PROBABILITY
AND
MATHEMATICAL STATISTICS
Vol. 17, Fasc. 1 (1997), pp. 21–28

THE RATE OF CONVERGENCE OF PMLEs

BY

X. SHEN AND K. HE* (LAWRENCE, KANSAS)

Abstract. The rate of convergence of penalized maximum likelihood estimation will be developed based on Hellinger entropy with bracketing which measures the size of underlying spaces.

1. Introduction. In this paper, we address issues associated with penalized criterion (log-likelihood) function in general parameter space: consistency and rate of convergence. We derive a large deviation inequality associated with penalized log-likelihood function based on which the consistency and rate of convergence can be established.

The use of criterion function with penalty has a long history, which goes back to Whittaker [13] and Tikhonov [10]. To overcome some undesirable properties such as inconsistency and non-smoothness, a penalty measuring such properties of parameters is often attached to the criterion function. The estimate is obtained by optimizing the penalized criterion function. This method is called the *method of regularization.* In statistics, using a penalty function can also be interpreted as formulating some prior knowledge about the parameter of interest (Good and Gaskins [5]).

The rate of convergence of penalized estimates has been studied by many authors. Among others, Wahba [12] gives a review and contains many references. The paper by Cox and O'Sullivan [3] contains some general results for the rate of convergence of the penalized estimates based on the Sobolev spaces and the L_2-penalty. Van de Geer [11] uses entropy approach to establish the rate of convergence for estimates of a regression function by using the squared loss function. Gu and Qiu [6] consider the rate of convergence of smoothing spline density estimation in terms of the symmetrized Kullback–Leibler distance.

In this paper, we will discuss the rate of convergence of penalized maximum likelihood estimation. More precisely, a general theory for consistency and the rate of convergence based on an index measuring sizes of parameter spaces (bracketing L_2 (local) metric entropy) is established. This is done in a similar spirit to the rate of convergence of the sieve estimates as in Wong and Shen [14].

* Supported by University of Kansas general research fund.

Let $Y_1, \ldots, Y_n$ be independently distributed with density $p(\theta, y)$. We estimate $\theta \in \Theta$ based on a criterion function $l(\theta, y)$, where Θ is a parameter space. Denote by $L_n(\theta)$ the scaled criterion function to be optimized, or $n^{-1}\sum_{i=1}^n l(\theta, Y_i)$. Write $\tilde{l}(\theta, y) = l(\theta, y) - \lambda_n J(\theta)$, where $J(\theta)$ is a non-negative penalty function and λ_n is the penalizing coefficient or degree of penalization. The corresponding penalized criterion function is

$$\tilde{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \tilde{l}(\theta, Y_i).$$

The maximizer of the penalized criterion function, denoted by $\hat{\theta}_n$, is called a *penalized estimate*

$$\tilde{L}_n(\hat{\theta}_n) \geqslant \sup_{\theta \in \Theta} \tilde{L}_n(\theta) - a_n, \tag{1.1}$$

where $a_n \to 0$ as $n \to \infty$. This procedure is the method of regularization, which is called the *penalized maximum likelihood estimation* (PMLE) if l is log-likelihood.

The consistency and the rate of convergence for PMLEs will be discussed in the next section. After that, an example, density estimation, will be presented to show the application of the method.

2. Convergence properties for PMLEs. We will present an exponential inequality for likelihood ratios with penalty. This inequality is based on appropriate left truncations of the lower tails of likelihood ratios as in Wong and Shen [14]. Based on this inequality, the consistency and the rate of convergence for PMLEs under the Hellinger distance can be established under one simple condition on the Hellinger metric entropy with bracketing.

Let $f: \Theta \times \mathcal{Y} \to \mathcal{R}$ with $Ef^2(\theta, Y) < \infty$ for all $\theta \in \Theta$ and let $\|\cdot\|_2$ be the usual L_2-norm, i.e.

$$\|f(\theta, Y)\|_2 = \left(\int f^2(\theta, y)\, d\mu(y)\right)^{1/2},$$

where $\mu(y)$ is the Lebesgue measure on $\mathcal{R}^1$. Let $\mathcal{F} = \{f(\theta, \cdot)\colon \theta \in \Theta, \|f\|_2 < \infty\}$. For any given $\varepsilon > 0$, if there exists

$$S(\varepsilon, n) = \{f_1^l, f_1^u, \ldots, f_n^l, f_n^u\} \subset \mathcal{L}_2 \quad \text{with} \max_{1 \leqslant j \leqslant n} \|f_j^u - f_j^l\|_2 \leqslant \varepsilon$$

such that for any $f \in \mathcal{F}$ there exists a j such that $f_j^l \leqslant f \leqslant f_j^u$ a.e., then $S(\varepsilon, n)$ is called a *bracketing ε-covering of $\mathcal{F}$ with respect to* $\|\cdot\|_2$. $H(\varepsilon, \mathcal{F}) = \log N(\varepsilon, \mathcal{F})$ is called the *L_2 metric entropy of $\mathcal{F}$ with bracketing*, where

$$N(\varepsilon, \mathcal{F}) = \min\{n\colon S(\varepsilon, n) \text{ is a bracketing } \varepsilon\text{-covering of } \mathcal{F}\}.$$

In this paper, f will be the square root density, and $H(\varepsilon, \cdot)$ is called the *Hellinger entropy with bracketing*. For more discussions about metric entropy, see, for example, Kolmogorov and Tihomirov [7], Birman and Solomjak [2], and Ossiander [8].

Let $p(\theta, y)$ be the underlying density. Let

$$h(\eta, \theta) = \frac{1}{n}\sum_{i=1}^{n} \|p^{1/2}(\eta, y_i) - p^{1/2}(\theta, y_i)\|_2^2 \quad \text{and} \quad A(k) = \{\eta \in \Theta\colon J(\eta) \leqslant k\},$$

where $\|\cdot\|_2$ is the usual L_2-norm, i.e.,

$$\|f\|_2 = \left(\int f^2(x)\, d\mu(x)\right)^{1/2}.$$

Let $P_\theta = n^{-1}\sum_{i=1}^{n} P_{\theta,i}$, where $P_{\theta,i}$ is the probability measure corresponding to the density function $p(\theta, y_i)$. In the following, k takes large values and c_i are some positive constants. Let

$$\mathscr{F}_1(k) = \{p^{1/2}(\eta, y)\colon \eta \in A(k)\},$$

and

$$\psi(\varepsilon, k) \geqslant \int_{L}^{L^{1/2}} H^{1/2}(u, \mathscr{F}_1(k))\, du/L \quad \text{for } 0 < L < 1,$$

where $L = c_1\varepsilon^2 + \lambda_n(k-1)$ and $0 < c_1 < 1$. Note that if the integral on the right-hand side is monotone of k, then we pick the integral as the $\psi(\varepsilon, k)$; otherwise, we can pick any upper bound function which is monotone in k.

ASSUMPTION 2.1. Assume that there exists $\psi(\varepsilon, k)$ such that, for all small $\varepsilon > 0$, $\psi(\varepsilon, k)$ is non-increasing with respect to large k and

$$\psi(\varepsilon, 1) \leqslant c_2 n^{1/2}. \tag{2.1}$$

THEOREM 2.1. *Suppose that Assumption* 2.1 *holds and* $J(\theta)\lambda_n \leqslant c_3\varepsilon^2$, *where* $0 < c_3 < c_1$. *Then for any* $\varepsilon > 0$ *satisfying* (2.1) *and for* $0 < c_4 < 1 - c_1$

$$P_\theta\Big(\sup_{\{\eta\in\Theta: h(\eta,\theta)\geqslant\varepsilon\}} \prod_{i=1}^{n} [p(\eta, Y_i)\exp(-\lambda_n J(\eta))/p(\theta, Y_i)\exp(-\lambda_n J(\theta))] \geqslant \exp(-c_4 n\varepsilon^2)\Big)$$
$$\leqslant 5\exp\big(-c_5 n(c_1\varepsilon^2 + \lambda_n J(\theta))\big) + 5\exp(-c_6 n\varepsilon^2),$$

where c_5 *and* c_6 *are positive constants depending on* c_i, $1 \leqslant i \leqslant 4$.

Proof. The following notation will be used. Let

$$v_n(\tilde{l}(\eta, Y) - \tilde{l}(\theta, Y)) = n^{-1/2}\sum_{i=1}^{n} (\tilde{l}(\eta, y_i) - \tilde{l}(\theta, y_i) - \mathrm{E}_\theta(\tilde{l}(\eta, y_i) - \tilde{l}(\theta, y_i)))$$
$$= n^{-1/2} v_n(l(\eta, Y) - l(\theta, Y)),$$

and

$$A_{ij} = \{\eta \in \Theta\colon 2^{i-1}\varepsilon \leqslant \varrho(\eta, \theta) < 2^i\varepsilon,\ 2^{j-1}J(\theta) \leqslant J(\eta) < 2^j J(\theta)\},$$
$$A_{i0} = \{\eta \in \Theta\colon 2^{i-1}\varepsilon \leqslant \varrho(\eta, \theta) < 2^i\varepsilon,\ J(\eta) < J(\theta)\}$$

for $i, j = 1, 2, \ldots$

Let $p^{(\tau)}(\eta, y)$ be the left truncation version of $p(\eta, y)$, i.e.,

$$p^{(\tau)}(\eta, y) = \begin{cases} \exp(-\tau) & \text{if } p(\eta, y) < \exp(-\tau)\, p(\theta, y), \\ p(\eta, y) & \text{otherwise} \end{cases}$$

for any $0 \leqslant \tau < \infty$.

The truncation techniques in Wong and Shen [14] will be used in the following derivations. Although the results in [14] are developed for independent and identically distributed observations, it can be seen that all results hold generally for equidistributed observations if the corresponding quantities in [14] are replaced by the average quantities based on each observation. By Lemma 4 of [14] we have

$$n^{-1} \sum_{i=1}^{n} \mathrm{E}_\theta \log\bigl(p(\theta, Y_i)/p^{(\tau)}(\eta, Y_i)\bigr) \geqslant (1-T)h^2(\theta, \eta),$$

$$\text{where } T = 2\exp(-\tau/2)/\bigl(1-\exp(-\tau/2)\bigr)^2.$$

Thus

$$P_\theta\Bigl(\sup_{\{\eta\in\Theta: h(\eta,\theta)\geqslant\varepsilon\}} \prod_{i=1}^{n} \bigl[p(\eta, Y_i)\exp\bigl(-\lambda_n J(\eta)\bigr)/p(\theta, Y_i)\exp\bigl(-\lambda_n J(\theta)\bigr)\bigr] \geqslant \exp(-c_4 n\varepsilon^2)\Bigr)$$

$$\leqslant P_\theta\Bigl(\sup_{\{\eta\in\Theta: h(\eta,\theta)\geqslant\varepsilon\}} n^{-1} \sum_{i=1}^{n} \log\bigl(p^{(\tau)}(\eta, Y_i)/p(\theta, Y_i)\bigr) \geqslant \lambda_n\bigl(J(\eta)-J(\theta)\bigr) - c_4\varepsilon^2\Bigr)$$

$$\leqslant \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} P_\theta\Bigl(\sup_{A_{ij}} n^{-1/2} v_n\bigl(\log\bigl(p^{(\tau)}(\eta, Y)/p(\theta, Y)\bigr)\bigr) \geqslant \inf_{A_{ij}}\bigl[(1-T)h^2(\theta, \eta) + \lambda_n\bigl(J(\eta)-J(\theta)\bigr)\bigr] - c_4\varepsilon^2\Bigr)$$

$$+ \sum_{i=1}^{\infty} P_\theta\Bigl(\sup_{A_{i0}} n^{-1/2} v_n\bigl(\log\bigl(p^{(\tau)}(\eta, Y)/p(\theta, Y)\bigr)\bigr) \geqslant \inf_{A_{i0}}\bigl[(1-T)h^2(\theta, \eta) + \lambda_n\bigl(J(\eta)-J(\theta)\bigr)\bigr] - c_4\varepsilon^2\Bigr)$$

$$= I_1 + I_2.$$

Let us consider I_1. Note that $\varepsilon^2 \leqslant h^2(\theta, \eta)$ by choosing τ such that $1-T-c_4 > c_1$; then

$$\inf_{A_{ij}}\bigl[(1-T)h^2(\theta, \eta) + \lambda_n\bigl(J(\eta)-J(\theta)\bigr) - c_4\varepsilon^2\bigr] \geqslant n^{-1/2} M(i, j),$$

where $M(i, j) = n^{1/2}\bigl[c_1(2^{i-1}\varepsilon)^2 + \lambda_n(2^j-1)J(\theta)\bigr]$. By Lemma 3 of Wong and

Shen [14] we have

$$\sup_{A_{ij}} \frac{1}{n} \sum_{i=1}^{n} \mathrm{Var}_\theta \left(\log p^{(\tau)}(\eta, Y_i)/p(\theta, Y_i)\right) \leqslant v^2(i, j),$$

where $v^2(i, j) = 4\exp(\tau)[(2^i\varepsilon)^2 + \lambda_n(2^j - 1)J(\theta)]$.

We now verify the conditions in Lemma 7 of [14]. With specified $M(i, j)$ and $v(i, j)$ condition (2.3) of Lemma 7 holds. By Assumption 2.1 and the monotonicity property of $H(u, \cdot)$,

$$\int_{M(i,j)/n^{1/2}}^{v(i,j)} H^{1/2}\left(u, \mathscr{F}_1(j)\right) du/M(i, j)$$

$$\leqslant \int_{M(1,j)/n^{1/2}}^{v(1,j)} H^{1/2}\left(u, \mathscr{F}_1(j)\right) du/M(1, j) \leqslant \psi(\varepsilon, 1) \leqslant c_2 n^{1/2};$$

thus condition (2.4) of Lemma 7 holds. Note that Lemma 7 of [14] continues to hold if $M(i, j) > an^{1/2}v^2(i, j)$ for any i, j, where $a = c_2^{1/2}$ is a constant. Therefore, by Lemma 7,

$$\begin{aligned} I_1 &\leqslant \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} P_\theta\left(\sup_{A_{ij}} v_n\left(\log\left(p^{(\tau)}(\eta, Y)/p(\theta, Y)\right)\right) \geqslant M(i, j)\right) \\ &\leqslant \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} 4\exp\left(-c_5 n\left[c_1(2^{i-1}\varepsilon)^2 + (2^j - 1)\lambda_n J(\theta)\right]\right) \\ &\leqslant 4\exp\left(-c_5 n\left[c_1\varepsilon^2 + \lambda_n J(\theta)\right]\right)/\left(1 - \exp\left(-c_5 n\left(c_1\varepsilon^2 + \lambda_n J(\theta)\right)\right)\right) \\ &= 5\exp\left(-c_5 n\left(c_1\varepsilon^2 + \lambda_n J(\theta)\right)\right), \end{aligned}$$

where

$$c_5 = (1 - a)\exp(-\tau)/8(a + 32T_1) \quad \text{with} \quad T_1 = \left(\exp(\tau/2) - 1 - \tau/2\right)/\left(1 - \exp(-\tau/2)\right)^2.$$

Now let us consider I_2. Since $\lambda_n J(\theta) \leqslant c_3\varepsilon^2$, we have

$$\begin{aligned} I_2 &\leqslant \sum_{i=1}^{\infty} P_\theta\left(\sup_{A_{i0}} n^{-1/2} v_n\left(\log p^{(\tau)}(\eta, Y)/p(\theta, Y)\right) \geqslant c_1(2^{i-1}\varepsilon)^2 - \lambda_n J(\theta)\right) \\ &\leqslant \sum_{i=1}^{\infty} P_\theta\left(\sup_{A_{i0}} v_n\left(\log p^{(\tau)}(\eta, Y)/p(\theta, Y)\right) \geqslant M(i)\right), \end{aligned}$$

where $M(i) = (c_1 - c_3)n^{1/2}(2^{i-1}\varepsilon)^2$. The result then follows from a similar argument to that for I_1. This completes the proof. ∎

Remark. With the optimal choice of $\lambda_n = c_3\varepsilon^2/J(\theta)$, for any $\varepsilon > 0$ satisfying

$$\int_{\varepsilon^2}^{\varepsilon} H^{1/2}\left(u, \mathscr{F}_1(1)\right) du \leqslant c_2 n^{1/2}\varepsilon^2, \tag{2.2}$$

we have

$$P_\theta\Big(\sup_{\{\eta\in\Theta:h(\eta,\theta)\geqslant\varepsilon\}}\prod_{i=1}^{n}[p(\eta, Y_i)\exp(-\lambda_n J(\eta))]/[p(\theta, Y_i)\exp(-\lambda_n J(\theta))] \geqslant \exp(-c_4 n\varepsilon^2)\Big) \leqslant 10\exp(-c_6 n\varepsilon^2).$$

COROLLARY 2.1. *Under Assumption* 2.1, *for the PMLE defined in* (1.1) *with* $a_n \leqslant c_7\varepsilon_n^2$, *where* $0 < c_7 < c_1$, *the rate of convergence under the Hellinger distance is* $\max(\varepsilon_n, \lambda_n^{1/2})$, *where* ε_n *is the smallest* ε *satisfying* (2.1). *The best possible rate of convergence for PMLE is then determined by the smallest* ε *satisfying* (2.2) *with* $\lambda_n J(\theta) = c_3\varepsilon_n^2$.

Proof. By (1.1), we have

$$P_\theta(h(\hat\theta_n, \theta) \geqslant \varepsilon) \leqslant P_\theta\Big(\sup_{\{\eta\in\Theta:h(\eta,\theta)\geqslant\varepsilon\}}(\tilde L_n(\eta)-\tilde L_n(\theta)) \geqslant -a_n\Big).$$

By Theorem 2.1, $h(\hat\theta_n, \theta) = O_p(\varepsilon_n)$ when $\lambda_n J(\theta) \leqslant c_3\varepsilon_n^2$. Note that ε_n is the smallest ε satisfying (2.2); then $h(\hat\theta_n, \theta) = O_p(\lambda_n^{1/2})$ if ε_n is replaced by $\lambda_n^{1/2}$ when $\varepsilon_n^2 < J(\theta)\lambda_n$. Then the result follows. ∎

3. Example: the density estimation. Let $Y_1, \ldots, Y_n$ be independently and identically distributed according to a density θ^2. We estimate the square root density

$$\theta\in\Theta = W^{m,q}[0, 1] = \{\eta\in C^{m-1}[0, 1]: \eta \geqslant 0, J(\eta) < +\infty, q \geqslant 2\}$$

by the PML method. For the case of $m > 1/2$,

$$J(\eta) = \|\eta^{([m])}\|_r + \Big[\iint(|\eta^{([m])}(x)-\eta^{([m])}(y)|/|x-y|^{m-[m]})^r\,dxdy\Big]^{1/r},$$

where $[m]$ is the integer part of m and $r \geqslant 1/m$. For the case of $m \leqslant 1/2$,

$$J(\eta) = \sup_{x,y}|\eta(x)-\eta(y)|/|x-y|^m.$$

Two results for metric entropies of different values of m will be used. For the case of $m > 1/2$, by the norm equivalence (the Theorem of Adams [1], p. 79) and Theorem 5.2 of Birman and Solomjak [2],

$$H(u, \mathscr{F}_1(k)) \leqslant c_0(k/u)^{1/m}.$$

For the case of $m < 1/2$, by Theorem 2 of Gabushin [4],

$$\|\eta\|_{\sup} \leqslant d[J(\eta)]^{(2m)/(2m+1)};$$

then

$$H(u, \mathscr{F}_1(k)) \leqslant \exp(d_1\log(k^{(2m)/(2m+1)}/u)+d_2\log(k/u)^{1/m}) \leqslant c_0(k/u)^{1/m}$$

for all small $u > 0$ and some constants d_i (see Theorem 15 of Kolmogorov and Tihomirov [7] for the corresponding constants for the sup-entropy). Take

$$\psi(k) = c_0^{1/2}(1-1/2m)^{-1}\left(\lambda_n(k-1)\right)^{-(1+2m)/(4m)} k^{1/(2m)} \quad \text{if } m > 1/2,$$

$$\psi(k) = -c_0^{1/2} k\left(\log\left(\lambda_n(k-1)\right)\right)/\lambda_n(k-1) \quad \text{if } m = 1/2,$$

and

$$\psi(k) = c_0^{1/2}(1-1/2m)^{-1}\left(\lambda_n(k-1)\right)^{-1/2m} k^{1/2m} \quad \text{if } m < 1/2;$$

then Assumption 2.1 is satisfied.

By Corollary 2.1, the rate of convergence of PMLE $h(\hat{\theta}_n, \theta) = \max(\varepsilon_n, \lambda_n^{1/2})$, where ε_n is the solution of the equation

$$\int_{c_1^2\varepsilon_n^2}^{c_1\varepsilon_n} H^{1/2}\left(u, \mathscr{F}_1(1)\right) du = c_2 n^{1/2}\varepsilon_n^2,$$

where $c_2 = c_0^{1/2}(1-1/2m)^{-1}$ if $m > 1/2$; $c_2 = c_0^{1/2}$ if $m = 1/2$; and $c_2 = c_0^{1/2}(1/2m-1)^{-1}$ if $m < 1/2$. The resulting rate ε_n is $O_p(n^{-m/(2m+1)})$ if $m > 1/2$; $O_p\left(n^{-1/4}(\log n)^{1/2}\right)$ if $m = 1/2$; $O_p(n^{-m/2})$ if $m < 1/2$. Hence, with the choice of $\lambda_n = \varepsilon_n^2$, the best possible rate ε_n for the PMLE under $h(\cdot, \cdot)$ is $O_p(n^{-m/(2m+1)})$ if $m > 1/2$; $O_p\left(n^{-1/4}(\log n)^{1/2}\right)$ if $m = 1/2$; and $O_p(n^{-m/2})$ if $m < 1/2$. The rate obtained is optimal (Stone [9]).

Acknowledgment. The authors thank Professors Larry Brown, Chong Gu and Jinchao Xu for many helpful discussions.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York 1975.

[2] M. S. Birman and M. Z. Solomjak, *Piecewise-polynomial approximation of functions of the classes W_p*, Mat. Sb. 73 (1967), pp. 295–317.

[3] D. D. Cox and F. O'Sullivan, *Asymptotic analysis of penalized likelihood and related estimators*, Ann. Statist. 18 (1990), pp. 1676–1695.

[4] V. N. Gabushin, *Inequalities for norms of functions and their derivatives in the L_p metric*, Mat. Zametki 1 (1967), pp. 291–298.

[5] I. J. Good and R. A. Gaskins, *Non-parametric roughness penalties for probability densities*, Biometrika 58 (1971), pp. 255–277.

[6] C. Gu and C. Qiu, *Smoothing spline density estimation: Theory*, Ann. Statist. 21 (1993), pp. 217–234.

[7] A. N. Kolmogorov and V. M. Tihomirov, *ε-entropy and ε-capacity of sets in function spaces*, Uspekhi Mat. Nauk 14 (1959), pp. 3–86. [In Russian. English translation: Amer. Math. Soc. Transl. (1961), pp. 277–364.]

[8] M. Ossiander, *A central limit theorem under metric entropy with L_2 bracketing*, Ann. Probab. 15 (1987), pp. 897–919.

[9] C. Stone, *Optimal global rates of convergence for nonparametric regression*, Ann. Statist. 10 (1982), pp. 1040–1053.

[10] A. Tikhonov, *Solution of incorrectly formulated problems and the regularization method*, Soviet. Math. Dokl. 5 (1963), pp. 1035–1038.

[11] S. van de Geer, *Estimating a regression function*, Ann. Statist. 18 (1990), pp. 907–924.

[12] G. Wahba, *Spline Models for Observational Data*, CBMS–NSF Regional Conference Series, Philadelphia 1990.
[13] E. Whittaker, *On a new method of graduation*, Proc. Edinburgh Math. Soc. (2) (1923), p. 41.
[14] W. H. Wong and X. Shen, *Probability inequalities for likelihood ratios and convergence rates of sieve MLEs*, Ann. Statist. 23 (1995), pp. 339–362.

University of Kansas
Department of Mathematics
405 Snow Hall
Lawrence, KS 66045, U.S.A.

Received on 5.8.1996