Chapter 5

Convexity

Very rough draft

5.1 Approximation of convex functions

Convexity::approximation

Expand Convexity simplifies arguments from Chapter 3. Reasons: local minima are global minima and pointwise convergence of a sequnce of convex functions implies uniform convergence on compacta. The first Lemma in this Section contains the standard result (Rockafellar 1970, Theorem 10.8) that pointwise convergence of a sequence of (deterministic) convex functions $\{H_n\}$ to a limit function B on an open set G implies uniform convergence on each compact subset of G.

For a subset H of $\mathbb{R}^k$ and $\delta > 0$, write H^δ for the set $\{y : d(y, H) \le \delta\}$.

For a bounded subset T of $\mathbb{R}^k$, call a finite subset $T_0 \subset T$ a δ-net for T if $d(t, T_0) \le \delta$ for ever t in T. Equivalently, the union of closed balls $\cup_{t\in T_0} B(t, \delta)$ covers T. Also equivalently, $B(t, \delta) \cap T_0 \ne \emptyset$ for each t in T.

Define $T^\delta := \{t \in \mathbb{R}^k : d(t, T) \le \delta\}$.

bound <1> **Lemma.** *(based on Pollard 1991) Let K be a compact,convex subset of $\mathbb{R}^d$. Let $H(\cdot)$ and $A(\cdot)$ be real-valued functions defined on $K^{4\delta}$, with H convex and A satisfying the uniform continuity requirement*

$$|A(s) - A(t)| \le \eta \qquad \text{whenever } |s - t| \le 2\delta \text{ and } s, t \in K^{4\delta}.$$

Let T_0 be a δ-net for $K^{3\delta}$. Then

$$\sup_{s\in K} |H(s) - A(s)| \le 4\eta + 3 \max_{t\in T_0} |H(t) - A(t)|$$

Proof Write Δ for the maximum of $|H(t) - A(t)|$ over the finite set T_0. The asserted inequality will follow from a pair of bounds for A:

<2> $H(s) \le A(s) + \eta + \Delta$ for all s in $K^{2\delta}$

<3> $H(s) \ge A(s) - 4\eta - 3\Delta$ for all s in K

Proof of upper bound (2).

Let s be a point of $K^{2\delta}$. First I prove that s lies in the convex hull of the finite set

$$T(s) := \{t \in T : |s - t| \le 2\delta\} = T_0 \cap B(s, 2\delta).$$

Argue by contradiction. If s is not in the convex hull then, by the separating hyperplane theorem, there exists some unit vector u for which

$$u's > u't \qquad \text{for all } t \text{ in } T(s)$$

The closed ball $B(s^*, \delta)$ around the point $s^* = s + \delta u$ lies inside $B(s, 2\delta) \cap K^{4\delta}$ and it contains no points from T_0, which contradicts the δ-net property of T_0.

It follows that s must be a convex combination of points in T_s. That is, $s = \sum_{t \in T(s)} \alpha_t t$ with $\alpha_t \ge 0$ and $\sum_t \alpha_t = 1$. Then

$$\begin{aligned} H(s) &\le \sum_t \alpha_t H(t) \qquad \text{by convexity of } H \\ &\le \sum_t \alpha_t \left(A(t) + \Delta\right) \qquad \text{definition of } \Delta \\ &\le \sum_t \alpha_t \left(A(s) + \eta + \Delta\right) \qquad \text{because } |s - t| \le 2\delta \text{ for } t \in T(s) \\ &\le A(s) + \eta + \Delta. \end{aligned}$$

Proof of lower bound <3>.

Let s be a point of K. Find a point of T_0 with $|s - t| \le \delta$. Define $s^* = t - (s - t)$. Then $s^* \in T(s) \subseteq K^{2\delta}$ and $t = (s + s^*)/2$. By convexity of H,

$$H(t) \le \left(H(s) + H(s^*)\right)/2$$

From <2>,

$$H(s^*) \le A(s^*) + \eta + \Delta,$$

and, by definition of Δ,

$$H(t) \ge A(t) - \Delta$$

Combine the last three inequalities to get

$$2A(t) - 2\Delta \le H(s) + A(s^*) + \eta + \Delta.$$

Both $A(s^*)$ and $A(t)$ lie within η of $A(s)$. Inequality <3> follows.

□

5.2 Minimizers of convex functions

Convexity::argmin

Suppose M is a strictly convex function on $\mathbb{R}^k$ satisfing the growth condition

growth<4>
$$M(t)/|t| \to \infty \qquad \text{as } |t| \to \infty.$$

For each z in $\mathbb{R}^k$, the function $M(t,z) := M(t) - z't$ achieves its minimum value $M^*(z) = \inf_{t\in\mathbb{R}^k} M(t) - z't$ at a unique point, $\psi_M(z)$.

I claim that the map $z \mapsto \psi_M(z)$ is continuous. I would also like to show that if $M(t^*,z) \le M^*(z)+\epsilon$ then t^* must be close to $\psi_M(z)$ in some suitably uniform sense. What else would be needed to establish a more general form of Lemma <5>?

5.3 A limit theorem for M-estimators

Convexity::limit

stoch.approximation <5> **Lemma.** *Suppose $\{H_n(t) : t \in \mathbb{R}^k\}$ is a stochastic process with convex sample paths. Assume*

(i) $H_n(\hat{t}_n) \le \inf_{t\in\mathbb{R}^k} H_n(t) + o_p(1)$ for some random $\hat{t}_n$

(ii) there exists a sequence of random vectors $\{Z_n\}$ of order $O_p(1)$ for which

$$H_n(t) + Z_n't \to \tfrac{1}{2}|t|^2 \qquad \text{in probability, for each fixed } t \in \mathbb{R}^k.$$

Then $\hat{t}_n = Z_n + o_p(1)$.

Proof Define

$$A_n(t) = \tfrac{1}{2}|t|^2 - Z_n't = \tfrac{1}{2}|Z_n - t|^2 - \tfrac{1}{2}|Z_n|^2,$$

so that $|H_n(t) - A_n(t)| \to 0$ in probability, for each fixed t. Note that A_n is minimized at Z_n, which will imply that H_n is minimized near Z_n, with high probability.

Suppose $\epsilon > 0$ and $\eta > 0$ are given, small quantities. By (ii), there exists an $R > 0$ for which $\mathbb{P}\{|Z_n| > R\} < \epsilon$ eventually. Choose a $\delta < 1/4$ such that

$$|A_n(s) - A_n(t)| \le \eta^2 \qquad \text{when } |s-t| \le 2\delta \text{ and } |Z_n| \le R \text{ and } s,t \in B(0,R+1)$$

Let T_0 be a δ-net for $B(0, R+1)$. Then there exists a set Ω_n with probability eventually greater than $1 - 2\epsilon$, on which

$$\begin{aligned} |Z_n| &\le R \\ \Delta_n := \max_{t\in T_0} |H_n(t) - A_n(t)| &\le \eta^2 \\ H_n(\hat{t}_n) &\le H_n(Z_n) + \eta^2 \end{aligned}$$

On Ω_n, Lemma <1> implies that

$$|H_n(t) - A_n(t)| \le 7\eta^2 \qquad \text{for all } |t| \le R.$$

In particular, for all unit vectors v,

$$H_n(Z_n) - 7\eta^2 \le A_n(Z_n) = A_n(Z_n + 6\eta) - 18\eta^2 \le H_n(Z_n + 6\eta v) - 11\eta^2$$

Convexity of $r \mapsto H_n(Z_n + rv)$ then implies that

$$H_n(Z_n) + 4\eta^2 \le H_n(Z_n + rv) \qquad \text{for all } r \ge 6\eta \text{ and all unt vectors } v,$$

which forces $\hat{t}_n$ to lie within the ball of radius 6η around Z_n.

□

I would like to strengthen the Lemma by replacing (ii) by: there exists a sequence of random vectors $\{Z_n\}$ of order $O_p(1)$ for which

$$H_n(t) + Z_n' t \to M(t) \qquad \text{in probability, for each fixed } t \in \mathbb{R}^k,$$

where M is a strictly convex function satisfying the growth condition <4>.

I would like to conclude something like

$$\hat{t}_n = \Psi_M(Z_n) + o_p(1)$$

If we had $Z_n \rightsquigarrow Z$ then we would get $\hat{t}_n \rightsquigarrow \psi_M(Z)$, a very clean limit theorem.

---

random.sum <6> **Theorem.** *Let $G_n(\theta) = \sum_{i\le n} g_{n,i}(\theta)$ be a random process with convex sample paths defined on $\mathbb{R}^k$. Suppose there exist points θ_n in $\mathbb{R}^k$ and positive definite matrices J_n for which:*

(i) $\widehat{\theta}_n$ is an estimator for which $G_n(\widehat{\theta}_n) \le \inf_{\theta\in\Theta} G_n(\theta) + o_p(1)$

(ii) *the function* $\theta \mapsto \mathbb{P}G_n(\theta)$ *is minimized at* θ_n

(iii) *each* $\{g_{n,i}\}$ *has a linear approximation near* θ_n*,*

$$g_{n,i}(\theta) = g_{n,i}(\theta_n) + (\theta - \theta_n)' D_{n,i} + r_{n,i}(\theta),$$

where the $D_{n,i}$ *are random vectors with zero expected value*

(iv) $\mathbb{P}G_n(\theta_n + J_n^{-1/2}t) = \frac{1}{2}|t|^2 + o(1)$ *for each* t

(v) $var\left(\sum_{i\le n} r_{n,i}(\theta_n + J_n^{-1/2}t)\right) = o(1)$ *for each* t

(vi) $Z_n := -J_n^{-1/2}\sum_{i\le n} D_{n,i} = O_p(1)$

Then $J_n^{1/2}(\widehat{\theta}_n - \theta_n) = Z_n + o_p(1)$.

Try to replace (iv) by $\mathbb{P}G_n(\theta_n + J_n^{-1/2}t) = M(t) + o(1)$ for each t, with M as in Section 2. Maybe the Theorem will then be strong enough to cover the lasso (Example <17>).

PROOF The standardized random vector $\hat{t}_n = J_n^{1/2}(\widehat{\theta}_n - \theta_n)$ comes within $o_p(1)$ of minimizing the process

$$\begin{aligned} H_n(t) &:= G_n(\theta_n + J_n^{-1/2}t) - G_n(\theta_n) \\ &= -t'Z_n + \textstyle\sum_{i\le n} r_{n,i}(\theta_n + J_n^{-1/2}t). \end{aligned}$$

For a fixed t, assumption (v) ensures that the sum contributed by the $\{r_{n,i}\}$ lies within $o_p(1)$ of its expectation,

$$\begin{aligned} &\mathbb{P}\textstyle\sum_{i\le n} r_{n,i}(\theta_n + J_n^{-1/2}t) \\ &\qquad = \mathbb{P}G_n(\theta_n + J_n^{-1/2}t) - \mathbb{P}G_n(\theta_n) \qquad \text{because } \mathbb{P}D_{n,i} = 0 \\ &\qquad = \tfrac{1}{2}|t|^2 + o(1) \qquad \text{by (iv)} \end{aligned}$$

Thus $H_n(t) + t'Z_n = \frac{1}{2}|t|^2 + o_p(1)$ for each fixed t. An appeal to Lemma <??> completes the proof. □

Remarks. The theorem is almost true if θ ranges over only a subset Θ of $\mathbb{R}^d$, instead of over the whole of $\mathbb{R}^d$.

5.4 Examples

Convexity::examples

median <7> **Example.** Suppose $X_1, X_2, \ldots$ are independent observations from a probability distribution P on the real line with sample median θ_0. Suppose also that there is a neighborhood of θ_0 in which P has a continuous, strictly positive density (with respect to Lebesgue measure). Let $\widehat{\theta}_n$ be the sample median. Show that $2f(\theta_0)(\widehat{\theta}_n - \theta_0) \rightsquigarrow N(0,1)$.

□

spatial.median <8> **Example.** Generalize Example <7> to cover argmin $\sum_i |t - X_i|$ in higher dimensions.

□

LAD.regression <9> **Example.** Theorem 1 of Pollard (1991) for L1 regression.

□

exp.fam <10> **Example.** Modify the following result from Dou, Pollard, and Zhou (2009), but with improved argument from talk at Wellner conference:

Let $\{Q_\lambda : \lambda \in \mathbb{R}\}$ be an exponential family of probability measures with densities $dQ_\lambda/dQ_0 = f_\lambda(y) = \exp(\lambda y - \psi(\lambda))$. Remember that $e^{\psi(\lambda)} = Q_0 e^{\lambda y}$ and that the distribution Q_λ has mean $\psi^{(1)}(\lambda)$ and variance $\psi^{(2)}(\lambda)$.

We assume:

psi3 ($\psi 3$) There exists an increasing real function G on $\mathbb{R}^+$ such that

$$|\psi^{(3)}(\lambda + h)| \le \psi^{(2)}(\lambda) G(|h|) \qquad \text{for all } \lambda \text{ and } h$$

Without loss of generality we assume $G(0) \ge 1$.

psi2 ($\psi 2$) For each $\epsilon > 0$ there exists a finite constant C_ϵ for which $\psi^{(2)}(\lambda) \le C_\epsilon \exp(\epsilon\lambda^2)$ for all $\lambda \in \mathbb{R}$. Equivalently, $\psi^{(2)}(\lambda) \le \exp\big(o(\lambda^2)\big)$ as $|\lambda| \to \infty$.

As shown in Section **??**, these assumptions on the ψ function imply that

hell<11> $$h^2(Q_\lambda, Q_{\lambda+\delta}) \le \delta^2 \psi^{(2)}(\lambda)\,(1 + |\delta|)\, G(|\delta|) \qquad \text{for all } \lambda, \delta \in \mathbb{R}.$$

Remark. We may assume that $\psi^{(2)}(\lambda) > 0$ for every real λ. Otherwise we would have $0 = \psi^{(2)}(\lambda_0) = \text{var}_{\lambda_0}(y) = \nu f_{\lambda_0}(y)(y - \psi^{(1)}(\lambda_0))^2$ for some λ_0, which would make $y = \psi^{(1)}(\lambda_0)$ for ν almost all y and $Q_\lambda \equiv Q_{\lambda_0}$ for every λ.

The theory in this section combine ideas from Portnoy (1988) and from Hjort and Pollard (1993). We write our results in a notation that makes the applications in Section **??** and **??** more straightforward. The notational cost is that the parameters are indexed by $\{0, 1, \ldots, N\}$. To avoid an excess of parentheses we write N_+ for $N+1$. In the applications N changes with the sample size n and $\mathbb{Q}$ is replaced by $\mathbb{Q}_{n,a,\mathbb{B},N}$ or $\widetilde{\mathbb{Q}}_{n,a,\mathbb{B},N}$.

Suppose $\xi_1, \ldots, \xi_n$ are (nonrandom) vectors in $\mathbb{R}^{N_+}$. Suppose $\mathbb{Q} = \otimes_{i\le n} Q_{\lambda_i}$ with $\lambda_i = \xi_i'\gamma$ for a fixed $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_N)$ in $\mathbb{R}^{N_+}$. Under $\mathbb{Q}$, the coordinate maps $y_1, \ldots, y_n$ are independent random variables with $y_i \sim Q_{\lambda_i}$.

The log-likelihood for fitting the model is

$$L_n(g) = \sum\nolimits_{i\le n} (\xi_i' g) y_i - \psi(\xi_i' g) \qquad \text{for } g \in \mathbb{R}^{N_+},$$

which is maximized (over $\mathbb{R}^{N_+}$) at the MLE $\widehat{g}$ $(= \widehat{g}_n)$.

Remark. As a small amount of extra bookkeeping in the following argument would show, we do not need $\widehat{g}$ to exactly maximize L_n. It would suffice to have $L_n(\widehat{g})$ suitably close to $\sup_g L_n(g)$. In particular, we need not be concerned with questions regarding existence or uniqueness of the argmax.

Define

(i) $J_n = \sum_{i\le n} \xi_i \xi_i' \psi^{(2)}(\lambda_i)$, an $N_+ \times N_+$ matrix

(ii) $w_i := J_n^{-1/2} \xi_i$, an element of $\mathbb{R}^{N_+}$

(iii) $W_n = \sum_{i\le n} w_i \big(y_i - \psi^{(1)}(\lambda_i)\big)$, an element of $\mathbb{R}^{N_+}$

Notice that $\mathbb{Q}W_n = 0$ and $\operatorname{var}_{\mathbb{Q}}(W_n) = \sum_{i\le n} w_i w_i' \psi^{(2)}(\lambda_i) = I_{N_+}$ and

$$\mathbb{Q}|W_n|^2 = \operatorname{trace}\left(\operatorname{var}_{\mathbb{Q}}(W_n)\right) = N_+.$$

mle.approx<12> **Lemma.** *Suppose* $0 < \epsilon_1 \le 1/2$ *and* $0 < \epsilon_2 < 1$ *and*

$$\max\nolimits_{i\le n} |w_i| \le \frac{\epsilon_1 \epsilon_2}{2G(1) N_+} \qquad \textit{with } G \textit{ as in Assumption } (\psi 3).$$

Then $\widehat{g} = \gamma + J_n^{-1/2}(W_n + r_n)$ *with* $|r_n| \le \epsilon_1$ *on the set* $\{|W_n| \le \sqrt{N_+/\epsilon_2}\}$, *which has* $\mathbb{Q}$*-probability greater than* $1 - \epsilon_2$.

PROOF The equality $\mathbb{Q}|W_n|^2 = N_+$ and Tchebychev give $\mathbb{Q}\{|W_n| > \sqrt{N_+/\epsilon_2}\} \leq \epsilon_2$.

Reparametrize by defining $t = J_n^{1/2}(g - \gamma)$. The concave function

$$\mathcal{L}_n(t) := L_n(\gamma + J_n^{-1/2}t) - L_n(\gamma) = \sum\nolimits_{i\leq n} y_i w_i' t + \psi(\lambda_i) - \psi(\lambda_i + w_i' t)$$

is maximized at $\widehat{t}_n = J_n^{1/2}(\widehat{g} - \gamma)$. It has derivative

$$\dot{\mathcal{L}}_n(t) = \sum\nolimits_{i\leq n} w_i \left(y_i - \psi^{(1)}(\lambda_i + w_i' t)\right).$$

For a fixed unit vector $u \in \mathbb{R}^{N_+}$ and a fixed $t \in \mathbb{R}^{N_+}$, consider the real-valued function of the real variable s,

$$H(s) := u'\dot{\mathcal{L}}_n(st) = \sum\nolimits_{i\leq n} u'w_i \left(y_i - \psi^{(1)}(\lambda_i + sw_i' t)\right),$$

which has derivatives

$$\begin{aligned}
\dot{H}(s) &= -\sum\nolimits_{i\leq n} (u'w_i)(w_i' t)\psi^{(2)}(\lambda_i + sw_i' t) \\
\ddot{H}(s) &= -\sum\nolimits_{i\leq n} (u'w_i)(w_i' t)^2\psi^{(3)}(\lambda_i + sw_i' t).
\end{aligned}$$

Notice that $H(0) = u'W_n$ and $\dot{H}(0) = -u'\sum_{i\leq n} w_i w_i'\psi^{(2)}(\lambda_i)t = -u't$.

Write M_n for $\max_{i\leq n}|w_i|$. By virtue of Assumption (ψ3),

$$\begin{aligned}
|\ddot{H}(s)| &\leq \sum\nolimits_{i\leq n} |u'w_i|(w_i' t)^2\psi^{(2)}(\lambda_i)G\left(|sw_i' t|\right) \\
&\leq M_n G\left(M_n|st|\right) t' \sum\nolimits_{i\leq n} w_i w_i'\psi^{(2)}(\lambda_i)t \\
&= M_n G\left(M_n|st|\right)|t|^2.
\end{aligned}$$

By Taylor expansion, for some $0 < s^* < 1$,

$$|H(1) - H(0) - \dot{H}(0)| \leq \tfrac{1}{2}|\ddot{H}(s^*)| \leq \tfrac{1}{2} M_n G\left(M_n|t|\right)|t|^2.$$

That is,

u'll<13>
$$\left|u'\left(\dot{\mathcal{L}}_n(t) - W_n + t\right)\right| \leq \tfrac{1}{2} M_n G\left(M_n|t|\right)|t|^2.$$

Approximation <13> will control the behavior of $\widetilde{\mathcal{L}}(s) := \mathcal{L}_n(W_n + su)$, a concave function of the real argument s, for each unit vector u. By concavity, the derivative

$$\dot{\widetilde{\mathcal{L}}}(s) = u'\dot{\mathcal{L}}_n(W_n + su) = -s + R(s)$$

is a decreasing function of s with

$$|R(s)| \le \tfrac{1}{2} M_n G\left(M_n|W_n + su|\right)|W_n + su|^2$$

On the set $\{|W_n| \le \sqrt{N_+/\epsilon_2}\}$ we have

$$|W_n \pm \epsilon_1 u| \le \sqrt{N_+/\epsilon_2} + \epsilon_1.$$

Thus

$$M_n|W_n \pm \epsilon_1 u| \le \frac{\epsilon_1\epsilon_2}{2G(1)N_+}\left(\sqrt{N_+/\epsilon_2} + \epsilon_1\right) < 1,$$

implying

$$\begin{aligned}|R(\pm\epsilon_1)| &\le \tfrac{1}{2} M_n G(1)|W_n \pm \epsilon_1 u|^2 \\ &\le \frac{\epsilon_1\epsilon_2}{G(1)N_+}\left(N_+/\epsilon_2 + \epsilon_1^2\right) \\ &\le \epsilon_1\left(1 + \epsilon_1^2\epsilon_2/N_+\right) < \tfrac{5}{8}\epsilon_1.\end{aligned}$$

Deduce that

$$\begin{aligned}\dot{\tilde{\mathcal{L}}}(\epsilon_1) &= -\epsilon_1 + R(\epsilon_1) \le -\tfrac{3}{8}\epsilon_1 \\ \dot{\tilde{\mathcal{L}}}(-\epsilon_1) &= \epsilon_1 + R(-\epsilon_1) \ge \tfrac{3}{8}\epsilon_1\end{aligned}$$

The concave function $s \mapsto \mathcal{L}_n(W_n + su)$ must achieve its maximum for some s in the interval $[-\epsilon_1, \epsilon_1]$, for each unit vector u. It follows that $|\widehat{t}_n - W_n| \le \epsilon_1$.

□

AnBn<14> **Corollary.** *Suppose $\xi_i = D\eta_i$ for some nonsingular matrix D, so that*

$$J_n = nDA_nD \qquad \text{where} A_n := \frac{1}{n}\sum\nolimits_{i\le n} \eta_i\eta_i'\psi^{(2)}(\lambda_i).$$

If B_n is another nonsingular matrix for which

assumptionA<15>
$$\|A_n - B_n\|_2 \le \left(2\left\|B_n^{-1}\right\|_2\right)^{-1}$$

and if

max.eta<16>
$$\max_{i\le n}|\eta_i| \le \frac{\epsilon\sqrt{n}/N_+}{G(1)\sqrt{32\left\|B_n^{-1}\right\|_2}} \qquad \text{for some } 0 < \epsilon < 1$$

then for each set of vectors $\kappa_0, \ldots, \kappa_N$ in $\mathbb{R}^{N_+}$ there is a set $\mathcal{Y}_{\kappa,\epsilon}$ with $\mathbb{Q}\mathcal{Y}_{\kappa,\epsilon}^c < 2\epsilon$ on which

$$\sum\nolimits_{0\le j\le N}|\kappa_j'(\widehat{g} - \gamma)|^2 \le \frac{6\left\|B_n^{-1}\right\|_2}{n\epsilon}\sum\nolimits_{0\le j\le N}|D^{-1}\kappa_j|^2.$$

Remark. For our applications of the Corollary in Sections **??** and **??**, we need $D = \text{diag}(D_0, D_1, \ldots, D_N)$ and $\kappa_j = e_j$, the unit vector with a 1 in its jth position, for $j \le m$ and $\kappa_j = 0$ for $j > m$. In our companion paper we will need the more general κ_j's.

PROOF First we establish a bound on the spectral distance between A_n^{-1} and B_n^{-1}. Define $H = B_n^{-1}A_n - I$. Then$\|H\|_2 \le \left\|B_n^{-1}\right\|_2 \|A_n - B_n\|_2 \le 1/2$, which justifies the expansion

$$\left\|A_n^{-1} - B_n^{-1}\right\|_2 = \left\|\left((I+H)^{-1} - I\right) B_n^{-1}\right\|_2 \le \sum\nolimits_{j\ge 1} \|H\|_2^k \left\|B_n^{-1}\right\|_2 \le \left\|B_n^{-1}\right\|_2 .$$

As a consequence, $\left\|A_n^{-1}\right\|_2 \le 2\left\|B_n^{-1}\right\|_2$.

Choose $\epsilon_1 = 1/2$ and $\epsilon_2 = \epsilon$ in Lemma 12. The bound on $\max_{i\le n}|\eta_i|$ gives the bound on $\max_{i\le n}|w_i|$ needed by the Lemma:

$$n|w_i|^2 = \eta_i' D(J_n/n)^{-1} D\eta_i = \eta_i' A_n^{-1}\eta_i \le \left\|A_n^{-1}\right\|_2 |\eta_i|^2.$$

Define $K_j := J_n^{-1/2}\kappa_j$, so that $|\kappa_j'(\widehat{g}-\gamma)|^2 \le 2(K_j'W_n)^2 + 2(K_j'r_n)^2$. By Cauchy-Schwarz,

$$\sum\nolimits_j (K_j'r_n)^2 \le \sum\nolimits_j |K_j|^2 |r_n|^2 = U_\kappa |r_n|^2$$

where

$$\begin{aligned} U_\kappa := \sum\nolimits_j \kappa_j' J_n^{-1}\kappa_j &= \sum\nolimits_j n^{-1}(D^{-1}\kappa_j)' A_n^{-1} D^{-1}\kappa_j \\ &\le 2n^{-1}\left\|B_n^{-1}\right\|_2 \sum\nolimits_j |D^{-1}\kappa_j|^2. \end{aligned}$$

For the contribution $V_\kappa := \sum_j |K_j'W_n|^2$ the Cauchy-Schwarz bound is too crude. Instead, notice that $\mathbb{Q}V_\kappa = U_\kappa$, which ensures that the complement of the set

$$\mathcal{Y}_{\kappa,\epsilon} := \{|W_n| \le \sqrt{N_+/\epsilon}\} \cap \{V_\kappa \le U_\kappa/\epsilon\}$$

has $\mathbb{Q}$ probability less that 2ϵ. On the set $\mathcal{Y}_{\kappa,\epsilon}$,

$$\sum\nolimits_{0\le j\le N} |\kappa_j'(\widehat{g}-\gamma)|^2 \le 2V_\kappa + 2U_\kappa |r_n|^2 \le 3U_\kappa/\epsilon.$$

The asserted bound follows. □

□

lasso <17> **Example.** Try to apply Theorem <6> to cover Theorem 2 of Knight and Fu (2000). Comment that more recent results with $p = p_n$ are much closer to the way the lasso is used.

□

Notes

Convexity: Huber (1964)
Brown (1985) and Appendix A of Maritz (1981).
Andersen and Gill (1982)
Haberman (1989)
Niemiro (1992)
Bickel, Klaassen, Ritov, and Wellner (1993, pp 328, 473, 519)
Ritov (1987)
Jurečková (1977)

References

Andersen, P. K. and R. D. Gill (1982). Cox's regression model for counting processes: a large sample study. *Annals of Statistics 10*, 1100–1120.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Brown, B. M. (1985). Multiparameter linearization theorems. *Journal of the Royal Statistical Society, Series B 47*, 323–331.

Dou, W., D. Pollard, and H. H. Zhou (2009). Functional regression for general exponential families. Technical report, Yale University Statistics Department.

Haberman, S. J. (1989). Concavity and estimation. *Annals of Statistics 17*, 1631–1661.

Hjort, N. L. and D. Pollard (1993). Asymptotics for minimisers of convex processes. Technical report, Yale University.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics 35*, 73–101.

Jurečková, J. (1977). Asymptotic relations of m-estimates and r-estimates in linear regression model. *Annals of Statistics 5*, 464–472.

Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *Annals of Statistics 28*(5), 1356–1378.

Maritz, J. S. (1981). *Distribution-Free Statistical Methods*. Chapman and Hall.

Niemiro, W. (1992). Asymptotics for m-estimators defined by convex minimization. *Annals of Statistics 20*, 1514–1533.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory 7*, 186–199.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Annals of Statistics 16*(1), 356–366.

Ritov, Y. (1987). Tightness of monotone random fields. *Journal of the Royal Statistical Society, Series B 49*, 331–333.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton, New Jersey: Princeton Univ. Press.