A PREDICTIVE APPROACH TO BAYESIAN NONPARAMETRICS

PATRIZIA BERTI, EMANUELA DREASSI, LUCA PRATELLI, AND PIETRO RIGO

Abstract. In a Bayesian framework, to make predictions on a sequence $X_1, X_2, \ldots$ of random observations, the inferrer needs to assign the predictive distributions $\sigma_n(\cdot) = P\big(X_{n+1} \in \cdot \mid X_1, \ldots, X_n\big)$. 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 is to be assessed. The data sequence (X_n) is requested to be conditionally identically distributed (c.i.d.) in the sense of [4]. 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. Further, 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.

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 set S. Assuming that $(X_1, \ldots, X_n) = x$, for some $n \geq 1$ and $x \in S^n$, the problem is to predict X_{n+1} based on the observed data x. In a Bayesian framework, this means to assess the *predictive distribution*, say

$$\sigma_n(x)(B) = P\big(X_{n+1} \in B \mid (X_1, \ldots, X_n) = x\big) \qquad \text{for all measurable } B \subset S.$$

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

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

for each $n \geq 1$ and each $(s_1, \ldots, s_n, \ldots) \in S^\infty$. Also, we assume that S is a Borel subset of a Polish space.

Let $\mathcal{B}$ denote the Borel σ-field on S and $\mathcal{P}$ the collection of all probability measures on $\mathcal{B}$. Following Dubins and Savage, a *strategy* is a sequence

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

such that

- $\sigma_0 \in \mathcal{P}$ 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 $B \in \mathcal{B}$.

2010 *Mathematics Subject Classification.* 62G99, 62F15, 62M20, 60G25, 60G57.

Key words and phrases. Bayesian nonparametrics, Conditional identity in distribution, Exchangeability, Predictive distribution, Random probability measure, Sequential predictions, Strategy.

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 $(X_1, \ldots, X_n) = x$.

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

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

Such a P is denoted P_σ in the sequel.

1.1. **Standard and non-standard approach for exchangeable data.** The sequence (X_n) is usually requested to be exchangeable. In that case, the standard approach to Bayesian prediction problems is quite involved. First, a prior probability on $\mathcal{P}$, say π, is to be selected. Then, the posterior $\pi_n(x)$ of π is to be evaluated. And finally the predictive distribution is

$$\sigma_n(x)(B) = \int_{\mathcal{P}} p(B)\, \pi_n(x)(dp) \qquad \text{for all } B \in \mathcal{B}.$$

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

A non-standard approach (henceforth, NSA) is to assign σ_n directly, without passing through π and π_n. In other terms, instead of 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], [6], [9], [11], [12], [14], [17], [18]; see also [15], [22], [23], [25] and references therein.

NSA is in line with de Finetti, Dubins and Savage, among others. Recently, NSA has been used to obtain a fast online Bayesian prediction via copulas; see [17]. In addition, NSA is quite implicit in most of the machine learning literature. From our point of view, NSA has essentially two merits. Firstly, it requires to place 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, NSA is much more direct than the standard approach. In fact, if the main goal is to predict future observations, why to select the prior π explicitly ? Rather than wondering about π, it looks reasonable to reflect on how the next observation X_{n+1} is affected by $(X_1, \ldots, X_n)$.

However, if (X_n) is requested to be exchangeable, 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 [14, Theorem 3.1]. However, the conditions on σ for making (X_n) exchangeable are quite hard to be checked in real problems. This is the main reason for NSA has not developed so far.

1.2. **Conditionally identically distributed data.** Trivially, a way to bypass the gap mentioned in the above paragraph is to weaken the exchangeability assumption. One option is to request (X_n) to be *conditionally identically distributed* (c.i.d.), namely

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

where $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ 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 is stationary and c.i.d.

We refer to Subsection 2.1 for more 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 2. Therefore, unlike the exchangeable case, NSA can be easily implemented.
- The asymptotic theory of c.i.d. sequences is basically the same as that of exchangeable sequences.
- A number of meaningful strategies can not be used if (X_n) is requested to be exchangeable, but are available if (X_n) is only asked to be c.i.d. Examples are in Sections 4-6.

1.3. **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 by a simple recursive update of σ_n.

To introduce Σ, some further notation is needed. In the sequel, a *kernel* 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 $B \in \mathcal{B}$. If $x = (x_1, \ldots, x_n) \in S^n$ and $y \in S$, we write (x, y) to denote

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

In addition, for any strategy σ, we let

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

Then, each $\sigma \in \Sigma$ can be described as follows. Fix $\sigma_0 \in \mathcal{P}$ and a sequence of measurable functions $f_n : S^{n+2} \to [0, 1]$ satisfying

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

In addition, fix a kernel α on $(S, \mathcal{B})$ such that

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

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

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

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

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 5. In particular, conditions (a)-(b) trivially hold if

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

where δ_x denotes the point mass at x.

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

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

for all $n \geq 0$, $B \in \mathcal{B}$, $x \in S^n$ and $y \in S$. We define Σ to be the collection of all such strategies σ.

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, $\sigma_{n+1}(x,y)$ can be written explicitly (and not only in recursive form) as

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

for all $(x,y) \in A^{n+1}$, where A is the set involved in condition (b) and q_i a shorthand notation to denote

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

Specifying f_n and α suitably, many other examples are possible. For instance, letting $\alpha(x) = \delta_x$, various well known strategies are actually members of Σ, including the predictive distributions of Dirichlet sequences, species sampling sequences and generalized Polya urns. In addition, to our knowledge, Σ includes some meaningful strategies not proposed so far.

We also note that various strategies $\sigma \in \Sigma$ are such that $\sigma_n(x)$ is diffuse for all $n \geq 0$ and $x \in S^n$. (A probability measure is said to be *diffuse* if vanishes on singletons). The possibility of working with diffuse strategies is useful in real problems.

Our main results are Theorems 3-4, which state that (X_n) is c.i.d. under P_σ for each $\sigma \in \Sigma$, and Theorems 15-17 dealing with the asymptotics of σ_n. We spend a few words on Theorem 15.

Let $X_1^*, X_2^*, \ldots$ denote the (finite or infinite) sequence of distinct values corresponding to the observations $X_1, X_2, \ldots$ 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

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

where "a.s." stands for "P_σ-a.s."; see Subsection 2.1. Theorem 15 states that

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

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

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

Furthermore, W_k admits the representation

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

Applying Theorem 15 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 $X_1^*, X_2^*, \ldots$ to be i.i.d. and independent of the weights $W_1, W_2, \ldots$ are given as well.

It is worth noting that Theorem 15 holds true for any strategy σ which makes (X_n) c.i.d. Hence, Theorem 15 extends to all c.i.d. sequences a known fact concerning the exchangeable case; see e.g. [21].

In addition to the results quoted above, another main contribution of this paper are the examples included in Sections 4-6. In our intentions, these examples should support 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 [4] and [20] and then investigated in various papers; see e.g. [1], [2], [6], [7], [8], [9], [10], [16]. 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 is adapted to $(\mathcal{G}_n)$ and

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

When $(\mathcal{G}_n)$ is the canonical filtration of (Y_n), i.e., $\mathcal{G}_0$ is the trivial σ-field and $\mathcal{G}_n = \sigma(Y_1, \ldots, Y_n)$, the filtration is not mentioned at all and (Y_n) is just called c.i.d. By a result in [20], (Y_n) is exchangeable if and only if is stationary and c.i.d.

Let (Y_n) be c.i.d. with respect to $(\mathcal{G}_n)$. Under various respects, the asymptotic behavior of (Y_n) is similar to that of an exchangeable sequence. We support this claim by two facts.

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

$$(Y_n, Y_{n+1}, \ldots) \rightarrow (Z_1, Z_2, \ldots) \text{ in distribution, as } n \rightarrow \infty,$$

where $(Z_1, Z_2, \ldots)$ is an exchangeable sequence.

To state the second fact, let

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

be the empirical measure. Then, there is a random probability measure μ on $(S, \mathcal{B})$ satisfying

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

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

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

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

$$P\big(Y_{n+1} \in B \mid \mathcal{G}_n\big) = E\big\{\mu(B) \mid \mathcal{G}_n\big\} \xrightarrow{a.s.} \mu(B) \qquad \text{for every fixed } B \in \mathcal{B}.$$

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
$$\int \alpha(x)(B)\,\tau(dx) = \tau(B) \qquad \text{for all } B \in \mathcal{B};$$
- α is *reversible* with respect to τ if
$$\int_A \alpha(x)(B)\,\tau(dx) = \int_B \alpha(x)(A)\,\tau(dx) \qquad \text{for all } A,\, B \in \mathcal{B};$$
- 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
$$\int_A \alpha(x)(B)\,\tau(dx) = \tau(A \cap B) \quad \text{for all } A \in \mathcal{G} \text{ and } B \in \mathcal{B}.$$

Note 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 1. ([14, 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 2. ([7, 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 2 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 2 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 enough to assign at each step a kernel with a given stationary distribution.

3. A sequential updating rule

Our starting point is the following simple fact.

Theorem 3. *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*

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

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 $\phi(x) \in (0,1]$,

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

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 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 the 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.

As in Subsection 1.3, fix $\sigma_0 \in \mathcal{P}$, a kernel α on $(S, \mathcal{B})$ satisfying conditions (a)-(b), and a sequence of measurable functions $f_n : S^{n+2} \to [0,1]$ such that

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

Then, define the strategy σ according to

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

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

Note that, when a new observation $y \in S$ 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 4. *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*

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

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

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 2, 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

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

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

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

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

$$\sigma_{n+1}(x,y)(C) = \int_C \alpha(z)(C)\, f_n(x,y,z)\, \sigma_n(x)(dz) + \alpha(y)(C) \Big\{ 1 - \int f_n(x,y,z)\, \sigma_n(x)(dz) \Big\}$$

$$= \int_C f_n(x,y,z)\, \sigma_n(x)(dz) + 1 - \int f_n(x,y,z)\, \sigma_n(x)(dz) = 1 \qquad \text{for all } (x,y) \in C^{n+1}.$$

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

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

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

$$\int \alpha(v)(B)\,\sigma_{n+1}(x,y)(dv)=\int_C\int \alpha(v)(B)\,\alpha(z)(dv)\,f_n(x,y,z)\,\sigma_n(x)(dz)+$$
$$+\Big\{1-\int f_n(x,y,z)\,\sigma_n(x)(dz)\Big\}\int \alpha(v)(B)\,\alpha(y)(dv)$$
$$=\int_C \alpha(z)(B)\,f_n(x,y,z)\,\sigma_n(x)(dz)+\Big\{1-\int f_n(x,y,z)\,\sigma_n(x)(dz)\Big\}\,\alpha(y)(B)$$
$$=\sigma_{n+1}(x,y)(B)\qquad\text{for all } B\in\mathcal{B}.$$

Therefore, $\sigma_{n+1}(x,y)$ is a stationary distribution for the kernel α. By Theorem 3, $\sigma_{n+1}(x,y)$ 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 (1) is a direct consequence of Theorem 3 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,

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

□

Condition (1) is stronger than the c.i.d. condition. As an example, (1) implies

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

and this may fail for an arbitrary c.i.d. sequence; see e.g. [7, Example 3]. Therefore, when $\alpha(x)=\delta_x$, the updating procedure of this section yields a special type of c.i.d. sequences.

Finally, we turn to conditions (a)-(b). The next result is helpful to find a kernel α satisfying (a)-(b).

Lemma 5. *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 Subsection 2.2. In turn, the proof of (b) essentially agrees with that of [5, 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

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

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

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

□

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

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

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

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

one obtains

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

The strategy (2) is connected to Beta-GOS processes, as meant in [1], and is analogous to formula (10) of [17]. Further, if σ_0 is diffuse, the q_i have the following interpretation. Let $x = (x_1, \ldots, x_n)$. Since $\sigma_0\big(\{x_1, \ldots, x_n\}\big) = 0$ and $\delta_{x_i}\big(\{x_1, \ldots, x_n\}\big) = 1$ for $i \leq n$, it follows that

$$P_\sigma\Big(X_{n+1} = X_i \text{ for some } i \leq n \mid (X_1, \ldots, X_n) = x\Big) = \sigma_n(x)\big(\{x_1, \ldots, x_n\}\big)$$
$$= (1 - q_{n-1}) + \sum_{i=1}^{n-1}(1 - q_{i-1}) \prod_{j=i}^{n-1} q_j = 1 - \prod_{i=0}^{n-1} q_i.$$

More importantly, specifying the q_i suitably, a lot of meaningful predictive distributions can be obtained from (2).

Example 6. (Vague a priori knowledge). If $q_i = q$ for all $i \geq 0$, where $q \in [0,1]$ is any constant, formula (2) 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 4.

Example 7. (Dirichlet sequences). If $q_i = \frac{i+c}{i+1+c}$ for some constant $c > 0$, formula (2) 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 8. (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

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

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

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

where $\sigma_n(x)$ is given by (2). Also, up to replacing S with $S \times T$, Theorem 4 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 next Example 9, is the following. 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

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

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

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

where $\sigma_n(x)$ is given by (2).

Example 9. (Generalized Polya urns). An urn contains $a > 0$ white balls and $b > 0$ black balls. At each time $n \geq 1$, a ball is drawn 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 [4, Example 1.3], it is natural to let

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

Assuming D_1 constant, this is a special case of Example 8. 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

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

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). We denote by $\sigma\in\Sigma$ the strategy induced by σ_0, α and $(q_n:n\geq 0)$.

Two remarks are in order.

First, σ is diffuse whenever σ_0 and α are diffuse. (Here, a collection of probability measures is said to be diffuse if each of its members is diffuse). Having a diffuse strategy may be useful in applications. Instead, the strategy (2), as well as many other popular strategies, has a discrete part in correspondence with the observed data.

Second, 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 $x\in A^n$, where $q_i=q_i(x_1,\ldots,x_i)$ and A is the set involved in condition (b). For instance, if $x\in A^n$, the strategies of Examples 6 and 7 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 another example, take a countable class G of measurable maps $g:S\to S$ and say that σ_0 is G-invariant if

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

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

Example 10. (Invariant strategies). Suppose σ_0 is G-invariant and

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

Since S is nice (it is in fact a Borel subset of a Polish space) there is a regular conditional distribution $\alpha=\{\alpha(x):x\in S\}$ for σ_0 given $\mathcal{G}$. Because of Lemma 5, α satisfies conditions (a)-(b).

By standard arguments, since G is countable and $\mathcal{B}$ countably generated, it can be shown that $\alpha(x)$ is G-invariant for σ_0-almost all $x\in S$. Thus, the set A in condition (b) can be taken such that $\alpha(x)$ is G-invariant for every $x\in A$. In turn, this implies that $\sigma_n(x)$ is G-invariant for all $n\geq 0$ and $x\in A^n$.

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

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

Then,

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

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

As a further example, let $S = T^d$, where T is a Borel subset of a Polish space, and assume σ_0 exchangeable. Take $A = S$, G the set of all permutations of S, 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 next example does not belong to Σ. However, λ comes from essentially the same idea of Σ.

Example 11. (A strategy dominated by σ_0). For each $n \geq 1$ and $x \in S$, take a countable partition $\mathcal{H}_n$ of S and denote by $H_n(x)$ the unique $H \in \mathcal{H}_n$ such that $x \in H$. We assume

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

To avoid trivialities, we also assume $q_n > 0$ 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

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

(Here, we tacitly assumed $\tau(H) > 0$ for all $H \in \mathcal{H}_n$, but this assumption can be easily removed).

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

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

By Theorem 3, λ_0 is a stationary distribution for the kernel $\{\lambda_1(x) : x \in S\}$. Next, for every $(x, y) \in S^2$, define

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

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

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

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

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

In general, for every $n \geq 1$ and $x = (x_1, \ldots, x_n) \in S^n$, 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 $q_i(x_1, \ldots, x_i)$. Arguing as above, it is easily seen that, for 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 2 implies that (X_n) is c.i.d. under P_λ.

The strategy λ is reminiscent of (2). In fact, λ agrees with (2) up to replacing δ_{x_i} with $\sigma_0\big(\cdot \mid H_i(x_i)\big)$. Furthermore, the partitions $\mathcal{H}_n$ can be chosen such that

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

Unlike (2), however, $\lambda_n(x)$ is absolutely continuous with respect to σ_0 for all $n \geq 1$ and $x \in S^n$. In particular, $\lambda_n(x)$ is diffuse if σ_0 is diffuse.

6. Examples: Other choices of f_n

In the examples given so far, $f_n(x, y, z) = q_n(x)$ does not depend on (y, z). This is not so in the present section. We denote by $\sigma \in \Sigma$ the strategy induced by σ_0, α and $(f_n : n \geq 0)$ and we let $\alpha(x) = \delta_x$ for all $x \in S$.

Example 12. (Separating sets). For each $n \geq 0$ 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 can not, 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 $\sigma_n(x)\Big(A_n(x)\Big) = 0$. 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 13. (Decreasing functions of the distance). In the spirit of Example 12, 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

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

Then, σ can be attached an interpretation similar to Example 12. 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 14. (Ehrenfest-like models). Theorem 3 still works if the assumption $f \leq 1$ is weakened. Precisely, define β according to Theorem 3 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

$$f_n(x, 0, 1) - 1 \leq f_n(x, 0, 1)\, \sigma_n(x)(\{1\}) \leq 1 \qquad \text{for all } n \geq 0 \text{ and } x \in S^n, \tag{3}$$

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 (3) holds. On the other hand, if $f_n(x, 0, 1) > 1$, then

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

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 makes sense in some problems.

Think of two water-containers C_0 and C_1. At each time $n \geq 1$, one of C_0 and C_1 is selected and a part of its water is transferred into the other. 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

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

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

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

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

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

7. Discreteness of the limit of σ_n

This section is split into two subsections. The first exhibits a sequence of random variables whose predictive distributions are given by (2). The second, which includes the main results, deals with the limit of σ_n.

7.1. An explicit construction.

Let σ be the strategy (2). 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

$$\begin{aligned} P\big(Y_{n+1} \in \cdot \mid Y_1, \ldots, Y_n\big) &= \sigma_n(Y_1, \ldots, Y_n) \\ &= \sigma_0 \prod_{i=0}^{n-1} q_i + \delta_{Y_n}\,(1 - q_{n-1}) + \sum_{i=1}^{n-1} \delta_{Y_i}\,(1 - q_{i-1}) \prod_{j=i}^{n-1} q_j \quad \text{a.s. for } n \geq 1 \end{aligned} \tag{4}$$

where $q_i = q_i(Y_1, \ldots, Y_i)$. One such (Y_n) is provided by [9].

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 $U_{0,1} \le q_0$ or $U_{0,1} > q_0$. At step $n+1$, after $Y_1, \ldots, Y_n$ have been defined, let

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

It is not hard to verify that $Y_1 \sim \sigma_0$ and condition (4) holds; see [9, Lemma 3].

7.2. **Asymptotics.** Let $s = (s_1, \ldots, s_n, \ldots)$ 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 $B \in \mathcal{B}$,

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

see Subsection 2.1.

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 (2). 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 $X_1^*, X_2^*, \ldots$ the (finite or infinite) sequence of distinct observations corresponding to $X_1, X_2, \ldots$ Precisely, if N is the cardinality of the (random) set $\{X_1, X_2, \ldots\}$, we let

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

Theorem 15. *Suppose (X_n) is c.i.d. under P_σ, where σ is any strategy. Then,*

$$\mu \overset{a.s.}{=} \sum_k W_k \, \delta_{X_k^*}, \tag{5}$$

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

$$\lim_n P_\sigma\big(X_n \neq X_i \text{ for each } i < n\big) = 0. \tag{6}$$

In addition,

$$W_k \overset{a.s.}{=} \lim_n \frac{1}{n} \sum_{i=1}^n 1_{\{X_i = X_k^*\}}. \tag{7}$$

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

We first note a simple fact. Let

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

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

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

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 Subsection 2.1. Therefore,

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

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

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

It follows that

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

This proves the equivalence between (5) and (6). In fact,

$$\text{condition (5)} \quad \Leftrightarrow \quad \mu\big(\{X_1^*, X_2^*, \ldots\}\big) \overset{a.s.}{=} 1 \quad \Leftrightarrow \quad E\Big(\mu\big(\{X_1^*, X_2^*, \ldots\}\big)\Big) = 1.$$

We finally turn to (7). As noted in Subsection 2.1, μ also satisfies

$$\mu_n(B) \overset{a.s.}{\longrightarrow} \mu(B) \qquad \text{for every fixed } B \in \mathcal{B},$$

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

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

If condition (6) holds, then

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

where the first equation has been proved above and the second is trivial. Hence, under (6), μ_n converges to μ in total variation norm with probability 1, i.e.

$$\sup_{B \in \mathcal{B}} \Big|\mu_n(B) - \mu(B)\Big| \overset{a.s.}{\longrightarrow} 0.$$

In particular,

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

□

Theorem 15 extends to the c.i.d. case a result concerning exchangeability. In fact, the equivalence between (5) and (6) is already known if (X_n) is exchangeable under P_σ; see e.g. [21].

Finally, we focus on the special case where σ is assessed according to (2). Then, Theorem 15 provides conditions for μ to be a.s. discrete.

Theorem 16. *Suppose the strategy σ is given by* (2) *and*

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

Then, μ admits representation (5) *and the weights W_k are given by* (7)*.*

Proof. Just note that

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

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

$$P_\sigma\big(X_{n+1}\neq X_i \text{ for each } i\le n\big)=E_{P_\sigma}\Big\{\sigma_0\big(\{X_1,\ldots,X_n\}^c\big)\prod_{i=0}^{n-1} q_i(X_1,\ldots,X_i)\Big\}$$
$$\le E_{P_\sigma}\Big\{\prod_{i=0}^{n-1} q_i(X_1,\ldots,X_i)\Big\}\longrightarrow 0.$$

An application of Theorem 15 concludes the proof. □

Various popular random probability measures ν admit the representation

$$\nu \overset{a.s.}{=} \sum_k D_k\,\delta_{Z_k}, \tag{8}$$

where (Z_k) is an i.i.d. sequence of random variables and the weights (D_k) are independent of (Z_k). A well known example is the Dirichlet random probability measure; see e.g. [19] and [24]. Our last result is that μ often admits representation (8) provided σ is given by (2) and the q_i are constant.

Theorem 17. *Suppose the strategy σ is given by* (2) *and σ_0 is diffuse. Suppose also that q_i is constant for every $i\ge 0$, 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 (5) *and the weights W_k are given by* (7)*. 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 Subsection 7.1. Since the predictive distributions of (Y_n) are given by (2), we can replace (X_n) with (Y_n). In addition, since

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

the Borel-Cantelli lemma yields

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

Hence, one can define

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

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

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

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

Since $\prod_{i=0}^{n-1} q_i \to 0$, Theorem 16 implies

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

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 $Y_n \notin \{Y_1, \ldots, Y_{n-1}\}$ if and only if $Y_n = T_n$. Further, $P(A) = 1$ 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

$$P\big(H \cap \{Y_1^* \in B_1, \ldots, Y_k^* \in B_k\}\big) =$$

$$= \sum_{m_1, \ldots, m_k} P\big(H \cap \{\rho_1 = m_1, \ldots, \rho_k = m_k, T_{m_1} \in B_1, \ldots, T_{m_k} \in B_k\}\big)$$

$$= \sum_{m_1, \ldots, m_k} P\big(T_{m_1} \in B_1, \ldots, T_{m_k} \in B_k\big)\, P\big(H \cap \{\rho_1 = m_1, \ldots, \rho_k = m_k\}\big)$$

$$= \prod_{i=1}^k \sigma_0(B_i) \sum_{m_1, \ldots, m_k} P\big(H \cap \{\rho_1 = m_1, \ldots, \rho_k = m_k\}\big) = P(H) \prod_{i=1}^k \sigma_0(B_i).$$

This concludes the proof. □

Incidentally, if σ_0 is diffuse, Theorem 17 applies to Dirichlet sequences; see Example 7. In that case, as already noted, it is well known that μ admits representation (8). However, Theorem 17 says something more. Not only (8) holds, but one can take $Z_k = X_k^*$ and $D_k = W_k$, namely, the sequence (X_k^*) of distinct observations is i.i.d. and independent of (W_k). In addition, the weights W_k can be written according to (7). Most probably, in the special case of Dirichlet sequences, all these facts are already known, but we are not aware of any explicit reference; see e.g. [19] and references therein.

References

[1] Airoldi E.M., Thiago Costa, Bassetti F., Leisen F., Guindani M. (2014) Generalized species sampling priors with latent beta reinforcements, *J.A.S.A.*, 109, 1466-1480.

[2] Bassetti F., Crimaldi I., Leisen F. (2010) Conditionally identically distributed species sampling sequences, *Adv. in Appl. Probab.*, 42, 433-459.

[3] Berti P., Regazzini E., Rigo P. (1997) Well-calibrated, coherent forecasting systems, *Theory Probab. Appl.*, 42, 82-102.

[4] Berti P., Pratelli L., Rigo P. (2004) Limit theorems for a class of identically distributed random variables, *Ann. Probab.*, 32, 2029-2052.

[5] Berti P., Rigo P. (2007) 0-1 laws for regular conditional distributions, *Ann. Probab.*, 35, 649-662.
[6] Berti P., Crimaldi I., Pratelli L., Rigo P. (2009) Rate of convergence of predictive distributions for dependent data, *Bernoulli*, 15, 1351-1367.
[7] Berti P., Pratelli L., Rigo P. (2012) Limit theorems for empirical processes based on dependent data, *Electronic J. Probab.*, 17, 1-18.
[8] Berti P., Pratelli L., Rigo P. (2013) Exchangeable sequences driven by an absolutely continuous random measure, *Ann. Probab.*, 41, 2090-2102.
[9] Berti P., Dreassi E., Pratelli L., Rigo P. (2019) Asymptotics of certain conditionally identically distributed sequences, *Submitted*, currently available at: http://www-dimat.unipv.it/ rigo/approach.pdf
[10] Cassese A., Zhu W., Guindani M., Vannucci M. (2019) A Bayesian nonparametric spiked process prior for dynamic model selection, *Bayesian Analysis*, 14, 553-572.
[11] Cifarelli D.M., Regazzini E. (1996) De Finetti's contribution to probability and statistics, *Statist. Science*, 11, 253-282.
[12] de Finetti B. (1937) La prevision: ses lois logiques, ses sources subjectives, *Ann. Inst. H. Poincare*, 7, 1-68.
[13] Dubins L.E., Savage L.J. (1965) *How to gamble if you must: Inequalities for stochastic processes*, McGraw Hill.
[14] Fortini S., Ladelli L., Regazzini E. (2000) Exchangeability, predictive distributions and parametric models, *Sankhya* A, 62, 86-109.
[15] Fortini S., Petrone S. (2012) Predictive construction of priors in Bayesian nonparametrics, *Brazilian J. Probab. Statist.*, 26, 423-449.
[16] Fortini S., Petrone S., Sporysheva P. (2018) On a notion of partially conditionally identically distributed sequences, *Stoch. Proc. Appl.*, 128, 819-846.
[17] Hahn P.R., Martin R., Walker S.G. (2018) On recursive Bayesian predictive distributions, *J.A.S.A.*, 113, 1085-1093.
[18] Hill B.M. (1993) Parametric models for A_n: splitting processes and mixtures, *J. Royal Stat. Soc.* B, 55, 423-433.
[19] Ishwaran H., Zarepour M. (2002) Exact and approximate sum representations for the Dirichlet process, *Canadian J. Stat.*, 30, 269-283.
[20] Kallenberg O. (1988) Spreading and predictable sampling in exchangeable sequences and processes, *Ann. Probab.*, 16, 508-534.
[21] Krasker W.S., Pratt J.W. (1986) Discussion: On the consistency of Bayes estimates, *Ann. Statist.*, 14, 55-58.
[22] Martin R., Tokdar S.T. (2011) Semiparametric inference in mixture models with predictive recursion marginal likelihhod, *Biometrika*, 98, 567-582.
[23] Newton M.A., Zhang Y. (1999) A recursive algorithm for nonparametric analysis with missing data, *Biometrika*, 86, 15-26.
[24] Sethuraman J. (1994) A constructive definition of Dirichlet priors, *Stat. Sinica*, 4, 639-650.
[25] Tokdar S.T., Martin R., Ghosh J.K. (2009) Consistency of a recursive estimate of mixing distributions, *Ann. Statist.*, 37, 2502-2522.

Patrizia Berti, Dipartimento di Matematica Pura ed Applicata "G. Vitali", Universita' di Modena e Reggio-Emilia, via Campi 213/B, 41100 Modena, Italy
E-mail address: patrizia.berti@unimore.it

Emanuela Dreassi, Dipartimento di Statistica, Informatica, Applicazioni "G. Parenti", Universita' di Firenze, viale Morgagni 59, 50134 Firenze, Italy
E-mail address: emanuela.dreassi@unifi.it

Luca Pratelli, Accademia Navale, viale Italia 72, 57100 Livorno, Italy
E-mail address: pratel@mail.dm.unipi.it

Pietro Rigo (corresponding author), Dipartimento di Matematica "F. Casorati", Universita' di Pavia, via Ferrata 1, 27100 Pavia, Italy
E-mail address: pietro.rigo@unipv.it