5 Rates of convergence of M-estimators

Let (Θ, d) be a semimetric space. As usual, we are given i.i.d. observations $X_1, X_2, \ldots, X_n$ from a probability distribution P on $\mathcal{X}$. Let $\{\mathbb{M}_n(\theta) : \theta \in \Theta\}$ denote a stochastic process and let $\{M(\theta) : \theta \in \Theta\}$ denote a deterministic process. Suppose $\hat{\theta}_n$ maximizes $\mathbb{M}_n(\theta)$ and suppose θ_0 maximizes $M(\theta)$, i.e.,

$$\hat{\theta}_n = \operatorname*{argmax}_{\theta \in \Theta} \mathbb{M}_n(\theta), \qquad \text{and} \qquad \theta_0 = \operatorname*{argmax}_{\theta \in \Theta} M(\theta).$$

We assume that $\mathbb{M}_n(\theta)$ gets close to $M(\theta)$ as n increases and under this setting want to know how close $\hat{\theta}_n$ is to θ_0. If the metric d is chosen appropriately we may expect that the asymptotic criterion decreases *quadratically* when θ moves away from θ_0:

Taylor expansion

$$M(\theta) - M(\theta_0) \lesssim -d^2(\theta, \theta_0) \tag{35}$$

for all $\theta \in \Theta$. We want to find the rate δ_n of the convergence of $\hat{\theta}_n$ to θ_0 in the metric d i.e., $d(\hat{\theta}_n, \theta_0)$. A *rate of convergence*[32] of δ_n means that

$$\delta_n^{-1} d(\hat{\theta}_n, \theta_0) = O_{\mathbb{P}}(1).$$

Consider the probability $\mathbb{P}\big(d(\hat{\theta}_n, \theta_0) > 2^M \delta_n\big)$ for a large M. We want to understand for which δ_n this probability becomes small as M grows large. Write

$$\mathbb{P}\Big(d(\hat{\theta}_n, \theta_0) > 2^M \delta_n\Big) = \sum_{j > M} \mathbb{P}\Big(2^{j-1}\delta_n < d(\hat{\theta}_n, \theta_0) \le 2^j \delta_n\Big).$$

Let us define the "shells" $S_j := \{\theta \in \Theta : 2^{j-1}\delta_n < d(\theta, \theta_0) \le 2^j \delta_n\}$ so that

$$\mathbb{P}\Big(2^{j-1}\delta_n < d(\hat{\theta}_n, \theta_0) \le 2^j \delta_n\Big) = \mathbb{P}\Big(\hat{\theta}_n \in S_j\Big).$$

As $\hat{\theta}_n$ maximizes $\mathbb{M}_n(\theta)$, it is obvious that $\subseteq$ local

global

$$\mathbb{P}\Big(\hat{\theta}_n \in S_j\Big) \le \mathbb{P}\Big(\sup_{\theta \in S_j} (\mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0)) \ge 0\Big).$$

[32]Recall that a sequence of random variables $\{Z_n\}$ is said to be *bounded in probability* or $O_{\mathbb{P}}(1)$ if

$$\lim_{T \to \infty} \limsup_{n \to \infty} \mathbb{P}(|Z_n| > T) = 0.$$

In other words, $Z_n = O_{\mathbb{P}}(1)$, if for any given $\epsilon > 0$, there exists $T_\epsilon, N_\epsilon > 0$ such that

$$\mathbb{P}(|Z_n| > T_\epsilon) < \epsilon \qquad \text{for all } n \ge N_\epsilon.$$

Now $d(\theta, \theta_0) > 2^{j-1}\delta_n$ for $\theta \in S_j$ which implies, by (35), that Semi

$$M(\theta) - M(\theta_0) \lesssim -d^2(\theta, \theta_0) \lesssim -2^{2j-2}\delta_n^2 \qquad \text{for } \theta \in S_j \tag{36}$$

or $\sup_{\theta \in S_j} [M(\theta) - M(\theta_0)] \lesssim -2^{2j-2}\delta_n^2$. Thus, the event $\sup_{\theta \in S_j}[\mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0)] \geq 0$ can only happen if $\mathbb{M}_n$ and M are not too close. Let

$$U_n(\theta) := \mathbb{M}_n(\theta) - M(\theta), \qquad \text{for } \theta \in \Theta.$$

It follows from (36) that

$$\begin{aligned}
\mathbb{P}\Big(\sup_{\theta \in S_j}[\mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0)] \geq 0\Big) &\leq \mathbb{P}\Big(\sup_{\theta \in S_j}[U_n(\theta) - U_n(\theta_0)] \gtrsim 2^{2j-2}\delta_n^2\Big) \\
&\leq \mathbb{P}\left(\sup_{\theta : d(\theta,\theta_0) \leq 2^j\delta_n} [U_n(\theta) - U_n(\theta_0)] \gtrsim 2^{2j-2}\delta_n^2\right) \\
&\lesssim \frac{1}{2^{2j-2}\delta_n^2}\mathbb{E}\left[\sup_{\theta : d(\theta,\theta_0) \leq 2^j\delta_n} (U_n(\theta) - U_n(\theta_0))\right].
\end{aligned}$$

$P(Z > t) \leq \frac{E(Z)}{t}$

Markov ine

Suppose that there is a function $\phi_n(\cdot)$ such that

$$\mathbb{E}\left[\sup_{\theta : d(\theta,\theta_0) \leq u} \sqrt{n}(U_n(\theta) - U_n(\theta_0))\right] \lesssim \phi_n(u) \qquad \text{for every } u > 0. \tag{37}$$

We thus get

$$\mathbb{P}\Big(2^{j-1}\delta_n < d(\hat{\theta}_n, \theta_0) \leq 2^j\delta_n\Big) \lesssim \frac{\phi_n(2^j\delta_n)}{\sqrt{n}2^{2j}\delta_n^2}$$

$\sqrt{n}\,2^{2j-2}$

for every j. As a consequence,

$$\mathbb{P}\Big(d(\hat{\theta}_n, \theta_0) > 2^M\delta_n\Big) \lesssim \frac{1}{\sqrt{n}}\sum_{j > M}\frac{\phi_n(2^j\delta_n)}{2^{2j}\delta_n^2}.$$

The following assumption on $\phi_n(\cdot)$ is usually made to simplify the expression above: there exists $\alpha < 2$ such that

$$\phi_n(cx) \leq c^\alpha\phi_n(x) \qquad \text{for all } c > 1 \text{ and } x > 0. \tag{38}$$

Under this assumption, we get

$$\mathbb{P}\Big(d(\hat{\theta}_n, \theta_0) > 2^M\delta_n\Big) \lesssim \frac{\phi_n(\delta_n)}{\sqrt{n}\delta_n^2}\sum_{j > M} 2^{j(\alpha-2)}.$$

The quantity $\sum_{j>M} 2^{j(\alpha-2)}$ converges to zero as $M \to \infty$. Observe that if we further assume that

$$\phi_n(\delta_n) \lesssim \sqrt{n}\delta_n^2, \qquad \text{as } n \text{ varies}, \tag{39}$$

then

$$\mathbb{P}\Big(d(\hat{\theta}_n, \theta_0) > 2^M \delta_n\Big) \le c \sum_{j>M} 2^{j(\alpha-2)},$$

for a constant $c > 0$ (which does not depend on n, M). Let u_M denote the right side of the last display. It follows therefore that, under assumptions (38) and (39), we get

$$d(\hat{\theta}_n, \theta_0) \le 2^M \delta_n \qquad \text{with probability at least } 1 - u_M, \quad \text{for all } n.$$

Further note that $u_M \to 0$ as $M \to \infty$. This gives us the following non-asymptotic rate of convergence theorem.

Theorem 5.1. *Let (Θ, d) be a semi-metric space. Fix $n \ge 1$. Let $\{\mathbb{M}_n(\theta) : \theta \in \Theta\}$ be a stochastic process and $\{M(\theta) : \theta \in \Theta\}$ be a deterministic process. Assume condition (35) and that the function $\phi_n(\cdot)$ satisfies (37) and (38). Then for every $M > 0$, we get $d(\hat{\theta}_n, \theta_0) \le 2^M \delta_n$ with probability at least $1 - u_M$ provided (39) holds. Here $u_M \to 0$ as $M \to \infty$.*

$M(\theta) - M(\theta_0) \le -d^2(\theta, \theta_0)$

$E[\sup(V_n - V)]$

$\le \phi(u)$

Suppose now that condition (35) holds only for θ in a neighborhood of θ_0 and that (37) holds only for small u. Then one can prove the following asymptotic result under the additional condition that $\hat{\theta}_n$ is consistent (i.e., $d(\hat{\theta}_n, \theta_0) \xrightarrow{\mathbb{P}} 0$).

Theorem 5.2 (Rate theorem). *Let Θ be a semi-metric space. Let $\{\mathbb{M}_n(\theta) : \theta \in \Theta\}$ be a stochastic process and $\{M(\theta) : \theta \in \Theta\}$ be a deterministic process. Assume that (35) is satisfied for every θ in a neighborhood of θ_0. Also, assume that for every n and sufficiently small u condition (37) holds for some function ϕ_n satisfying (38), and that (39) holds. If the sequence $\hat{\theta}_n$ satisfies $\mathbb{M}_n(\hat{\theta}_n) \ge \mathbb{M}_n(\theta_0) - O_{\mathbb{P}}(\delta_n^2)$ and if $\hat{\theta}_n$ is consistent in estimating θ_0, then $d(\hat{\theta}_n, \theta_0) = O_{\mathbb{P}}(\delta_n)$.*

Proof. The above result is Theorem 3.2.5 in [van der Vaart and Wellner, 1996] where you can find its proof. The proof is very similar to the proof of Theorem 5.1. The crucial observation is to realize that: for any $\eta > 0$,

$$\mathbb{P}\Big(d(\hat{\theta}_n, \theta_0) > 2^M \delta_n\Big) = \sum_{j>M, 2^j\delta_n \le \eta} \mathbb{P}\Big(2^{j-1}\delta_n < d(\hat{\theta}_n, \theta_0) \le 2^j \delta_n\Big) + \mathbb{P}\Big(2d(\hat{\theta}_n, \theta_0) > \eta\Big). \quad \to 0$$

$\le \sum_{j \ge M} 2^{2j(\alpha-2)}$

The first term can be tackled as before while the second term goes to zero by the consistency of $\hat{\theta}_n$. □

Remark 5.1. *In the case of i.i.d. data and criterion functions of the form $\mathbb{M}_n(\theta) = \mathbb{P}_n[m_\theta]$ and $M(\theta) = P[m_\theta]$, the centered and scaled process $\sqrt{n}(\mathbb{M}_n - M)(\theta) = \mathbb{G}_n[m_\theta]$*

def

equals the empirical process at m_θ. Condition (37) involves the suprema of the empirical process indexed by classes of functions

$$\mathcal{M}_u := \{m_\theta - m_{\theta_0} : d(\theta, \theta_0) \leq u\}.$$

$E(\sup(\quad))$

Thus, we need to find the existence of $\phi_n(\cdot)$ *such that* $\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{M}_u} \lesssim \phi_n(u)$.

Remark 5.2. *The above theorem gives the correct rate in fair generality, the main problem being to derive sharp bounds on the modulus of continuity of the empirical process. A simple, but not necessarily efficient, method is to apply the maximal inequalities (with and without bracketing). These yield bounds in terms of the uniform entropy integral* $J(1, \mathcal{M}_u, M_u)$ *or the bracketing integral* $J_{[\,]}(\|M_u\|_{P,2}, \mathcal{M}_u, L_2(P))$ *of the class* $\mathcal{M}_u$ *given by*

$\sup_{m \in \mathcal{M}} |\mathbb{G}_n m - G m|$

$$\mathbb{E}\big[\|\mathbb{G}_n\|_{\mathcal{M}_u}\big] \lesssim J(1, \mathcal{M}_u, M_u)[P(M_u^2)]^{1/2}$$

Theorem 4.8

where

$$J(1, \mathcal{M}_u, M_u) = \int_0^1 \sup_Q \sqrt{\log N(\epsilon\|M_u\|_{Q,2}, \mathcal{M}_u, L_2(Q))}\, d\epsilon$$

$(\int M_u^2\, dQ)^{\frac{1}{2}}$

and

$$\mathbb{E}\big[\|\mathbb{G}_n\|_{\mathcal{M}_u}\big] \lesssim J_{[\,]}(\|M_u\|, \mathcal{M}_u, L_2(P)),$$

where

$$J_{[\,]}(\delta, \mathcal{M}_u, L_2(P)) = \int_0^\delta \sqrt{\log N_{[\,]}(\epsilon, \mathcal{M}_u, L_2(P))}\, d\epsilon.$$

Theorem 4.11

Here M_u *is the envelope function of the class* $\mathcal{M}_u$. *In this case, we can take* $\phi_n^2(u) = P[M_u^2]$ *and this leads to a rate of convergence* δ_n *of at least the solution of*

Assumed: $\phi_n(u) \leq n\delta_n^2 \Rightarrow \phi_n(\delta)$

$$P[M_{\delta_n}^2] \sim n\delta_n^4.$$

Theorem 4.12

$[P(M_u^2)]^{\frac{1}{2}}$

Observe that the rate of convergence in this case is driven by the sizes of the envelope functions as $u \downarrow 0$, *and the size of the classes is important only to guarantee a finite entropy integral.*

Remark 5.3. *In genuinely infinite-dimensional situations, this approach could be less useful, as it is intuitively clear that the precise entropy must make a difference for the rate of convergence. In this situation, the the maximal inequalities obtained in Section 4 may be used.*

Remark 5.4. *For a Euclidean parameter space, the first condition of the theorem is satisfied if the map* $\theta \mapsto Pm_\theta$ *is twice continuously differentiable at the point of maximum* θ_0 *with a nonsingular second-derivative matrix.*

So Typhon works

5.1 Some examples

5.1.1 Euclidean parameter

Let $X_1, \ldots, X_n$ be i.i.d. random elements on $\mathcal{X}$ with a common law P, and let $\{m_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ be a class of measurable maps. Suppose that $\Theta \subset \mathbb{R}^d$, and that, for every $\theta_1, \theta_2 \in \Theta$ (or just in a neighborhood of θ_0),

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \le F(x)\|\theta_1 - \theta_2\| \qquad (40)$$

Lip constant

for some measurable function $F : \mathcal{X} \to \mathbb{R}$ with $PF^2 < \infty$. Then the class of functions $\mathcal{M}_\delta := \{m_\theta - m_{\theta_0} : \|\theta - \theta_0\| \le \delta\}$ has envelope function δF and bracketing number (see Theorem 2.14) satisfying

$$N_{[\,]}(2\epsilon\|F\|_{P,2}, \mathcal{M}_\delta, L_2(P)) \le N(\epsilon, \{\theta : \|\theta - \theta_0\| \le \delta\}, \|\cdot\|) \le \left(\frac{C\delta}{\epsilon}\right)^d,$$

where the last inequality follows from Lemma 2.7 coupled with the fact that the ϵ-covering number of δB (for any set B) is the ϵ/δ-covering number of B. In view of the maximal inequality with bracketing (see Theorem 11.4),

?

$$\mathbb{E}_P\big[\|\mathbb{G}_n\|_{\mathcal{M}_\delta}\big] \lesssim \int_0^{\delta\|F\|_{P,2}} \sqrt{\log N_{[\,]}(\epsilon, \mathcal{M}_\delta, L_2(P))}\, d\epsilon \lesssim \delta.$$

Thus Theorem 8.1 applies with $\phi_n(\delta) \asymp \delta$, and the inequality $\phi_n(\delta_n) \le \sqrt{n}\delta_n^2$ is solved by $\delta_n = 1/\sqrt{n}$. We conclude that the rate of convergence of $\hat\theta_n$ is $n^{-1/2}$ as soon as $P(m_\theta - m_{\theta_0}) \le -c\|\theta - \theta_0\|^2$, for every $\theta \in \Theta$ in a neighborhood of θ_0.

Example 5.3 (Least absolute deviation regression). *Given i.i.d. random vectors $Z_1, \ldots, Z_n$, and $e_1, \ldots, e_n$ in $\mathbb{R}^d$ and $\mathbb{R}$, respectively, let*

$$Y_i = \theta_0^\top Z_i + e_i.$$

The least absolute-deviation estimator $\hat\theta_n$ minimizes the function

$$= \frac{1}{n}\sum \operatorname{sign}[(\theta_0 - \theta)^T z_i + e_i]$$

$$\theta \mapsto \frac{1}{n}\sum_{i=1}^n |Y_i - \theta^\top Z_i| = \mathbb{P}_n m_\theta,$$

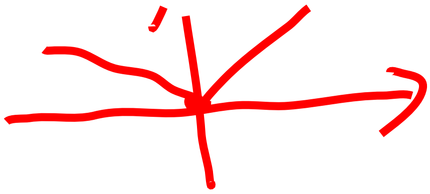
$$\frac{\partial|\theta^T z|}{\partial\theta} = \operatorname{sign}(\theta)$$

where $\mathbb{P}_n$ is the empirical measure of $X_i := (Z_i, Y_i)$, and $m_\theta(x) = |y - \theta^\top z|$.

Exercise (HW2): Show that the parameter θ_0 is a point of minimum of the map $\theta \mapsto P|Y - \theta^\top Z|$ if the distribution of the error e_1 has median zero. Furthermore, show that the maps $\theta \mapsto m_\theta$ satisfies condition (40):

$$\Big||y - \theta_1^\top z| - |y - \theta_2^\top z|\Big| \le \|\theta_1 - \theta_2\|\|z\|.$$

$$\text{LHS} \le \|y - \theta_1^T z - y + \theta_2^T z\| = \|(\theta_1 - \theta_2)^T z\| \le \text{RHS}.$$



$d(\theta, \theta_0) \to 0$

Argue the consistency of the least-absolute-deviation estimator from the convexity of the map $\theta \mapsto |y - \theta^\top z|$. Moreover, show that the map $\theta \mapsto P|Y - \theta^\top Z|$ is twice differentiable at θ_0 if the distribution of the errors has a positive density at its median. Furthermore, derive the rate of convergence of $\hat{\theta}_n$ in this situation. $\frac{1}{n}$

5.1.2 A non-standard example

Example 5.4 (Analysis of the shorth)**.** *Suppose that $X_1, \ldots, X_n$ are i.i.d. P on $\mathbb{R}$ with a differentiable density p with respect to the Lebesgue measure. Let F_X be the distribution function of X. Suppose that p is a unimodal (bounded) symmetric density with mode θ_0 (with $p'(x) > 0$ for $x < \theta_0$ and $p'(x) < 0$ for $x > \theta_0$). We want to estimate θ_0.*

θ_0

$|x-\theta| \le 1$

Exercise (HW2): Let

$$\mathbb{M}(\theta) := Pm_\theta = \mathbb{P}(|X - \theta| \le 1) = F_X(\theta + 1) - F_X(\theta - 1)$$

where $m_\theta(x) = \mathbf{1}_{[\theta-1,\theta+1]}(x)$. Show that $\theta_0 = \operatorname{argmax}_{\theta \in \mathbb{R}} \mathbb{M}(\theta)$. Thus, θ_0 is the center of an interval of length 2 that contains the largest possible (population) fraction of data points. We can estimate θ_0 by

$$\hat{\theta}_n := \operatorname*{argmax}_{\theta \in \mathbb{R}} \mathbb{M}_n(\theta), \qquad \textit{where} \qquad \mathbb{M}_n(\theta) = \mathbb{P}_n[m_\theta].$$

G-C Thm: $\sup |(\mathbb{M}_n - \mathbb{M})\theta| \xrightarrow{P} 0$. if θ_0 is well-separated

Show that $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$? The functions $m_\theta(x) = \mathbf{1}_{[\theta-1,\theta+1]}(x)$ are not Lipschitz in the parameter $\theta \in \Theta \equiv \mathbb{R}$. Nevertheless, the classes of functions $\mathcal{M}_\delta$ satisfy the conditions of Theorem 5.2. These classes have envelope function

$$\sup_{|\theta - \theta_0| \le \delta} \left| \mathbf{1}_{[\theta-1,\theta+1]} - \mathbf{1}_{[\theta_0-1,\theta_0+1]} \right| \le \mathbf{1}_{[\theta_0-1-\delta,\theta_0-1+\delta]} + \mathbf{1}_{[\theta_0+1-\delta,\theta_0+1+\delta]}.$$

3.5.1

The $L_2(P)$-norm of these functions is bounded above by a constant times $\sqrt{\delta}$. Thus, the conditions of the rate theorem are satisfied with $\phi_n(\delta) = c\sqrt{\delta}$ for some constant c, leading to a rate of convergence of $n^{-1/3}$. We will show later that $n^{1/3}(\hat{\theta}_n - \theta_0)$ converges in distribution to a non-normal limit as $n \to \infty$.

Example 5.5 (A toy change point problem)**.** *Suppose that we have i.i.d. data $\{X_i = (Z_i, Y_i) : i = 1, \ldots, n\}$ where $Z_i \sim \text{Unif}(0, 1)$ and*

$$Y_i = \mathbf{1}_{[0,\theta_0]}(Z_i) + \epsilon_i, \qquad \textit{for } i = 1, \ldots, n.$$

Here, ϵ_i's are the unobserved errors assumed to be i.i.d. $N(0, \sigma^2)$. Further, for simplicity, we assume that ϵ_i is independent of Z_i. The goal is to estimate the unknown

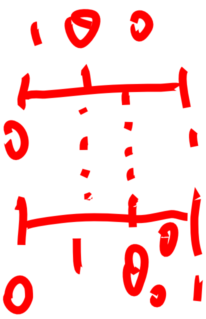
$\left(I_{[0,\theta_0]}(X) - I_{[0,\theta)}(X) \right)^2$

parameter $\theta_0 \in (0, 1)$. *A natural procedure is to consider the least squares estimator:*

$$\hat{\theta}_n := \operatorname*{argmin}_{\theta \in [0,1]} \mathbb{P}_n[(Y - \mathbf{1}_{[0,\theta]}(X))^2].$$

Exercise (HW2): Show that $\hat{\theta}_n := \mathrm{argmax}_{\theta\in[0,1]} \mathbb{M}_n(\theta)$ *where*

point-wise compact set $[0,1]$



$$\mathbb{M}_n(\theta) := \mathbb{P}_n[(Y - 1/2)\{\mathbf{1}_{[0,\theta]}(X) - \mathbf{1}_{[0,\theta_0]}(X)\}].$$

Prove that $\mathbb{M}_n$ *converges uniformly to* $\Leftarrow$ G-C entropy

$$M(\theta) := P[(Y - 1/2)\{\mathbf{1}_{[0,\theta]}(X) - \mathbf{1}_{[0,\theta_0]}(X)\}].$$

Show that $M(\theta) = |\theta - \theta_0|/2$*. As a consequence, show that* $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$. if θ_0 is well-sep 3.5.1

To find the rate of convergence of $\hat{\theta}_n$ *we consider the metric* $d(\theta_1, \theta_2) := \sqrt{|\theta_1 - \theta_2|}$*. Show that the conditions needed to apply Theorem 5.2 hold with this choice of* $d(\cdot, \cdot)$*. Using Theorem 5.2 derive that* $n(\hat{\theta}_n - \theta_0) = O_{\mathbb{P}}(1)$.

$M(\theta) - M(\theta_0) \le -d^2(\theta, \theta_0) = -|\theta - \theta_0|$

$-\frac{1}{2}|\theta - \theta_0| =$

5.1.3 Persistency in high-dimensional regression

Let $Z^i := (Y^i, X^i_1, \ldots, X^i_p)$, $i = 1, \ldots, n$, be i.i.d. random vectors, where $Z^i \sim P$. It is desired to predict Y by $\sum_j \beta_j X_j$, where $(\beta_1, \ldots, \beta_p) \in B_n \subset \mathbb{R}^p$, under a prediction loss. We assume that $p = n^\alpha$, $\alpha > 0$, that is, there could be many more explanatory variables than observations. We consider sets B_n restricted by the maximal number of non-zero coefficients of their members, or by their l_1-radius. We study the following asymptotic question: how 'large' may the set B_n be, so that it is still possible to select empirically a predictor whose risk under P is close to that of the best predictor in the set?

We formulate this problem using a triangular array setup, i.e., we model the observations $Z^1_n, \ldots, Z^n_n$ as i.i.d. random vectors in $\mathbb{R}^{p_n+1}$, having distribution P_n (that depends on n). In the following we will hide the dependence on n and just write $Z^1, \ldots, Z^n$. We will consider B_n of the form

set of non-zero coef

$$B_{n,b} := \{\beta \in \mathbb{R}^{p_n} : \|\beta\|_1 \le b\}, \tag{41}$$

where $\|\cdot\|_1$ denotes the l_1-norm. For any $Z := (Y, X_1, \ldots, X_p) \sim P$, we will denote the expected prediction error by

$$L_P(\beta) := \mathbb{E}_P\Big[(Y - \sum_{j=1}^p \beta_j X_j)^2\Big] = \mathbb{E}_P\Big[(Y - \beta^\top X)^2\Big]$$

where $X = (X_1, \ldots, X_p)$. The best linear predictor, where $Z \sim P_n$, is given by

$$\beta_n^* := \arg\min_{\beta \in B_{n,b_n}} L_{P_n}(\beta),$$

for some sequence of $\{b_n\}_{n\geq 1}$. We estimate the best linear predictor β_n^* from the sample by

mse

$$\hat{\beta}_n := \arg\min_{\beta \in B_{n,b_n}} L_{\mathbb{P}_n}(\beta) = \arg\min_{\beta \in B_{n,b_n}} \underline{\frac{1}{n}\sum_{i=1}^{n}(Y^i - \beta^\top X^i)^2},$$

where $\mathbb{P}_n$ is the empirical measure of the Z^i's. We say that $\hat{\beta}_n$ is *persistent* (relative to B_{n,b_n} and P_n) ([Greenshtein and Ritov, 2004]) if and only if

$$L_{P_n}(\hat{\beta}_n) - L_{P_n}(\beta_n^*) \xrightarrow{\mathbb{P}} 0.$$

This is certainly a weak notion of "risk-consistency" — we are only trying to consistently estimate the expected predictor error. However, note that this notion does not require any modeling assumptions on the (joint) distribution of Z (in particular, we are not assuming that there is a 'true' linear model). The following theorem is a version of Theorem 3 in [Greenshtein and Ritov, 2004].

Theorem 5.6. *Suppose that $p_n = n^\alpha$, where $\alpha > 0$. Let*

$$F(Z^i) := \max_{0\leq j\leq p} |X_j^i X_k^i - \mathbb{E}_{P_n}(X_j^i X_k^i)|, \quad \textit{where we take } X_0^i = Y^i, \textit{ for } i = 1, \ldots, n.$$

Suppose that $\mathbb{E}_{P_n}[F^2(Z^1)] \leq M < \infty$, for all n. Then for $b_n = o((n/\log n)^{1/4})$, $\hat{\beta}_n$ is persistent relative to B_{n,b_n}.

Proof. From the definition of β_n^* and $\hat{\beta}_n$ it follows that

$$L_{P_n}(\hat{\beta}_n) - L_{P_n}(\beta_n^*) \geq 0, \qquad \text{and} \qquad L_{\mathbb{P}_n}(\hat{\beta}_n) - L_{\mathbb{P}_n}(\beta_n^*) \leq 0.$$

Thus,

<?

$$\begin{aligned}
0 &\leq L_{P_n}(\hat{\beta}_n) - L_{P_n}(\beta_n^*) \\
&= \Big(L_{P_n}(\hat{\beta}_n) - L_{\mathbb{P}_n}(\hat{\beta}_n)\Big) + \Big(L_{\mathbb{P}_n}(\hat{\beta}_n) - L_{\mathbb{P}_n}(\beta_n^*)\Big) + \Big(L_{\mathbb{P}_n}(\beta_n^*) - L_{P_n}(\beta_n^*)\Big) \\
&\leq 2 \sup_{\beta \in B_{n,b_n}} |L_{\mathbb{P}_n}(\beta) - L_{P_n}(\beta)|,
\end{aligned}$$

Intercept

where we have used the fact that $L_{\mathbb{P}_n}(\hat{\beta}_n) - L_{\mathbb{P}_n}(\beta_n^*) \leq 0$. To simply our notation, let $\gamma = (-1, \beta) \in \mathbb{R}^{p_n+1}$. Then $L_{P_n}(\beta) = \gamma^\top \Sigma_{P_n} \gamma$ and $L_{\mathbb{P}_n}(\beta) = \gamma^\top \Sigma_{\mathbb{P}_n} \gamma$ where $\Sigma_{P_n} = \Big(E_{P_n}(X_j^1 X_k^1)\Big)_{0\leq j,k\leq p_n}$ and $\Sigma_{\mathbb{P}_n} = \Big(\frac{1}{n}\sum_{i=1}^n X_j^i X_k^i\Big)_{0\leq j,k\leq p_n}$. Thus,

$$|L_{\mathbb{P}_n}(\beta) - L_{P_n}(\beta)| \leq |\gamma^\top(\Sigma_{\mathbb{P}_n} - \Sigma_{P_n})\gamma| \leq \|\Sigma_{\mathbb{P}_n} - \Sigma_{P_n}\|_\infty \|\gamma\|_1^2,$$

?

where $\|\Sigma_{\mathbb{P}_n} - \Sigma_{P_n}\|_\infty = \sup_{0 \le j,k \le p_n} \left| \frac{1}{n} \sum_{i=1}^n X_j^i X_k^i - E_{P_n}(X_j^1 X_k^1) \right|$. Therefore,

$$
\begin{aligned}
\mathbb{P}\big(L_{P_n}(\hat\beta_n) - L_{P_n}(\beta_n^*) > \epsilon\big) &\le \mathbb{P}\Big(2 \sup_{\beta \in B_{n,b_n}} |L_{\mathbb{P}_n}(\beta) - L_{P_n}(\beta)| > \epsilon\Big) \\
&\le \mathbb{P}\Big(2(b_n+1)^2 \|\Sigma_{\mathbb{P}_n} - \Sigma_{P_n}\|_\infty > \epsilon\Big) \\
&\le \frac{2(b_n+1)^2}{\epsilon} \mathbb{E}\Big[\|\Sigma_{\mathbb{P}_n} - \Sigma_{P_n}\|_\infty\Big]. \qquad (42)
\end{aligned}
$$

def. $r = (b, \beta)$ $|\beta| \le b$ $\Longrightarrow$

Markov ineq.

Let $\mathcal{F} = \{f_{j,k} : 0 \le j,k \le p_n\}$ where $f_{j,k}(z) := x_j x_k - E_{P_n}(X_j^1 X_k^1)$ and $z = (x_0, x_1, \ldots, x_{p_n})$. Observe that $\|\Sigma_{\mathbb{P}_n} - \Sigma_{P_n}\|_\infty = \|\mathbb{P}_n - P_n\|_{\mathcal{F}}$. We will now use the following maximal inequality with bracketing entropy:

$$
\mathbb{E}\|\sqrt{n}(\mathbb{P}_n - P)\|_{\mathcal{F}} \lesssim J_{[\,]}(1, \mathcal{F}, L_2(P_n)) \|F_n\|_{P_n,2},
$$

$= \sqrt{\int F_n^2 dP}$ $\le \sqrt{M}$

where F_n is an envelope of $\mathcal{F}$. Note that F_n can be taken as F (defined in the statement of the theorem). We can obviously cover $\mathcal{F}$ with the ϵ-brackets $[f_{j,k} - \epsilon/2, f_{j,k} + \epsilon/2]$, for every $\epsilon > 0$, and thus, $\log N_{[\,]}(\epsilon, \mathcal{F}, L_2(P_n)) \le 2\log(p_n + 1)$. Therefore, using (42) and the maximal inequality above,

def

$$
\mathbb{P}\big(L_{P_n}(\hat\beta_n) - L_{P_n}(\beta_n^*) > \epsilon\big) \lesssim \frac{2(b_n+1)^2}{\epsilon} \frac{\sqrt{2\log(p_n+1)}}{\sqrt{n}} \sqrt{M} \lesssim \frac{b_n^2 \sqrt{\alpha \log n}}{\sqrt{n}} \to 0,
$$

as $n \to \infty$, by the assumption on b_n. $\square$

$\int_0^1 \sqrt{\log N_{[\,]}(\epsilon, \mathcal{F}, L_2(P_n))}\, d\epsilon$