ASYMPTOTIC PREDICTIVE INFERENCE
WITH EXCHANGEABLE DATA

PATRIZIA BERTI, LUCA PRATELLI, AND PIETRO RIGO

Abstract. Let \((X_n)\) be a sequence of random variables, adapted to a filtration \((G_n)\), and let \(\mu_n = (1/n) \sum_{i=1}^{n} \delta_{X_i}\) and \(a_n(\cdot) = P(X_{n+1} \in \cdot \mid G_n)\) be the empirical and the predictive measures. We focus on
\[
\|\mu_n - a_n\| = \sup_{B \in D} |\mu_n(B) - a_n(B)|
\]
where \(D\) is a class of measurable sets. Conditions for \(\|\mu_n - a_n\| \to 0\), almost surely or in probability, are given. Also, to determine the rate of convergence, the asymptotic behavior of \(r_n \|\mu_n - a_n\|\) is investigated for suitable constants \(r_n\). Special attention is paid to \(r_n = \sqrt{n}\) and \(r_n = \sqrt{\log \log n}\). The sequence \((X_n)\) is exchangeable or, more generally, conditionally identically distributed.

1. Introduction

Throughout, \(S\) is a Borel subset of a Polish space and \(X = (X_n : n \geq 1)\) a sequence of \(S\)-valued random variables on a probability space \((\Omega, \mathcal{A}, P)\). Further, \(\mathcal{G} = (\mathcal{G}_n : n \geq 0)\) is a filtration on \((\Omega, \mathcal{A}, P)\) and \(\mathcal{B}\) is the Borel \(\sigma\)-field on \(S\) (thus, \(\mathcal{B}\) is generated by the relative topology that \(S\) inherits as a subset of a Polish space).

We fix a subclass \(D \subset \mathcal{B}\) and we let \(\|\cdot\|\) denote the sup-norm over \(D\), namely
\[
\|\alpha - \beta\| = \sup_{B \in D} |\alpha(B) - \beta(B)|
\]
whenever \(\alpha\) and \(\beta\) are probability measures on \(\mathcal{B}\).

Let
\[
\mu_n = (1/n) \sum_{i=1}^{n} \delta_{X_i} \quad \text{and} \quad a_n(\cdot) = P(X_{n+1} \in \cdot \mid \mathcal{G}_n).
\]
Both \(\mu_n\) and \(a_n\) are random probability measures on \(\mathcal{B}\); \(\mu_n\) is the empirical measure and (if \(X\) is \(\mathcal{G}\)-adapted) \(a_n\) is the predictive measure.

Under some conditions, \(\mu_n(B) - a_n(B) \xrightarrow{a.s.} 0\) for fixed \(B \in \mathcal{B}\). In that case, a question is whether \(D\) is such that \(\|\mu_n - a_n\| \xrightarrow{a.s.} 0\). As discussed in Section 2, such a question naturally arises in several frameworks, including Bayesian consistency and frequentistic approximation of Bayesian procedures.

In this paper, conditions for \(\|\mu_n - a_n\| \to 0\), almost surely or in probability, are given. Also, to determine the rate of convergence, the limit behavior of \(r_n \|\mu_n - a_n\|\) is investigated for suitable constants \(r_n\). Special attention is paid to \(r_n = \sqrt{n}\) and

2010 Mathematics Subject Classification. 60G09, 60B10, 60A10, 60G57, 62F15.

\( r_n = \sqrt{\frac{n}{\log \log n}} \). Various new results are proved. In addition, to get a reasonably complete picture, a few known facts from [2]-[5] are connected and unified.

The sequence \( X \) is assumed to be exchangeable or, more generally, conditionally identically distributed. We refer to Section 3 for conditionally identically distributed sequences, and we recall that \( \mu_n \) identically distributed. We refer to Section 3 for conditionally identically complete picture, a few known facts from [2]-[5] are connected and unified.

We next briefly state some results. We assume a mild measurability condition \( \mathcal{D} \), called countable determinacy and introduced in Section 3. For the sake of simplicity, we take \( X \) exchangeable and \( \mathcal{G} = \mathcal{G}^X \), where

\[
\mathcal{G}^0_X = \{\emptyset, \Omega\} \quad \text{and} \quad \mathcal{G}^X_n = \sigma(X_1, \ldots, X_n), \quad n \geq 1,
\]
is the filtration induced by \( X \). We also recall that, since \( X \) is exchangeable, there is a (a.s. unique) random probability measure \( \mu \) on \( \mathcal{B} \) such that \( \mu_n(B) \xrightarrow{\text{a.s.}} \mu(B) \) for each \( B \in \mathcal{B} \); see e.g. [1].

Then, \( \|\mu_n - a_n\| \xrightarrow{\text{a.s.}} 0 \) with \( \mathcal{D} = \mathcal{B} \) provided \( \mu \) is a.s. discrete; see Example 4.

This simple fact may be useful in Bayesian nonparametrics, for \( \mu \) is a.s. discrete under most popular priors. Indeed, examples of nonparametric priors which lead to a discrete \( \mu \) are: Dirichlet [26], two-parameter Poisson-Dirichlet [24], normalized completely random measures [20], Gibbs-type priors [12] and beta-stacy [23].

Another useful fact (Theorem 2 and Corollary 3) is that

\[
\limsup_n \sqrt{\frac{n}{\log \log n}} \|\mu_n - a_n\| \leq \sqrt{2 \sup_{B \in \mathcal{D}} \mu(B) (1 - \mu(B))} \quad \text{a.s.}
\]

provided \( \mathcal{D} \) is a VC-class. Unlike the i.i.d. case, inequality (1) is not sharp. If \( X \) is exchangeable, it may be even that \( n \|\mu_n - a_n\| \) converges a.s. to a finite limit. This happens, for instance, when the probability distribution of \( X \) is of the Ferguson-Dirichlet type, as defined in Subsection 4.2; see also forthcoming Theorem 6. Even if not sharp, however, inequality (1) provides a meaningful information on the rate of convergence of \( \|\mu_n - a_n\| \) when \( X \) is exchangeable and \( \mathcal{D} \) a VC-class.

The notion of VC-class is recalled in Subsection 4.1 (before Corollary 3). VC-classes are quite popular in frameworks such as empirical processes and statistical learning, and in real problems \( \mathcal{D} \) is often a VC-class. If \( S = \mathbb{R}^k \), for instance, \( \mathcal{D} = \{(-\infty, t_1] \times \cdots \times (-\infty, t_k] : (t_1, \ldots, t_k) \in \mathbb{R}^k\} \), \( \mathcal{D} = \{\text{half spaces}\} \) and \( \mathcal{D} = \{\text{closed balls}\} \) are VC-classes.

A further result (Corollary 8) concerns \( r_n = \sqrt{n} \). Let

\[
a_n^*(B) = P\{X_{n+1} \in B \mid I_B(X_1), \ldots, I_B(X_n)\}
\]

where \( I_B(X_i) \) denotes the indicator of the set \( \{X_i \in B\} \). Roughly speaking, \( a_n^*(B) \) is the conditional probability that the next observation falls in \( B \) given only the history of \( B \) in the previous observations. Suppose that the random variable \( \mu(B) \) has an absolutely continuous distribution (with respect to Lebesgue measure) for those \( B \in \mathcal{D} \) satisfying \( 0 < P(X_1 \in B) < 1 \). Then, for fixed \( B \in \mathcal{D} \),

\[
\sqrt{n} \{\mu_n(B) - a_n(B)\} \xrightarrow{P} 0 \iff \sqrt{n} \{a_n(B) - a_n^*(B)\} \xrightarrow{P} 0.
\]

In addition, under some assumptions on the empirical processes \( W_n = \sqrt{n} (\mu_n - \mu) \) (satisfied in several real situations), one obtains

\[
\sqrt{n} \|\mu_n - a_n\| \xrightarrow{P} 0 \iff \sqrt{n} \{a_n(B) - a_n^*(B)\} \xrightarrow{P} 0 \text{ for each } B \in \mathcal{D}.
\]
However, $\sqrt{n} \left\{ a_n(B) - a_n^*(B) \right\}$ may fail to converge to 0 in probability even if $\mu(B)$ has an absolutely continuous distribution; see Example 9.

We finally mention a result (Theorem 10) which, though in the spirit of this paper, is quite different from those described above. Such a result has been inspired by [22]. Let $S = \{0, 1\}$ and $C$ the Borel $\sigma$-field on $[0, 1]$. For $C \in C$, define

$$\pi_n(C) = P(\mu_n(1) \in C) \quad \text{and} \quad \pi_n^*(C) = P(a_n(1) \in C)$$

and denote by $\rho$ the bounded Lipschitz metric between probability measures on $C$. Then,

$$\rho(\pi_n, \pi_n^*) \leq \frac{1}{n} \left( 1 + \frac{c}{3} \right)$$

provided the limit frequency $\mu(1)$ has an absolutely continuous distribution with Lipschitz density $f$. Here, $c$ is the Lipschitz constant of $f$. This rate of convergence cannot be improved.

2. Motivations

There are various (non-independent) reasons for investigating how close $\mu_n$ and $a_n$ are. We now list a few of them under the assumption that

$$(\Omega, \mathcal{A}) = (S^\infty, \mathcal{B}^\infty), \quad X_n = n\text{-th coordinate projection}, \quad \mathcal{G} = \mathcal{G}^X.$$

Most remarks, however, apply to any filtration $\mathcal{G}$ which makes $X$ adapted.

Similarly, in most of the subsequent comments, $\|\cdot\|$ could be replaced by some other distance $\rho$ between probability measures. For instance, in [10], the asymptotics of $\rho(\mu_n, a_n)$ is taken into account with $\rho$ the bounded Lipschitz metric and $\rho$ the Wasserstein distance.

For a general background of Bayesian nonparametrics, often mentioned in what follows, we refer to [18]-[19]; see also [11].

2.1. Bayesian predictive inference. In a number of frameworks, mainly in Bayesian nonparametrics and discrete time filtering, one main goal is to evaluate $a_n$. Quite frequently, however, the latter cannot be obtained in closed form. For some nonparametric priors, for instance, no closed form expression of $a_n$ is known. In these situations, there are essentially two ways out: to compute $a_n$ numerically (MCMC) or to estimate it by the available data. If we take the second route, and if data are exchangeable or conditionally identically distributed, $\mu_n$ is a reasonable estimate of $a_n$. Then, the asymptotic behavior of the error $\mu_n - a_n$ plays a role. In a sense, this is the basic reason for investigating $\|\mu_n - a_n\|$.

2.2. Bayesian consistency. In the spirit of Subsection 2.1, with $\mu_n$ regarded as an estimate of $a_n$, it makes sense to say that $\mu_n$ is consistent if $\|\mu_n - a_n\| \to 0$ a.s. or in probability. In this brief discussion, to fix ideas, we focus on a.s. convergence.

Suppose $X$ is exchangeable. Let $\mathcal{P}$ be the set of all probability measures on $\mathcal{B}$ and $\mu$ the random probability measure on $\mathcal{B}$ introduced in Section 1. For each $\nu \in \mathcal{P}$, let $P_\nu$ denote the probability measure on $\mathcal{B}^\infty$ which makes $X$ i.i.d. with common distribution $\nu$. By de Finetti’s theorem, conditionally on $\mu$, the sequence $X$ is i.i.d. with common distribution $\mu$; see e.g. [1]. It follows that

$$P(\cdot) = \int_{\mathcal{P}} P_\nu(\cdot) \pi(d\nu)$$
where $\pi$ is the probability distribution of $\mu$. Such a $\pi$ is usually called the prior distribution.

In the standard approach to consistency, after Diaconis and Freedman [13], the asymptotic behavior of any statistical procedure is investigated under $P_\nu$ for each $\nu \in \mathcal{P}$. The procedure is consistent provided it behaves properly for each $\nu \in \mathcal{P}$ (or at least for each $\nu$ in some known subset of $\mathcal{P}$); see e.g. [18], [19] and references therein. In particular, $\mu_n$ is a consistent estimate of $a_n$ if

$$P_\nu(\|\mu_n - a_n\| \to 0) = 1 \quad \text{for each } \nu \in \mathcal{P}.$$ 

A different point of view is taken in this paper. Indeed, $\|\mu_n - a_n\|$ is investigated under $P$ and $\mu_n$ is a consistent estimate of $a_n$ if

$$P(\|\mu_n - a_n\| \to 0) = 1.$$ 

In a sense, in the first approach, consistency of Bayesian procedures is evaluated from a frequentistic point of view. Regarding $P$ as a parameter space, in fact, $\mu_n$ is demanded to approximate $a_n$ for each possible value of the parameter $\nu$. This request is certainly admissible. Furthermore, the first notion of consistency is technically stronger than the second. On the other hand, it is not so clear why a Bayesian inferrer should take a frequentistic point of view. Even if $P$ is a mixture of $\{P_\nu : \nu \in \mathcal{P}\}$, when dealing with $X$ the relevant probability measure is $P$ and not $P_\nu$. Furthermore, according to de Finetti, any probability statement should concern “observable” facts, while $P_\nu$ is conditional on the “unobservable” fact $\mu = \nu$. Thus, according to us, the second approach to consistency is in line with the foundations of Bayesian statistics. A similar opinion is in [10] and [16].

2.3. Frequentistic approximation of Bayesian procedures. In Subsection 2.1, $\mu_n$ is viewed as an estimate of $a_n$. A similar view, developed in [10], is to regard $\mu_n$ as a frequentistic approximation of the Bayesian procedure $a_n$. For instance, such an approximation makes sense within the empirical Bayes approach, where the orthodox Bayesian reasoning is combined in various ways with frequentistic elements; see e.g. [15] and [25]. We also note that, historically, one reason for introducing exchangeability (possibly, the main reason) was to justify observed frequencies as predictors of future events; see [9] and [28]. In this sense, to focus on $\|\mu_n - a_n\|$ is in line with de Finetti’s ideas.

2.4. Predictive distributions of exchangeable sequences. If $X$ is exchangeable, just very little is known on the general form of $a_n$ for given $n$; see e.g. [16]. Indeed, a representation theorem for $a_n$ would be a major breakthrough. Failing the latter, to fix the asymptotic behavior of $\|\mu_n - a_n\|$ contributes to fill the gap.

2.5. Empirical processes for non-ergodic data. Slightly abusing terminology, say that $X$ is ergodic if $P$ is 0-1 valued on the sub-$\sigma$-field

$$\sigma(\limsup_n \mu_n(B) : B \in \mathcal{B}).$$

In real problems, $X$ is often non-ergodic. Most stationary sequences, for instance, fail to be ergodic. Or else, an exchangeable sequence is ergodic if and only if is i.i.d. Now, if $X$ is i.i.d., the empirical process is defined as $G_n = \sqrt{n}(\mu_n - \mu_0)$ where $\mu_0$ is the probability distribution of $X_1$. But this definition has various drawbacks when $X$ is not ergodic; see [6]. In fact, unless $X$ is i.i.d., the probability distribution of $X$
is not determined by that of $X_1$. More importantly, if $G_n$ converges in distribution in $l^\infty(\mathcal{D})$ (the metric space $l^\infty(\mathcal{D})$ is recalled before Corollary 8) then

$$\|\mu_n - \mu_0\| = n^{-1/2}\|G_n\| \xrightarrow{P} 0.$$ 

But $\|\mu_n - \mu_0\|$ typically fails to converge to 0 in probability when $X$ is not ergodic. Thus, empirical processes for non-ergodic data should be defined in some different way. At least in the exchangeable case, a meaningful option is to center $\mu_n$ by $a_n$, namely, to let $G_n = \sqrt{n}(\mu_n - a_n)$.

3. Assumptions

Let $\mathcal{D} \subset \mathcal{B}$. To avoid measurability problems, $\mathcal{D}$ is assumed to be countably determined. This means that there is a countable subclass $\mathcal{D}_0 \subset \mathcal{D}$ such that

$$\|\alpha - \beta\| = \sup_{B \in \mathcal{D}_0} |\alpha(B) - \beta(B)| \text{ for all probability measures } \alpha, \beta \text{ on } \mathcal{B}.$$ 

A sufficient condition is that there is a countable subclass $\mathcal{D}_0 \subset \mathcal{D}$ such that, for each $B \in \mathcal{D}$ and each probability measure $\alpha$ on $B$, one obtains

$$\lim_n \alpha(B \Delta B_n) = 0 \text{ for some sequence } B_n \in \mathcal{D}_0.$$ 

Most classes $\mathcal{D}$ involved in applications are countably determined. For instance, $\mathcal{D} = \mathcal{B}$ is countably determined (for $\mathcal{B}$ is countably generated). Or else, if $S = \mathbb{R}^k$, then $\mathcal{D} = \{\text{closed convex sets}\}$, $\mathcal{D} = \{\text{half spaces}\}$, $\mathcal{D} = \{\text{closed balls}\}$ and

$$\mathcal{D} = \left\{(-\infty, t_1] \times \cdots \times (-\infty, t_k] : (t_1, \ldots, t_k) \in \mathbb{R}^k\right\}$$

are countably determined.

We next recall the notion of conditionally identically distributed (c.i.d.) random variables. The sequence $X$ is c.i.d. with respect to $\mathcal{G}$ if it is $\mathcal{G}$-adapted and

$$P\left(X_k \in \cdot \mid G_n\right) = P\left(X_{n+1} \in \cdot \mid G_n\right) \text{ a.s. for all } k > n \geq 0.$$ 

Roughly speaking, at each time $n \geq 0$, the future observations $(X_k : k > n)$ are identically distributed given the past $G_n$. When $\mathcal{G} = G^X$, the filtration $\mathcal{G}$ is not mentioned at all and $X$ is just called c.i.d. Then, $X$ is c.i.d. if and only if

$$\left(X_1, \ldots, X_n, X_{n+2}\right) \sim \left(X_1, \ldots, X_n, X_{n+1}\right) \text{ for all } n \geq 0.$$ 

Exchangeable sequences are c.i.d., for they meet (2), while the converse is not true. Indeed, $X$ is exchangeable if and only if it is stationary and c.i.d. We refer to [4] for more on c.i.d. sequences. Here, it suffices to mention the strong law of large numbers and some of its consequences.

If $X$ is c.i.d., there is a random probability measure $\mu$ on $\mathcal{B}$ satisfying

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

As a consequence, if $X$ is c.i.d. with respect to $\mathcal{G}$, for each $n \geq 0$ and $B \in \mathcal{B}$ one obtains

$$E\{\mu(B) \mid G_n\} = \lim_m E\{\mu_m(B) \mid G_n\} = \lim_m \frac{1}{m} \sum_{k=n+1}^{m} P(X_k \in B \mid G_n)$$

$$= P(X_{n+1} \in B \mid G_n) = a_n(B) \text{ a.s.}$$

In particular, $a_n(B) = E\{\mu(B) \mid G_n\} \xrightarrow{a.s.} \mu(B)$ so that $\mu_n(B) - a_n(B) \xrightarrow{a.s.} 0$. 

\textbf{ASYMPTOTIC PREDICTIVE INFERENCE 5}
From now on, \( X \) is c.i.d. with respect to \( \mathcal{G} \). In particular, \( X \) is identically distributed and \( \mu_0 \) denotes the probability distribution of \( X_1 \). We also let

\[
W_n = \sqrt{n}(\mu_n - \mu).
\]

Note that, if \( X \) is i.i.d., then \( \mu = \mu_0 \) a.s. and \( W_n \) reduces to the usual empirical process.

4. Results

Our results can be sorted into three subsections.

4.1. Two general criterions. Since \( a_n(B) = E\{\mu(B) \mid \mathcal{G}_n\} \) a.s. and \( \mathcal{D} \) is countably determined, one obtains

\[
\|\mu_n - a_n\| = \sup_{B \in \mathcal{D}_n} |\mu_n(B) - a_n(B)|
= \sup_{B \in \mathcal{D}_0} |E\{\mu_n(B) - \mu(B) \mid \mathcal{G}_n\}| \leq E\{\|\mu_n - \mu\| \mid \mathcal{G}_n\} \text{ a.s.}
\]

This simple inequality has some nice consequences. Recall that \( \mathcal{D} \) is a universal Glivenko-Cantelli class if \( \|\mu_n - \mu_0\| \xrightarrow{a.s.} 0 \) whenever \( X \) is i.i.d.; see e.g. [14], [17], [27].

**Theorem 1**. ([3] and [5]). Suppose \( \mathcal{D} \) is countably determined and \( X \) is c.i.d. with respect to \( \mathcal{G} \). Then,

(i) \( \|\mu_n - a_n\| \xrightarrow{a.s.} 0 \) if \( \|\mu_n - \mu\| \xrightarrow{a.s.} 0 \) and \( \|\mu_n - a_n\| \xrightarrow{P} 0 \) if \( \|\mu_n - \mu\| \xrightarrow{P} 0 \).

In particular, \( \|\mu_n - a_n\| \xrightarrow{a.s.} 0 \) provided \( X \) is exchangeable, \( \mathcal{G} = \mathcal{G}^X \) and \( \mathcal{D} \) is a universal Glivenko-Cantelli class.

(ii) \( r_n\|\mu_n - a_n\| \xrightarrow{P} 0 \) whenever the constants \( r_n \) satisfy \( r_n/\sqrt{n} \rightarrow 0 \) and \( \sup_n E\{\|W_n\|^p\} < \infty \) for some \( p \geq 1 \).

**Proof.** Since \( \|\mu_n - \mu\| \leq 1 \), if \( \|\mu_n - \mu\| \xrightarrow{a.s.} 0 \) then

\[
\|\mu_n - a_n\| \leq E\{\|\mu_n - \mu\| \mid \mathcal{G}_n\} \xrightarrow{a.s.} 0
\]

because of the martingale convergence theorem in the version of [8]. Similarly, \( \|\mu_n - \mu\| \xrightarrow{P} 0 \) implies \( E\{\|\mu_n - \mu\| \mid \mathcal{G}_n\} \xrightarrow{P} 0 \) by an obvious argument based on subsequences. Next, let \( X \) be exchangeable. By de Finetti’s theorem, conditionally on \( \mu \), the sequence \( X \) is i.i.d. with common distribution \( \mu \). If \( \mathcal{D} \) is a universal Glivenko-Cantelli class, it follows that

\[
P(\|\mu_n - \mu\| \rightarrow 0) = \int P(\|\mu_n - \mu\| \rightarrow 0 \mid \mu) \, dP = \int 1 \, dP = 1.
\]

This concludes the proof of (i). As to (ii), just note that

\[
E\{r_n\|\mu_n - a_n\|^p\} \leq r_n^p E\{E\{\|\mu_n - \mu\| \mid \mathcal{G}_n\}^p\}
\leq r_n^p E\{\|\mu_n - \mu\|^p\} = (r_n/\sqrt{n})^p E\{\|W_n\|^p\}.
\]

\( \square \)

While Theorem 1 is essentially known (the proof has been provided for completeness only) the next result is new.
Theorem 2. Suppose $\mathcal{D}$ is countably determined and $X$ is c.i.d. with respect to $\mathcal{G}$. Fix the constants $r_n > 0$ and define
\[ M_k = \sup_{n \geq k} r_n \| \mu_n - \mu \|. \]
If $E(M_k) < \infty$ for some $k$, then
\[ \limsup_n r_n \| \mu_n - a_n \| \leq \limsup_n r_n \| \mu_n - \mu \| < \infty \quad \text{a.s.} \]
Moreover, if $X$ is exchangeable, then $E(M_k) < \infty$ for some $k$ whenever
\begin{align*}
(iii) & \quad r_n = \frac{\sqrt{n}}{(\log n)^{1/2}} \text{ and } \sup_n E\left\{ \| W_n \|^p \right\} < \infty \text{ for some } p > 1 \text{ and } 0 < c < p; \\
(iv) & \quad r_n = \sqrt{\frac{n}{\log \log n}} \text{ and } \sup_n E\left\{ \exp (a \| W_n \|) \right\} \leq a \exp (b u^2) \quad \text{for all } u > 0 \text{ and some } a, b > 0.
\end{align*}
Proof. Fix $j \geq k$. Since $E(M_j) \leq E(M_k) < \infty$, then
\[ \limsup_n r_n \| \mu_n - a_n \| \leq \limsup_n E\left\{ r_n \| \mu_n - \mu \| \mid \mathcal{G}_n \right\} \leq \limsup_n E\left\{ M_j \mid \mathcal{G}_n \right\} = M_j \quad \text{a.s.} \]
where the last equality is due to the martingale convergence theorem. Hence,
\[ \limsup_n r_n \| \mu_n - a_n \| \leq \inf_{j \geq k} M_j = \limsup_n r_n \| \mu_n - \mu \| \quad \text{a.s.} \]
Further, $E(M_k) < \infty$ obviously implies $\limsup_n r_n \| \mu_n - \mu \| \leq M_k < \infty$ a.s.
Next, suppose $X$ exchangeable. Then,
\[ S_n = n \| \nu_n - \mu \| = \sqrt{n} \| W_n \| \]
is a submartingale with respect to the filtration $\mathcal{U}_n = \sigma \left[ \mathcal{G}_n^X \cup \sigma(\mu) \right]$. In fact,
\begin{align*}
(n + 1) E\left\{ \mu_{n+1} (B) \mid \mathcal{U}_n \right\} &= n \mu_n (B) + P \{ X_{n+1} \in B \mid \mathcal{U}_n \} \\
&= n \mu_n (B) + P \{ X_{n+1} \in B \mid \sigma(\mu) \} = n \mu_n (B) + \mu (B) \quad \text{a.s.}
\end{align*}
Therefore,
\[ E(S_{n+1} \mid \mathcal{U}_n) \geq (n + 1) \sup_{B \in \mathcal{D}} E\left\{ \mu_{n+1} (B) \mid \mathcal{U}_n \right\} - \mu (B) = n \sup_{B \in \mathcal{D}} \mu_n (B) - \mu (B) = S_n \quad \text{a.s.} \]
(iii) Let $r_n = \frac{\sqrt{n}}{(\log n)^{1/2}}$ and $\sup_n E\| W_n \|^p < \infty$, where $p > 1$ and $0 < c < p$. Then,
\[ E(M_j^p) = E\left\{ \left( \sup_{n \geq 1} \max_{2^n < j \leq 2^{n+1}} r_j \| \mu_j - \mu \| \right)^p \right\} \leq \sum_{n=1}^{\infty} E\left\{ \max_{2^n < j \leq 2^{n+1}} r_j^p \| \mu_j - \mu \|^p \right\}. \]
If $2^n < j \leq 2^{(n+1)}$, then
\[ r_j \| \mu_j - \mu \| = j^{-1/2} (\log j)^{-1/2} S_j \leq (2^n)^{-1/2} (\log 2^n)^{-1/2} S_j. \]
By such inequality and since \((S_j)\) is a submartingale, one obtains
\[
E(M^p_4) \leq \sum_n (2^n)^{-p/2} (\log 2^n)^{-p/c} E\left\{ \max_{j \leq 3^{(n+1)}} S^p_j \right\}
\]
\[
\leq \left( p/(p-1) \right)^p \sum_n (2^n)^{-p/2} (\log 2^n)^{-p/c} E\left\{ S^p_{3^{(n+1)}} \right\}
\]
\[
= \left( p/(p-1) \right)^p 2^{p/2} \sum_n (\log 2^n)^{-p/c} E\{\|W_{3^{(n+1)}}\|^p\}
\]
\[
\leq \left( \sup_j E\{\|W_j\|^p\} \right) \left( p/(p-1) \right)^p 2^{p/2} (\log 2)^{-p/c} \sum_n n^{-p/c} < \infty.
\]

(iii) Let \(r_n = \sqrt{n \log \log n}\) and \(\sup_n E\{\exp(u\|W_n\|)\} \leq a \exp(bu^2)\) for all \(u > 0\) and some \(a, b > 0\). We aim to prove that
\[
P(M_4 > t) \leq c \exp\left(-vt^2\right)
\]
for large \(t\) and suitable constants \(c, v > 0\).

In this case, \(E(M_4) = \int_0^\infty P(M_4 > t)\,dt < \infty\).

First note that
\[
P(M_4 > t) = P\left( \bigcup_{n \geq 1} \left\{ \max_{j \leq 3^n} |r_j| > t \right\} \right) \leq \sum_{n=1}^\infty P\left( \max_{j \leq 3^{(n+1)}} |S_j| > m_n t \right)
\]
where \(m_n = \sqrt{3^n \log \log 3^n} = \sqrt{3^n} (\log n + \log \log 3)\).

Let \(\theta > 0\). On noting that \(\exp(\theta S_n)\) is still a submartingale, one also obtains
\[
P\left( \max_{j \leq 3^{(n+1)}} |S_j| > m_n t \right) = P\left( \max_{j \leq 3^{(n+1)}} \exp(\theta S_j) > \exp(\theta m_n t) \right)
\]
\[
\leq \exp(-\theta m_n t) E\left\{ \exp(\theta S_{3^{(n+1)}}) \right\}
\]
\[
= \exp(-\theta m_n t) E\left\{ \exp\left( \theta \sqrt{3^{(n+1)}} \|W_{3^{(n+1)}}\| \right) \right\}
\]
\[
\leq a \exp(-\theta m_n t + \theta^2 b 3^{(n+1)}).
\]

The minimum over \(\theta\) is attained at \(\theta = \frac{m_n t}{b 3^{(n+1)}}\). Thus,
\[
P\left( \max_{j \leq 3^{(n+1)}} |S_j| > m_n t \right) \leq a \exp\left( -\frac{m_n^2 t^2}{12 b 3^{(n)}} \right) = a \exp\left( -\frac{(t^2 \log \log 3}{12 b} \right) \, n^{-t^2/12 b}.
\]

If \(t \geq \sqrt{24 b}\), then \(t^2 > 12 b\) and \(\frac{t^2}{t^2 - 12 b} \leq 2\). Thus, one finally obtains
\[
P(M_4 > t) \leq a \exp\left( -\frac{t^2 \log \log 3}{12 b} \right) \sum_n n^{-t^2/12 b}
\]
\[
\leq a \exp\left( -\frac{t^2 \log \log 3}{12 b} \right) \frac{t^2}{t^2 - 12 b}
\]
\[
\leq 2 a \exp\left( -\frac{t^2 \log \log 3}{12 b} \right)\text{ for every } t \geq \sqrt{24 b}.
\]

\qed

Some remarks are in order. In the sequel, if \(\alpha\) and \(\beta\) are measures on a \(\sigma\)-field \(\mathcal{E}\), we write \(\alpha \ll \beta\) to mean that \(\alpha\) is absolutely continuous with respect to \(\beta\), namely, \(\alpha(A) = 0\) whenever \(A \in \mathcal{E}\) and \(\beta(A) = 0\).
• Sometimes, the condition of Theorem 1-(i) is necessary as well, namely, 
\[ \|\mu_n - a_n\| \overset{a.s.}{\to} 0 \] if and only if \[ \|\mu_n - \mu\| \overset{a.s.}{\to} 0. \] For instance, this happens when \( G = G^X \) and \( \mu \ll \lambda \) a.s., where \( \lambda \) is a (non-random) \( \sigma \)-finite measure on \( B \). In this case, in fact, \( \|a_n - \mu\| \overset{a.s.}{\to} 0 \) by [7, Theorem 1].

• Several examples of universal Glivenko-Cantelli classes are available; see [14], [17], [27] and references therein. Moreover, for many choices of \( \mathcal{D} \) and \( p \) there is a universal constant \( c(p) \) such that \( \sup_n E \{\|W_n\|^p\} \leq c(p) \) provided \( X \) is i.i.d.; see e.g. [27, Sect. 2.14.1-2.14.2]. For such \( \mathcal{D} \) and \( p \), de Finetti’s theorem yields \( \sup_n E \{\|W_n\|^p\} \leq c(p) \) even if \( X \) is exchangeable. In fact, conditionally on \( \mu \), the sequence \( X \) is i.i.d. with common distribution \( \mu \). Hence, \( E \{\|W_n\|^p | \mu\} \leq c(p) \) a.s. for all \( n \). By the same argument, if there are \( a, b > 0 \) such that

\[ \sup_n E \{\exp (u \|W_n\|)\} \leq a \exp (b u^2) \] for all \( u > 0 \) if \( X \) is i.i.d., such inequality is still true (with the same \( a \) and \( b \)) if \( X \) is exchangeable.

• A straightforward consequence of the law of iterated logarithm is that convergence in probability can not be replaced by a.s. convergence in Theorem 1-(ii). Take in fact \( r_n = \sqrt{\frac{n}{\log \log n}}; \ G = G^X \) and \( X \) i.i.d. Then, for each \( B \in \mathcal{D} \), the law of iterated logarithm yields

\[
\limsup_n r_n \|\mu_n - a_n\| \geq \limsup_n r_n \{\mu_n(B) - a_n(B)\} = \limsup_n \frac{\sum_{i=1}^n I_B(X_i) - \mu_0(B)}{\sqrt{n \log \log n}} = \frac{2 \mu_0(B)(1 - \mu_0(B))}{a.s.}
\]

• Let \( \mathcal{D} \) be countably determined, \( X \) exchangeable and \( G = G^X \). In view of Theorem 2, for \( r_n \|\mu_n - a_n\| \overset{a.s.}{\to} 0 \), it suffices that \( \sup_n E \{\|W_n\|^p\} < \infty \) and \( r_n (\log(n))^{1/p} \overset{a.s.}{\to} 0 \), for some \( p > 1 \) and \( 0 < c < p \), or that \( E \{\exp (u \|W_n\|)\} \) can be estimated as in (iv) and \( r_n \frac{\log \log n}{n} \overset{a.s.}{\to} 0 \). For instance,

\[
\sqrt{\frac{n}{\log n}} \|\mu_n - a_n\| \overset{a.s.}{\to} 0
\]

whenever \( \sup_n E \{\|W_n\|^p\} < \infty \) for some \( p > 2 \). Another example is provided by Corollary 3. To state it, a definition is to be recalled.

Say that \( \mathcal{D} \) is a Vapnik-Cervonenkis class, or simply a VC-class, if

\[ \text{card} \{ B \cap I : B \in \mathcal{D} \} < 2^n \]

for some integer \( n \geq 1 \) and all subsets \( I \subset S \) with \( \text{card} (I) = n \); see e.g. [14], [17], [21], [27]. In other terms, the power set of \( I \) can not be written as \( \{ B \cap I : B \in \mathcal{D} \} \) for each collection \( I \) of \( n \) points from \( S \). As noted in Section 1, VC-classes are instrumental to empirical processes and statistical learning. If \( S = \mathbb{R}^k \), for instance, \( \mathcal{D} = \{ (-\infty, t_1] \times \ldots \times (-\infty, t_k] : (t_1, \ldots, t_k) \in \mathbb{R}^k \} \), \( \mathcal{D} = \{ \text{half spaces} \} \) and \( \mathcal{D} = \{ \text{closed balls} \} \) are (countably determined) VC-classes.
Corollary 3. Let $\mathcal{D}$ be a countably determined VC-class. If $X$ is exchangeable and $\mathcal{G} = G^X$, then
\[
\limsup_n \frac{n}{\log \log n} \|\mu_n - a_n\| \leq \sqrt{\frac{2}{\sup_{B \in \mathcal{D}} \mu(B)(1 - \mu(B))}} \quad \text{a.s.}
\]

Proof. Just note that, if $X$ is i.i.d. and $\mathcal{D}$ is a countably determined VC-class, then $E\{\exp(u \|W_n\|)\}$ can be estimated as in Theorem 2-(iv) and
\[
\limsup_n \frac{n}{\log \log n} \|\mu_n - a_n\| = \sqrt{\frac{2}{\sup_{B \in \mathcal{D}} \mu_0(B)(1 - \mu_0(B))}} \quad \text{a.s.}
\]
See e.g. [14, Sect. 9.5], [21, Corollary 2.4] and [27, page 246].

We finally give a couple of examples concerning Theorem 1.

Example 4. Let $\mathcal{D} = \mathcal{B}$. If $X$ is i.i.d., then $\|\mu_n - \mu_0\| \overset{a.s.}{\longrightarrow} 0$ if and only if $\mu_0$ is discrete. By de Finetti’s theorem, it follows that $\|\mu_n - \mu\| \overset{a.s.}{\longrightarrow} 0$ whenever $X$ is exchangeable and $\mu$ is a.s. discrete. Thus, under such assumptions and $\mathcal{G} = G^X$, Theorem 1-(i) implies $\|\mu_n - a_n\| \overset{a.s.}{\longrightarrow} 0$. This result has a possible practical interest in Bayesian nonparametrics. As noted in Section 1, in fact, most nonparametric priors are such that $\mu$ is a.s. discrete.

Example 5. Let $S = \mathbb{R}^k$ and $\mathcal{D} = \{\text{closed convex sets}\}$. If $X$ is i.i.d. and $\mu_0 \ll \lambda$, where $\lambda$ is a $\sigma$-finite product measure on $\mathcal{B}$, then $\|\mu_n - \mu_0\| \overset{a.s.}{\longrightarrow} 0$; see [17, page 198]. Applying Theorem 1-(i) again, one obtains $\|\mu_n - a_n\| \overset{a.s.}{\longrightarrow} 0$ provided $X$ is exchangeable, $\mathcal{G} = G^X$ and $\mu \ll \lambda$ a.s. While “morally true”, this argument does not work for $\mathcal{D} = \{\text{Borel convex sets}\}$ since the latter choice of $\mathcal{D}$ is not countably determined.

4.2. The dominated case. In the sequel, as in Section 2, it is convenient to work on the coordinate space. Accordingly, from now on, we let
\[
(\Omega, \mathcal{A}) = (S^\infty, \mathcal{B}^\infty), \quad X_n = n\text{-th coordinate projection,} \quad \mathcal{G} = G^X.
\]
Further, $Q$ is a probability measure on $(\Omega, \mathcal{A})$ and
\[
b_n(\cdot) = Q(X_{n+1} \in \cdot | \mathcal{G}_n)
\]
is the predictive measure under $Q$. We say that $Q$ is a Ferguson-Dirichlet law if
\[
b_n(\cdot) = cQ(X_1 \in \cdot) + n \mu_n(\cdot), \quad Q\text{-a.s. for some constant } c > 0.
\]
If $P \ll Q$, the asymptotic behavior of $\|\mu_n - a_n\|$ under $P$ should be affected by that of $\|\mu_n - b_n\|$ under $Q$. This (rough) idea is realized by the next result.

Theorem 6. (Theorems 1 and 2 of [5]). Suppose $\mathcal{D}$ is countably determined, $X$ is c.i.d., and $P \ll Q$. Then,
\[
\sqrt{n} \|\mu_n - a_n\| \overset{P}{\longrightarrow} 0
\]
whenever $\sqrt{n} \|\mu_n - b_n\| \overset{Q}{\longrightarrow} 0$ and the sequence $(W_n)$ is uniformly integrable under both $P$ and $Q$. In addition,
\[
n \|\mu_n - a_n\| \text{ converges a.s. to a finite limit}
\]
provided \( Q \) is a Ferguson-Dirichlet law, \( \sup_{n} E_{Q}\{\|W_n\|^2\} < \infty \), and
\[
\sup_{n} n \left\{ E_Q(f^2) - E_Q\{E_Q(f|G_n)\} \right\} < \infty \quad \text{where} \quad f = dP/dQ.
\]

To make Theorem 6 effective, the condition \( P \ll Q \) should be given a simple characterization. This happens at least when \( S \) is finite.

As an example, suppose \( S = \{0,1\}, \) an exchangeable \( X \) and a Ferguson-Dirichlet. Then, for all \( n \geq 1 \) and \( x_1, \ldots, x_n \in \{0,1\}, \)
\[
P(\eta = x_1, \ldots, \eta = x_n) = \int_{[0,1]} \theta^k (1 - \theta)^{n-k} \pi_P(d\theta),
\]
where \( k = \sum_{i=1}^{n} x_i \) and \( \pi_P \) are the probability distributions of \( \mu\{1\} \) under \( P \) and \( Q \). Thus, \( P \ll Q \) if and only if \( \pi_P \ll \pi_Q \). In addition, \( \pi_Q \) is known to be a beta distribution. Let \( m \) denote the Lebesgue measure on the Borel \( \sigma \)-field on \([0,1]\). Since any beta distribution has the same null sets as \( m \), one obtains \( P \ll Q \) if and only if \( \pi_p \ll m \). This fact is behind the next result.

**Theorem 7.** (Corollaries 4 and 5 of [5]). Suppose \( S = \{0,1\} \) and \( X \) exchangeable. Then, \( \sqrt{n} \left( \mu_n\{1\} - a_n\{1\} \right) \xrightarrow{P} 0 \) whenever the distribution of \( \mu\{1\} \) is absolutely continuous. Moreover, \( n (\mu_n\{1\} - a_n\{1\}) \) converges a.s. (to a finite limit) provided the distribution of \( \mu\{1\} \) is absolutely continuous with an almost Lipschitz density.

In Theorem 7, a real function \( f \) on \((0,1)\) is said to be *almost Lipschitz* in case \( x \to f(x)x^u(1-x)^v \) is Lipschitz on \((0,1)\) for some reals \( u, v < 1 \).

A consequence of Theorem 7 is to be stressed. For each \( B \in \mathcal{B} \), define
\[
\mathcal{G}_n^B = \sigma(\eta \cap X_1, \ldots, \eta \cap X_n) \quad \text{and} \quad T_n(B) = \sqrt{n} \left\{ a_n(B) - P\{X_{n+1} \in B \mid \mathcal{G}_n^B \} \right\}.
\]
Also, let \( l^\infty(\mathcal{D}) \) be the set of real bounded functions on \( \mathcal{D} \), equipped with uniform distance. In the next result, \( W_n \) is regarded as a random element of \( l^\infty(\mathcal{D}) \) and convergence in distribution is meant in Hoffmann-Jørgensen’s sense; see [27].

**Corollary 8.** Let \( \mathcal{D} \) be countably determined and \( X \) exchangeable. Suppose that
- \( \mu(B) \) has an absolutely continuous distribution for each \( B \in \mathcal{D} \) such that \( 0 < P(X_1 \in B) < 1 \);
- the sequence \( \{\|W_n\|\} \) is uniformly integrable;
- \( W_n \) converges in distribution, in the space \( l^\infty(\mathcal{D}) \), to a tight limit.

Then,
\[
\sqrt{n} \|\mu_n - a_n\| \xrightarrow{P} 0 \iff \quad T_n(B) \xrightarrow{P} 0 \text{ for each } B \in \mathcal{D}.
\]

**Proof.** Let \( U_n(B) = \sqrt{n} \left\{ \mu_n(B) - P\{X_{n+1} \in B \mid \mathcal{G}_n^B \} \right\} \). Then, \( U_n(B) \xrightarrow{P} 0 \) for each \( B \in \mathcal{D} \). In fact, \( U_n(B) = 0 \) a.s. if \( P(X_1 \in B) \in \{0,1\} \). Otherwise, \( U_n(B) \xrightarrow{P} 0 \) follows from Theorem 7, since \( \{\eta \cap X_n\} \) is an exchangeable sequence of indicators and \( \mu(B) \) has an absolutely continuous distribution. Next, suppose
prove that $\parallel T_b \parallel \rightarrow 0$ for each $b \in \mathcal{D}$. Letting $C_n = \sqrt{n} (\mu_n - a_n)$, we have to prove that $\parallel C_n \parallel \rightarrow 0$. Equivalently, regarding $C_n$ as a random element of $l^\infty (\mathcal{D})$, we have to prove that $C_n (B) \overset{P}{\rightarrow} 0$ for fixed $B \in \mathcal{D}$ and the sequence $(C_n)$ is asymptotically tight; see e.g. [27, Section 1.5]. Given $B \in \mathcal{D}$, since both $U_n (B)$ and $T_n (B)$ converge to 0 in probability, then $C_n (B) = U_n (B) - T_n (B) \overset{P}{\rightarrow} 0$. Moreover, since $C_n (B) = E \{ W_n (B) \mid \mathcal{G}_n \}$ a.s., the asymptotic tightness of $(C_n)$ follows from (jj)- (jjj); see [4, Remark 4.4]. Hence, $\parallel C_n \parallel \overset{P}{\rightarrow} 0$. Conversely, if $\parallel C_n \parallel \overset{P}{\rightarrow} 0$, one trivially obtains

$$
\vert T_n (B) \vert = \vert U_n (B) - C_n (B) \vert \leq \vert U_n (B) \vert + \parallel C_n \parallel \overset{P}{\rightarrow} 0 \quad \text{for each } B \in \mathcal{D}.
$$

If $X$ is exchangeable, it frequently happens that $\sup_n E \{ \parallel W_n \parallel^2 \} < \infty$, which in turn implies condition (jj). Similarly, (jjj) is not unusual. As an example, conditions (jj)-(jjj) hold if $S = \mathbb{R}$, $\mathcal{D} = \{(x,t) : t \in \mathbb{R}\}$ and $\mu_0$ is discrete or $P (X_1 = X_2) = 0$; see [4, Theorem 4.5].

Unfortunately, as shown by the next example, $T_n (B)$ may fail to converge to 0 in probability even if $\mu (B)$ has a absolutely continuous distribution. This suggests the following general question. In the exchangeable case, in addition to $\mu_n (B)$, which further information is required to evaluate $a_n (B)$? Or at least, are there reasonable conditions for $T_n (B) \overset{P}{\rightarrow} 0$? Even if intriguing, to our knowledge, such a question does not have a satisfactory answer.

**Example 9.** Let $S = \mathbb{R}$ and $X_n = Y_n Z^{-1}$, where $Y_n$ and $Z$ are independent real random variables, $Y_n \sim N (0, 1)$ for all $n$, and $Z$ has an absolutely continuous distribution supported by $[1, \infty)$. Conditionally on $Z$, the sequence $X = (X_1, X_2, \ldots)$ is i.i.d. with common distribution $N (0, Z^{-2})$. Thus, $X$ is exchangeable and

$$
\mu (B) = P (X_1 \in B \mid Z) = f_B (Z) \quad \text{a.s. for each } B \in \mathcal{B}
$$

where $f_B (z) = (2 \pi)^{-1/2} z \int_B \exp \left(-x^2 / 2\right) dx$ for $z \geq 1$.

Fix $B \in \mathcal{B}$, with $B \subset [1, \infty)$ and $P (X_1 \in B) > 0$, and set $C = \{-x : x \in B\}$. Since $f_B = f_C$, then $\mu (B) = \mu (C)$ and $a_n (B) = a_n (C)$ a.s. Further, $\mu (B)$ has an absolutely continuous distribution, for $f_B$ is differentiable and $f_B' \neq 0$. Nevertheless, one between $T_n (B)$ and $T_n (C)$ does not converge to 0 in probability. Define in fact $g = I_B - I_C$ and $R_n = n^{-1/2} \sum_{i=1}^n g (x_i)$. Since $\mu (g) = \mu (B) - \mu (C) = 0$ a.s., then $R_n$ converges stably to the kernel $N (0, 2 \mu (B))$; see [4, Theorem 3.1]. On the other hand, since $a_n (B) = a_n (C)$ a.s., one obtains

$$
R_n = \sqrt{n} \left\{ \mu_n (B) - \mu_n (C) \right\} = T_n (C) - T_n (B) + \sqrt{n} \left\{ \mu_n (B) - P \{ X_{n+1} \in B \mid \mathcal{G}_n^B \} \right\} - \sqrt{n} \left\{ \mu_n (C) - P \{ X_{n+1} \in C \mid \mathcal{G}_n^C \} \right\} \quad \text{a.s.}
$$

Therefore, if $T_n (B) \overset{P}{\rightarrow} 0$ and $T_n (C) \overset{P}{\rightarrow} 0$, Theorem 7 implies the contradiction $R_n \overset{P}{\rightarrow} 0$.

**4.3. Exchangeable sequences of indicators.** Let $\mathcal{P}$ be the set of all probability measures on $\mathcal{B}$, equipped with the topology of weak convergence. Since $\mu_n$ and $a_n$ are $\mathcal{P}$-valued random variables, we can define their probability distributions on the
Borel σ-field on \( P \), say \( \pi_n(\cdot) = P(\mu_n \in \cdot) \) and \( \pi^*_n(\cdot) = P(a_n \in \cdot) \). Another way to compare \( \mu_n \) and \( a_n \), different from the one adopted so far, is to focus on \( \rho(\pi_n, \pi^*_n) \) where \( \rho \) is a suitable distance between the Borel probability measures on \( P \). In this subsection, we actually take this point of view.

Let \( C \) be the Borel σ-field on \([0, 1]\) and \( \rho \) the bounded Lipschitz metric between probability measures on \( C \). We recall that \( \rho \) is defined as

\[
\rho(\pi, \pi^*) = \sup_{\phi} |\pi(\phi) - \pi^*(\phi)|
\]

where \( \pi \) and \( \pi^* \) are probability measures on \( C \) and sup is over those functions \( \phi \) on \([0, 1]\) such that \( \phi \) is 1-Lipschitz and \(-1 \leq \phi \leq 1\).

Suppose \( S = \{0, 1\} \) and \( X \) exchangeable. Define \( \pi_n(C) = P(\mu_n \{1\} \in C) \) and \( \pi^*_n(C) = P(a_n \{1\} \in C) \) for \( C \in C \). Because of Theorem 7, \( n(\mu_n \{1\} - a_n \{1\}) \) converges a.s. whenever the distribution of \( \mu \{1\} \) is absolutely continuous with a Lipschitz density \( f \). Our last result, inspired by [22], provides a sharp estimate of \( \rho(\pi_n, \pi^*_n) \) under the assumption that \( f \) is Lipschitz (and not only almost Lipschitz).

**Theorem 10.** Suppose \( S = \{0, 1\} \), \( X \) exchangeable, and the distribution of \( \mu \{1\} \) absolutely continuous with a Lipschitz density \( f \). Then,

\[
\rho(\pi_n, \pi^*_n) \leq \frac{1}{n} (1 + \frac{c}{3})
\]

for all \( n \geq 1 \), where \( c \) is the Lipschitz constant of \( f \).

**Proof.** Let \( \overline{X}_n = (1/n) \sum_{i=1}^n X_i \) and \( V = \lim\sup_n \overline{X}_n \). Since the \( X_n \) are indicators, \( \mu_n \{1\} = \overline{X}_n, \mu \{1\} = V \) and \( a_n \{1\} = E(V \mid \mathcal{G}_n) \) a.s.

Take \( Q \) to be the Ferguson-Dirichlet law such that

\[
b_n \{1\} = E_Q(V \mid \mathcal{G}_n) = \frac{1 + n \overline{X}_n}{n + 2}, \quad Q\text{-a.s.}
\]

Then, \(|\overline{X}_n - E_Q(V \mid \mathcal{G}_n)| \leq 1/(n + 2)\). Further, since \( V \) is uniformly distributed on \([0, 1]\) under \( Q \),

\[
P(X_1 = x_1, \ldots, X_n = x_n) = \int_0^1 \theta^k (1 - \theta)^{n-k} f(\theta) \, d\theta
\]

\[
= \int V^k (1 - V)^{n-k} f(V) \, dQ = \int_{\{X_1=x_1,\ldots,X_n=x_n\}} f(V) \, dQ
\]

for all \( n \geq 1 \) and \( x_1, \ldots, x_n \in \{0, 1\} \), where \( k = \sum_{i=1}^n x_i \). Hence, \( f(V) \) is a density of \( P \) with respect to \( Q \). In particular,

\[
E(V \mid \mathcal{G}_n) = \frac{E_Q(V f(V) \mid \mathcal{G}_n)}{E_Q(f(V) \mid \mathcal{G}_n)} \quad \text{a.s.}
\]

Note also that

\[
E_Q((\overline{X}_n - V)^2) = E_Q\{E_Q((\overline{X}_n - V)^2 \mid V) \} = E_Q\{V(1 - V) \} = \frac{1}{6n}.
\]

Next, define \( U_n = f(V) - E_Q(f(V) \mid \mathcal{G}_n) \). Then,

\[
E_Q(X_n U_n \mid \mathcal{G}_n) = \overline{X}_n E_Q(U_n \mid \mathcal{G}_n) = 0, \quad Q\text{-a.s.}
\]
Since $P \ll Q$, then $E_Q\{\mathbb{X}_n \mid G_n\} = 0$ a.s. with respect to $P$ as well. Hence,
\[
\begin{align*}
|\mathbb{X}_n - E(V \mid G_n)| &\leq |\mathbb{X}_n - E_Q(V \mid G_n)| + |E_Q(V \mid G_n) - E(V \mid G_n)| \\
&\leq \frac{1}{n+2} + \left| \frac{E_Q(V \mid G_n) - E_Q\{V f(V) \mid G_n\}}{E_Q\{f(V) \mid G_n\}} \right| \\
&= \frac{1}{n+2} + \left| \frac{E_Q\{V U_n \mid G_n\}}{E_Q\{f(V) \mid G_n\}} \right| \\
&= \frac{1}{n+2} + \left| \frac{E_Q\{(V - \mathbb{X}_n) U_n \mid G_n\}}{E_Q\{f(V) \mid G_n\}} \right| \\
&\leq \frac{1}{n+2} + \frac{E_Q\{|(V - \mathbb{X}_n) U_n \mid G_n\}}{E_Q\{f(V) \mid G_n\}} \quad \text{a.s.}
\end{align*}
\]

Since $f$ is Lipschitz, one also obtains
\[
E_Q(U_n^2) = E_Q\left\{ f(V) - f(\mathbb{X}_n) - E_Q\{f(V) - f(\mathbb{X}_n) \mid G_n\} \right\}^2 \leq 4 E_Q\{(f(V) - f(\mathbb{X}_n))^2\} \leq 4 c^2 E_Q\{(\mathbb{X}_n - V)^2\}.
\]

We are finally in a position to estimate $\rho(\pi_n, \pi_n^*)$. In fact, if $\phi$ is a function on $[0, 1]$, with $\phi$ 1-Lipschitz and $-1 \leq \phi \leq 1$, then
\[
|\pi_n(\phi) - \pi_n^*(\phi)| = \left| E\{\phi(\mathbb{X}_n)\} - E\{\phi(E(V \mid G_n))\} \right| \leq E|\mathbb{X}_n - E(V \mid G_n)|
\]
\[
\leq \frac{1}{n+2} + E\left\{ \frac{E_Q(|\mathbb{X}_n - V) U_n| \mid G_n\}}{E_Q\{f(V) \mid G_n\}} \right\}
\]
\[
= \frac{1}{n+2} + E_Q\{ f(V) \frac{E_Q(|\mathbb{X}_n - V) U_n| \mid G_n\}}{E_Q\{f(V) \mid G_n\}} \right\}
\]
\[
= \frac{1}{n+2} + E_Q\{ |\mathbb{X}_n - V| U_n\}
\]
\[
\leq \frac{1}{n+2} + \sqrt{E_Q\{(|\mathbb{X}_n - V|^2\} E_Q(U_n^2)}
\]
\[
\leq \frac{1}{n+2} + 2 c E_Q\{(|\mathbb{X}_n - V|^2\}
\]
\[
= \frac{1}{n+2} + \frac{c}{3n} < \frac{1}{n} (1 + \frac{c}{3}).
\]

\[\square\]

The rate provided by Theorem 10 cannot be improved. Take in fact $\phi(x) = x^2/2$ and suppose $P$ a Ferguson-Dirichlet law with $a_n\{1\} = \frac{1+n \mu_n(1)}{n+2}$ a.s. Then, since $\mu\{1\}$ is uniformly distributed on $[0, 1]$, one obtains
\[
2 (n+2) \rho(\pi_n, \pi_n^*) \geq 2 (n+2) |\pi_n(\phi) - \pi_n^*(\phi)|
\]
\[
= (n+2) \left\{ E(\mu_n\{1\}^2) - E(a_n\{1\}^2) \right\}
\]
\[
= (n+2) E(\mu_n\{1\}^2) - \frac{1+n^2 E(\mu_n\{1\}^2) + 2 n E(\mu_n\{1\})}{n+2}
\]
\[
= \frac{4 (n+1) E(\mu_n\{1\}^2) - 2 n E(\mu_n\{1\}) - 1}{n+2} \rightarrow 4 E(\mu\{1\}^2) - 2 E(\mu\{1\}) = \frac{1}{3}.
\]
REFERENCES
