Entropy and Convergence Rates: the test-function and the information-theoretic approach

Willem Kruijer (Paris Dauphine), joint work with Aad van der Vaart (VU Amsterdam)

Rennes, 29 August 2008

Notation

$$X = (X_1, \ldots, X_n) \sim q(x) = \prod_{i=1}^{n} q_i(x_i)$$

The convergence rate is the fastest sequence $\epsilon_n \to 0$ such that

$$\Pi(p : \frac{1}{n} \sum_{i=1}^{n} d^2(p_i, q_i) \geq \epsilon_n^2 | X_1, \ldots, X_n) \xrightarrow{q} \quad 0.$$

Alternatively, the rate is said to be ϵ_n if

$$E_{X \sim q} \int \frac{1}{n} \sum_{i=1}^{n} d^2(p_i, q_i) d\Pi(p \mid X) \lesssim \epsilon_n^2.$$

Notation (2)

- For $0 < \beta < 1$,

$$\rho_\beta(p_1, p_2) = \int p_1^\beta p_2^{1-\beta} d\mu = E_{X \sim p_2}(\frac{p_1}{p_2}(X))^\beta \leq 1$$

  denotes the Hellinger affinity between p_1 and p_2.
- The Renyi-affinity is defined as

$$D_\beta(p_2|p_1) = -\log \rho_\beta(p_1, p_2)$$

- $D_{1/2}(p_2|p_1) \geq h^2(p_1, p_2)$.

Obtaining rates (1)

Let the packing number $D(\epsilon, A, d)$ of a set $A \subset \mathcal{P}$ be defined as the cardinality of the largest ϵ-separated set of points in A.

Define $KL(q, \epsilon_n) = \{p : E_q \log \frac{q}{p} \leq \epsilon_n^2, E_q \left(\log \frac{q}{p}\right)^2 \leq \epsilon_n^2\}$.

Theorem (Ghosal, Ghosh and van der Vaart, 2000)

Let $\epsilon_n \to 0$ be a sequence such that $n{\epsilon_n}^2 \to \infty$. Suppose that for a semi-metric d and constants k_1 and k_2,

$$\begin{aligned} \log D(\epsilon_n, \{p : \epsilon_n < d(p, q) < 2\epsilon_n\}, d) &\leq k_1 n\epsilon_n^2 \qquad (1) \\ \Pi_n\big(KL(q, \epsilon_n)\big) &\geq e^{-k_2 n\epsilon_n^2}. \end{aligned}$$

Then for $M > 0$ sufficiently large,

$$E_{q^n}\Pi(p : d(p, q) \leq M\epsilon_n \mid X_1, \ldots, X_n) \longrightarrow 1. \qquad (2)$$

Obtaining rates (2)

Theorem (Zhang, 2006)

For $\beta \in (0,1)$, $\gamma \geq 1$, *and* $\lambda' = \frac{\gamma-1}{\gamma-\beta} \in (0,1)$,

$$E_X \int \frac{1}{n} D_\beta^{Re}(q \mid p) d\Pi(p \mid X) \leq \big((\gamma-\beta)k_1 + \gamma(1+k_2)\big)\epsilon_n^2, \quad (3)$$

provided that for a partition $\mathcal{P} = \cup_j \mathcal{P}_j$

$$\log\Big[\sum_j \Pi(\mathcal{P}_j)^{\lambda'} r_n(\mathcal{P}_j)\Big] \leq k_1 n\epsilon_n^2, \quad (4)$$

$$\Pi\big(\{p \in \mathcal{P} : D_{KL}(q \parallel p) < n\epsilon_n^2\}\big) \geq e^{-k_2 n\epsilon_n^2}, \quad (5)$$

where the upper-bracketing radius of $\mathcal{P}_j \subset \mathcal{P}$ *is defined as*

$$r_n(\mathcal{P}_j) = \int \sup_{p \in \mathcal{P}_j} p(x) d\mu(x).$$

How to verify (4) ? How different are these theorems ?

Heuristics: L_1-brackets

- Let $\mathcal{P} \subset \mathcal{H}^\alpha$ be a class of densities on a compact interval
- Estimation of $q \in \mathcal{P}$: can we obtain $\epsilon_n = n^{-\alpha/(1+2\alpha)}$?
- Choose $\mathcal{P}_j = [l_j, u_j]$, L_1-brackets of size ϵ_n^2.
- $j = 1, \ldots, N(\epsilon_n^2)$, where $N(\epsilon_n^2) = N_{[\,]}(\epsilon_n^2, \mathcal{P}, L_1)$.
- $\log\left[\sum_j \Pi(\mathcal{P}_j)^{\lambda'} r_n(\mathcal{P}_j)\right] \leq \log r_n(\mathcal{P}_j) + \log N_{[\,]}(\epsilon_n^2, \mathcal{P}, L_1)$.
- Indications that it is not bounded by a multiple of $n\epsilon_n^2 = n^{1/(1+2\alpha)}$:

$$N_{[\,]}(\epsilon_n^2, \mathcal{P}, L_1) \geq N(\epsilon_n^2, \mathcal{P}, L_1)$$

$$\log N(\epsilon_n^2, \mathcal{P}, L_1) \leq \log N(\epsilon_n^2, \mathcal{P}, \|\cdot\|_\infty) \approx \left(\frac{1}{\epsilon_n^2}\right)^{\frac{1}{\alpha}} \approx n^{2/(1+2\alpha)}$$

Introducing the distance between the $\mathcal{P}_j$ and q

For $\delta \in (0, 1]$, redefine the upper-bracketing radius:

$$r_{\delta,n}(\mathcal{P}_j) = \int \left(\sup_{p \in \mathcal{P}_j} p(x) \right)^{\delta} q^{1-\delta}(x) d\mu(x). \tag{6}$$

The entropy condition

$$\log \left[\sum_j \Pi(\mathcal{P}_j)^{\lambda'} r_n(\mathcal{P}_j) \right] \leq k_1 n \epsilon_n^2, \tag{7}$$

can be replaced by

$$\log \left[\sum_j \Pi(\mathcal{P}_j)^{\delta\lambda'} r_{\delta,n}(\mathcal{P}_j) \right] \leq k_1 n \epsilon_n^2. \tag{8}$$

Further improvements are possible, but for regression models with Gaussian or exponential-errors this is good enough.

proof of Theorem 2

From the information inequality it follows that if $\gamma > 1$, $0 < \beta < 1$, $\lambda' = (\gamma - \beta)/(\gamma - 1) \in (0, 1)$ and $X \sim q$,

$$
\begin{aligned}
& E_X \int \frac{1}{n} D_\beta^{Re}(q \mid p) d\Pi(p \mid X) \\
& \quad \leq -\frac{\gamma}{n} \log \int \exp\Big(-D_{KL}(q \parallel p)\Big) d\Pi(p) \\
& \qquad + \frac{(\gamma - \beta)\lambda'}{n} E_X \log \left(\int \left(\frac{p}{q}(X) \right)^{1/\lambda'} d\Pi(p) \right).
\end{aligned}
\tag{9}
$$

Localizing the entropy term

$$
\begin{aligned}
&\frac{\lambda'}{n} E_X \log \left(\int \left(\frac{p}{q}(X) \right)^{1/\lambda'} d\Pi(p) \right) \\
&\quad \leq \frac{1}{\delta n} \log E_X \left(\int \left(\frac{p}{q}(X) \right)^{1/\lambda'} d\Pi(p) \right)^{\delta\lambda'} \\
&\quad \leq \frac{1}{\delta n} \log E_X \left[\sum_j \Pi(\mathcal{P}_j) \left(\sup_{p \in \mathcal{P}_j} \frac{p}{q}(X) \right)^{1/\lambda'} \right]^{\delta\lambda'} \\
&\quad \leq \frac{1}{\delta n} \log E_X \sum_j \Pi(\mathcal{P}_j)^{\delta\lambda'} \left(\sup_{p \in \mathcal{P}_j} \frac{p}{q}(X) \right)^{\delta} \\
&\quad = \frac{1}{\delta n} \log \sum_j \Pi(\mathcal{P}_j)^{\delta\lambda'} r_{\delta,n}(\mathcal{P}_j).
\end{aligned}
$$

Fixed design Gaussian regression

- Given covariates $x_1, \ldots, x_n$, let $q(y) = \prod_{i=1}^n \phi_\sigma(y_i - f_0(x_i))$
- $p = p_f = (\phi_\sigma(y_1 - f(x_1)), \ldots, \phi_\sigma(y_n - f(x_n)))$, with $f \in \mathcal{F}$.
- $\|f_1 - f_2\|_n := \frac{1}{2\sigma^2 n} \sum_{i=1}^n (f_1(x_i) - f_2(x_i))^2$

Theorem (Ghosal and Van der Vaart (2007))

Assume that

$$\Pi\big(\{f \in \mathcal{F} : \|f - f_0\|_n < \epsilon_n^2\}\big) \geq e^{-k_1 n \epsilon_n^2}, \tag{10}$$

and that there exists a nondecreasing function $N(\epsilon)$ such that

$$\log N\big(\epsilon_n, \{\epsilon_n < \|f - f_0\|_n < 2\epsilon_n\}, \|\cdot\|_n\big) \leq N(\epsilon_n) \leq c_2 n \epsilon_n^2. \tag{11}$$

Then

$$E_Y \int \|f - f_0\|_n^2 d\Pi(f \mid Y) \lesssim \epsilon_n^2.$$

Proof (application of Theorem 2)

The Kullback-Leibler divergence and Renyi-entropy between p_f and $p_{f_0} = q$ are multiples of $\|f - f_0\|_n$.

To verify the entropy condition, choose the following subsets:
$\mathcal{F}_0 = \{f \in \mathcal{F} : \|f - f_0\|_n < \epsilon_n\}$
For $k \geq 1$, cover

$$\mathcal{F}_k = \{f \in \mathcal{F} : k\epsilon_n \leq \|f - f_0\|_n < 2k\epsilon_n\}$$

with $\|\cdot\|_n$-balls of radius $Lk\epsilon_n$, where $0 < L < \frac{1}{2}$.
Removing the overlapping parts, we find a partition
$\mathcal{F}_0, \mathcal{F}_{k,j}$, $k \geq 1$, $j = 1, \ldots, N_k$.

...

If $\mathcal{F}_{k,j}$ is contained in $B(f_{k,j}, Lk\epsilon_n, \|\cdot\|_n)$, there exists a $C > 0$ s.t.

$$r_{\delta,n}(\mathcal{F}_{k,j}) = \int \prod_{i=1}^{n} \sqrt{\phi_\sigma(y_i - f_0(x_i))}\, sup_{f\in\mathcal{F}_{k,j}} \prod_{i=1}^{n} \sqrt{\phi_\sigma(y_i - f(x_i))} dy$$
$$\leq \exp\{-Cnk^2\epsilon_n^2\},$$

see also Le Cam (1983, 1984 ??).

$$\log\Big[\sum_j \Pi(\mathcal{P}_j)^{\delta\lambda'} r_{\delta,n}(\mathcal{F}_{k,j})\Big] \leq \log \sum_{k\geq 0}\sum_{j=1}^{N_k} r_{\delta,n}(\mathcal{F}_{k,j})$$
$$\leq \log N(k\epsilon_n) + \log \sum_{k\geq 0} \sup_j r_{\delta,n}(\mathcal{F}_{k,j})$$
$$\leq \log N(\epsilon_n) + \log \sum_{k\geq 0} exp\{-Cnk^2\epsilon_n^2\} \leq k_1 n\epsilon_n^2.$$

Misspecification

If $q \notin \mathcal{P}$, define $\tilde{q} = argmin_{p \in \mathcal{P}} D_{KL}(q \parallel p)$, and

$$r^*_{\delta,n}(A) = \int \left(\frac{\sup_{p \in A} p(x)}{\tilde{q}(x)} \right)^{\delta} q(x) d\mu(x).$$

Theorem

For $\beta \in (0,1)$ and $\gamma \geq 1$, let $\lambda' = \frac{\gamma - 1}{\gamma - \beta} \in (0,1)$. Assume that

$$\log \left[\sum_j \Pi(\mathcal{P}_j)^{\delta \lambda'} r^*_{\delta,n}(\mathcal{P}_j) \right] \leq k_1 n \epsilon_n^2,$$

$$\Pi\left(\{ p \in \mathcal{P} : \int q(x) \log \frac{\tilde{q}}{p}(x) d\mu(x) < n\epsilon_n^2 \} \right) \geq e^{-k_2 n \epsilon_n^2},$$

where $\{\mathcal{P}_j\}$ is an arbitrary countable cover of $\mathcal{P}$. Then

$$E_X \int \frac{1}{n} D^{Re}_{\beta}(\tilde{q} \mid p) d\Pi(p \mid X) \leq \big((\gamma - \beta) k_1 + \gamma(1 + k_2) \big) \epsilon_n^2.$$

Work in progress

A different discretization:

- Given $\{\mathcal{P}_j\}_{j\in J}$, define $p_j(X) = \int_{\mathcal{P}_j} p(X)d\Pi(X)/\Pi(\mathcal{P}_j)$.
- Apply (9) to the model $\{p_j\}_{j\in J}$ with prior $\{\Pi(\mathcal{P}_j)\}_{j\in J}$.
- The posterior is $\{\Pi(\mathcal{P}_j|X)\}_{j\in J}$.
- Take the supremum and infimum over the convex hulls of the $\mathcal{P}_j$'s, to achieve sub-additivity.

Conclusion

- Often, optimal rates can only be obtained if the entropy condition is 'localized'.
- For comparison with the result of Ghosal, Ghosh and van der Vaart (2000), a different discretization is necessary.

References

[1] Kleijn, B.J.K. and van der Vaart, A.W. (2006) Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics* **34**(2), 837–877.

[2] Zhang, T. (2006) Information-theoretic upper and lower bounds for statistical estimation. *IEEE Trans. Inform. Theory* **52** 1307–1321.

[3] Zhang, T. (2006) From ϵ-entropy to KL-entropy: analysis of minimum information complexity density estimation. *Annals of Statistics* **34**(5), 2180–2210.