代数幾何と学習理論 6.4節

Loading [MathJax]/jax/output/HTML-CSS/jax.js

6.4 ML and MAP

前節まではBayes推定量を扱った.

この節は ML(最尤法; Maximum likelihood) と MAP(事後確率最大化法; Maximum a posteriori) を含むクラスの推定量を考察する：

$$\hat{w}_n \in \underset{w \in W}{\arg\min} \sum_{i=1}^{n} f(X_i, w) + a_n \sigma(w),$$

ここで

$$\{a_n\}_{n=1}^{\infty} : \text{positive, non-decreasing},$$

$$\sigma(w) \geq 0, \ \forall w \in W.$$

特に

$$\begin{cases} a_n = 0 & ; \quad \sigma = (\text{なんでも}) & \rightsquigarrow \text{ML}, \\ a_n = 1 & ; \quad \sigma(w) = -\log \varphi(w) - (\min_{w'} \log \varphi(w')) & \rightsquigarrow \text{MAP} \end{cases}$$

(φ
は事前分布.$\sigma(w) \geq 0$
を仮定できるのはK
がコンパクトだから)

6.4 ML and MAP 道筋

まずはともあれ一致性を示す.

Def(参考). (一致性)

推定量 $\hat{w}_n$ が $w_0 \in W_0$ に対して一致性を持つ

$\overset{\Delta}{\Leftrightarrow}$

$$\hat{w}_n \xrightarrow{p} W_0$$

i.e.

$$\forall M > 0, P(|\hat{w}_n - w_0| \geq M) \to 0(n \to \infty)$$

6.4 ML and MAP 道筋

まずはともあれ一致性を示す.

Def. (集合への一致性; ここだけ？)

推定量 $\hat{w}_n$ が $W_0 \subset W$ に対して一致性を持つ

$\overset{\Delta}{\Leftrightarrow}$

$$\hat{w}_n \xrightarrow{p} W_0. \text{ (注意: ここだけ？)}$$

i.e.

$$P(\hat{w}_n \notin W_0) \to 0(n \to \infty)$$

6.4 ML and MAP 道筋

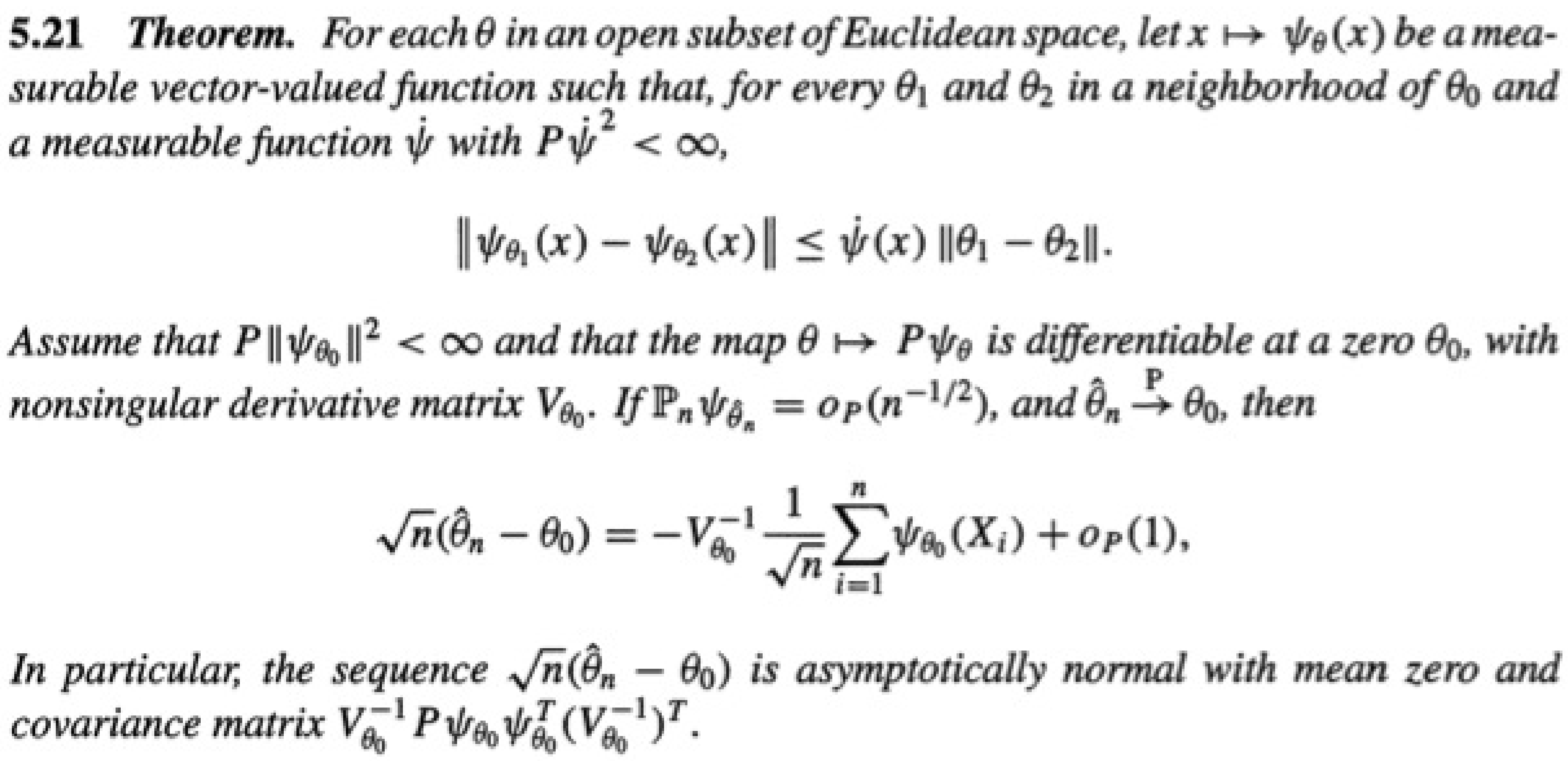
一致性を示したら，近づき方を見る.

一致性から，推定量の変動は0に近づくので，そのまま極限に飛ばすと変動の様子は取り出せない

$\rightsquigarrow$ 収束先との差をn倍や$\sqrt{n}$倍などした後に，$n \to \infty$して収束を調べる(確率収束や分布収束，とくに分布収束).

6.4 ML and MAP 道筋

Thm(参考). (正則な場合の，漸近正規性の定理の例)



5.21 Theorem. *For each θ in an open subset of Euclidean space, let $x \mapsto \psi_\theta(x)$ be a measurable vector-valued function such that, for every θ_1 and θ_2 in a neighborhood of θ_0 and a measurable function $\dot\psi$ with $P\dot\psi^2 < \infty$,*

$$\left\|\psi_{\theta_1}(x) - \psi_{\theta_2}(x)\right\| \leq \dot\psi(x)\,\|\theta_1 - \theta_2\|.$$

Assume that $P\|\psi_{\theta_0}\|^2 < \infty$ and that the map $\theta \mapsto P\psi_\theta$ is differentiable at a zero θ_0, with nonsingular derivative matrix V_{θ_0}. If $\mathbb{P}_n\psi_{\hat\theta_n} = o_P(n^{-1/2})$, and $\hat\theta_n \xrightarrow{\mathrm{P}} \theta_0$, then

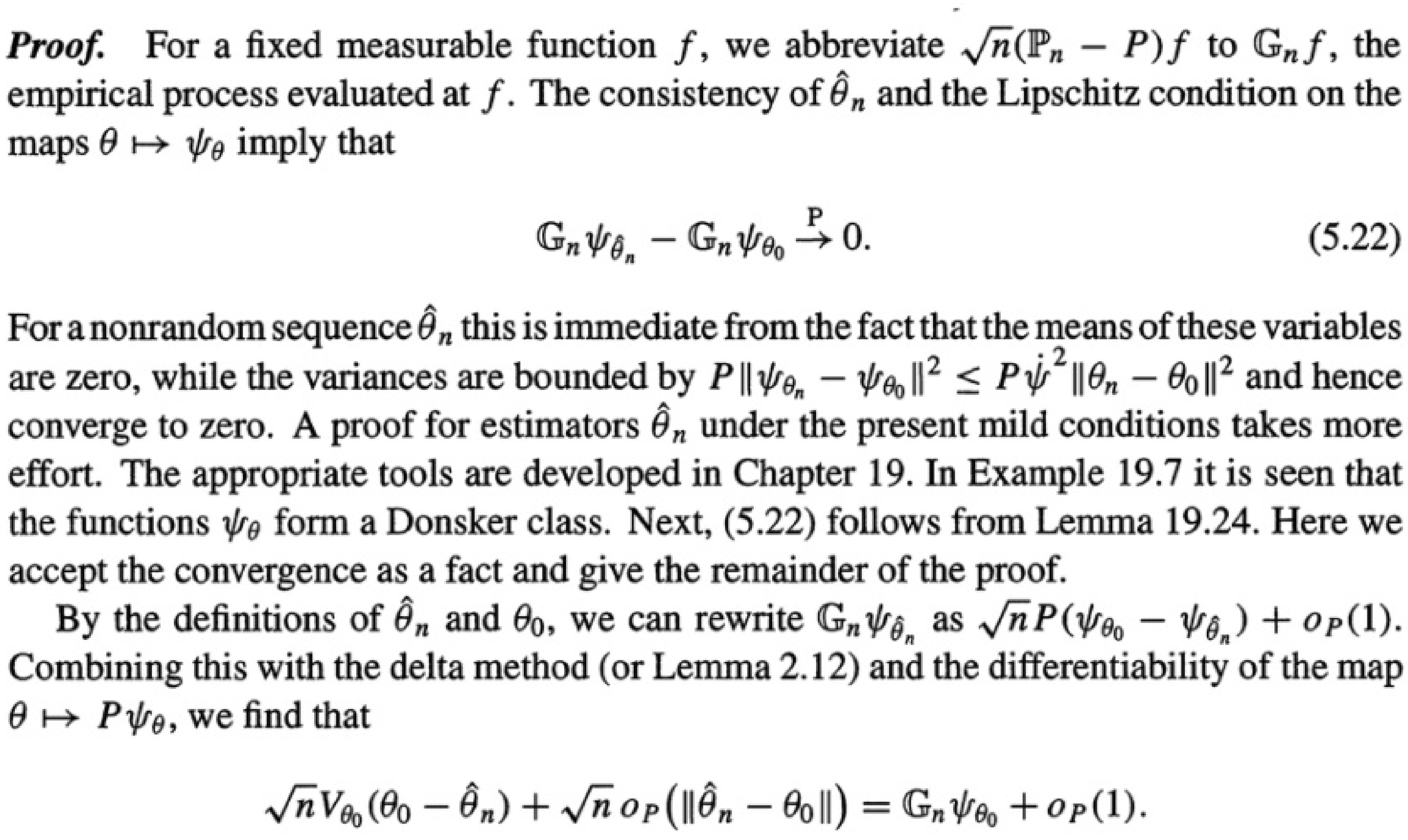
$$\sqrt{n}(\hat\theta_n - \theta_0) = -V_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{\theta_0}(X_i) + o_P(1),$$

In particular, the sequence $\sqrt{n}(\hat\theta_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix $V_{\theta_0}^{-1}P\psi_{\theta_0}\psi_{\theta_0}^T(V_{\theta_0}^{-1})^T$.

(source: [1])

6.4 ML and MAP 道筋

Thm(参考). (正則な場合の，漸近正規性の定理の例)



Proof. For a fixed measurable function f, we abbreviate $\sqrt{n}(\mathbb{P}_n - P)f$ to $\mathbb{G}_n f$, the empirical process evaluated at f. The consistency of $\hat\theta_n$ and the Lipschitz condition on the maps $\theta \mapsto \psi_\theta$ imply that

$$\mathbb{G}_n \psi_{\hat\theta_n} - \mathbb{G}_n \psi_{\theta_0} \xrightarrow{\mathrm{P}} 0. \tag{5.22}$$

For a nonrandom sequence $\hat\theta_n$ this is immediate from the fact that the means of these variables are zero, while the variances are bounded by $P\|\psi_{\theta_n} - \psi_{\theta_0}\|^2 \le P\dot\psi^2 \|\theta_n - \theta_0\|^2$ and hence converge to zero. A proof for estimators $\hat\theta_n$ under the present mild conditions takes more effort. The appropriate tools are developed in Chapter 19. In Example 19.7 it is seen that the functions ψ_θ form a Donsker class. Next, (5.22) follows from Lemma 19.24. Here we accept the convergence as a fact and give the remainder of the proof.

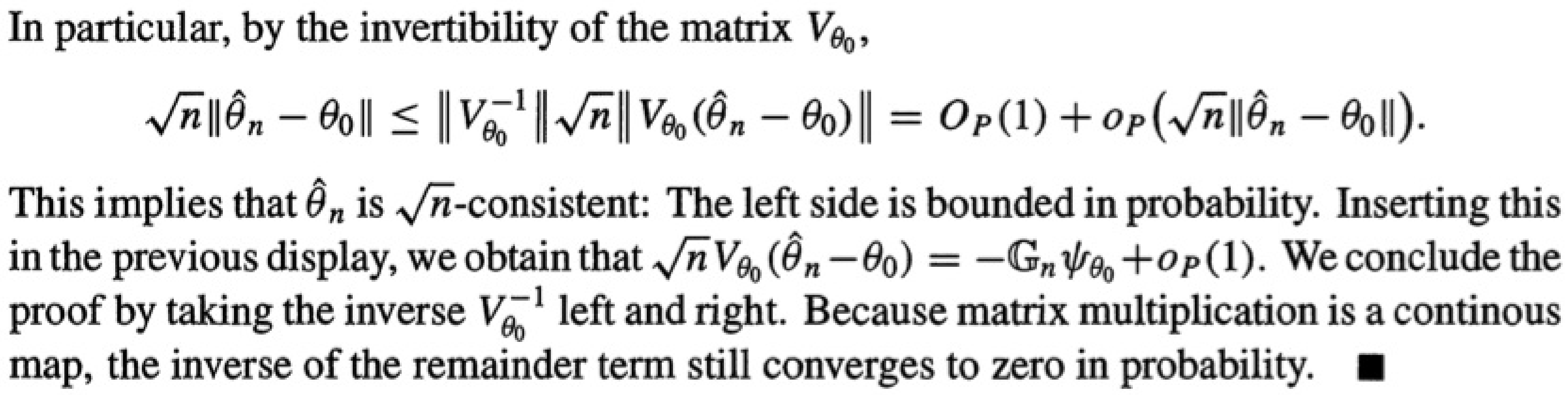
By the definitions of $\hat\theta_n$ and θ_0, we can rewrite $\mathbb{G}_n \psi_{\hat\theta_n}$ as $\sqrt{n}P(\psi_{\theta_0} - \psi_{\hat\theta_n}) + o_P(1)$. Combining this with the delta method (or Lemma 2.12) and the differentiability of the map $\theta \mapsto P\psi_\theta$, we find that

$$\sqrt{n}V_{\theta_0}(\theta_0 - \hat\theta_n) + \sqrt{n}\, o_P\big(\|\hat\theta_n - \theta_0\|\big) = \mathbb{G}_n \psi_{\theta_0} + o_P(1).$$

(source: [1])

6.4 ML and MAP 道筋

Thm(参考). (正則な場合の，漸近正規性の定理の例)



In particular, by the invertibility of the matrix V_{θ_0},

$$\sqrt{n}\|\hat{\theta}_n - \theta_0\| \leq \left\|V_{\theta_0}^{-1}\right\|\sqrt{n}\left\|V_{\theta_0}(\hat{\theta}_n - \theta_0)\right\| = O_P(1) + o_P\left(\sqrt{n}\|\hat{\theta}_n - \theta_0\|\right).$$

This implies that $\hat{\theta}_n$ is $\sqrt{n}$-consistent: The left side is bounded in probability. Inserting this in the previous display, we obtain that $\sqrt{n}V_{\theta_0}(\hat{\theta}_n - \theta_0) = -\mathbb{G}_n\psi_{\theta_0} + o_P(1)$. We conclude the proof by taking the inverse $V_{\theta_0}^{-1}$ left and right. Because matrix multiplication is a continous map, the inverse of the remainder term still converges to zero in probability. ■

(source: [1])

6.4 ML and MAP 道筋

今の場合，一致性は$\forall \epsilon > 0,\ P(K(\hat{w}_n) > \epsilon) \to 0$をみる．　漸近挙動は$R_g := K(\hat{w})$や$R_t := K_n(\hat{w})$の挙動を見る[*]

$$E_q[R_g], E_q[R_t]$$

を調べる．これらは漸近的には0に近づく．そのためn倍に拡大した挙動を調べる．

[*] $\sqrt{n}(\hat{w}_n - w_0)$ の漸近挙動ではなくR_gの挙動を知りたいのは，w_0に興味がないからだろう．

6.4 ML and MAP 道筋

$$nE_q[R_g],$$
$$nE_q[R_t]$$

これは(サンプルの分布による積分ひいては)$\hat{w}_n$ という特異点を持つ集合(W_0)に確率収束をしていくような確率変数の分布での積分になっている．(check → @todo)

$\hat{w}_n$ 自身が分布収束しない[1]ので，積分の極限操作ができない．そのため，特異点を正規交差型に解消した方のパラメーター空間で

$$\xi_n(u) := \sum_{i=1}^{n} \frac{K(w) - f(X_i, w)}{\sqrt{nK(w)}}$$

を調べる(ξ_n なら分布収束し，しかも積分の極限移行ができる[2])．

[1] 特異点のせいで？check @todo

[2] Why?

Reference

[1] van der Vaart, A. W. Asymptotic Statistics. (Cambridge University Press, 2000).

$$E[nK(\hat{w})]_{\{\|\psi_n\|^2>n\epsilon\}} \leq \frac{C'}{n^2}, \tag{6.77}$$

C' C

$$E[nK_n(\hat{w})]_{\{\|\psi_n\|^2>n\epsilon\}} \leq \frac{1}{2}E[3nK(\hat{w}) + \|\psi_n\|^2]_{\{\|\psi_n\|^2>n\epsilon\}} \leq \frac{C''}{n^2}. \tag{6.78}$$

$$E[nK_n(\hat{u})]_{\{\|\xi_n\|^2>n^\delta\}} \leq (1/2)E[3nK(\hat{w}) + \|\xi_n\|^2]_{\{\|\xi_n\|^2>n^\delta\}} \tag{6.79}$$

$$\leq \frac{c_6}{n^{\delta(s/2-1)}}. \tag{6.80}$$

$$|nR_\mathrm{g}| \leq 2\|R_1\|, \tag{6.82}$$

$$\underset{\text{(LHS)}}{|nR_\mathrm{t}|} \leq 2\|R_1\|. \tag{6.83}$$

$\forall v : \xi_n(0,v)>0$, $$\xi_n(0, v)^2/4 - 2\|R_1\| \leq nR_\mathrm{g} \leq \xi_n(0, v)^2/4 + 2\|R_1\|, \tag{6.86}$$

$|nR_g - \frac{1}{4}\xi_n(0,v)^2$

$\forall v : \xi_n(0,v)>0$, $$-\xi_n(0, v)^2/4 - 2\|R_1\| \leq nR_\mathrm{t} \leq -\xi_n(0, v)^2/4 + 2\|R_1\|. \tag{6.87}$$

$|nR_t + \frac{1}{4}\xi_n$