On consistency of Bayesian parameter estimators for a class of ergodic Markov models

A.I. Nurieva
•
National Research University
Higher School of Economics
ai_nurieva@mail.ru

A.Yu. Veretennikov
•
Institute for Information
Transmission Problems
ayv@iitp.ru

Abstract

The consistency of the Bayesian estimation of a parameter is shown for a class of ergodic discrete Markov chains. J.L. Doob's method was used, offered earlier for the i.i.d. situation. The result may be useful in the reliability theory for models with unknown parameters, in the risk management in financial mathematics, and in other applications.

Keywords: Bayesian estimator; consistency; ergodic Markov chain

MSC2020: 62F12; 62F15; 62M05

1. Introduction

Parameter estimation plays a significant and in some cases possibly even a crucial role in quite a few applications such as the reliability theory for models with unknown parameters, see [4, chapter 3], in the Extreme Values theory for Markov processes, in the risk management in financial mathematics, et al. In the asymptotic sense, one of the basic desirable properties of any estimator in the long run is its consistency, weak or strong, as it shows that the estimation is close to the "true" parameter if the classical setting is accepted. Similarly, in the Bayesian setting consistency means literally the same – convergence to the sample value of the parameter, even though there is no such thing as a "true parameter value" because it is to be sampled from the prior distribution. Also, as it is well-known, Bayesian estimators often work well in the classical setting, too, assuming some fictitious prior distribution for the parameter is chosen.

In this paper the problem of strong consistency is tackled for a certain class of Markov models in the Bayesian setting, and, as was already mentioned, in the classical situation with a fixed nonrandom "true" parameter value. Assume that there is a family of distributions $\{\mathbb{P}^\theta\}$ parameterised by some variable $\theta \in \Theta$, where $\Theta \subset R^m$ is a given parametric space. Any estimator is a measurable function of the observations, or, a bit more generally, a mapping from the space of outcomes Ω, say, to the space $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$ which is Borel measurable with respect to the sigma-algebra of the observations $\mathcal{F}^X$; here $\mathcal{B}(\mathbb{R}^m))$ is the Borel sigma-algebra in $\mathbb{R}^m$.

In the Bayesian setting it is assumed that there is some prior distribution for θ on the set Θ; the latter is usually a topological space, and in this paper, it will be assumed that Θ is a domain in $\mathbb{R}^m$ which is not necessarily bounded. S.N. Bernstein and R. von Mises were the first to establish consistency and the first steps towards the asymptotic normality of the Bayesian estimator for some particular i.i.d. cases, see [1, Chapter IV, p.271], [18, pp. 188-192]. The general theory about asymptotic normality was developed later by Le Cam [13] and Ibragimov and Khasmisnky [5]; for more recent results see, for example, [12], [15]. Another direction related to the problem was asymptotic singularity of measures for large observation samples based on martingale theory and developed in [6, 7, 8, 14, 17], et al. Naturally asymptotic normality requires more restrictive assumptions. On the other hand, "just" consistency may often be used for constructions of more efficient estimations by certain modifications. Also, in a situation where the conditions for asymptotic normality are not met, it may be even more desirable to know whether the applied estimator is consistent. Hence, it makes sense to separate the studies of sufficiency conditions for both properties, asymptotic normality and consistency.

In this paper the approach offered for the i.i.d. observations in [3] is used, adjusted for a class of markovian models. An important point in [3] was a Strong Law of Large Numbers for the sample distribution functions (d.f. in what follows) as the number of observations tends to infinity. Also essential was an assumption that theoretical d.f. are different for different parameters. In this paper discrete densities on a finite or countable state space are used. This restriction looks not crucial and likely may be relaxed. At the level of ideas, the most close to this study is the paper [17], where the earlier basic results from [6, 7, 8] are applied precisely to the problem of parameter estimators' consistency. However, formally conditions in for this property [17] and in what follows are different. Also, in a way, this paper is based on a more simple background than that in [6, 7, 8, 17].

The paper consists of this Introduction, The setting, Auxiliary lemmata, Main result (theorem 4), and Proof of theorem 4.

2. The setting

Let$\{X_t\}$ be a homogeneous Markov chain (MC) in discrete time $T = \{0, 1, \ldots\}$ with a finite or countable (denumerable) state space $\mathcal{X} \subset \mathbb{R}^1$ (it will be clear in what follows why it is convenient to work on $\mathbb{R}^1$: although it is not a restriction, but it may be desirable that the elements of the state space are linearly ordered). The transition probabilities are denoted by $p_{ij}(s,t) = \mathbb{P}(X_t = j | X_s = i)) = \mathbb{P}(X_{t-s} = j | X_0 = i)) = p_{ij}(t-s)$ for $s \le t$, and let $\mathcal{P}(t) = (p_{ij}(t))$ be the transition probability matrix over time t; furthermore, they will all depend on a parameter θ. The notion of ergodicity of a MC is not uniquely determined in the literature; in the present paper we understand it as follows.

Definition 1. *A homogeneous MC $(X_n, n = 0, 1, \ldots)$ is called ergodic if there exists a limiting invariant probability measure μ which does not depend on the initial distribution – say, μ_0 – and to which there is a convergence in total variation for each μ_0:*

$$\lim_{t\to\infty} \|p_{\mu_0,\cdot}(t) - \mu_\cdot\|_{TV} = 0, \tag{1}$$

where $p_{\mu_0,j}(t) = \mathbb{P}_{\mu_0}(X_t = j)$. Recall that the total variation metric, or distance is given by the formula

$$\|\mu - \nu\|_{TV} := 2 \sup_{A\in\mathcal{F}(\mathcal{X})} (\mu(A) - \nu(A)).$$

As it was said, the transition probabilities depend on a parameter and the problem under consideration is estimation of this parameter given observations on the time interval $[1, n]$ where $n \to \infty$. It is assumed that $\theta \in \Theta \subset \mathbb{R}^m$; Θ is a domain, not necessarily bounded. Naturally, a stationary measure, generally speaking, also depends on θ: denote it from now on by $\mu^\theta(dx)$ and note that under the assumption of convergence (1) it is necessarily unique. We will need the extended process $Y_n = (X_n, X_{n+1})$ which is also a MC on the state space $\mathcal{X} \times \mathcal{X}$. The symbol $\mu^\theta(dx, dx')$ will denote the stationary measure for the MC (Y_n); it is easy to see that such an invariant measure does exist. Assume that the functions $p^\theta(\cdot, \cdot)$ are Borel measurable with respect to the variable θ. Then due to the ergodicity (see (1)) the invariant probabilities are also Borel measurable in θ. Following Doob's approach, *suppose* that a (weak) Law of Large Numbers (LLN) holds true for the MC (X_n) with respect to the corresponding measure $\mathbb{P}^\theta$, for each θ . In this case, LLN is also valid for the MC (Y_n), where $Y_n := (X_n, X_{n+1})$. It is easy to see that these two conditions – LLN for the MC (X_n) and for the MC (Y_n) – are equivalent. Hence, the following assumption will be accepted in what follows.

Assumption 2. *It is assumed that for each $\theta \in \Theta$ and any measurable A, B a convergence holds true,*

$$\left| \frac{1}{T} \sum_{s=0}^{T-1} 1(X_s \in A, X_{s+1} \in B) - \mu^\theta(A \times B) \right| \overset{\mathbb{P}^\theta}{\to} 0, \quad T \to \infty.$$

This assumption is equivalent also to the condition

$$\left|\frac{1}{T}\sum_{s=0}^{T-1} g(Y_s) - \int g(y)\mu^\theta(dy)\right| \overset{\mathbb{P}^\theta}{\to} 0, \quad T \to \infty,$$

for any bounded measurable function $g(y)$, where $y = (x, x')$.

Let us collect the comments made earlier in the form of a proposition.

Proposition 3. *Under the assumptions made above the following statements hold:*

1. *If (X_n) is a homogeneous MC then Y_n is also a homogeneous MC.*
2. *If the MC (X_n) is ergodic then the MC (Y_n) is also ergodic, and vice versa.*

Note that, as usual, all sigma-algebras in the text are regarded as completed with respect to the corresponding probability measures.

3. Main result

The Bayesian setting assumes that the parameter θ is random; let it have a prior probability distribution $\mathbb{Q}$ on Θ. Recall that here Θ is a domain in $\mathbb{R}^m$, not necessarily bounded. It is assumed that

$$\mathbb{E}\theta < \infty. \tag{2}$$

Any estimator of the parameter given observations is represented by some Borel measurable function $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$. As it is well-known (cf., for example, [2, chapter 19]), there exists a Borel measurable function ϕ_n such that the Bayesian estimator reads,

$$\mathbb{E}(\theta|X_1, \ldots, X_n) \overset{\mathbb{P}^\theta\text{-a.s.}}{=} \phi_n(X_1, \ldots, X_n).$$

So, the statistic $\hat{\theta}_n := \phi_n(X_1, \ldots, X_n) = \mathbb{E}(\theta|X_1, \ldots, X_n)$ is necessarily $(\mathcal{F}^X_N, \mathcal{B}(\mathbb{R}^m))$-measurable; hence, also $(\mathcal{F}^X_\infty, \mathcal{B}(\mathbb{R}^m))$-measurable, and ϕ_n is measurable with respect to the pair of σ-algebras $(\mathcal{B}(\mathcal{X})^n, \mathcal{B}(\mathbb{R}^m)), \forall n \in \mathbb{N}$, where $\mathcal{B}(\mathcal{X})$ is the set of all subsets of the state space $\mathcal{X}$, that is, $\mathcal{B}(\mathcal{X}) = 2^{\mathcal{X}}$. Recall that a pointwise limit of measurable functions is also measurable.

Theorem 4. *Let the following conditions be satisfied:*

1. *Transition probability matrices of the MC (X_n) for different values of θ are different, that is, for any $\theta \neq \theta'$ there exist i, j such that $p^\theta_{ij} \neq p^{\theta'}_{ij}$.*
2. *Let MC (Y_n) be ergodic for each θ under the measure $\mathbb{P}^\theta$ in the sense of the definition (1), and let the (weak) LLN hold for the process Y for each θ in the sense of the assumption (2). Then there is a convergence*

$$\hat{\theta}_n \to \theta, \quad n \to \infty, \quad \mathbb{P} - a.s. \tag{3}$$

Here, as usual in the Bayesian setting,

$$\mathbb{P}(d\theta, d\omega) = \mathbb{Q}(d\theta)\mathbb{P}^\theta(d\omega).$$

Remark 5. *Recall that similar results under different conditions were established in [17, theorems 1-2]. Formally, those conditions in [17] may be applicable, or not applicable in our situation because the assumption of the absolute continuity for the projection measures on the sigma-algebra $\mathcal{F}^X_n$ for any two values of the parameter is not assumed, see [17, Theorem 1, condition (C)] and [17, Theorem 2, condition (b)]. In markovian examples in [7, §13] a similar condition to [17, Theorem 1, condition (C)] was assumed as well, see theorem 22, condition (b). In the present paper such a condition is neither assumed, nor it follows from the other assumptions. Intuitively, the lack of continuity should only help consistency; nevertheless, even if so, it apparently does require some calculus. In any case, the proof of the theorem 4 in what follows does not distinguish between the cases tackled in [17] and the cases not covered by this cited paper.*

Remark 6. *As in the setting of Doob in [3], this result may also be used in the classical setting where θ is not random and there exists a unique "true" parameter value. For that, an artificial prior density should be introduced on Θ which must be everywhere positive. Then, as in [3], the analogous assertion will hold true about an almost sure convergence of the artificial Bayesian estimator under the product measure on $\Theta \times \mathcal{X}^\infty$.*

In particular, what is usually highlighted about Bernstein and von Mises theorem is that if the measure $\mathbb{Q}$ has a densty $q(\theta)$ which is everywhere positive, then convergence of the Bayesian estimator towards θ will take place almost everywhere in Θ with respect to the Lebesgue measure. Actually, it suffices for this property that the measure $\mathbb{Q}$ were absolutely continuous with respect to the latter. However, in either case there is no way to know for which particular values of θ this convergence is valid and for which maybe not; it may only be claimed that the set of "bad" values of θ with no convergence has measure zero.

4. Auxiliary results

Let us define the sample distribution function

$$F_N(x,x') := \frac{1}{N}\sum_{t=0}^{N-1} 1(X_t \le x, X_{t+1} \le x').$$

Denote by $\mathcal{S} = \{F(x,x'), x, x' \in \mathbb{R}\}$ the space of all functions of two variables (x,x') with the following properties:

1. $0 \le F(x,x') \le 1$ for each $x, x' \in \mathbb{R}$.
2. If $x \le z$, $x' \le z'$, then $F(x,x') \le F(z,z')$ (monotonicity).
3. For each $x, x' \in R$
$$\lim_{z\downarrow x, z'\downarrow x'} F(z,z') = F(x,x').$$
4. For each $x, x' \in \mathbb{R}$ there exists a limit
$$\lim_{z\uparrow x, z'\uparrow x'} F(z,z') =: F(x,x')_-.$$
(NB: Actually, the latter notation will not be used in what follows; it is just an analogue of the one-dimensional property of "làg" – possessing "limites à gauche" – for the one-dimensional case. Respectively, the property 3 is the analogue of the "càd" – being "continue à droite" for a function of one variable.)
5. $$\lim_{z\uparrow+\infty, z'\uparrow+\infty} F(z,z') = 1.$$
6. $$\lim_{z\downarrow-\infty, z'\downarrow-\infty} F(z,z') = 0.$$

In fact, in the situation under the consideration we deal with some proper subset of all distribution functions of two variables, because all corresponding measures on $\mathbb{R}^2$ have atoms in our setting. However, all we need is that this more general space of distribution functions with a certain metric is a Polish space, and this will be guaranteed by proposition 16 in what follows.

Denote by $\Sigma(\mathcal{S})$ the sigma-algebra on $\mathcal{S}$ generated by all finite cylinders, i.e.,

$$\Sigma(\mathcal{S}) := \sigma(F \in \mathcal{S} : F(x_1,x_1') \le a_1, \ldots F(x_n,x_n') \le a_n))$$

for any $(x_1,x_1'), \ldots, (x_n,x_n') \in \mathbb{R}^2$ and $a_1, \ldots, a_n \in [0,1]$.

Note that the distribution function of any two-dimensional random vector belongs to the space $\mathcal{S}$, and all sample d.f. F_N belong to this space, too.

Lemma 7. *The function $F_N : \Omega \mapsto \mathbb{R}$ is a measurable map with respect to the corresponding pair of sigma-algebras $(\mathcal{F}^X_N; \Sigma(S))$.*

Proof. The proof is elementary and is shown here only for the convenience of the reader. Indeed, for any couple (x, x') the mapping

$$F_N(x,x') := \frac{1}{N}\sum_{t=0}^{N-1} 1(X_t \le x, X_{t+1} \le x')$$

is measurable as a function of ω, as a finite sum of random variables (indicators), which may be expressed by the relation

$$(\omega : \ F_N(x,x') \le a) \in \mathcal{F}^X_N$$

for any $a \in R$. Then, for any finite sets of $(x_1, x'_1), \ldots, (x_n, x'_n)$ and $a_1, \ldots, a_n$ we have,

$$(\omega : \ F_N(x_1,x'_1) \le a_1, \ldots F_N(x_n,x'_n) \le a_n)) \in \mathcal{F}^X_N,$$

by the definition of what is a sigma-algebra. Therefore, $F_N(\cdot,\cdot)$ as a function of ω is, indeed, $(\mathcal{F}^X_N; \Sigma(S))$-measurable, as required. ■

In the next lemma it is assumed that the distribution of $Y_0 = (X_0, X_1)$ is invariant. In this case its distribution function is denoted by $\hat{F}^\theta(x,x')$; recall that due to ergodicity it is unique. It may be presented by the formula

$$\hat{F}^\theta(x,x') = \hat{F}^\theta(x) p^\theta_{x,x'}, \tag{4}$$

where, in turn, $\hat{F}^\theta(x)$ is the (unique) invariant distribution function of the MC X_n with respect to the probability measure P^θ, which is simultaneously the limiting distribution function for the (X_n).

Lemma 8. *Under the assumption that all transition probabilities p^θ_{ij}, $i, j \in \mathcal{X}$ are Borel measurable in θ, the invariant distribution function $\hat{F}^\theta(x,x')$ is Borel measurable in θ for each pair (x,x').*

Proof. Indeed, invariant probabilities $p^\theta_{inv}(i)$, $i \in \mathcal{X}$ are measurable in θ as limits of measurable n-step transition probabilities. So, the "double" invariant probabilities $p^\theta_{inv}(i)p^\theta(ij)$, $i, j \in \mathcal{X}$ also have the same property. Hence, the theoretical d.f.

$$\mathbb{P}^\theta_{x_0}(X_n \le x, X_{n+1} \le x') = \sum_{i \le x} p^\theta_{x_0,i}(n) \sum_{j \le x'} p^\theta_{i,j}$$

is clearly Borel measurable in θ, too. So is its limit at $n \to \infty$ which equals $\hat{F}^\theta(x,x')$, as required. ■

Let us recall the Lèvy–Doob theorem on convergence of conditional expectations.

Proposition 9. *(see, e.g., [11, Theorem 4.3.10]) Let $E|\xi| < \infty$ and let $\mathcal{F}_n, n = 0, 1, \ldots$ be an increasing sequence of σ-algebras, $\mathcal{F}_n \subset \mathcal{F}_{n+1}$, and let $\mathcal{F}_\infty$ be the minimal σ-algebra which contans all $\mathcal{F}_n$, that is, $\mathcal{F}_\infty = \bigvee_n \mathcal{F}_n$ (that is the minimal sigma-algebra generated by all $\mathcal{F}_n$). Then*

$$\lim_{n\to\infty} \mathbb{E}(\xi|\mathcal{F}_n) = \mathbb{E}(\xi|\mathcal{F}_\infty), \quad a.s.$$

and

$$\lim_{n\to\infty} \mathbb{E}|\mathbb{E}(\xi|\mathcal{F}_\infty) - \mathbb{E}(\xi|\mathcal{F}_n)| = 0.$$

In our setting due to the proposition 9 we have,

$$\lim_{n\to\infty} \mathbb{E}(\theta|X_1, \ldots, X_n) = \lim_{n\to\infty} \phi_n(X) = \mathbb{E}(\theta|\mathcal{F}^X_\infty) \quad \text{a.s.}$$

This implies that the limit in the left hand side in the latter double equality is $\mathcal{F}^X_\infty$-measurable.

Lemma 10. *Assume that the transition probability matrices are different for different parameter values, that is, $\theta \neq \theta'$ implies that there exist i, j such that $p^{\theta}_{ij} \neq p^{\theta}_{ij}$. Then the mapping $\theta \mapsto \hat{F}^{\theta}(j, j'), j, j' \in \mathcal{X}$ is one-to-one. Moreover, the mapping*

$$G: \ \theta \mapsto \hat{F}^{\theta}(x, x'), \quad x, x' \in \mathbb{R}, \tag{5}$$

is also one-to-one.

Proof. The proof follows from the formula (4). Indeed, if for $\theta \neq \theta'$ the one-dimensional invariant distribution functions $\hat{F}_{\theta}(\cdot)$ are different, then two-dimensional are different, too. If for some pair $\theta \neq \theta'$ the one-dimensional d.f. coincide, $\hat{F}^{\theta}(\cdot) = \hat{F}^{\theta'}(\cdot)$, then the two-dimensional one are yet different due to the formula (4) and by virtue of the distinguishability assumption of transition probabilities for different parameter values. The same property for the mapping G follows straightforwardly. ■

Further, due to the assumed LLN the following convergence of relative frequencies holds,

$$\frac{1}{n}\sum_{t=0}^{n-1} 1(X_t \leq j) \overset{\mathbb{P}^{\theta}}{\to} \hat{F}^{\theta}(j) = \mathbb{E}^{\theta}_{inv} 1(X_0 \leq j), \quad n \to \infty,$$

where $\mathbb{E}^{\theta}_{inv}$ is expectation with respect to the corresponding invariant measure. A similar convergence holds true for two-dimensional relative frequencies,

$$\frac{1}{n}\sum_{t=0}^{n-1} 1(X_t \leq j, X_{t+1} \leq j') \overset{P^{\theta}}{\to} \hat{F}^{\theta}(j, j') = \mathbb{E}^{\theta}_{inv} 1(X_0 \leq j, X_1 \leq j'), \quad n \to \infty.$$

Since two-dimensional invariant d.f. $\hat{F}^{\theta}(\cdot, \cdot)$ are different for any two different parameter values, the value θ is uniquely determined by the infinite trajectory of observations $X = (X_n, n = 1, \ldots)$. In other words, the mapping $\theta \mapsto \hat{F}^{\theta}(\cdot, \cdot)$ is one-to-one. This mapping is measurable due to the LLN and because the limit of measurable mappings is also measurable. Moreover, as it follows from proposition 13 (see below; it is not linked to this lemma), the inverse mapping is also measurable.

Let us recall some further definitions; it is necessary because one of them is not standard in most of mathematics areas (see definition 12 in what follows).

Definition 11. *Borel measurable sets in a Polish (& more generally, in any topological) space $\mathbb{X}$ are the sets of the minimal σ-algebra $\mathcal{B}(\mathbb{X})$ of subsets in $\mathbb{X}$ which contains all open subsets in $\mathbb{X}$.*

Definition 12. *Let X, Y be Borel measurable sets in Polish spaces $\mathbb{X}, \mathbb{Y}$, respectively. The mapping $f : X \to Y$ is called:*

1. *Borel iff its graph $\Gamma_f = \{(x, y) : x \in X, f(x) = y\}$ is a Borel set in the space $\mathbb{X} \times \mathbb{Y}$;*
2. *B-measurable iff the image of any Borel set from the space Y under the inverse mapping f^{-1} is a Borel set in $\mathbb{X}$.*

Note that the "usual" definition of a Borel function in the majority of areas of mathematics coincides with 10.2.

The next result may be found in [10, Theorem 2.4.3] (we only state the part of this theorem which will be used in what follows).

Proposition 13 ([10, Theorem 2.4.3]). *Let X, Y be Borel sets in Polish spaces and $f : X \to Y$ be some mapping. Then:*

1. *If f is Borel measurable then the images of all Borel sets from Y under the inverse mapping f^{-1} are also Borel, so that the mapping f is B-measurable;*
2. *Vice versa, if f is B-measurable then it is a Borel function.*

Corollary 14. *If the mapping f is Borel measurable and one-to-one, then its inverse f^{-1} is B-measurable and, hence, Borel one in the sense of definition 12.*

In order to apply proposition 13 in the proof of our main result in the next section, let us show that both proposition and its corollary 14 are applicable to the mapping G (see (5)).

Lemma 15. *Under the assumptions of theorem 4 the mapping G^{-1} is Borel and B-measurable.*

Proof. Firstly, the mapping $G : \theta \mapsto \hat{F}^{\theta}(\cdot, \cdot)$ is B-measurable in the sense of the definition 12. Indeed, the element $\hat{F}^{\theta}(\cdot, \cdot)$ is a limit in probability $\mathbb{P}^{\theta}$ of the sequence of functions $\mathbb{E}^{\theta} F_N(\cdot, \cdot)$, which are all B-measurable in θ; therefore, so is their limit.

Secondly, according to lemma 10, the mapping G is one-to one; hence, so is its inverse is G^{-1}. The claim of lemma 15 now follows from corollary 14. ■

Further, it is desirable that the parametric space Θ and the space of invariant distribution functions were complete and separable metric spaces. It is trivial with $\Theta \subset \mathbb{R}^m$ with the Euclidean metric; for the space of "double" distribution functions a suitable matric should be chosen which is, of course, not unique. To each distribution function there correspond a probability distribution on $\mathbb{R}^2$. Let us accept that the distance between two distribution functions is defined as a distance between their corresponding measures. Let us choose Prokhorov's metric $d_p(\nu_1, \nu_2)$ for them: if α-neighbourhood of a set $A \subset \mathbb{R}^2$ is denoted by

$$A_\alpha := \{\bar{a} := (a_1, a_2) \in \mathbb{R}^2 : \ d(\bar{a}, A) < \alpha\}, \text{ if } A \neq \emptyset, \ \emptyset_\alpha = \emptyset \quad \forall \alpha > 0,$$

then the distance between probability measures ν_1, ν_2 on $\mathbb{R}^2$ is defined by the formula

$$d_p(\nu_1, \nu_2) := \inf\{\alpha > 0 : \nu_2(A) \le \nu_1(A_\alpha) + \alpha \ \& \ \nu_1(A) \le \nu_2(A_\alpha) + \alpha, \ \forall A \in \mathcal{B}(S)\}.$$

The same formula provides the distance between two distribution functions, namely, as a distance $d_p(\cdot, \cdot)$ between the corresponding measures on $\mathbb{R}^2$.

Proposition 16 ([16, Lemma 1.4]). *Let a metric space be complete and separable. Then the space of probability measures on it with the Prokhorov metric is also complete and separable.*

5. Proof of theorem 4

Proof. By virtue of Lèvy–Doob's theorem (see proposition 9) we have,

$$\hat{\theta}_n \equiv \mathbb{E}(\theta | X_1, \ldots, X_n) = \mathbb{E}(\theta | \mathcal{F}_n^X) \to \mathbb{E}(\theta | \mathcal{F}_\infty^X) =: \hat{\theta}_\infty, \quad n \to \infty \quad P\text{-a.s..} \tag{6}$$

Due to its definition, the random variable $\hat{\theta}_\infty$ is $\mathcal{F}_\infty^X$-measurable; being a conditional expectation, it is a Bayesian estimator of θ constructed upon the infinite sequence of observations $X_1, X_2, \ldots$ For the proof of the theorem, it suffices to establish the equality

$$\hat{\theta}_\infty \overset{\mathbb{P}-\text{a.s.}}{=} \theta. \tag{7}$$

The basis for thsi equality is the empirical fact that θ is uniquely deterined by the infinite sequence of observations due to the assumed LLN and because of the one-to-one correspondence between θ and the invariant distribution of the pair (X_0, X_1). Let us provide more rigorous considerations related, in particular, to the measurability.

By virtue of the LLN assumptions, we have

$$\mathcal{F}_{N-1}^X \ni F_N(x) = \frac{1}{N} \sum_{i=0}^{N-1} I(X_i \le x) \overset{\mathbb{P}^\theta}{\to} \hat{F}^\theta(x) = \mathbb{E}_{inv}^\theta I(X_0 \le x),$$

and also

$$\mathcal{F}_N^X \ni F_N(x, x') = \frac{1}{N} \sum_{i=0}^{N-1} I(X_i \le x, X_{i+1} \le x') \overset{\mathbb{P}^\theta}{\to} \hat{F}^\theta(x, x') = \mathbb{E}_{inv}^\theta I(X_0 \le x, X_1 \le x').$$

The random variable $F_N(x,x')$ is $(\mathcal{F}_N^X, \mathcal{B}(\mathbb{R}^2))$-measurable for any pair $(x,x') \in \mathbb{R}^2$. According to lemma 7, the mapping $F_N(\cdot,\cdot)$ is $(\mathcal{F}_N^X, \Sigma(\mathcal{S}))$-measurable, hence, it is a random variable in the space of distribution functions.

Now, according to lemma 10 the following equality holds true,

$$\theta \overset{\mathbb{P}-\text{a.s.}}{=} G^{-1}(\underbrace{\hat{F}^\theta(\cdot,\cdot)}_{\in \mathcal{F}_\infty^X}).$$

By virtue of lemma 15, the mapping G^{-1} is Borel and B-measurable. Therefore,

$$G^{-1}(\hat{F}^\theta(\cdot,\cdot)) \in \mathcal{F}_\infty^X.$$

Therefore, by virtue of (6),

$$\hat{\theta}_n \to \mathbb{E}(\theta|\mathcal{F}_\infty^X) \overset{\mathbb{P}-\text{a.s.}}{=} \mathbb{E}(G^{-1}(\hat{F}^\theta(\cdot,\cdot))|\mathcal{F}_\infty^X) \overset{\mathbb{P}-\text{a.s.}}{=} G^{-1}(\hat{F}^\theta(\cdot,\cdot)) \overset{\mathbb{P}-\text{a.s.}}{=} \theta.$$

This means that (7) holds true. implies the desired convergence (3). Theorem 4 is proved. ■

Acknowledgements

For both authors this study is supported by the Russian Foundation for Basic Research, grant 20-01-00575a. Theorem 4 and lemma 8 are established by the first author; lemmata 7 and 10 are proved by the second author.

References

[1] S.N. Bernstein, The theory of probability, 1927 (In Russian). (Chapter IV, p.271)

[2] A.A. Borovkov, Mathematical statistics, CRC Press, 1999 (The first edition in Russian 1984).

[3] J.L. Doob, Application of the theory of martingale, in: B. Locker, Doob at Lyon, Electr. Journal Of History of Probability and Statistics, June 2009, 1-28. https://eudml.org/doc/130498

[4] B.V. Gnedenko, Yu.K. Belyaev, A.D. Solovyev, Mathematical methods of reliability theory, Academic Press, 2014 (The first edition in Russian 1965, Eng. translation: AP 1969).

[5] I.A. Ibragimov, R.Z. Khasminsky, Statistical Estimation. Asymptotic Theory. Springer, 1981.

[6] Yu.M. Kabanov, R.Sh. Liptser, A.N. Shiryaev, Absolute continuity and singularity of locally absolutely continuous probability distributions. I, Math. USSR-Sb., 35:5 (1979), 631-680 https://doi.org/10.1070/SM1979v035n05ABEH001615

[7] Yu.M. Kabanov, R.Sh. Liptser, A.N. Shiryaev, Absolute continuity and singularity of locally absolutely continuous probability distributions. II Sb. Math. 1980, 36(1), 31-58 https://doi.org/10.1070/SM1980v036n01ABEH001760

[8] Yu.M. Kabanov, R.Sh. Liptser, A.N. Shiryaev, On the question of absolute continuity and singularity of probability measures, Math. USSR-Sb., 33:2 (1977), 203-221. https://doi.org/10.1070/SM1977v033n02ABEH002421

[9] S. Kakutani, On equivalence of infinite product measures, Ann. Math., 49, No 1 (1948), 214-224 https://doi.org/10.2307/1969123

[10] V.G. Kanovei, V.A. Lyubetsky, The modern set theory: Borel and projective sets, MCNMO, Moscow, 2010 (In Russian) https://elibrary.ru/item.asp?id=19462694

[11] N.V. Krylov, Introduction to the theory of random processes, AMS, Providece, Rhode Island, 2002.

[12] Yu.A. Kutoyants, Statistical Inference for Ergodic Diffusion Processes, London, 2004. DOI: 10.1007/978-1-4471-3866-2

[13] L. Le Cam, On some asymptotic properties of maximum likelihood estimates and related Bayes estimates, Univ. California Publ. Stat., 1953, 1, 277-330. http://mi.mathnet.ru/mat131

[14] R.Sh. Liptser, F. Pukelsheim, A.N. Shiryaev, Necessary and sufficient conditions for contiguity and entire asymptotic separation of probability measures. Russ. Math. Surv. 37, No. 6, 107-136 (1982). https://doi.org/10.1070/RM1982v037n06ABEH004025

[15] M. Panov, V. Spokoiny, Finite sample Bernstein – von Mises theorem for semiparametric problems, Bayesian Analysis, 2015, 10 (3), 665-710. DOI: 10.1214/14-BA926

[16] Yu.V. Prokhorov, Convergence of Random Processes and Limit Theorems in Probability Theory, Theory Probab. Appl. 1(2) (1956), 157-214. https://doi.org/10.1137/1101016

[17] A.I. Yashin, On Consistency of Bayesian Parameter Estimation, Problems Inform. Transmission, 17:1 (1981), 42-49. Yashin, A. I. http://mi.mathnet.ru/ppi1381

[18] R. von Mises, Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik. Leipzig & Wien, Franz Deuticke, 1931. (pp. 188-192)