sigma_n(x)$ is G-invariant for all $n \geq 0$ and $x \in S^n$.

As a simple example, let $S = \mathbb{R}$ and σ_0 symmetric. Take $G = \{g\}$ where $g(x) = -x$, and

$$\alpha(x) = \frac{\delta_x + \delta_{-x}}{2}.$$

Then,

$$\sigma_n(x) = \sigma_0 \prod_{i=0}^{n-1} q_i + \frac{1}{2}\left(\delta_{x_n}(1 - q_{n-1}) + \sum_{i=1}^{n-1} \delta_{x_i}(1 - q_{i-1}) \prod_{j=i}^{n-1} q_j\right)$$
$$+ \frac{1}{2}\left(\delta_{-x_n}(1 - q_{n-1}) + \sum_{i=1}^{n-1} \delta_{-x_i}(1 - q_{i-1}) \prod_{j=i}^{n-1} q_j\right)$$

is a symmetric strategy which makes (X_n) c.i.d.

As a further example, let $S = T^d$ and σ_0 exchangeable, where T is a Borel subset of a Polish space. Take G as the set of all permutations of T^d, and

$$\alpha(x) = \frac{\sum_{g \in G} \delta_{g(x)}}{d!}.$$

Then, σ is an exchangeable strategy which makes (X_n) c.i.d.

The strategy λ obtained in the last example does not belong to Σ. However, λ comes from essentially the same idea of Σ and (X_n) is c.i.d. under P_λ.

Example 13 (Another strategy dominated by σ_0). For each $n \geq 1$, take a countable partition $\mathcal{H}_n$ of S and assume

$$\mathcal{H}_n \subset \mathcal{B}, \qquad \mathcal{H}_{n+1} \text{ finer than } \mathcal{H}_n \text{ and } \sigma_0(H) > 0 \quad \text{for all } H \in \mathcal{H}_n.$$

To avoid trivialities, assume also that $q_n > 0$ for all $n \geq 0$.

For every $n \geq 1$ and $\tau \in \mathcal{P}$, a kernel $\alpha_n = \{\alpha_n(x) : x \in S\}$ which admits τ as a stationary distribution is

$$\alpha_n(x) = \sum_{H \in \mathcal{H}_n} 1_H(x)\tau(\cdot \mid H) = \tau\bigl(\cdot \mid H_n(x)\bigr).$$

(Here, $H_n(x)$ is the unique $H \in \mathcal{H}_n$ such that $x \in H$ and we tacitly assumed $\tau(H) > 0$ for all $H \in \mathcal{H}_n$.)

Let us define a strategy λ as follows. Let $\lambda_0 = \sigma_0$ and

$$\lambda_1(x) = q_0\sigma_0 + (1-q_0)\sigma_0\big(\cdot \mid H_1(x)\big) \quad \text{for all } x \in S.$$

By Theorem 4, λ_0 is a stationary distribution for the kernel $\{\lambda_1(x) : x \in S\}$. Next, for every $(x, y) \in S^2$, define

$$\lambda_2(x, y) = q_1(x)\lambda_1(x) + \big(1 - q_1(x)\big)\lambda_1(x)\big(\cdot \mid H_2(y)\big).$$

The kernel $\{\lambda_2(x, y) : y \in S\}$ admits $\lambda_1(x)$ as a stationary distribution. Moreover, since $\mathcal{H}_2$ is finer than $\mathcal{H}_1$, one obtains

$$\lambda_1(x)\big(B \mid H_2(y)\big) = \sigma_0\big(B \mid H_2(y)\big) \quad \text{for all } B \in \mathcal{B}.$$

Therefore, $\lambda_2(x, y)$ can be written as

$$\lambda_2(x, y) = q_0q_1(x)\sigma_0 + (1-q_0)q_1(x)\sigma_0\big(\cdot \mid H_1(x)\big) + \big(1-q_1(x)\big)\sigma_0\big(\cdot \mid H_2(y)\big).$$

In general, for every $n \ge 1$ and $x = (x_1, \dots, x_n) \in S^n$, define

$$\lambda_n(x) = \sigma_0 \prod_{i=0}^{n-1} q_i + \sigma_0\big(\cdot \mid H_n(x_n)\big)(1-q_{n-1}) + \sum_{i=1}^{n-1} \sigma_0\big(\cdot \mid H_i(x_i)\big)(1-q_{i-1}) \prod_{j=i}^{n-1} q_j$$

where q_i stands for $q_i(x_1, \dots, x_i)$. Arguing as above, it is easily seen that, for a fixed $x \in S^n$, the kernel $\{\lambda_{n+1}(x, y) : y \in S\}$ admits $\lambda_n(x)$ as a stationary distribution. Hence, Theorem 3 implies that (X_n) is c.i.d. under P_λ.

The strategy λ is reminiscent of (4). As a matter of fact, λ agrees with (4) up to replacing $\sigma_0(\cdot \mid H_i(x_i))$ with δ_{x_i}. Furthermore, the partitions $\mathcal{H}_n$ can be chosen such that

$$\{x\} = \bigcap_n H_n(x) \quad \text{for each } x \in S.$$

Unlike (4), however, $\lambda_n(x)$ is absolutely continuous with respect to σ_0 for all $n \ge 0$ and $x \in S^n$.

6. Examples: Other choices of f_n

In the previous examples, $f_n(x, y, z) = q_n(x)$ does not depend on (y, z). This is not so in the present section. We let $\alpha(x) = \delta_x$ and we denote by $\sigma \in \Sigma$ the strategy induced by σ_0, α and $(f_n : n \ge 0)$. Once again, to our knowledge, the strategies obtained in this section have never been proposed before.

Example 14 (Separating sets). For each $n \ge 0$ and $x \in S^n$, take a set $A_n(x) \in \mathcal{B}$ and define

$$f_n(x, y, z) = 1_{A_n(x)}(y)1_{A_n(x)}(z) + 1_{A_n^c(x)}(y)1_{A_n^c(x)}(z)$$

where $A_n^c(x)$ is the complement of $A_n(x)$. Thus, $f_n(x, y, z) = 0$ or $f_n(x, y, z) = 1$ according to whether y and z can, or cannot, be separated by the set $A_n(x)$. A direct calculation shows that

$$\sigma_{n+1}(x, y) = \sigma_n(x)\big(A_n(x)\big)\sigma_n(x)\big(\cdot \mid A_n(x)\big) + \sigma_n(x)\big(A_n^c(x)\big)\delta_y \quad \text{if } y \in A_n(x),$$

where the first summand on the right is meant to be 0 in case $\sigma_n(x)(A_n(x)) = 0$. Similarly,

$$\sigma_{n+1}(x, y) = \sigma_n(x)\big(A_n^c(x)\big)\sigma_n(x)\big(\cdot \mid A_n^c(x)\big) + \sigma_n(x)\big(A_n(x)\big)\delta_y \quad \text{if } y \notin A_n(x).$$

According to the heuristic interpretation of Section 3, such a strategy σ can be described as follows. At time $n+1$, after observing $(x, y) \in S^{n+1}$, the inferrer selects a new state z according to $\sigma_n(x)$. Then, he/she remains in y or goes to z according to whether y and z are, or are not, separated by $A_n(x)$. This could be reasonable, for instance, if the inferrer has some reason to request

$$\sigma_{n+1}(x, y)\big(A_n(x)\big) = 1_{A_n(x)}(y).$$

Example 15 (Decreasing functions of the distance). In the spirit of Example 14, let

$$f_n(x, y, z) = g_n\big[x, d(y, z)\big]$$

where d is the distance on S and $g_n : S^n \times [0, \infty) \to [0, 1]$ a measurable function such that

$$g_n(x, t) < g_n(x, s) < g_n(x, 0) = 1 \quad \text{for all } x \in S^n \text{ and } 0 < s < t.$$

Then, σ can be attached an interpretation similar to Example 14. Again, after observing $(x, y) \in S^{n+1}$, the inferrer selects a new state z according to $\sigma_n(x)$. Then, he/she goes to z with probability $f_n(x, y, z)$ or remains in y with probability $1 - f_n(x, y, z)$. Moreover, the chance of reaching z starting from y is a decreasing function of $d(y, z)$ and is 1 if and only if $y = z$.

Example 16 (Ehrenfest-like models). Theorem 4 still works if the assumption $f \leq 1$ is weakened. More specifically, define β according to Theorem 4 with $\alpha(x) = \delta_x$ and f a measurable symmetric function such that $0 \leq f \leq c$, where c is any constant. Then, β is a reversible kernel provided $\beta(x)(B) \geq 0$ for all $x \in S$ and $B \in \mathcal{B}$. Note that the latter condition is trivially true if $c \leq 1$.

As an example, take $S = \{0, 1\}$ and f_n a non-negative function on S^{n+2} such that $f_n(x, y, z) = f_n(x, z, y)$. If

$$f_n(x, 0, 1) - 1 \leq f_n(x, 0, 1)\sigma_n(x)\big(\{1\}\big) \leq 1 \quad \text{for all } n \geq 0 \text{ and } x \in S^n, \tag{6}$$

then $\sigma_n(x)(B) \geq 0$ for all n, x and B. Hence, (X_n) is c.i.d. under P_σ whenever condition (6) holds. On the other hand, if $f_n(x, 0, 1) > 1$, then

$$\sigma_{n+1}(x, y)\big(\{y\}\big) = \sigma_n(x)\big(\{y\}\big) + \big(1 - f_n(x, 0, 1)\big)\sigma_n(x)\big(\{1 - y\}\big) < \sigma_n(x)\big(\{y\}\big).$$

In other terms, observing y at step $n + 1$ makes the probability of y at step $n + 2$ strictly less than the probability of y at step $n + 1$. This may look counterintuitive but it makes sense in some problems.

Think of two water-containers C_0 and C_1. At each time $n \geq 1$, either C_0 or C_1 is selected and a part of its water is transferred into the other one. The total quantity of water, say w, remains constant in time. The data are the selected containers. To model this situation, it is quite natural to let $S = \{0, 1\}$ and

$$\lambda_n(x)\big(\{y\}\big) = \frac{\text{quantity of water in } C_y \text{ after observing } x}{w}$$

for all $n \geq 0$, $x \in S^n$ and $y \in S$. Such a strategy λ belongs to Σ under some assumptions on the quantity of water moving from one container to the other. For instance suppose that, after observing (x, y) for some $x \in S^n$ and $y \in S$, the quantity of water transferred from C_y into C_{1-y} is

$$\lambda_n(x)\big(\{1-y\}\big)^2 \lambda_n(x)\big(\{y\}\big)w.$$

Then, $\lambda \in \Sigma$. In fact, λ is induced by λ_0, $\{\delta_x : x \in S\}$ and

$$f_n(x, 0, 1) = 1 + \lambda_n(x)\big(\{0\}\big)\lambda_n(x)\big(\{1\}\big).$$

7. Discreteness of the limit of σ_n

This section investigates the limit μ of σ_n as $n \to \infty$. It is split into two subsections. The first introduces a sequence of random variables the predictive distributions of which are given by (4), while the second includes the main results.

7.1. An explicit construction

Let σ be the strategy (4). To better understand the meaning of σ, it may be useful to build a sequence (Y_n) of random variables satisfying $Y_1 \sim \sigma_0$ and

$$\begin{aligned} P(Y_{n+1} \in \cdot \mid Y_1, \dots, Y_n) &= \sigma_n(Y_1, \dots, Y_n) \\ &= \sigma_0 \prod_{i=0}^{n-1} q_i + \delta_{Y_n}(1 - q_{n-1}) + \sum_{i=1}^{n-1} \delta_{Y_i}(1 - q_{i-1}) \prod_{j=i}^{n-1} q_j \quad \text{a.s. for } n \geq 1 \end{aligned} \tag{7}$$

where $q_i = q_i(Y_1, \dots, Y_i)$. One such (Y_n) is provided by [4].

Let $(T_n : n \geq 1)$ and $(U_{i,j} : j \geq 1, 0 \leq i < j)$ be random variables such that:

(j) (T_n) is an i.i.d. sequence of S-valued random variables with $T_1 \sim \sigma_0$;

(jj) $(U_{i,j})$ is an i.i.d. array of $[0, 1]$-valued random variables with $U_{0,1}$ uniformly distributed on $[0, 1]$;

(jjj) (T_n) is independent of $(U_{i,j})$.

Using (T_n) and $(U_{i,j})$ as building blocks, the sequence (Y_n) is obtained as follows.

Let $Y_1 = T_1$. Then, define $Y_2 = T_2$ or $Y_2 = Y_1$ according to whether $U_{0,1} \leq q_0$ or $U_{0,1} > q_0$. At step $n+1$, after $Y_1, \dots, Y_n$ have been defined, let

$$\begin{aligned} Y_{n+1} = T_{n+1} \quad & \text{if } U_{i,n} \leq q_i(Y_1, \dots, Y_i) \text{ for all } 0 \leq i < n, \\ Y_{n+1} = Y_{i+1} \quad & \text{if } U_{i,n} > q_i(Y_1, \dots, Y_i) \text{ and } U_{j,n} \leq q_j(Y_1, \dots, Y_j) \\ & \text{for some } 0 \leq i < n \text{ and all } j > i. \end{aligned}$$

It is not hard to verify that $Y_1 \sim \sigma_0$ and condition (7) holds; see [4], Lemma 3.

7.2. Asymptotics

Let $s = (s_1, \dots, s_n, \dots)$ denote a point of S^∞. For any strategy σ which makes (X_n) c.i.d., there is a random probability measure μ on $(S, \mathcal{B})$ such that, for every fixed $B \in \mathcal{B}$,

$$\sigma_n(s_1, \dots, s_n)(B) \longrightarrow \mu(s)(B) \quad \text{for } P_\sigma\text{-almost all } s \in S^\infty.$$

As noted in Section 2.1, the role played by μ is not as crucial as in the exchangeable case, as P_σ is not completely determined by μ.

Example 17. Take (X_n) c.i.d. but not exchangeable under P_σ and define

$$Q(A) = E_{P_\sigma}\big\{\mu^\infty(A)\big\} \quad \text{for all } A \in \mathcal{B}^\infty.$$

By definition, (X_n) is exchangeable under Q. Also, since S is nice, $Q = P_{\sigma^*}$ for some strategy σ^*. Thus, $P_\sigma \neq P_{\sigma^*}$. However, for every fixed $B \in \mathcal{B}$,

$$\sigma_n^*(s_1, \dots, s_n)(B) \longrightarrow \mu(s)(B) \quad \text{for } P_{\sigma^*}\text{-almost all } s \in S^\infty.$$

Despite Example 17, μ is an important random parameter for (X_n) and a (natural) question is: What kind of random probability measures μ can be obtained if $\sigma \in \Sigma$? We address this question when σ is given by (4). To this end, we first prove a general result.

In the next statement, we write "a.s." to mean "P_σ-a.s." and we denote by $X_1^*, X_2^*, \dots$ the (finite or infinite) sequence of distinct observations corresponding to $X_1, X_2, \dots$ More specifically, if N is the cardinality of the (random) set $\{X_1, X_2, \dots\}$, we let

$$X_n^* = X_{\tau_n} \quad \text{for all integers } n \text{ such that } 1 \le n \le N,$$
$$\text{where } \tau_1 = 1 \text{ and } \tau_n = \inf\big\{j : X_j \notin \{X_1^*, \dots, X_{n-1}^*\}\big\}.$$

Theorem 18. *Suppose (X_n) is c.i.d. under P_σ, where σ is any strategy. Then,*

$$\mu \overset{a.s.}{=} \sum_k W_k \delta_{X_k^*}, \tag{8}$$

for some random variables $W_k \ge 0$ such that $\sum_k W_k = 1$, if and only if

$$\lim_n P_\sigma(X_n \neq X_i \text{ for each } i < n) = 0. \tag{9}$$

In addition,

$$W_k \overset{a.s.}{=} \lim_n \frac{1}{n} \sum_{i=1}^n 1_{\{X_i = X_k^*\}}. \tag{10}$$

Proof. To make the notation easier, write $P = P_\sigma$, $E = E_{P_\sigma}$ and $I_{n-1} = (X_1, \dots, X_{n-1})$.

We first note a simple fact. Let

$$\gamma_1 = \delta_{I_{n-1}} \times \delta_{X_n}, \qquad \gamma_2 = \delta_{I_{n-1}} \times \mu, \quad \text{and}$$
$$H = \big\{(s_1, \dots, s_n) \in S^n : s_n = s_i \text{ for some } i < n\big\}.$$

Then, γ_1 and γ_2 are random probability measures on $(S^n, \mathcal{B}^n)$ such that

$$\gamma_1(H) = \delta_{X_n}\big(\{X_1, \dots, X_{n-1}\}\big) \quad \text{and} \quad \gamma_2(H) = \mu\big(\{X_1, \dots, X_{n-1}\}\big).$$

Next, define two (non-random) probability measures on $(S^n, \mathcal{B}^n)$ as

$$\gamma_1^*(C) = E\big\{\gamma_1(C)\big\} \quad \text{and} \quad \gamma_2^*(C) = E\big\{\gamma_2(C)\big\} \quad \text{for all } C \in \mathcal{B}^n.$$

Since (X_n) is c.i.d. under P, then $P(X_n \in B \mid I_{n-1}) = E(\mu(B) \mid I_{n-1})$ a.s. for each $B \in \mathcal{B}$; see Section 2.1. Therefore,

$$\begin{aligned}\gamma_1^*(A \times B) &= P(I_{n-1} \in A, X_n \in B)\\ &= E\big\{1_A(I_{n-1})P(X_n \in B \mid I_{n-1})\big\}\\ &= E\big\{1_A(I_{n-1})E\big(\mu(B) \mid I_{n-1}\big)\big\}\\ &= E\big\{1_A(I_{n-1})\mu(B)\big\} = \gamma_2^*(A \times B)\end{aligned}$$

for all $A \in \mathcal{B}^{n-1}$ and $B \in \mathcal{B}$. Hence, $\gamma_1^* = \gamma_2^*$ on $\mathcal{B}^n$, which in turn implies

$$\begin{aligned}P(X_n = X_i \text{ for some } i < n) &= E\big(\delta_{X_n}\big(\{X_1, \dots, X_{n-1}\}\big)\big)\\ &= \gamma_1^*(H) = \gamma_2^*(H) = E\big(\mu\big(\{X_1, \dots, X_{n-1}\}\big)\big).\end{aligned}$$

It follows that

$$E\big(\mu\big(\{X_1^*, X_2^*, \dots\}\big)\big) = \lim_n E\big(\mu\big(\{X_1, \dots, X_{n-1}\}\big)\big) = \lim_n P(X_n = X_i \text{ for some } i < n).$$

This proves the equivalence between (8) and (9). In fact,

$$\text{condition (8)} \quad \Leftrightarrow \quad \mu\big(\{X_1^*, X_2^*, \dots\}\big) \overset{\text{a.s.}}{=} 1 \quad \Leftrightarrow \quad E\big(\mu\big(\{X_1^*, X_2^*, \dots\}\big)\big) = 1.$$

We finally turn to (10). As noted in Section 2.1, μ also satisfies

$$\mu_n(B) \xrightarrow{\text{a.s.}} \mu(B) \quad \text{for every fixed } B \in \mathcal{B},$$

where $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is the empirical measure. Hence,

$$P(\mu_n \overset{\text{weakly}}{\longrightarrow} \mu) = 1.$$

If condition (9) holds, then

$$\mu\big(\{X_1^*, X_2^*, \dots\}\big) \overset{\text{a.s.}}{=} 1 \quad \text{and} \quad \mu_n\big(\{X_1^*, X_2^*, \dots\}\big) = 1 \quad \text{for each } n,$$

where the first equation has been proved above and the second is trivial. Hence, under (9), μ_n converges to μ in total variation norm with probability 1, that is,

$$\sup_{B \in \mathcal{B}} \big|\mu_n(B) - \mu(B)\big| \xrightarrow{\text{a.s.}} 0.$$

In particular,

$$W_k = \mu\big(\{X_k^*\}\big) = \lim_n \mu_n\big(\{X_k^*\}\big) = \lim_n \frac{1}{n}\sum_{i=1}^n 1_{\{X_i = X_k^*\}} \quad \text{a.s.}$$

□

Theorem 18 extends a result concerning exchangeability to the c.i.d. case. In fact, the equivalence between (8) and (9) is already known if (X_n) is exchangeable under P_σ; see e.g. [23].

Finally, we focus on the special case where σ is assessed according to (4). Then, Theorem 18 provides conditions for μ to be a.s. discrete.

Theorem 19. *Suppose the strategy σ is given by* (4) *and*

$$\prod_{i=0}^{n-1} q_i(X_1,\dots,X_i) \xrightarrow{P_\sigma} 0.$$

Then, μ admits representation (8) *and the weights W_k are given by* (10).

Proof. Just note that

$$\begin{aligned} P_\sigma\big(X_{n+1} \notin \{X_1,\dots,X_n\} \mid (X_1,\dots,X_n)=x\big) &= \sigma_n(x)\big(\{x_1,\dots,x_n\}^c\big) \\ &= \sigma_0\big(\{x_1,\dots,x_n\}^c\big)\prod_{i=0}^{n-1} q_i \end{aligned}$$

where $n \geq 1$, $x=(x_1,\dots,x_n)\in S^n$ and $q_i = q_i(x_1,\dots,x_i)$. Hence,

$$\begin{aligned} P_\sigma(X_{n+1} \neq X_i \text{ for each } i \leq n) &= E_{P_\sigma}\left\{\sigma_0\big(\{X_1,\dots,X_n\}^c\big)\prod_{i=0}^{n-1} q_i(X_1,\dots,X_i)\right\} \\ &\leq E_{P_\sigma}\left\{\prod_{i=0}^{n-1} q_i(X_1,\dots,X_i)\right\} \longrightarrow 0. \end{aligned}$$

An application of Theorem 18 concludes the proof. □

Various popular random probability measures ν admit the representation

$$\nu \overset{\text{a.s.}}{=} \sum_k D_k \delta_{Z_k}, \tag{11}$$

where (Z_k) is an i.i.d. sequence of random variables and the weights (D_k) are independent of (Z_k). Our last result is that μ often admits representation (11) provided σ is given by (4) and the q_i are constant.

Theorem 20. *Suppose the strategy σ is given by* (4) *and σ_0 is diffuse. Suppose also that q_i is constant for every $i \geq 0$, and*

$$\prod_{i=0}^{n-1} q_i \to 0 \quad \textit{and} \quad \sum_{n=1}^{\infty}\prod_{i=0}^{n-1} q_i = \infty.$$

Then, μ admits representation (8) *and the weights W_k are given by* (10). *Moreover, the sequence (X_k^*) is i.i.d., $X_1^* \sim \sigma_0$, and (X_k^*) is independent of (W_k).*

Proof. Take (T_n) and $(U_{i,j})$ satisfying conditions (j)–(jjj) and define (Y_n) as in Section 7.1. Since the predictive distributions of (Y_n) are given by (4), we can replace (X_n) with (Y_n). In addition, since

$$\sum_n P\big(Y_{n+1} \notin \{Y_1,\dots,Y_n\} \mid Y_1,\dots,Y_n\big) \overset{\text{a.s.}}{=} \sum_n \prod_{i=0}^{n-1} q_i = \infty,$$

the Borel–Cantelli lemma yields

$$P\big(Y_{n+1} \notin \{Y_1, \dots, Y_n\} \text{ for infinitely many } n\big) = 1.$$

Hence, one can define

$$Y_n^* = Y_{\rho_n} \quad \text{for all } n \geq 1,$$

where $\rho_1 = 1$ and $\rho_n = \inf\{j : Y_j \notin \{Y_1^*, \dots, Y_{n-1}^*\}\}$.

Let ν be a random probability measure on $(S, \mathcal{B})$ such that

$$P(Y_{n+1} \in B \mid Y_1, \dots, Y_n) \xrightarrow{\text{a.s.}} \nu(B) \quad \text{for each fixed } B \in \mathcal{B}.$$

Since $\prod_{i=0}^{n-1} q_i \to 0$, Theorem 19 implies

$$\nu \overset{\text{a.s.}}{=} \sum_k D_k \delta_{Y_k^*} \quad \text{where } D_k \overset{\text{a.s.}}{=} \lim_n \frac{1}{n} \sum_{i=1}^n 1_{\{Y_i = Y_k^*\}}.$$

We now prove that (Y_k^*) is i.i.d., $Y_1^* \sim \sigma_0$, and (Y_k^*) is independent of (D_k).

Let $\mathcal{U}$ be the σ-field generated by $U_{i,j}$ for all i and j and

$$A = \{T_i \neq T_j \text{ for all } i \neq j\}.$$

On the set A, one obtains $Y_n \notin \{Y_1, \dots, Y_{n-1}\}$ if and only if $Y_n = T_n$. Furthermore, $P(A) = 1$ for (T_n) is i.i.d. and σ_0 diffuse. Thus, up to a negligible set, ρ_k is $\mathcal{U}$-measurable for each k. Similarly, up to a negligible set, D_k is $\mathcal{U}$-measurable for each k. Since (T_k) is independent of $\mathcal{U}$, it follows that (T_k) is independent of (D_k, ρ_k). Therefore, for each event H in the σ-field generated by (D_k), one obtains

$$\begin{aligned}
&P\big(H \cap \{Y_1^* \in B_1, \dots, Y_k^* \in B_k\}\big) \\
&\quad = \sum_{m_1, \dots, m_k} P\big(H \cap \{\rho_1 = m_1, \dots, \rho_k = m_k, T_{m_1} \in B_1, \dots, T_{m_k} \in B_k\}\big) \\
&\quad = \sum_{m_1, \dots, m_k} P(T_{m_1} \in B_1, \dots, T_{m_k} \in B_k) P\big(H \cap \{\rho_1 = m_1, \dots, \rho_k = m_k\}\big) \\
&\quad = \prod_{i=1}^k \sigma_0(B_i) \sum_{m_1, \dots, m_k} P\big(H \cap \{\rho_1 = m_1, \dots, \rho_k = m_k\}\big) = P(H) \prod_{i=1}^k \sigma_0(B_i).
\end{aligned}$$

This concludes the proof. □

If σ_0 is diffuse, Theorem 20 applies to Dirichlet sequences of Example 8. In this special case, however, and more generally in case of exchangeable species sampling models, the conclusions of Theorem 20 are due to Pitman [27], Proposition 11.

8. Concluding remarks

In this paper, only a very few indications on how to select f_n and α have been given. Are there some general criterions which help the inferrer in the choice of f_n and α? Which problems are appropriate for $f_n(x, y, z) = q_n(x)$ and which require more elaborate choices of f_n? Or else, when is it reasonable to take $\alpha(y) = \delta_y$?

These questions are certainly crucial from a practical point of view, but to answer them is beyond the scope of this paper. However, we make three brief remarks.

First, in the subjective view of probability, any general criterion to select σ is possibly useful but never mandatory. At least in principle, the choice of σ depends only on the knowledge/feelings of the inferrer about the dependence structure of (X_n).

Second, the kernel α controls how much the next outcome is affected by the observed data. This is quite evident in equation (1), where $f_n(x, y, z) = q_n(x)$ and the updating rule reduces to $\sigma_{n+1}(x, y) = q_n(x)\sigma_n(x) + \{1 - q_n(x)\}\alpha(y)$. In this case, the prediction of X_{n+2} is affected by the last observation $X_{n+1} = y$ only through $\alpha(y)$. Thus, in a sense, the choice of α has to do with how much the last observation should be "reinforced". In this respect, $\alpha(y) = \delta_y$ is the more extreme choice of α.

Third, as suggested by an anonymous referee, the inferrer could apply a sort of empirical Bayes procedure and select $\sigma \in \Sigma$ based on the available data. As a simple example, the strategy of Example 7 depends on a single parameter $q \in [0, 1]$. Such a q could be estimated by the data. For instance, a prior for q could be assessed and q could be estimated by the mean of the corresponding posterior.

Acknowledgements

This paper has been improved by the helpful remarks of three anonymous referees.

References

[1] Airoldi, E.M., Costa, T., Bassetti, F., Leisen, F. and Guindani, M. (2014). Generalized species sampling priors with latent beta reinforcements. *J. Amer. Statist. Assoc.* **109** 1466–1480. MR3293604 https://doi.org/10.1080/01621459.2014.950735

[2] Bassetti, F., Crimaldi, I. and Leisen, F. (2010). Conditionally identically distributed species sampling sequences. *Adv. in Appl. Probab.* **42** 433–459. MR2675111 https://doi.org/10.1239/aap/1275055237

[3] Berti, P., Crimaldi, I., Pratelli, L. and Rigo, P. (2009). Rate of convergence of predictive distributions for dependent data. *Bernoulli* **15** 1351–1367. MR2597596 https://doi.org/10.3150/09-BEJ191

[4] Berti, P., Dreassi, E., Pratelli, L. and Rigo, P. (2020). Asymptotics of certain conditionally identically distributed sequences. *Statist. Probab. Lett.* To appear.

[5] Berti, P., Pratelli, L. and Rigo, P. (2004). Limit theorems for a class of identically distributed random variables. *Ann. Probab.* **32** 2029–2052. MR2073184 https://doi.org/10.1214/009117904000000676

[6] Berti, P., Pratelli, L. and Rigo, P. (2012). Limit theorems for empirical processes based on dependent data. *Electron. J. Probab.* **17** no. 9, 18. MR2878788 https://doi.org/10.1214/EJP.v17-1765

[7] Berti, P., Pratelli, L. and Rigo, P. (2013). Exchangeable sequences driven by an absolutely continuous random measure. *Ann. Probab.* **41** 2090–2102. MR3098068 https://doi.org/10.1214/12-AOP786

[8] Berti, P., Regazzini, E. and Rigo, P. (1997). Well-calibrated, coherent forecasting systems. *Teor. Veroyatn. Primen.* **42** 144–168. MR1453335 https://doi.org/10.1137/S0040585X97975988

[9] Berti, P. and Rigo, P. (2007). 0–1 laws for regular conditional distributions. *Ann. Probab.* **35** 649–662. MR2308591 https://doi.org/10.1214/009117906000000845

[10] Cassese, A., Zhu, W., Guindani, M. and Vannucci, M. (2019). A Bayesian nonparametric spiked process prior for dynamic model selection. *Bayesian Anal.* **14** 553–572. MR3934097 https://doi.org/10.1214/18-BA1116

[11] Cifarelli, D.M. and Regazzini, E. (1996). De Finetti's contribution to probability and statistics. *Statist. Sci.* **11** 253–282. MR1445983 https://doi.org/10.1214/ss/1032280303

[12] de Finetti, B. (1937). La prévision : Ses lois logiques, ses sources subjectives. *Ann. Inst. Henri Poincaré* **7** 1–68. MR1508036

[13] Diaconis, P. and Freedman, D.A. (1990). Cauchy's equation and de Finetti's theorem. *Scand. J. Stat.* **17** 235–249. MR1092946

[14] Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7** 269–281. MR0520238

[15] Dubins, L.E. and Savage, L.J. (1965). *How to Gamble If You Must. Inequalities for Stochastic Processes*. New York: McGraw-Hill. MR0236983

[16] Fortini, S., Ladelli, L. and Regazzini, E. (2000). Exchangeability, predictive distributions and parametric models. *Sankhyā Ser. A* **62** 86–109. MR1769738

[17] Fortini, S. and Petrone, S. (2012). Predictive construction of priors in Bayesian nonparametrics. *Braz. J. Probab. Stat.* **26** 423–449. MR2949087 https://doi.org/10.1214/11-BJPS176

[18] Fortini, S., Petrone, S. and Sporysheva, P. (2018). On a notion of partially conditionally identically distributed sequences. *Stochastic Process. Appl.* **128** 819–846. MR3758339 https://doi.org/10.1016/j.spa.2017.06.008

[19] Hahn, P.R., Martin, R. and Walker, S.G. (2018). On recursive Bayesian predictive distributions. *J. Amer. Statist. Assoc.* **113** 1085–1093. MR3862341 https://doi.org/10.1080/01621459.2017.1304219

[20] Hill, B.M. (1993). Parametric models for A_n: Splitting processes and mixtures. *J. Roy. Statist. Soc. Ser. B* **55** 423–433. MR1224406

[21] Kallenberg, O. (1988). Spreading and predictable sampling in exchangeable sequences and processes. *Ann. Probab.* **16** 508–534. MR0929061

[22] Kallenberg, O. (2002). *Foundations of Modern Probability*, 2nd ed. *Probability and Its Applications* (*New York*). New York: Springer. MR1876169 https://doi.org/10.1007/978-1-4757-4015-8

[23] Krasker, W.S. and Pratt, J.W. (1986). Discussion: On the consistency of Bayes estimates. *Ann. Statist.* **14** 55–58.

[24] Lee, J., Quintana, F.A., Müller, P. and Trippa, L. (2013). Defining predictive probability functions for species sampling models. *Statist. Sci.* **28** 209–222. MR3112406 https://doi.org/10.1214/12-sts407

[25] Martin, R. and Tokdar, S.T. (2011). Semiparametric inference in mixture models with predictive recursion marginal likelihood. *Biometrika* **98** 567–582. MR2836407 https://doi.org/10.1093/biomet/asr030

[26] Newton, M.A. and Zhang, Y. (1999). A recursive algorithm for nonparametric analysis with missing data. *Biometrika* **86** 15–26. MR1688068 https://doi.org/10.1093/biomet/86.1.15

[27] Pitman, J. (1996). Some developments of the Blackwell–MacQueen urn scheme. In *Statistics, Probability and Game Theory. Institute of Mathematical Statistics Lecture Notes – Monograph Series* **30** 245–267. Hayward, CA: IMS. MR1481784 https://doi.org/10.1214/lnms/1215453576

[28] Pitman, J. (2006). *Combinatorial Stochastic Processes: Lectures from the 32nd Summer School on Probability Theory Held in Saint-Flour, July 7–24, 2002. Lecture Notes in Math.* **1875**. Berlin: Springer. With a foreword by Jean Picard. MR2245368

[29] Tokdar, S.T., Martin, R. and Ghosh, J.K. (2009). Consistency of a recursive estimate of mixing distributions. *Ann. Statist.* **37** 2502–2522. MR2543700 https://doi.org/10.1214/08-AOS639

[30] Walker, S. and Muliere, P. (1997). A characterisation of Polya tree distributions. *Statist. Probab. Lett.* **31** 163–168. MR1440632 https://doi.org/10.1016/S0167-7152(96)00028-4

Received March 2020 and revised June 2020