Consistency of the Bayes Estimator of a Regression Curve

A.G. Nogales

Dpto. de Matemáticas, Universidad de Extremadura
Avda. de Elvas, s/n, 06006–Badajoz, SPAIN.
e-mail: nogales@unex.es

Abstract. Strong consistency of the Bayes estimator of a regression curve for the $L^1$-squared loss function is proved. It is also shown the convergence to 0 of the Bayes risk of this estimator both for the $L^1$ and $L^1$-squared loss functions. The Bayes estimator of a regression curve is the regression curve with respect to the posterior predictive distribution.

AMS Subject Class. (2020): 62F15, 62G08, 62G20.

Key words and phrases: Bayesian estimation of a regression curve, posterior predictive distribution.

1. Introduction.

Nogales (2022a) addresses the problem of estimation of a density from a Bayesian point of view and, under mild conditions, shows that the posterior predictive density is the Bayes estimator for the $L^1$-squared loss function, and Nogales (2022c) shows the strong consistency of this estimator. Nogales (2022b) deals, among others, with the problem of estimation of a regression curve and proves that the regression curve with respect to the posterior predictive distribution is the Bayes estimator and, here, we wonder about its consistency. This is the main goal of the paper and the Theorem 1 below answers the question in the affirmative.

The interested reader can find in the papers mentioned above, and the references therein, more information on the problems of estimating a density or a regression curve, both from a frequentist and a Bayesian perspective, or about the usefulness of the posterior predictive distribution in Bayesian Inference and its calculation. We place special emphasis on the monographs Geisser (1993), Gelman et al. (2014) and Ghosal et al. (2017).

Some examples are provided to illustrate the main result of the paper. For ease of reading we reproduce here an appendix from Nogales (2022a) to recall the basic concepts of Bayesian inference but, mainly, to explain the (rather unusual) notation used in the paper.

2. The framework.

We recall from Nogales (2022b) the appropriate framework to address the problem, and update it to incorporate the required asymptotic flavor.

Let $(\Omega, \mathcal{A}, \{P_\theta : \theta \in (\Theta, T, Q)\})$ be a Bayesian statistical experiment and $X_i : (\Omega, \mathcal{A}, \{P_\theta : \theta \in (\Theta, T, Q)\}) \rightarrow (\Omega_i, \mathcal{A}_i), i = 1, 2$, two statistics. Consider the Bayesian experiment image of $(X_1, X_2)$

$$(\Omega_1 \times \Omega_2, \mathcal{A}_1 \times \mathcal{A}_2, \{P_\theta^{(X_1, X_2)} : \theta \in (\Theta, T, Q)\}).$$
In the next, we shall suppose that $P^{(X_1,X_2)}(\theta, A_{12}) := P^{(X_1,X_2)}_\theta(A_{12})$, $\theta \in \Theta$, $A_{12} \in A_1 \times A_2$, is a Markov kernel.

Let us write $R_\theta = P^{(X_1,X_2)}_\theta$ and $p_j(x) := x_j$ for $j = 1, 2$, $x := (x_1, x_2) \in \Omega_1 \times \Omega_2$. Hence

$$
P^{X_1}_\theta = R^{x_1}_\theta, \quad P^{X_2|X_1=x_1}_\theta = R^{p_2|p_1=x_1}_\theta \quad \text{and} \quad E_{R_\theta}(X_2 | X_1 = x_1) = E_{R_\theta}(p_2 | p_1 = x_1).
$$

Given an integer $n$, for $m = n$ (resp. $m = \mathbb{N}$), the Bayesian experiment corresponding to a $n$-sized sample (resp. an infinite sample) of the joint distribution of $(X_1, X_2)$ is

$$(\Omega_1 \times \Omega_2)^m, (A_1 \times A_2)^m, \{R^m_\theta : \theta \in (\Theta, \mathcal{T}, Q)\} \quad (1)$$

We write $R^m(\theta, A'_{12,m}) = R^m_\theta(A'_{12,m})$ for $A'_{12,m} \in (A_1 \times A_2)^m$ and

$$
\Pi_{12,m} := Q \otimes R^m,
$$

for the joint distribution of the parameter and the sample, i.e.

$$
\Pi_{12,m}(A'_{12,m} \times T) = \int_TR^m_\theta(A'_{12,m})dQ(\theta), \quad A'_{12,m} \in (A_1 \times A_2)^m, T \in \mathcal{T}.
$$

The corresponding prior predictive distribution $\beta^*_{12,m,Q}$ on $(\Omega_1 \times \Omega_2)^m$ is

$$
\beta^*_{12,m,Q}(A'_{12,m}) = \int_\Theta R^m_\theta(A'_{12,m})dQ(\theta), \quad A'_{12,m} \in (A_1 \times A_2)^m.
$$

The posterior distribution is a Markov kernel

$$
R^*_m : ((\Omega_1 \times \Omega_2)^m, (A_1 \times A_2)^m) \rightarrow (\Theta, \mathcal{T})
$$

such that, for all $A'_{12,m} \in (A_1 \times A_2)^m$ and $T \in \mathcal{T}$,

$$
\Pi_{12,m}(A'_{12,m} \times T) = \int_TR^*_m(A'_{12,m})dQ(\theta) = \int_{A'_{12,m}}R^*_m(x', T)d\beta^*_{12,m,Q}(x').
$$

Let us write $R^*_m(x', T) := R^*_m(x', T)$.

The posterior predictive distribution on $A_1 \times A_2$ is the Markov kernel

$$
R^R_m : ((\Omega_1 \times \Omega_2)^m, (A_1 \times A_2)^m) \rightarrow (\Omega_1 \times \Omega_2, A_1 \times A_2)
$$

defined, for $x' \in (\Omega_1 \times \Omega_2)^m$, by

$$
R^R_m(x', A_{12}) := \int_\Theta R_\theta(A_{12})dR^*_m(x', \theta)
$$

It follows that, with obvious notations,

$$
\int_{\Omega_1 \times \Omega_2} f(x)dR^*_m(x', \theta) = \int_\Theta \int_{\Omega_1 \times \Omega_2} f(x)dR_\theta(x)dR^*_m(x', \theta)
$$

for any non-negative or integrable real random variable (r.r.v. for short) $f$.

We can also consider the posterior predictive distribution on $(A_1 \times A_2)^m$ defined as the Markov kernel

$$
R^R_m : ((\Omega_1 \times \Omega_2)^m, (A_1 \times A_2)^m) \rightarrow ((\Omega_1 \times \Omega_2)^m, (A_1 \times A_2)^m)
$$

2
such that

\[ R^*_m R^n(x', A'_{1,2,m}) := \int_{\Theta} R^n_\theta(A'_{1,2,m})dR^*_m(x', \theta) \]

We introduce some notations for \((x', x, \theta) \in (\Omega_1 \times \Omega_2)^m \times (\Omega_1 \times \Omega_2) \times \Theta:\)

\[
\pi'_m(x', x, \theta) := x', \quad \pi_m(x', x, \theta) := x, \quad \pi_{j,m}(x', x, \theta) := x_j, \quad j = 1, 2, \quad q_m(x', x, \theta) := \theta
\]

\[
\pi'_{i,m}(x', x, \theta) := x'_i := (x'_{i1}, x'_{i2}), \quad \pi_{(i),m}(x', x, \theta) := (x'_1, \ldots, x'_i),
\]

for \(1 \leq i \leq m\) (read \(i \in \mathbb{N}\) if \(m = \mathbb{N}\)).

Let us consider the probability space

\[(\Omega_1 \times \Omega_2)^m \times (\Omega_1 \times \Omega_2) \times \Theta, (A_1 \times A_2)^m \times (A_1 \times A_2) \times \mathcal{T}, \Pi_m,\]

(2)

where

\[\Pi_m(A'_{1,2,m} \times A_{12} \times \mathcal{T}) = \int_\mathcal{T} R_\theta(A_{12})R^n_\theta(A'_{1,2,m})dQ(\theta),\]

when \(A'_{1,2,m} \in (A_1 \times A_2)^m, A_{12} \in A_1 \times A_2\) and \(T \in \mathcal{T}\). So, for a r.r.v. \(f\) on \((\Omega_1 \times \Omega_2)^m \times (\Omega_1 \times \Omega_2) \times \Theta, (A_1 \times A_2)^m \times (A_1 \times A_2) \times \mathcal{T},\)

\[
\int f d\Pi_m = \int_{\Theta} \int_{(\Omega_1 \times \Omega_2)^m} \int_{\Omega_1 \times \Omega_2} f(x', x, \theta)dR_\theta(x)dR^n_\theta(x')dQ(\theta)
\]

provided that the integral exists. Moreover, for a r.r.v. \(h\) on \((\Omega_1 \times \Omega_2) \times \Theta, (A_1 \times A_2) \times \mathcal{T},\)

\[
\int h d\Pi_m = \int_{\Theta} \int_{\Omega_1 \times \Omega_2} h(x, \theta)dR_\theta(x)dQ(\theta) = \int_{\Omega_1 \times \Omega_2} \int_{\Theta} h(x, \theta)dR^n_{1,x}(\theta)d\beta^*_m_{12,1,Q}(x).
\]

The following result is taken from Nogales (2022b).

**Proposition 1.** For \(n \in \mathbb{N},\)

\[\Pi_{(\pi'_m, \pi_{1,n})}^{(\pi'_m, \pi_{1,n})} = \Pi_n, \quad \Pi_{(\pi'_m, \pi_{1,n})}^{(\pi'_m, \pi_{1,n})} = \Pi_n^{(\pi'_m, \pi_{1,n})}\]

\[\Pi_m^{\pi_m} = Q, \quad \Pi_{m, q_m}^{\pi_m} = \Pi_{12,m}, \quad \Pi_{m, q_m}^{\pi_m} = \beta^*_{12,m, Q} , \quad \Pi_{m, q_m}^{\pi_m} = \Pi_{12,1}, \quad \Pi_{m, q_m}^{\pi_m} = \beta^*_{12,1, Q}\]

\[\Pi_{m, q_m}^{\pi_m} = \Pi_m^{\pi'_m} = \Pi_m^{\pi_{1,n}} = \Pi_n^{(\pi'_m, \pi_{1,n})}\]

In particular, the probability space (2) contains all the basic ingredients of the Bayesian experiment (1), i.e., the prior distribution, the sampling probabilities, the posterior distributions and the prior predictive distribution. When \(m = \mathbb{N},\) (2) becomes the natural framework to address the asymptotic problem considered in this paper.

### 3. Consistency of the Bayes estimator of the regression curve

Now suppose \((\Omega_2, A_2) = (\mathbb{R}, \mathcal{R}).\) Let \(X_2\) be an squared-integrable r.r.v. such that \(E_\theta(X_2^2)\) has a finite prior mean; in particular, \(E_\theta(X_2)\) also has a finite prior mean.

The regression curve of \(X_2\) given \(X_1\) is the map \(x_1 \in \Omega_1 \mapsto r_\theta(x_1) := E_\theta(X_2|X_1 = x_1).\) An estimator of the regression curve \(r_\theta\) from a sample of size \(n\) of the joint distribution of \((X_1, X_2)\) is a statistic

\[m : (x', x_1) \in (\Omega_1 \times \mathbb{R})^n \times \Omega_1 \mapsto m(x', x_1) \in \mathbb{R},\]

3
so that, being observed \( x' \in (\Omega_1 \times \mathbb{R})^n, m(x', \cdot) \) is the estimation of \( r_\theta \).

From a classical point of view, the simplest way to evaluate the error in estimating an unknown regression curve is to use the expectation of the quadratic deviation (see Nadaraya (1989, p. 120)):

\[
E_{\theta} \left[ \int_{\Omega_1} (m(x', x_1) - r_\theta(x_1))^2 dP_\theta^X(x_1) \right] = \int_{(\Omega_1 \times \mathbb{R})^n} \int_{\Omega_1} (m(x', x_1) - r_\theta(x_1))^2 dR^p_{\theta} (x_1) dR^0_{\theta} (x').
\]

From a Bayesian point of view, the Bayes estimator of the regression curve \( r_\theta \) should minimize the Bayes risk (i.e. the prior mean of the expectation of the quadratic deviation)

\[
\int_{\Theta} \int_{(\Omega_1 \times \mathbb{R})^n} \int_{\Omega_1} (m(x', x_1) - r_\theta(x_1))^2 dR^p_{\theta} (x_1) dR^0_{\theta} (x') dQ(\theta) = E_{\Pi_n} \left[ (m(x', x_1) - r_\theta(x_1))^2 \right].
\]

Recall from Nogales (2022) that the regression curve of \( p_2 \) on \( p_1 \) with respect to the posterior predictive distribution \( R_{n,x'}^x \)

\[
m^*_n (x', x_1) := E_{R_{n,x'}^x} (p_2 | p_1 = x_1)
\]
is the Bayes estimator of the regression curve \( r_\theta (x_1) := E_\theta (X_2 | X_1 = x_1) \) for the squared error loss function, i.e.,

\[
E_{\Pi_n} [(m^*_n (x', x_1) - r_\theta(x_1))^2] \leq E_{\Pi_n} [(m_n (x', x_1) - r_\theta(x_1))^2]
\]

for any other estimator \( m_n \) of the regression curve \( r_\theta \).

We wonder about the consistency of this Bayes estimator. Another question of interest is whether the Bayes risk \( E_{\Pi_n} [(m^*_n (x', x_1) - r_\theta(x_1))^2] \) converges to 0 when \( n \) goes to \( \infty \).

The following result is key to solving the problem.

**Lemma 1.** Let \( Y(x', x, \theta) := E_\theta (p_2 | p_1 = x_1) \).

(i) For \( n \in \mathbb{N} \) and we have that

\[
E_{R_{n,x'}^x} (p_2 | p_1 = x_1) = E_{\Pi_{n}} (Y | (\pi'_n)_{N}, \pi_{1,N}) = (x'_n, x_1))
\]

(ii) \( E_{R_{n,x'}^x} (p_2 | p_1 = x_1) = E_{\Pi_{n}} (Y | (\pi'_n, \pi_{1,N}) = (x', x_1)) \)

**Proof.** (i) According to Lemma 1 of Nogales (2022), we have that, for all \( A'_{12,n} \in (\mathcal{A}_1 \times \mathcal{A}_2)^n \) and all \( A_i \in \mathcal{A}_i, i = 1, 2, \)

\[
\int_{A'_{12,n} \times A_1 \times \Omega_2 \times \Theta} R_{\theta}^{p_2 | p_1 = x_1} (A_2) d\Pi_n (x', x, \theta) = \int_{A'_{12,n} \times A_1} [R_{n,x'}^{x}]^{p_2 | p_1 = x_1} (A_2) d\Pi_n (\pi'_n) (x', x_1). \tag{14}
\]

The proof of (i) follows in a standard way from this and Proposition 1 as

\[
E_{\theta} (p_2 | p_1 = x_1) = \int_{\mathbb{R}} x_2 dR_{\theta}^{p_2 | p_1 = x_1} (x_2) \quad \text{and} \quad E_{R_{n,x'}^x} (p_2 | p_1 = x_1) = \int_{\mathbb{R}} x_2 dR_{n,x'}^x (x_2).
\]

(ii) The proof of (ii) follows from (6) as (i) derives from (14). \( \square \)
When \( A'_n := (\pi'_n, \pi_{1,n})^{-1}((A_1 \times A_2)^n \times A_1) \), we have that \((A'_n)_n\) is an increasing sequence of sub-\(\sigma\)-fields of \((A_1 \times A_2)^N \times A_1\) such that \((A_1 \times A_2)^N \times A_1 = \sigma(\bigcup_n A'_n)\). According to the martingale convergence theorem of Lévy, if \( Y \) is \((A_1 \times A_2)^N \times A_1 \times \mathcal{T}\)-measurable and \(\Pi_N\)-integrable, then

\[
E_{\Pi_N}(Y|A'_n(n))
\]

converges \(\Pi_N\)-a.e. and in \(L^1(\Pi_N)\) to \(E_{\Pi_N}(Y|(A_1 \times A_2)^N \times A_1)\).

Let us consider the measurable function

\[
Y(x', x, \theta) := E_\theta(X_2|X_1 = x_1).
\]

Notice that \(E_{\Pi_N}(Y) = \int_\Theta E_\theta(X_2) dQ(\theta)\), so \(Y\) is \(\Pi_N\)-integrable. Hence, it follows from the aforementioned theorem of Lévy that

\[
\lim_{n} E_{R^\ast_{n,x'(n)}} R(p_2|p_1 = x_1) = E_{R^\ast_{N,x'}} R(p_2|p_1 = x_1), \quad \Pi_N - \text{a.e.} \quad (15)
\]

and

\[
\lim_{n} \int_{(\Omega_1 \times \Omega_2)^N \times (\Omega_1 \times \Omega_2) \times \Theta} \left| E_{R^\ast_{n,x'(n)}} R(p_2|p_1 = x_1) - E_{R^\ast_{N,x'}} R(p_2|p_1 = x_1) \right| d\Pi_N(x', x, \theta) = 0. \quad (16)
\]

In the next we will assume the following additional regularity conditions:

(i) \((\Omega_1, \mathcal{A}_1)\) is a standard Borel space,

(ii) \(\Theta\) is a Borel subset of a Polish space and \(\mathcal{T}\) is its Borel \(\sigma\)-field, and

(iii) \(\{R_\theta : \theta \in \Theta\}\) is identifiable.

As a consequence of a known theorem of Doob (see Theorem 6.9 and Proposition 6.10 from Ghosal et al. (2017, p. 129, 130)) we have that, for every \(x_1 \in \Omega_1\),

\[
\lim_{n} \int_{\Theta} E_\theta(X_2|X_1 = x_1) d\Pi_N^{(\pi', \pi_{1,n}) = (x'(n), x_1)}(\theta') = E_\theta(X_2|X_1 = x_1), \quad R^N_{\theta} - \text{a.e.}
\]

for \(Q\)-almost every \(\theta\). Hence, according to Lemma (i),

\[
\lim_{n} E_{R^\ast_{n,x'(n)}} R(p_2|p_1 = x_1) = E_\theta(X_2|X_1 = x_1), \quad R^N_{\theta} - \text{a.e.}
\]

for \(Q\)-almost every \(\theta\).

In particular,

\[
\lim_{n} E_{R^\ast_{n,x'(n)}} R(p_2|p_1 = x_1) = E_\theta(X_2|X_1 = x_1), \quad \Pi_N - \text{a.e.}
\]

In this sense we can say that the predictive posterior regression curve \( E_{R^\ast_{n,x'(n)}} R(p_2|p_1 = x_1) \) of \(X_2\) given \(X_1 = x_1\) is a strongly consistent estimator of the sampling regression curve \( E_\theta(X_2|X_1 = x_1) \) of \(X_2\) given \(X_1 = x_1\).

From this and (15) we obtain that

\[
E_{R^\ast_{N,x'}} R(p_2|p_1 = x_1) = E_\theta(X_2|X_1 = x_1), \quad \Pi_N - \text{a.e.}
\]
According to (16) we obtain that
\[
\lim_n \int_{(\Omega_1 \times \Omega_2)^{n} \times (\Omega_1 \times \Omega_2) \times \Theta} \left| E_{\pi_{n,x'(n)}}^{R} (p_2 | p_1 = x_1) - E_{\theta}(X_2 | X_1 = x_1) \right| d\Pi_{n}(x', x, \theta) = 0, \quad (17)
\]
which proves that the Bayes risk of the optimal estimator \( E_{\pi_{n,x'(n)}}^{R} (p_2 | p_1 = x_1) \) of the regression curve \( E_{\theta}(X_2 | X_1 = x_1) \) converges to 0 for the \( L^1 \)-loss function.

We wonder if that also happens for the \( L^1 \)-squared loss function, i.e., if the Bayes risk
\[
E_{\Pi_n}[(m_{n}^{s}(x', x_1) - r_{\theta}(x_1))^2]
\]
converges vers 0 when \( n \) goes to \( \infty \). Theorem 6.6.9 of Ash et al. (2000) shows that the answer is affirmative because
\[
m_{n}^{s}(x', x_1) = E_{\Pi_n}(Y | A_{(n)})
\]
and, by Jensen’s inequality,
\[
E_{\Pi_n}(E_{\Pi_n}(Y | A_{(n)'})^2) \leq E_{\Pi_n}(E_{\Pi_n}(Y^2 | A_{(n)})) = E_{\Pi_n}(Y^2) = \int_{\Theta} X_2^2 dQ(\theta) < \infty.
\]
So, we have proved the following result (in fact part (i) is shown in Nogales (2022a) and its statement is reproduced here for the sake of completeness).

**Theorem 1.** Let \( (\Omega, \mathcal{A}, \{ P_{\theta} : \theta \in (\Theta, T, Q) \}) \) be a Bayesian statistical experiment, and \( X_1 : (\Omega, \mathcal{A}, \{ P_{\theta} : \theta \in (\Theta, T, Q) \}) \rightarrow (\Omega_1, \mathcal{A}_1) \) and \( X_2 : (\Omega, \mathcal{A}, \{ P_{\theta} : \theta \in (\Theta, T, Q) \}) \rightarrow (\mathbb{R}, \mathcal{R}) \) two statistics such that \( E_{\theta}(X_2^2) \) has finite prior mean. Let us suppose that conditions (i)-(iii) above hold. Then:

(i) The regression curve of \( p_2 \) on \( p_1 \) with respect to the posterior predictive distribution \( R_{n,x'}^{R} \)
\[
m_{n}^{s}(x', x_1) := E_{\pi_{n,x'(n)}}^{R} (p_2 | p_1 = x_1)
\]
is the Bayes estimator of the regression curve \( r_{\theta}(x_1) := E_{\theta}(X_2 | X_1 = x_1) \) for the squared error loss function, i.e.,
\[
E_{\Pi_n}[(m_{n}^{s}(x', x_1) - r_{\theta}(x_1))^2] \leq E_{\Pi_n}[(m_{n}(x', x_1) - r_{\theta}(x_1))^2]
\]
for any other estimator \( m_{n} \) of the regression curve \( r_{\theta} \).

(ii) Moreover, \( m_{n}^{s} \) is a strongly consistent estimator of the regression curve, in the sense that
\[
\lim_n \int_{(\Omega_1 \times \Omega_2)^{n} \times (\Omega_1 \times \Omega_2) \times \Theta} | E_{\pi_{n,x'(n)}}^{R} (p_2 | p_1 = x_1) - E_{\theta}(X_2 | X_1 = x_1) | d\Pi_{n}(x', x, \theta) = 0, \quad \Pi_{n} \text{ - a.e.}
\]

(iii) Finally, the Bayes risk of \( m_{n}^{s} \) converges to 0 both for the \( L^1 \)-loss function and the \( L^1 \)-squared loss function, i.e.,
\[
\lim_n E_{\Pi_n}[(m_{n}^{s}(x', x_1) - r_{\theta}(x_1))^k] = 0, \quad k = 1, 2.
\]
4. Examples.

Example 1. Let us suppose that, for \( \theta, \lambda, x_1 > 0 \), \( P_{X_1} = G(1, \theta^{-1}) \), \( P_{X_2|X_1=x_1} = G(1, (\theta x_1)^{-1}) \) and \( Q = G(1, \lambda^{-1}) \), where \( G(\alpha, \beta) \) denotes the gamma distribution of parameters \( \alpha, \beta > 0 \).

Hence the joint density of \( X_1 \) and \( X_2 \) is

\[
f_{\theta}(x_1, x_2) = \theta^2 x_1 \exp\{-\theta x_1(1 + x_2)\} I_{[0, \infty]}(x_1, x_2).
\]

It is shown in Nogales (2022b), Example 1, the Bayes estimator of the regression curve \( r_\theta(x_1) := E_\theta(X_2|X_1 = x_1) = \frac{1}{\theta x_1} \) is, for \( x_1 > 0 \),

\[
m_{n}^*(x', x_1) = \int_{0}^{\infty} x_2 \cdot f_{n,x'}^*X_2|X_1=x_1(x_2)dx_2 = \frac{\lambda + x_1 + \sum_{i=1}^{n} x_{i1}'(1 + x_{i2}')}{(2n + 1)x_1}.
\]

Theorem 1 shows that this a strongly consistent estimator of the regression curve \( r_\theta(x_1) \).

Example 2. Let us suppose that \( X_1 \) has a Bernoulli distribution of unknown parameter \( \theta \in]0,1[ \) (i.e. \( P_{X_1} = Bi(1, \theta) \)) and, given \( X_1 = k_1 \in \{0, 1 \} \), \( X_2 \) has distribution \( Bi(1, 1 - \theta) \) when \( k_1 = 0 \) and \( Bi(1, \theta) \) when \( k_1 = 1 \), i.e. \( P_{X_2|X_1=k_1} = Bi(1, k_1 + (1 - 2k_1)(1 - \theta)) \). We can think of tossing a coin with probability \( \theta \) of getting heads \((=1)\) and making a second toss of this coin if it comes up heads on the first toss, or tossing a second coin with probability \( 1 - \theta \) of making heads if the first toss is tails \((=0)\). Consider the uniform distribution on \( ]0,1[ \) as the prior distribution \( Q \).

So, the joint probability function of \( X_1 \) and \( X_2 \) is

\[
f_{\theta}(k_1, k_2) = \theta^{k_1}(1 - \theta)^{1-k_1}[k_1 + (1 - 2k_1)(1 - \theta)]^{k_2}[1 - k_1 - (1 - 2k_1)(1 - \theta)]^{1-k_2}
\]

\[
= \begin{cases} 
\theta(1 - \theta) \text{ if } k_2 = 0, \\
(1 - \theta)^2 \text{ if } k_1 = 0, k_2 = 1, \\
\theta^2 \text{ if } k_1 = 1, k_2 = 1. 
\end{cases}
\]

It is shown in Nogales (2022b), Example 2, that the Bayes estimator of the conditional mean \( r_\theta(k_1) := E_\theta(X_2|X_1 = k_1) = \theta^{k_1}(1 - \theta)^{1-k_1} \) is, for \( k_1 = 0, 1, \)

\[
m_{n}^*(k', k_1) = f_{n,k'}^*X_2|X_1=k_1(1) = \begin{cases} 
\frac{n_{10}(k') + 2n_{01}(k') + 1}{2n + n_{00}(k') + 2n_{01}(k') + 3} & \text{if } k_1 = 0, \\
\frac{n_{10}(k') + 2n_{01}(k') + 1}{2n + n_{00}(k') + 2n_{01}(k') + 4} & \text{if } k_1 = 1,
\end{cases}
\]

being \( n_{ji}(k') \) the number of indices \( i \in \{1, \ldots, n\} \) such that \((k'_{i1}, k'_{i2}) = (j_1, j_2)\) and \( n_{+j} = n_{0j} + n_{1j} \) for \( j = 0, 1 \).

Theorem 1 proves that it is a strongly consistent estimator of the conditional mean \( r_\theta(k_1) \).

Example 3. Let \((X_1, X_2)\) have bivariate normal distribution

\[
N_2 \left( \begin{pmatrix} \theta \\ \theta \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right),
\]

and consider the prior distribution \( Q = N(\mu, \tau^2) \). It is shown in Nogales (2022b), Example 3, that that the conditional mean

\[
E_{R_n^*}(p_2|p_1 = x_1) = (1 - p_1)m_1(x') + p_1 x_1
\]
is the Bayes estimator of the regression curve
\[ E_\theta(X_2 | X_1 = x_1) = (1 - \rho) \theta + \rho x_1 \]
for the squared error loss function, where
\[ \rho_1 = \frac{a_n(\rho, \sigma, \tau)}{a_n(\rho, \sigma, \tau)} - \frac{1 - \rho}{\alpha + \rho} \cdot \rho, \quad m_1(x') = \frac{s_1(x') + (1 + \rho) \frac{\sigma^2}{\tau^2} \mu}{2(1 - \rho_1)(1 + \rho)^2 \sigma^2 a_n(\rho, \sigma, \tau)}, \]
being
\[ s_1(x') := \sum_{i} (x'_{1i} + x'_{2i}), \quad a_n(\rho, \sigma, \tau) := 2(n + 1)(1 + \rho) + \frac{\sigma^2}{\tau^2}. \]

Theorem 1 proves that it is a strongly consistent estimator of this regression curve.

5. Appendix.

We recover here the Appendix of Nogales (2022a) to briefly recall some basic concepts about Markov kernels, mainly to fix the notations. In the next, (Ω, A), (Ω₁, A₁) and so on will denote measurable spaces.

**Definition 1.**
1) (Markov kernel) A Markov kernel \( M_1 : (Ω, A) \rightarrow (Ω₁, A₁) \) is a map \( M_1 : Ω \times A₁ \rightarrow [0, 1] \) such that: (i) \( \forall \omega \in Ω, M_1(\omega, \cdot) \) is a probability measure on \( A₁ \), (ii) \( \forall A₁ \in A₁, M_1(\cdot, A₁) \) is \( A \)-measurable.

2) (Image of a Markov kernel) The image (or probability distribution) of a Markov kernel \( M_1 : (Ω, A, P) \rightarrow (Ω₁, A₁) \) on a probability space is the probability measure \( P^{M_1} \) on \( A₁ \) defined by \( P^{M_1}(A₁) := \int_{Ω₁} M₁(\omega, A₁) dP(\omega). \)

3) (Composition of Markov kernels) Given two Markov kernels \( M_1 : (Ω₁, A₁) \rightarrow (Ω₂, A₂) \) and \( M_2 : (Ω₂, A₂) \rightarrow (Ω₃, A₃) \), its composition is defined as the Markov kernel \( M_2M₁ : (Ω₁, A₁) \rightarrow (Ω₃, A₃) \) given by
\[ M_2M₁(ω₁, A₃) = \int_{Ω₂} M₂(ω₂, A₃)M₁(ω₁, dω₂). \]

**Remarks.**
1) (Markov kernels as extensions of the concept of random variable) The concept of Markov kernel extends the concept of random variable (or measurable map). A random variable \( T₁ : (Ω, A, P) \rightarrow (Ω₁, A₁) \) will be identified with the Markov kernel \( M_{T₁} : (Ω, A, P) \rightarrow (Ω₁, A₁) \) defined by \( M_{T₁}(\omega, A₁) = \delta_{T₁(\omega)}(A₁) = I_{A₁}(T₁(\omega)) \), where \( \delta_{T₁(\omega)} \) denotes the Dirac measure -the degenerate distribution- at the point \( T₁(\omega) \), and \( I_{A₁} \) is the indicator function of the event \( A₁ \). In particular, the probability distribution \( P_{T₁} \) of \( M_{T₁} \) coincides with the probability distribution \( P^{T₁}(A₁) := P(T₁ \in A₁) \).

2) Given a Markov kernel \( M₁ : (Ω₁, A₁) \rightarrow (Ω₂, A₂) \) and a random variable \( X₂ : (Ω₂, A₂) \rightarrow (Ω₃, A₃) \), we have that \( M_{X₂}M₁(ω₁, A₃) = M₁(ω₁, X₂⁻¹(A₃)) = M₁(ω₁, X₂(A₃)) \). We write \( X₂M₁ := M_{X₂}M₁ \).

3) Given a Markov kernel \( M₁ : (Ω₁, A₁, P₁) \rightarrow (Ω₂, A₂) \) we write \( P₁ \otimes M₁ \) for the only probability measure on the product \( σ \)-field \( A₁ \times A₂ \) such that
\[ (P₁ \otimes M₁)(A₁ \times A₂) = \int_{A₁} M₁(ω₁, A₂)dP₁(ω₁), \quad A_i \in A_i, i = 1, 2. \]
4) Given two r.v. $X_i : (\Omega, A, P) \rightarrow (\Omega_i, A_i), i = 1, 2$, we write $P_{X_2|X_1}$ for the conditional distribution of $X_2$ given $X_1$, i.e. for the Markov kernel $P_{X_2|X_1} : (\Omega_1, A_1) \rightarrow (\Omega_2, A_2)$ such that

$$P^{(X_1, X_2)}(A_1 \times A_2) = \int_{A_1} P_{X_2|X_1} = x_1(A_2) dP^{X_1}(x_1), \quad A_i \in A_i, i = 1, 2.$$ 

So $P^{(X_1, X_2)} = P^{X_1} \otimes P^{X_2|X_1}$. □

Let $(\Omega, A, \{P_0 : \theta \in (\Theta, T, Q)\})$ be a Bayesian statistical experiment, where $Q$ denotes the prior distribution on the parameter space $(\Theta, T)$. We suppose that $P(\theta, A) := P_0(A)$ is a Markov kernel $P : (\Theta, T) \rightarrow (\Omega, A)$. When needed we shall suppose that $P_0$ has a density (Radon-Nikodym derivative) $p_0$ with respect to a $\sigma$-finite measure $\mu$ on $A$ and that the likelihood function $L(\omega, \theta) := p_0(\omega)$ is $A \times T$-measurable (this is sufficient to prove that $P$ is a Markov kernel).

Let $\Pi := Q \otimes P$, i.e.

$$\Pi(A \times T) = \int_T P_0(A) dQ(\theta), \quad A \in A, T \in T.$$ 

The prior predictive distribution is $\beta^*_Q := \Pi^T$ (the distribution of $I$ with respect to $\Pi$), where $I(\omega, \theta) := \omega$. So

$$\beta^*_Q(A) = \int_\Theta P_0(A) dQ(\theta).$$ 

The posterior distribution is a Markov kernel $P^* : (\Omega, A) \rightarrow (\Theta, T)$ such that

$$\Pi(A \times T) = \int_T P_0(A) dQ(\theta) = \int_A P^*_T(T) d\beta^*_Q(\omega), \quad A \in A, T \in T,$n

i.e. such that $\Pi = Q \otimes P = \beta^*_Q \otimes P^*$. This way the Bayesian statistical experiment can be identified with the probability space $(\Omega \times \Theta, A \times T, \Pi)$, as proposed, for instance, in Florens et al. (1990).

It is well known that, for $\omega \in \Omega$, the posterior $Q$-density is proportional to the likelihood, i.e.

$$p^*_\omega(\theta) := \frac{dP^*_\omega}{dQ}(\theta) = C(\omega)p_0(\omega)$$

where $C(\omega) = [\int_\Theta p_0(\omega) dQ(\theta)]^{-1}$.

The posterior predictive distribution on $A$ given $\omega$ is

$$P^*_\omega(A) = \int_\Theta P_0(A) dP^*_\omega(\theta), \quad A \in A.$$ 

This is a Markov kernel

$$PP^*_\omega(\omega, A) := P^*_\omega(A).$$ 

It is readily shown that the posterior predictive density is

$$\frac{dP^*_\omega}{d\mu}(\omega') = \int_\Theta p_0(\omega') p^*_\omega(\theta) dQ(\theta).$$

We know from Nogales (2022a) that

$$\int_{\Omega \times \Theta} \sup_{A \in A} |P^*_\omega(A) - P_0(A)|^2 d\Pi(\omega, \theta) \leq \int_{\Omega \times \Theta} \sup_{A \in A} |M(\omega, A) - P_0(A)|^2 d\Pi(\omega, \theta),$$

p.n.
for every Markov kernel $M : (\Omega, \mathcal{A}) \rightarrow (\Omega, \mathcal{A})$ provided that $\mathcal{A}$ is separable (recall that a $\sigma$-field is said to be separable, or countably generated, if it contains a countable subfamily which generates it). We also have that, for a real statistic $X$ with finite mean, the posterior predictive mean

$$E_{(P_\omega)^*}(X) = \int_\Theta \int_\Omega X(\omega')dP_\theta(\omega')dP_\omega^*(\theta)$$

is the Bayes estimator of $f(\theta) := E_\theta(X)$, as $E_{(P_\omega)^*}(X) = E_{P_\omega^*}(E_\theta(X))$.

6. Acknowledgements.

This paper has been supported by the Junta de Extremadura (Spain) under the grant Gr21044.

7. References.

- Ash, R.B., Doleans-Dade, C. (2000) Probability and Measure Theory, 2nd Ed., Academic Press, San Diego USA.
- Geisser, S. (1993) Predictive Inference: An Introduction, Springer Science+ Business Media, Dordrecht.
- Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B. (2014) Bayesian Data Analysis, 3rd ed., CRC Press.
- Ghosal, S., Vaart, A.v.d. (2017) Fundamentals of Nonparametric Bayesian Inference, Cambridge University Press, Cambridge UK.
- Nadaraya, E.A. (1989), Nonparametric Estimation of Probability Densities and Regression Curves, Kluwer Academic Publishers, Dordrecht.
- Nogales, A.G. (2022a), On Bayesian Estimation of Densities and Sampling Distributions: the Posterior Predictive Distribution as the Bayes Estimator, Statistica Neerlandica 76(2), 236-250.
- Nogales, A.G. (2022b), Optimal Bayesian Estimation of a Regression Curve, a Conditional Density, and a Conditional Distribution, Mathematics 10(8), 1213.
- Nogales, A.G. (2022c), On consistency of the Bayes Estimator of the Density, Mathematics, 10(4), 636.