

Consistency of Bayes estimators in nonparametric regression

Steve Lalley
Department of Statistics
University of Chicago

Marc Coram
Department of Statistics
University of Chicago

September 23, 2005

“A foolish consistency is the hobgoblin of little minds, adored by little statesmen and philosophers and divines.”

-Ralph Waldo Emerson

Bayesian Nonparametric Regression

Regression Problem: Observe data $(X_1, Y_1), (X_2, Y_2), \dots$ where (under P_f)

1. X_1, X_2, \dots are i.i.d. uniform-[0,1];
2. Y_1, Y_2, \dots are conditionally independent given \mathbf{X} ;
3. (binary regression) $Y_n | \mathbf{X}$ is Bernoulli- $f(X_n)$;
4. (ordinary regression) $Y_n | \mathbf{X}$ is Normal- $(f(X_n), 1)$.

Bayes Procedure: Put a prior distribution π on the space of regression functions f , set $Q^\pi = \int P_f d\pi(f)$, and report posterior distribution $Q^\pi(\cdot | (\mathbf{X}, \mathbf{Y})_n)$.

(Weak) Consistency: For any L^2 -neighborhood \mathcal{N} of f , as $n \rightarrow \infty$, the P_f -probability that the posterior is concentrated in \mathcal{N} converges to 1.

Hierarchical Priors I

Let $\pi_m(df)$ be priors on finite-parameter subspaces of the set of regression functions f , and let $\{\nu_m\}_{m \geq 0}$ be a probability distribution on \mathbb{Z}_+ . Consider *hierarchical priors*

$$\pi = \pi^\nu := \sum_{m=0}^{\infty} \nu_m \pi_m.$$

Example 1: (Diaconis & Freedman) Probability π_m is concentrated on step functions with 2^{m-1} discontinuities, located at the dyadic rationals $k/2^{m-1}$. Heights of the steps are chosen uniformly at random.

Example 2: (M. Coram, dissertation) Probability π_m is concentrated on step functions with m discontinuities, located at m randomly chosen points of the unit interval. Heights of the steps are chosen uniformly at random.

Hierarchical Priors II

Example 3: (M. Coram, dissertation; for k -dimensional covariate X_n) Probability π_m is concentrated on step functions constant on the m cells of a *random Voronoi tessellation* of the k -cube $[0, 1]^k$.

Example 4: (For ordinary regression) Probability π_m is the distribution of the random function

$$\sum_{j=1}^m \zeta_j \varphi_j$$

where $\{\varphi_j\}_{j \geq 1}$ is a fixed orthonormal basis of $L^2[0, 1]$ and (say)

- (a) (S. Lalley) ζ_1, ζ_2, \dots are i.i.d. Normal- $(0, m^{-1})$.
- (b) (L. Zhao) ζ_1, ζ_2, \dots are independent Normal- $(0, \tau_j^2)$ with $\tau_j = j^{-2+\varepsilon}$.

Posterior Distributions for Hierarchical Priors

Hierarchical Prior: $\pi = \pi^\nu := \sum_{m=0}^{\infty} \nu_m \pi_m$.

Posterior:

$$Q^\pi(\cdot | \mathcal{F}_n) = \left\{ \sum_{m=0}^{\infty} \nu_m Z_{m,n} Q_m(\cdot | \mathcal{F}_n) \right\} / \left\{ \sum_{m=0}^{\infty} \nu_m Z_{m,n} \right\}$$

where \mathcal{F}_n is the σ -algebra generated by the first n data points and $Z_{m,n}$ are the *predictive probabilities* for the data $(\mathbf{X}, \mathbf{Y})_n$ based on the model Q_m :

$$Z_{m,n} = \int \text{Likelihood}((\mathbf{X}, \mathbf{Y})_n | f) \pi^m(df).$$

Strategy for Proving Consistency

1. Show that for $0 \ll m \ll n$, $Q_m(\cdot | (\mathbf{X}, \mathbf{Y})_n)$ concentrates near f .
2. Show that as $m, n \rightarrow \infty$ with $m/n \rightarrow \alpha$,

$$n^{-1} \log Z_{m,n} \rightarrow \psi(\alpha).$$

3. Show that $\psi(\alpha)$ is uniquely maximized at $\alpha = 0$.

Remarks:

- In examples 1–4, item (1) is not difficult: When the number of data points swamps the number of parameters to be estimated, WLLN usually does the trick.
- For the Diaconis-Freedman priors, the large deviations problem (2)-(3) reduces to the Cramer Theorem for sums of i.i.d. random variables.
- Always the case that (2) implies $\psi(0) \geq \psi(\alpha)$ (Jensen).

Consistency Theorem I

Theorem: ($d = 1$, Coram priors) If the hierarchy prior ν is not supported by any finite subset of \mathbb{Z}_+ then for every binary regression function f except $f \equiv 1/2$, the Bayes procedure is weakly consistent at f . (Coram–Lalley, *Annals of Statistics* 2117).

Notes:

- General results of Barron, Shervish, Wasserman (*Ann. Stat.* 1999) and SG Walker (*Ann. Stat.* 2004) seem to imply consistency for hierarchy priors ν with superexponentially decaying tails.
- (1)+ (2) in the general strategy proves consistency for hierarchy priors with tails that decay faster than *some* exponential.

Analysis of the Partition Function I

Each configuration (u_1, u_2, \dots, u_m) of m points in $[0, 1]$ induces a partition of $[0, 1]$ into $m + 1$ nonoverlapping intervals. Let N_i^S, N_i^F be the success-failure counts in the i th interval of the partition. Then

$$Z_{m,n} = \int_{\mathbf{u} \in (0,1)^m} \prod_{i=0}^m B(N_i^S, N_i^F) d(\mathbf{u})$$

where

$$B(m, n) = \left\{ (m + n + 1) \binom{m + n}{m} \right\}^{-1}$$

Analysis of the Partition Function II

Reformulation: $Z_{m,n} = E_U \prod_i B(\cdot, \cdot)$ where E_U denotes expectation w.r.t. uniform distribution on unit m -cube. Observe that

$$\begin{aligned} \prod_{i=0}^m B(N_i^S, N_i^F) &= B(N_{i_*}^S, N_{i_*}^F) \prod_{i'} B(N_{i'}^S, N_{i'}^F) \prod_{i''} B(N_{i''}^S, N_{i''}^F) \\ &\leq \prod_{i'} B(N_{i'}^S, N_{i'}^F) \prod_{i''} B(N_{i''}^S, N_{i''}^F) \end{aligned}$$

where

- $\prod_{i'}$ is over the factors for partition intervals contained in $[0, 1/2)$;
- $\prod_{i''}$ is over the factors for partition intervals contained in $(1/2, 1]$; and
- $B(N_{i_*}^S, N_{i_*}^F)$ is the factor for the interval that straddles $1/2$.

Analysis of the Partition Function III

Poissonization: Suppose E_U were changed to be expectation w.r.t. a Poisson mixture (mean αn) of uniform distributions on the m -cubes, for $m \geq 0$. Then (conditional on data!) under E_U ,

$$\prod_{i'} \quad \text{and} \quad \prod_{i''}$$

are independent.

Self-Similarity: Suppose now that $f \equiv p$ is constant, and that instead of fixed sample size n we have random sample size Λ_n where $\Lambda_n \sim \text{Poisson}(n)$. Then (under P_f)

$$E_U \prod_{i'} \stackrel{\mathcal{L}}{=} Z_{m/2, n/2} \quad \text{and} \quad E_U \prod_{i''} \stackrel{\mathcal{L}}{=} Z_{m/2, n/2}$$

Note: This is not quite true, because the intervals in i', i'' are constrained to exclude the interval in \prod_i that straddles the point $1/2$. However, the error is only $O_P(1)$.

Analysis of the Partition Function IV

Subadditivity: Hence,

$$Z_{m,n} \leq Z'_{m/2,n/2} Z''_{m/2,n/2} (O_P(1)\text{error})$$

where $Z'_{m/2,n/2}$ and $Z''_{m/2,n/2}$ are independent copies of $Z_{m/2,n/2}$. A subadditive Weak Law of Large Numbers now implies that under P_f (for *constant* f),

$$\frac{1}{n} \log Z_{m,n} \longrightarrow \psi(\alpha) \tag{1}$$

for some constant $\psi(\alpha)$ when $m/n \rightarrow \alpha$.

Subadditive WLLN

Theorem: Let S_n be real random variables. Suppose that for each pair $m, n \geq 1$ of positive integers there exist random variables $S'_{m,m+n}, S''_{n,m+n}$ and a nonnegative random variable $R_{m,n}$ such that

- (a) $S'_{m,m+n}$ and $S''_{n,m+n}$ are independent;
- (b) $S'_{m,m+n}$ has the same distribution as S_m ;
- (c) $S''_{n,m+n}$ has the same distribution as S_n ;
- (d) the random variables $\{R_{m,n}\}_{m,n \geq 1}$ are identically distributed;
- (e) $ER_{1,1} < \infty$ and $\{S_n/n\}_{n \geq 1}$ are uniformly integrable; and
- (f) for all $m, n \geq 1$,

$$S_{m+n} \leq S'_{m,m+n} + S''_{n,m+n} + R_{m,n}.$$

Then

$$\frac{S_n}{n} \xrightarrow{L^1} \gamma := \liminf_{n \rightarrow \infty} \frac{ES_n}{n}$$

The Rechargeable Polya Urn

What does the data sequence (X_j, Y_j) “look like” under the mixing measure Q_m ? Assume m, n large and $m/n \approx \alpha$. Order the covariates X_j , and let Y_j^* be the value of the response corresponding to the j th largest covariate in the sample. Then the distribution of Y_1^*, Y_2^*, \dots is approximately that of the successive draws from a *rechargeable Polya urn*:

RPU (α): This is the same as the ordinary Polya urn, *except* that before each draw, with probability $\alpha/(1 + \alpha)$, the urn is flushed and then reseeded with one red and one black ball.

The proof that $\psi(\alpha) < \psi(0)$ for $\alpha > 0$ follows from (i) exponential mixing of the RPU; and (ii) the RPU process has a different law from i.i.d. Bernoulli- p .

Ordinary Regression

Regression Problem: Observe data $(X_1, Y_1), (X_2, Y_2), \dots$ where (under P_f)

1. X_1, X_2, \dots are i.i.d. uniform-[0,1];
2. Y_1, Y_2, \dots are conditionally independent given \mathbf{X} ;
3. $Y_n | \mathbf{X}$ is Normal- $(f(X_n), 1)$.

Lalley Prior: $\pi = \sum_m \nu_m \pi_m$ where π_m is the distribution function of

$$m^{-1/2} \sum_{j=1}^m \zeta_j \varphi_j$$

and

- $\{\varphi_j\}$ is an ONB of $L^2[0, 1]$ and
- ζ_1, ζ_2, \dots are i.i.d. Normal- $(0, \tau^2)$

Consistency Theorem II

Theorem (?) Suppose that *either*

- $\{\varphi_j(X)\}_{j \geq 1}$ are independent (e.g., Rademacher functions), or
- $\varphi_j(x) = \cos \pi j x$.

Suppose also that the hierarchy prior $\{\nu_m\}$ is not supported by any finite subset of \mathbb{Z}_+ . Then the Bayes procedure based on the Lalley prior is weakly consistent at every f .

The Partition Function I

Recall:

$$Z_{m,n} = \int \text{Likelihood}((\mathbf{X}, \mathbf{Y})_n | f) \pi^m(df).$$

This is a Gaussian integral that may be evaluated in close form:

The Partition Function II

$$Z_{m,n} = \exp\{-\|\mathbf{Y}\|^2/2\} \\ \exp\{\mathbf{Y}^T \Phi^T (\Sigma + I)^{-1} \Phi \mathbf{Y} / 2\} \\ \det(\Sigma + I)^{-1/2}$$

where

$$\Sigma = \frac{\Phi \Phi^T}{m}, \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{pmatrix}, \Phi = \begin{pmatrix} \varphi_1(X_1) & \varphi_1(X_2) & \cdots & \varphi_1(X_n) \\ \varphi_2(X_1) & \varphi_2(X_2) & \cdots & \varphi_2(X_n) \\ \cdots & \cdots & \cdots & \cdots \\ \varphi_m(X_1) & \varphi_m(X_2) & \cdots & \varphi_m(X_n) \end{pmatrix}$$

Marchenko-Pastur Law

Theorem: If the random variables $\{\varphi_j(X)\}_{j \geq 1}$ are independent, mean zero, and uniformly bounded then as $m, n \rightarrow \infty$ in such a way that $m/n \rightarrow \alpha$, the empirical spectral distribution $F^{\Phi\Phi^T/n}$ of the matrix $\Phi\Phi^T/n$ converges to the M-C distribution $G_\alpha(dt)$ with density

$$g_\alpha(t) = \frac{1}{2\pi\alpha t} \sqrt{(b-t)(t-a)}$$

for $a \vee 0 < t < b$ where

$$b = b(\alpha) = (1 + \alpha^{1/2})^2$$

$$a = a(\alpha) = (1 - \alpha^{1/2})^2,$$

and with an additional point mass $1 - 1/\alpha$ at the origin if $\alpha > 1$.

Consequences

Corollary: As $m, n \rightarrow \infty$ so that $m/n \rightarrow \alpha$,

$$\begin{aligned}n^{-1} \log(\det(\Sigma + I)) &\longrightarrow \gamma(\alpha), \\n^{-1} \mathbf{Y}^T \Phi^T (\Sigma + I)^{-1} \Phi \mathbf{Y} &\longrightarrow \beta(\alpha),\end{aligned}$$

and so

$$n^{-1} \log Z_{m,n} \longrightarrow \psi(\alpha) = \beta(\alpha) + \gamma(\alpha).$$

Notes:

1. Convergence of the Empirical Spectral Distribution of Σ implies convergence also for the ESD of $\Phi^T (\Sigma + I)^{-1} \Phi / n$.
2. If D is a diagonal $m \times m$ matrix whose eigenvalues decrease, and if the ESD of D converges as $m \rightarrow \infty$, then so do the ESDs of $(\Sigma + D)$ and $\Phi^T (\Sigma + D)^{-1} \Phi$.

MP Law for the Cosine Basis

Generalization of the foregoing analysis to other Orthonormal Bases φ_j requires a substitute for the Marchenko-Pastur Theorem.

Theorem: Let $\varphi_m(x) = \sqrt{2} \cos(m\pi x)$, and let $\Sigma = \Phi\Phi^T/m$ with $\Phi_{ij} = \varphi_i(X_j)$, where X_1, X_2, \dots are i.i.d. uniform-[0, 1]. Then as $m, n \rightarrow \infty$ in such a way that $m/n \rightarrow \alpha$,

$$ESD(\Sigma) \xrightarrow{\mathcal{D}} H_\alpha.$$

The limit distribution H_α is *not* the MP Law G_α (I think).

Question: For which other ONBs is there a MP Law?