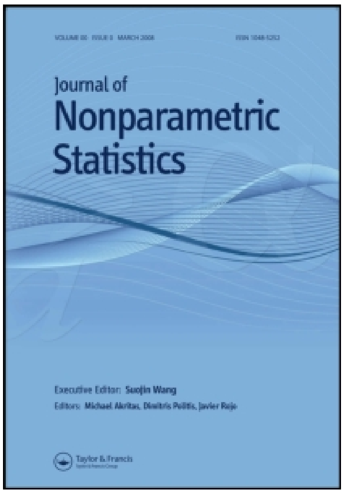

Journal of Nonparametric Statistics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713645758

Strong approximations for resample quantile processes and application to ROC methodology

Jiezhun Gu [a]; Subhashis Ghosal [b]
[a] Duke Clinical Research Institute, Duke University Medical Center, Durham, NC, USA [b] Department of Statistics, North Carolina State University, Raleigh, NC, USA

Online Publication Date: 01 April 2008

To cite this Article Gu, Jiezhun and Ghosal, Subhashis(2008)'Strong approximations for resample quantile processes and application to ROC methodology',Journal of Nonparametric Statistics,20:3,229 — 240

To link to this Article: DOI: 10.1080/10485250801954128

URL: http://dx.doi.org/10.1080/10485250801954128

Journal of Nonparametric Statistics
Vol. 20, No. 3, April 2008, 229–240

Strong approximations for resample quantile processes and application to ROC methodology

Jiezhun Gu[a]* and Subhashis Ghosal[b]

[a]*Duke Clinical Research Institute, Duke University Medical Center, Durham, NC, USA;* [b]*Department of Statistics, North Carolina State University, Raleigh, NC, USA*

(*Final version received 06 December 2007*)

The receiver operating characteristic (ROC) curve is defined as true positive rate versus false positive rate obtained by varying a decision threshold criterion. It has been widely used in medical sciences for its ability to measure the accuracy of diagnostic or prognostic tests. Mathematically speaking, ROC curve is the composition of survival function of one population with the quantile function of another population. In this paper, we study strong approximation for the quantile processes of the Bayesian bootstrap (BB) resampling distributions, and use this result to study strong approximations for the BB version of the ROC process in terms of two independent Kiefer processes. The results imply asymptotically accurate coverage probabilities for the confidence bands for the ROC curve and confidence intervals for the area under the curve functional of the ROC constructed using the BB method. Similar results follow for the bootstrap resampling distribution.

Keywords: Bayesian bootstrap; Kiefer process; ROC curve; strong approximation; quantile process

1. Introduction

Originally introduced in the context of electronic signal detection [1], the receiver operating characteristic (ROC) curve, which is a plot of the true positive rate versus the false positive rate by varying all the possible threshold values, has become an indispensable tool for measuring the accuracy of diagnostic tests since the 1970s [2]. The main attractions of ROC curve may be described by the following properties: (1) one ROC curve can display all the decision characteristic information of a diagnostic test in a unit graph, which can be used for classification purposes; (2) different diagnostic tests can be compared by their corresponding ROC curves. For example, the data set analysed by Wieand *et al.* [3] was based on 51 patients as the control group diagnosed as pancreatitis and 90 patients as the case group diagnosed as pancreatic cancer by two biomarkers, cancer antigen (CA 125) and a carbohydrate antigen (CA 19-9), respectively. The purpose was to decide which biomarker would better distinguish the case group from the control group on a certain range of false positive rate (0, 0.2). One of the solutions is to estimate and compare two ROC curves and their corresponding confidence bands for the two biomarkers. If one lies

*Corresponding author. Email: jiezhun.gu@duke.edu

ISSN 1048-5252 print/ISSN 1029-0311 online

DOI: 10.1080/10485250801954128
http://www.informaworld.com

above another, then we can conclude that one is better than another. The area under the curve (AUC) functional of ROC can be interpreted as the probability that the diagnostic value of a randomly chosen patient with the positive condition (usually referring to a disease) is greater than the diagnostic value of a randomly chosen patient without the positive condition [4]. A non-parametric estimator of ROC may be obtained by substituting into the empirical counterparts, and its variability may be estimated by bootstrap [5]. More recently, Gu *et al.* [6] proposed a smoother estimator and related confidence bands by using the Bayesian bootstrap (BB) method. These authors treated the BB distribution as the posterior distribution of the unknown quantities and make inference based on the 'BB posterior distribution', which actually corresponds to the posterior distribution with respect to Dirichlet prior with precision parameters converging to zero. Interestingly, the smoothness of their estimator comes from resampling and does not involve selecting a bandwidth. The BB estimator of ROC curve and its AUC performs well compared to some existing semiparametric and non-parametric methods.

We will first obtain strong approximations for quantile processes of the BB resampling distribution by a sequence of appropriate Gaussian processes (more specifically, Kiefer processes). The reason for employing the strong approximation theory for this study instead of the weak convergence theory is discussed in Remark 2 of Section 4. The existing strong approximation theories primarily include the following findings:

- Komlós *et al.* [7] showed that the uniform empirical process can be strongly approximated by a sequence of Brownian bridges obtained from a single Kiefer process;
- Csörgő and Révész [8] proved that under suitable conditions quantile process can be strongly approximated by a Kiefer process;
- Lo [9] studied the strong approximation theory for the cumulative distribution function (c.d.f.) of bootstrap and BB processes;
- Hsieh and Turnbull [10] studied the strong approximation for the empirical ROC estimate on a fixed subinterval of (0, 1).

The ROC function is $R(t) = \bar{G}(\bar{F}^{-1}(t))$, where $\bar{F}(x) = 1 - F(x)$ and $\bar{G}(y) = 1 - G(y)$ are the survival functions of independent variables $X \sim F$ and $Y \sim G$, respectively. Combining these results, we shall develop strong approximations for the BB version of the ROC process, its AUC and other functionals. In particular, these results imply Gaussian weak limits for the processes and asymptotically valid coverage probabilities for the resulting confidence bands and intervals. Interestingly, it will be seen that the forms of these Gaussian approximations are identical, and therefore the BB distribution of the ROC function, conditioned on the samples, is identical to the Gaussian approximation of the empirical ROC estimate. This means that frequentist variability of the empirical estimate of ROC can be asymptotically accurately estimated by resampling variability of the BB procedures, given that the samples and the 'posterior' mean (i.e. mean of the BB distribution of the ROC curve) are asymptotically equivalent to that of the empirical estimate up to the first order ($O(N^{-1/2})$), where N is the total sample size. Also, the result implies that for functionals like AUC, the empirical estimator is asymptotically normal and asymptotically equivalent to the BB estimator, and the corresponding confidence intervals have asymptotic frequentist validity. Graphical displays of these findings will be shown in the simulation studies in Section 5. Similar results hold for the bootstrap resampling distributions; see the technical report of Gu and Ghosal [11] for this and more detailed proofs.

Our strong approximation results imply that the estimator proposed by Gu *et al.* [6], namely, the mean of the BB distribution, is asymptotically equivalent to the empirical estimator of ROC, and converges to the true ROC curve at the rate $N^{-1/2}$. This, along with the smoothness of the estimator, provides strong justifications in favour of the estimator proposed there. Further, our results establish that confidence bands for the ROC function and confidence intervals for the AUC and other functionals of the ROC constructed using the BB method have asymptotically correct coverage probabilities. Moreover, our result on strong approximation of the quantile function of the BB and bootstrap distribution completes a gap in strong approximation theory and may be useful in various contexts other than the ROC methodology.

2. Notation

For $X_1, \dots, X_n \sim$ i.i.d. F, define the domain of X as $[a, b]$: $a = \sup\{x : F(x) = 0\}$, $b = \inf\{x : F(x) = 1\}$; the order statistic $X_{j:n}$, $j = 0, 1, \dots, n+1$, where $X_{0:n} = a$ and $X_{n+1:n} = b$, similarly define $V_{j:n}$ and $U_{j:n}$ for independent uniform samples $V_1, \dots, V_n$ and $U_1, \dots, U_n$; the quantile function $F^{-1}(t) = \inf\{x : F(x) \geq t\}, t \in [0, 1]$; for a given $0 < y < 1, k = \lceil ny \rceil$, where $\lceil \cdot \rceil$ is the ceiling function, that is, the smallest integer greater than or equal to x; $U(0, 1)$ denotes the uniform distribution on [0, 1]; abbreviate almost sure convergence by $a.s.$; define the error rate $l_n = n^{-1/4}(\log\log n)^{1/4}(\log n)^{1/2}$. Conditions A [8] and B are assumed, which hold for a large class of distribution functions.

Condition A The continuous distribution function F is twice differentiable on (a, b), $F' = f \neq 0$ on (a, b) and for some $\gamma > 0$, $\sup_{a<x<b} F(x)(1 - F(x))|f'(x)/f^2(x)| \leq \gamma$.

Condition B Let F and G satisfy Condition A and let the following conditions hold:

$$\sup_{a<x<b} F(x)(1-F(x))\left|\frac{g'(x)}{f^2(x)}\right|, \quad \sup_{a<x<b} F(x)(1-F(x))\left|\frac{g(x)}{f(x)}\right| \text{ are bounded.}$$

We define the following empirical, quantile functions and processes:

$$\text{empirical function: } \mathbb{F}_n(x) = \frac{j}{n}, \quad \text{if } X_{j:n} \leq x < X_{j+1:n}, \ j = 0, 1, \dots, n,$$

$$\text{empirical process: } \mathbb{J}_n(x) = \sqrt{n}(\mathbb{F}_n(x) - F(x)),$$

$$\text{quantile function: } \mathbb{F}_n^{-1}(y) = X_{k:n} = F^{-1}(U_{k:n}), \text{ where } k = \lceil ny \rceil,$$

$$\text{quantile process: } \mathbb{Q}_n(y) = \sqrt{n}(\mathbb{F}_n^{-1}(y) - F^{-1}(y)).$$

Based on sample $X_1, \dots, X_n$, BB's c.d.f. is defined as $\mathbb{F}_n^{\#}(x) = \sum_{1\leq j\leq n} \Delta_{j:n} 1(X_{j:n} \leq x)$, $x \in \mathbb{R}$, where $V_1, \dots, V_{n-1} \sim$ i.i.d. $U(0, 1)$, independent of X_is, $\Delta_{j:n} = V_{j:n-1} - V_{j-1:n-1}$, the quantile function, empirical and quantile processes of the BB resampling distributions are as follows:

$$\text{quantile function of BB: } \mathbb{F}_n^{\#-1}(y) = \begin{cases} X_{j:n}, & V_{j-1:n-1} < y \leq V_{j:n-1}, \\ X_{0:n}, & y = 0, \ j = 1, 2, \dots, n \end{cases}$$

$$\text{empirical process of BB: } \mathbb{J}_n^{\#}(x) = \sqrt{n}(\mathbb{F}_n^{\#}(x) - \mathbb{F}_n(x)),$$

$$\text{quantile process of BB: } \mathbb{Q}_n^{\#}(y) = \sqrt{n}(\mathbb{F}_n^{\#-1}(y) - \mathbb{F}_n^{-1}(y)).$$

Set $\bar{\mathbb{F}}_n^{\#}(x) = 1 - \mathbb{F}_n^{\#}(x)$ and $\bar{\mathbb{F}}_n^{\#-1}(y) = \mathbb{F}_n^{\#-1}(1-y)$, and similarly for G. Let $\mathbb{U}_n(x)$, $\mathbb{H}_n(x)$, $\mathbb{U}_n^{-1}(y)$, $\mathbb{W}_n(y)$, $\mathbb{H}_n^{\#}(x)$, $\mathbb{W}_n^{\#}(y)$, respectively, stand for $\mathbb{F}_n(x)$, $\mathbb{J}_n(x)$, $\mathbb{F}_n^{-1}(y)$, $\mathbb{Q}_n(y)$, $\mathbb{J}_n^{\#}(x)$, $\mathbb{Q}_n^{\#}(y)$ when F is replaced by $U(0, 1)$.

3. Strong approximations for the BB quantile processes

THEOREM 3.1 *Let $X_1, \ldots, X_n \sim$ i.i.d. F satisfying Condition A. Then the quantile process of BB resampling distribution $\mathbb{Q}_n^{\#}(y)$ can be strongly approximated by a Kiefer process K,*

$$\sup_{\delta_n^{\#} \le y \le 1-\delta_n^{\#}} \mid f(F^{-1}(y))\mathbb{Q}_n^{\#}(y) - n^{-1/2}K(y,n) \mid =_{a.s.} O(l_n), \tag{1}$$

where $\delta_n^{\#} = \delta_n + n^{-1/2}(\log\log n)^{1/2}$, $\delta_n = 25n^{-1}\log\log n$.

The proof requires the following Lemma whose proof is deferred to Section 6.

LEMMA 3.1 *Let $X_1, \ldots, X_n \sim$ i.i.d. F satisfying Condition A, $V_1, \ldots, V_{n-1} \sim$ i.i.d. $U(0,1)$, independent of $X_1, \ldots, X_n$, $V_1, \ldots, V_{n-1}$'s empirical function denoted as $\tilde{\mathbb{U}}_{n-1}(x)$. Then*

$$\sup_{0<y<1} \sqrt{n} \mid \mathbb{F}_n^{\#-1}(y) - \mathbb{F}_n^{-1}(\tilde{\mathbb{U}}_{n-1}(y)) \mid =_{a.s} O(n^{-1/2}\log n), \tag{2}$$

In addition, there exists a Kiefer process K, such that

$$\sup_{\delta_n^{\#} \le y \le 1-\delta_n^{\#}} \mid \sqrt{n} f(F^{-1}(y))(\mathbb{F}_n^{-1}(\tilde{\mathbb{U}}_{n-1}(y)) - \mathbb{F}_n^{-1}(y)) - n^{-1/2}K(y,n) \mid =_{a.s.} O(l_n). \tag{3}$$

Proof Besides Lemma 3.1, two important steps are shown as follows:

(1) We may represent $X_i = F^{-1}(U_i)$, where U_is $\sim$ i.i.d. $U(0,1)$, $i = 1, \ldots, n$, $V_1, \ldots, V_{n-1} \sim$ i.i.d. $U(0,1)$, independent of Xs. Then we have

$$\mathbb{F}_n^{\#-1}(y) = \begin{cases} \mathbb{F}_n^{-1}\left(\dfrac{1+(n-1)\tilde{\mathbb{U}}_{n-1}(y)}{n}\right), & V_{k:n-1} < y < V_{k+1:n-1}, k = 0, \ldots, n-1, \\ \mathbb{F}_n^{-1}\left(\dfrac{(n-1)\tilde{\mathbb{U}}_{n-1}(y)}{n}\right), & y = V_{k:n-1}, k = 1, \ldots, n-1. \end{cases}$$

(2) The BB quantile process can be written as $\mathbb{Q}_n^{\#}(y) = \sqrt{n}(\mathbb{F}_n^{\#-1}(y) - \mathbb{F}_n^{-1}(\tilde{\mathbb{U}}_{n-1}(y))) + \sqrt{n}(\mathbb{F}_n^{-1}(\tilde{\mathbb{U}}_{n-1}(y)) - \mathbb{F}_n^{-1}(y))$.

■

Remark 1 The strong approximation for quantile process of bootstrap resampling distribution can be obtained similarly with the same error rate but different ranges of t [11].

4. Limit theorem for the BB version of the ROC process and its functionals

Let $X_1, \ldots, X_m \sim$ i.i.d. F, and $Y_1, \ldots, Y_n \sim$ i.i.d. G. For example, in medical contexts, F may stand for the c.d.f. of a population without disease and G for the c.d.f. of a population with disease. Assume F and G satisfy Conditions A and B and Xs and Ys are independent. Let $N = m + n$, where $N \to +\infty$, $m/N \to \lambda$, $0 < \lambda < 1$, $\tau_m = l_m/\sqrt{m}$, $\alpha_m^{\#} = \alpha_m + (m-1)^{-1/2}(\log\log(m-1))^{1/2}$, and $\alpha_m^{-1} = O(m^{1/4-\epsilon})$, for some $\epsilon \in (0, 1/4)$.

Recall a Kiefer process defined as follows: $\{K(x,y) : 0 \le x \le 1, 0 \le y < \infty\}$ is determined by $K(x,y) = W(x,y) - xW(1,y)$, where $W(x,y)$ be a two-parameter standard Wiener process. Its covariance function is $E[K(x_1,y_1)K(x_2,y_2)] = (x_1 \wedge x_2 - x_1x_2)(y_1 \wedge y_2)$, where $\wedge$ stands for the minimum.

4.1. *Functional limit theorems for ROC curve*

The ROC curve is defined as $\{(\mathrm{P}(X > c), \mathrm{P}(Y > c)) : X \sim F, Y \sim G, c \in \mathbb{R}\}$, or alternatively as $R(t) = \bar{G}(\bar{F}^{-1}(t))$; its derivative is $R'(t) = g(\bar{F}^{-1}(t))/f(\bar{F}^{-1}(t))$. Empirical estimate of ROC can be obtained by plugging in their empirical counterparts. Based on the given samples, BB estimate can be obtained by doing the following two steps: (1) get one realisation from the posterior distribution of ROC curves by plugging the c.d.f. and quantile function of the BB resampling distributions into the ROC functional form, respectively; (2) get BB posterior mean by averaging over many realisations. The BB version of the ROC curve described above can be equivalently regarded by putting two independent non-informative Dirichlet priors on F and G. Hence, our BB estimate of ROC is essentially a non-parametric Bayesian estimate. Let $\mathbb{R}(t) = \mathbb{R}_{m,n}(t) = \bar{\mathbb{G}}_n(\bar{\mathbb{F}}_m^{-1}(t))$ and $\mathbb{R}^{\#}(t) = \mathbb{R}^{\#}_{m,n}(t) = \bar{\mathbb{G}}^{\#}_n(\bar{\mathbb{F}}_m^{\#-1}(t))$.

THEOREM 4.1 *Let $X_1, \ldots, X_m \sim$ i.i.d. F and $Y_1, \ldots, Y_n \sim$ i.i.d. G satisfies Conditions A and B. Then*

$$\mathbb{R}(t) = R(t) + R'(t)\frac{K_1(t,m)}{m} + \frac{K_2(R(t),n)}{n} + O(\alpha_m^{-1}\tau_m), \quad t \in (\alpha_m, 1-\alpha_m), \tag{4}$$

$$\mathbb{R}^{\#}(t) = \mathbb{R}(t) + R'(t)\frac{K_1(t,m)}{m} + \frac{K_2(R(t),n)}{n} + O(\alpha_m^{-1}\tau_m), \quad t \in (\alpha_m^{\#}, 1-\alpha_m^{\#}), \tag{5}$$

in the sense of a.s. for (5), *where K_1 and K_2 are independent generic Kiefer processes* (*not identical in each appearance*).

Proof Sketch of the proof of (5): By Theorem 3.1 and Theorem C of Appendix A, there exist two independent Kiefer processes K_1 and K_2, such that when $t \in (\alpha_m^{\#}, 1-\alpha_m^{\#}) \subset (\delta_m^{\#}, 1-\delta_m^{\#})$, the strong approximation for BB estimator $\mathbb{R}^{\#}(t)$ of ROC curve is given by $\bar{\mathbb{G}}_n(\bar{\mathbb{F}}_m^{-1}(t)) + I_1^{\#} + n^{-1}K_2(\bar{G}(\bar{F}^{-1}(t)), n) + I_2^{\#} + O(\tau_n)$, where

$$\begin{aligned}
I_1^{\#} &= \bar{\mathbb{G}}_n(\bar{\mathbb{F}}_m^{\#-1}(t)) - \bar{\mathbb{G}}_n(\bar{\mathbb{F}}_m^{-1}(t)) = I_{11}^{\#} + I_{12}^{\#},\\
I_{11}^{\#} &= \bar{\mathbb{G}}_n(\bar{\mathbb{F}}_m^{\#-1}(t)) - \bar{G}(\bar{\mathbb{F}}_m^{\#-1}(t)) - (\bar{\mathbb{G}}_n(\bar{\mathbb{F}}_m^{-1}(t)) - \bar{G}(\bar{\mathbb{F}}_m^{-1}(t))),\\
I_{12}^{\#} &= \bar{G}(\bar{\mathbb{F}}_m^{\#-1}(t)) - \bar{G}(\bar{\mathbb{F}}_m^{-1}(t)) = I_{121}^{\#} + I_{122}^{\#},\\
I_{121}^{\#} &= \bar{G}(\bar{\mathbb{F}}_m^{\#-1}(t)) - \bar{G}(\bar{\mathbb{F}}_m^{-1}(\tilde{\mathbb{U}}_{m-1}(t))),\\
I_{122}^{\#} &= \bar{G}(\bar{\mathbb{F}}_m^{-1}(\tilde{\mathbb{U}}_{m-1}(t))) - \bar{G}(\bar{\mathbb{F}}_m^{-1}(t)),\\
I_2^{\#} &= \{n^{-1}K_2(\bar{\mathbb{G}}_n(\bar{\mathbb{F}}_m^{\#-1}(t)), n) - n^{-1}K_2(\bar{\mathbb{G}}_n(\bar{\mathbb{F}}_m^{-1}(t)), n)\}\\
&\quad + \{n^{-1}K_2(\bar{\mathbb{G}}_n(\bar{\mathbb{F}}_m^{-1}(t)), n) - n^{-1}K_2(\bar{G}(\bar{F}^{-1}(t)), n)\},
\end{aligned}$$

and $\tilde{\mathbb{U}}_{m-1}(t)$ is the empirical function of $V_1, \ldots, V_{m-1} \sim$ i.i.d. $U(0,1)$.

First, we will show

$$I_{122}^{\#} = m^{-1}R'(t)K_1(t,m) + O(\alpha_m^{-1}\tau_m). \tag{6}$$

Proof of (6)

$$\begin{aligned}
I_{122}^{\#} &= \bar{G}(\bar{\mathbb{F}}_m^{-1}(\tilde{\mathbb{U}}_{m-1}(t))) - \bar{G}(\bar{\mathbb{F}}_m^{-1}(t))) = \{-g(\bar{\mathbb{F}}_m^{-1}(t))[\bar{\mathbb{F}}_m^{-1}(\tilde{\mathbb{U}}_{m-1}(t)) - \bar{\mathbb{F}}_m^{-1}(t)]\}\\
&\quad + \{-g'(\bar{\mathbb{F}}_m^{-1}(\xi))[\bar{\mathbb{F}}_m^{-1}(\tilde{\mathbb{U}}_{m-1}(t))) - \bar{\mathbb{F}}_m^{-1}(t)]^2\},
\end{aligned} \tag{7}$$

where ξ lies between t and $\tilde{\mathbb{U}}_{m-1}(t)$. The first and second terms of (7) can be strongly approximated a.s. by $m^{-1}R'(t)K_1(t,m) + O(\alpha_m^{-1}\tau_m)$ and $O(\alpha_m^{-1}\tau_m)$, respectively, after carefully splitting the process and applying Theorem 3.1, Lemma 3.1, similar argument in Csörgő and Révész [8]. The term $I^{\#}_{121}$ is bounded a.s. by $O(m^{-1}\log m)$ by Condition A and Lemma 3.1. Hence $I^{\#}_{12} = m^{-1}R'(t)K_1(t,m) + O(\alpha_m^{-1}\tau_m)$. The term $I^{\#}_{11}$ can be bounded a.s. by $O((m/n)^{1/2}\alpha_m^{-1/2}\tau_m)$, because by Theorem A, there exists Kiefer process K, such that $I^{\#}_{11}$ can be majorised by

$$\left|\frac{K(\bar{G}(\bar{\mathbb{F}}_m^{\#-1}(t)),n) - K(\bar{G}(\bar{\mathbb{F}}_m^{-1}(t)),n)}{n}\right| + O(n^{-1}\log^2 n)$$
$$\leq n^{-1/2} \mid I^{\#}_{12} \mid^{1/2} (\log(1/\mid I^{\#}_{12} \mid))^{1/2} + O(n^{-1}\log^2 n) = O((m/n)^{1/2}\alpha_m^{-1/2}\tau_m). \quad (8)$$

The last statement of (8) holds by the estimate of the modulus of continuity for Brownian motion on bounded interval. Therefore, $I^{\#}_1 = m^{-1}R'(t)K_1(t,m) + O(\alpha_m^{-1}\tau_m)$ and, subsequently, $I^{\#}_2 = O(\tau_m)$. ■

Remark 1 A result similar to (4) can be found in Hsieh and Turnbull [10]. The difference between our strong approximation for $\mathbb{R}_{m,n}(t)$ given by (4) and theirs is a different trade-off between interval of validity and the approximation error rate. They used any fixed subinterval of $[0,1]$ and obtained a better approximation error rate. From a practical point of view, inclusion of levels near 0 is important. Our result treats the domain of $R(t)$ as the whole of $[0,1]$ in the limiting sense. The strong approximation for ROC estimator of bootstrap resampling distribution can be obtained similarly with the same error rate but different ranges of t [11].

4.2. *Implication of the functional limit theorems for ROC curves*

From (4), then as $L^\infty[0,1]$ valued random function,

$$\sqrt{N}(\mathbb{R}(t) - R(t))1(\{\alpha_m < t < 1-\alpha_m\}) = \left\{\frac{\sqrt{N}}{\sqrt{m}}R'(t)\frac{K_1(t,m)}{\sqrt{m}} + \frac{\sqrt{N}}{\sqrt{n}}\frac{K_2(R(t),n)}{\sqrt{n}}\right.$$
$$\left. + O(\sqrt{N}\alpha_m^{-1}\tau_m)\right\} 1(\{\alpha_m < t < 1-\alpha_m\}) \rightsquigarrow \frac{1}{\sqrt{\lambda}}R'(t)B_1(t) + \frac{1}{\sqrt{1-\lambda}}B_2(R(t)), \quad (9)$$

in the sense of generalised weak convergence on non-separable spaces [12], where B_1 and B_2 are independent Brownian bridges. To see that (9) holds, observe that the realisations of $1/(\sqrt{\lambda})R'(t)B_1(t) + 1/(\sqrt{1-\lambda})B_2(R(t))$ are continuous and are tied to 0 at the two end points 0 and 1. Hence, as $m \to \infty$,

$$\frac{1}{\sqrt{\lambda}}R'(t)B_1(t) + \frac{1}{\sqrt{1-\lambda}}B_2(R(t))1(\{\alpha_m < t < 1-\alpha_m\}) \longrightarrow \frac{1}{\sqrt{\lambda}}R'(t)B_1(t) + \frac{1}{\sqrt{1-\lambda}}B_2(R(t)),$$

uniformly on $t \in [0,1]$. Similarly, from (5), we can get, a.s., conditionally on the samples,

$$\sqrt{N}(\mathbb{R}^{\#}(t) - \mathbb{R}(t))1(\{\alpha^{\#}_m < t < 1-\alpha^{\#}_m\}) \rightsquigarrow \frac{1}{\sqrt{\lambda}}R'(t)B_1(t) + \frac{1}{\sqrt{1-\lambda}}B_2(R(t)). \quad (10)$$

Remark 2 Li *et al.* [13] obtained a functional weak convergence result to a Gaussian limit for the empirical estimate ROC process using Donsker's theorem and functional Delta method. Although the weak convergence approach is very elegant, the functional delta method requires fixed mean and hence does not apply directly in the BB or bootstrap distribution. Further, the approach based on strong approximation lets us work with real-valued random variables and ordinary Taylor's series expansion, derive a stronger representation, and treat the whole domain $t \in [0,1]$ in the limit. In contrast, the weak convergence approach must leave out some neighbourhoods of 0 and 1 in the limit.

Result 1 If A is a continuity set for the process on the right hand side of (9), then

$$\Pr^{\#}\{\sqrt{N}(\mathbb{R}^{\#}(\cdot) - \mathbb{R}(\cdot)) \in A\} - \Pr\{\sqrt{N}(\mathbb{R}(\cdot) - R(\cdot)) \in A\} \longrightarrow 0, \tag{11}$$

where $\Pr^{\#}$ stands for BB resampling distribution conditional on the samples.
Result 2 [Equivalence of empirical and BB estimator of ROC] Conditionally on samples,

$$\sqrt{N} \parallel \mathrm{E}(\mathbb{R}^{\#}(\cdot) \mid X_1, \ldots, X_m, Y_1, \ldots, Y_n) - \mathbb{R}(\cdot) \parallel_{L^\infty} \longrightarrow 0. \text{ a.s.} \tag{12}$$

4.3. *Functional limit theorems for AUC*

COROLLARY 4.1 *Let $\psi : \mathbb{R} \to \mathbb{R}$ be a bounded function with continuous second derivatives. Then for any $q > 0$ being the continuity point of the limiting random variable*

$$\sup_{t\in(0,1)} \{\psi'(R(t))[\lambda^{-1/2}R'(t)B_1(t) + (1-\lambda)^{-1/2}B_2(R(t))]\},$$

we have that almost surely for the given samples,

$$\begin{aligned} &\mid Pr^{\#}\{\sup_{t\in(\alpha_m^{\#},1-\alpha_m^{\#})} \sqrt{N} \mid \psi(\mathbb{R}^{\#}(t)) - \psi(\mathbb{R}(t)) \mid\leq q\} \\ &- Pr\{\sup_{t\in(\alpha_m^{\#},1-\alpha_m^{\#})} \sqrt{N} \mid \psi(\mathbb{R}(t)) - \psi(R(t)) \mid\leq q\} \mid\longrightarrow 0, \end{aligned} \tag{13}$$

where $Pr^{\#}$ stands for BB probability conditional on the given samples.

Remark 3 Equivalently, a metric d (such as Levy's metric) which can characterise weak convergence can also be used to quantify the nearness of the distributions.

COROLLARY 4.2 *Let $\phi(\cdot) : C[0,1] \to \mathbb{R}$ be a bounded linear functional. Then conditionally on the samples, we have for all q*

$$Pr^{\#}\{\sqrt{N}(\phi(\mathbb{R}^{\#}) - \phi(\mathbb{R})) \leq q\} - Pr\{\sqrt{N}(\phi(\mathbb{R}) - \phi(R)) \leq q\} \longrightarrow 0 \;\; a.s.$$

COROLLARY 4.3 [For partial AUC (pAUC)] *Conditionally on the samples, a.s. for any $(e_1, e_2) \subset (0,1)$, we have $\sqrt{N}(\mathbb{A}^{\#}(e_1^*, e_2^*) - \hat{\mathbb{A}}(e_1^*, e_2^*)) \to_d N(0, \sigma^2(e_1, e_2))$, where $e_1^* = \max(e_1, \alpha_m^{\#})$, $e_2^* = \min(e_2, 1-\alpha_m^{\#})$*

$$\begin{aligned} \sigma^2(e_1, e_2) &= \lambda^{-1} \int_{e_1}^{e_2} \int_{e_1}^{e_2} R'(t)R'(s)(s \wedge t - st)\,\mathrm{d}t\mathrm{d}s \\ &\quad + (1-\lambda)^{-1} \int_{e_1}^{e_2} \int_{e_1}^{e_2} (R(t) \wedge R(s) - R(t)R(s))\,\mathrm{d}t\mathrm{d}s, \end{aligned}$$

$$\mathbb{A}^{\#}(e_1^*, e_2^*) = \int_{e_1^*}^{e_2^*} \mathbb{R}^{\#}(t)\,\mathrm{d}t, \quad \hat{\mathbb{A}}(e_1^*, e_2^*) = \int_{e_1^*}^{e_2^*} \mathbb{R}(t)\,\mathrm{d}t.$$

Remark 4 (1) It can be easily shown that Condition B ensures that $\sigma^2(0, 1) < \infty$. (2) Asymptotic form for AUC ($e_1 = 0, e_2 = 1$) can be obtained from Corollary 4.2. (3) Similar results hold for the empirical estimate and the bootstrap resampling distribution. Further, it follows that

$$\sup\{|\Pr^{\#}\{\sqrt{N}(\mathbb{A}^{\#}(e_1^*, e_2^*) - \hat{\mathbb{A}}(e_1^*, e_2^*)) \le q\} \\ - \Pr\{\sqrt{N}(\hat{\mathbb{A}}(e_1^*, e_2^*) - A(e_1^*, e_2^*)) \le q\}| : -\infty < q < \infty\} \longrightarrow 0 \;\; a.s.$$

where $A(e_1^*, e_2^*) = \int_{e_1^*}^{e_2^*} R(t)\,\mathrm{d}t$.

5. Simulation studies

5.1. *Comparison of the frequentist variability and the BB resampled variability*

To investigate the asymptotic equivalence of frequentist variability of the empirical estimate of AUC and resampled variability of AUC estimate based on the BB procedures, we simulated 6000 data sets of $X_1, \ldots, X_m \sim$ i.i.d. F = Normal (0,1), $Y_1, \ldots, Y_n \sim$ i.i.d. G = Normal(1.868, 1.5^2) to get 6000 frequentist samples of $\sqrt{N}(\hat{\mathbb{A}} - A)$, where $\hat{\mathbb{A}} = 1/(mn)\sum_{i=1}^{m}\sum_{j=1}^{n} 1(Y_j > X_i)$, A being the true AUC value, $N = m + n$, m and n chosen to be 50 and 300, shown in Figure 1a and b, respectively; based on one simulated data set defined above, BB's one random realisation of AUC denoted as $\mathbb{A}^{\#}$ can be obtained based on that BB's random realisation of the ROC curve denoted as $\mathbb{R}^{\#}(t)$; hence, one BB resample of $\sqrt{N}(\mathbb{A}^{\#} - \hat{\mathbb{A}})$ can be generated. Let the resample size be 6000. The density plot (Figure 1) of the asymptotic normal distribution $N(0, \sigma^2(0, 1))$, where σ can be calculated by Corollary 4.3, overlays those of $\sqrt{N}(\hat{\mathbb{A}} - A)$ and $(\sqrt{N}(\mathbb{A}^{\#} - \hat{\mathbb{A}}) \mid \text{data})$. The grid points on [0,1] are chosen with equal interval length 0.001 to calculate AUC.

The accuracy of the BB distribution of AUC in estimating the sampling distribution of the empirical estimate of the AUC is studied. The accuracy of the normal limit is also studied for the

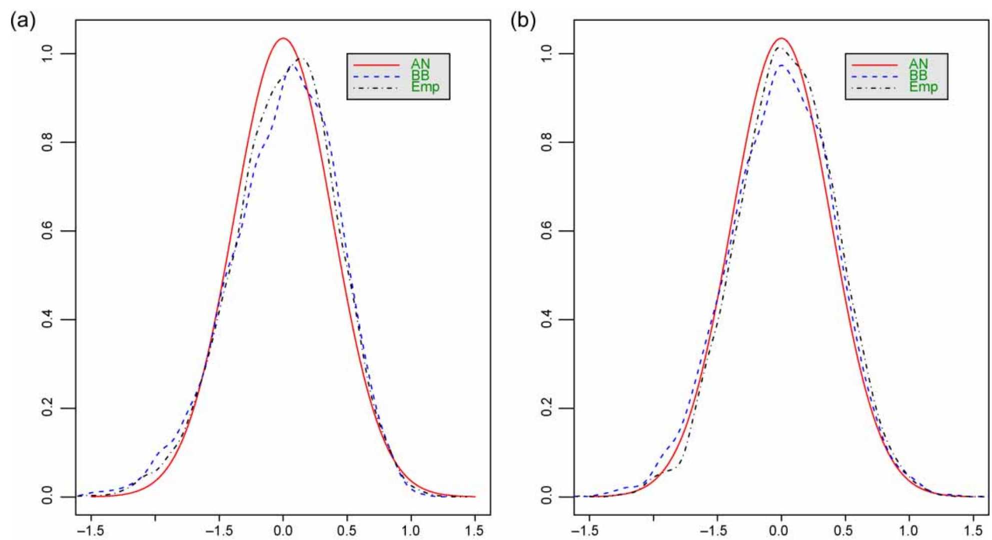

Figure 1. Comparison of density plots obtained by empirical $\sqrt{N}(\hat{\mathbb{A}} - A)$ based on 6000 simulated data sets, the BB's $(\sqrt{N}(\mathbb{A}^{\#} - \hat{\mathbb{A}})|\text{data})$ based on one simulated data set and corresponding 6000 resamples, and the asymptotic normal distribution N(0, $\sigma^2(0, 1)$) where σ can be calculated by Corollary 4.3.

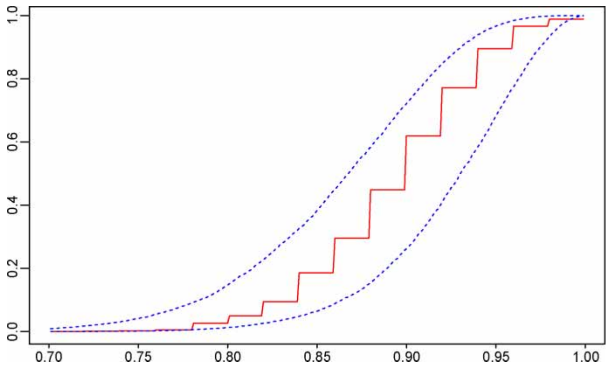

Figure 2. Comparison c.d.f. of empirical (solid curve) and the BB's interval of 25–75th percentile (dotted curves) the cdf of the empirical estimate of $R(0.5)$, based on sample size $m = n = 50$, 1000 simulated data sets, 1000 resample size, grid points at even intervals of length 0.001 on [0, 1].

same data. It is observed that both estimates are fairly close to the actual distribution obtained by simulation, but the BB estimate is somewhat more accurate than the normal limit.

5.2. *Comparison of the empirical and BB's estimates of ROC*

By simulating the data sets described above, with sample size $m = n = 50$ and $\psi = R(0.5)$, we compare the estimate of the ROC at the point $t = 0.5$. We obtain $P(\hat{\mathbb{R}}(0.5) \le q)$ based on 1000 simulated data sets, and $P^{\#}(\mathbb{R}^{\#}(0.5) \le q \mid \text{data})$ based on 1000 resamples. The grid points on [0,1] are chosen at equal intervals of length 0.001 to calculate ROC. The plot $\hat{P}(\hat{\mathbb{R}}(0.5) \le q)$ and the interval of 25–75th percentile of $P^{\#}(\mathbb{R}^{\#}(0.5) \le q \mid \text{data})$ are given in Figure 2.

6. Proof of the Lemma 3.1

6.1. *Proof of the statement (2) in Lemma 3.1*

Proof By Theorem M, we have

$$\begin{aligned}
&\sup_{0<y<1} \sqrt{n} \mid \mathbb{F}_n^{\#-1}(y) - \mathbb{F}_n^{-1}(\tilde{\mathbb{U}}_{n-1}(y)) \mid \\
&\le \sup_{0<y<1} \sqrt{n} \left| \mathbb{F}_n^{-1}\left(\frac{1 + (n-1)\tilde{\mathbb{U}}_{n-1}(y)}{n} \right) - \mathbb{F}_n^{-1}(\tilde{\mathbb{U}}_{n-1}(y)) \right| \\
&\quad + \sup_{0<y<1} \sqrt{n} \left| \mathbb{F}_n^{-1}\left(\frac{(n-1)\tilde{\mathbb{U}}_{n-1}(y)}{n} \right) - \mathbb{F}_n^{-1}(\tilde{\mathbb{U}}_{n-1}(y)) \right| \\
&\le \sup_{0\le k\le n} 2\sqrt{n} \mid X_{k+1:n} - X_{k:n} \mid =_{a.s.} O(n^{-1/2} \log n).
\end{aligned}$$

■

6.2. *Proof of the statement (3) in Lemma 3.1*

Proof Because $y \in [\delta_n^{\#}, 1 - \delta_n^{\#}] \subset [\epsilon_n, 1 - \epsilon_n]$, where $\epsilon_n = 0.236 n^{-1} \log\log n$, let $k = \lceil ny \rceil$, ξ lies between y and $\tilde{\mathbb{U}}_{n-1}(y)$. The following four steps are needed to finish the proof.

(1) By Theorems D, K and N, we have $\sup_{\delta_n^\# \le y \le 1-\delta_n^\#} \left| (\tilde{\mathbb{U}}_{n-1}(y) - y) \frac{f'(F^{-1}(\xi))}{f(F^{-1}(\xi))} \mathbb{Q}_n(\tilde{\mathbb{U}}_{n-1}(y)) \right| =_{\text{a.s}} O(l_n)$.

(2) By Theorem B, $\sup_{\delta_n^\# \le y \le 1-\delta_n^\#} | \sqrt{n} f(F^{-1}(y))(\mathbb{F}_n^{-1}(y) - F^{-1}(y)) - n^{-1/2} K(y, n) | =_{\text{a.s.}} O(l_n)$.

(3) $\sup_{\delta_n^\# \le y \le 1-\delta_n^\#} | \sqrt{n} f(F^{-1}(y))[F^{-1}(\tilde{\mathbb{U}}_{n-1}(y)) - F^{-1}(y)] - n^{-1/2} K(y, n) | \le \sup_{\delta_n^\# \le y \le 1-\delta_n^\#} | \sqrt{n} (\tilde{\mathbb{U}}_{n-1}(y) - y) - n^{-1/2} K(y, n) | + \sup_{\delta_n^\# \le y \le 1-\delta_n^\#} \left| (\tilde{\mathbb{U}}_{n-1}(y) - y)^2 \frac{f'(F^{-1}(\xi))}{f^2(F^{-1}(\xi))} \frac{f(F^{-1}(y))}{f(F^{-1}(\xi))} \right|$, which is $O(l_n)$ a.s. using Theorems A, G and H.

(4) By Theorems G and I, $\sup_{0<y<1} \left| \frac{K(\tilde{\mathbb{U}}_{n-1}(y), n) - K(y, n)}{\sqrt{n-1}} \right| =_{\text{a.s.}} O(l_n)$.

The proof can be completed by combining parts (1)–(4) above and splitting $\mathbb{F}_n^{-1}(\tilde{\mathbb{U}}_{n-1}(y)) - \mathbb{F}_n^{-1}(y)$ into $[\mathbb{F}_n^{-1}(\tilde{\mathbb{U}}_{n-1}(y)) - F^{-1}(\tilde{\mathbb{U}}_{n-1}(y))] - [\mathbb{F}_n^{-1}(y) - F^{-1}(y)] + [F^{-1}(\tilde{\mathbb{U}}_{n-1}(y)) - F^{-1}(y)]$; see Gu and Ghosal [11] for details. ■

Acknowledgements

Research of the authors is partially supported by NSF grant number DMS-0349111. The authors wish to thank the referee for pointing out an inaccuracy in the statement of a result.

References

[1] D.M. Green and J.A. Swets, *Signal Detection Theory and Psychophysics*, John Wiley & Sons, New York, 1966.
[2] C.E. Metz, *Basic principles of ROC analysis*, Seminars Nuc. Med. 8, pp. 283–298.
[3] S. Wieand, M.H. Gail, B.R. James, and K.L. James, *A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data*, Biometrika 76 (1989), pp. 585–592.
[4] D. Bamber, *The area above the ordinal dominance graph and the area below the receiver operating characteristic graph*, J. Math. Psychol. 12 (1975), pp. 387–415.
[5] M.S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, Oxford, 2003.
[6] J. Gu, S. Ghosal, and A. Roy, Non-parametric estimation of ROC curve, *Institute of Statistics Mimeo Series North Carolina State University*, 2592 (2006a), pp. 1–19.
[7] J. Komlós, P. Major, and G. Tusnády, *An approximation of partial sums of independent RV's and the sample DF. I.*, Z. Wahrsch. Verw. Gebiete, 32 (1975), pp. 111–131.
[8] M. Csörgő and P. Révész, *Strong approximations of the quantile process*, Ann. Statist. 6 (1978), pp. 882–894.
[9] A.Y. Lo, *A large sample study of the Bayesian bootstrap*, Ann. Statist. 15 (1987), pp. 360–375.
[10] F.S. Hsieh and B.W. Turnbull, *Nonparametric and semiparametric estimation of the receiver operating characteristic curve*, Ann. Statist. 24 (1996), pp. 25–40.
[11] J. Gu and S. Ghosal, Strong approximations for resample quantile processes and application to ROC methodology. *Institute of Statistics Mimeo Series North Carolina State University,* 2593 (2006b), pp. 1–33.
[12] A.W. Van der Vaart and J.A. Wellner, *Weak Convergence and Empirical Processes*, Springer-Verlag, New York, 1996.
[13] G. Li, R.C. Tiwari, and M.T. Wells, *Quantile comparison functions in two-sample problems, with application to comparisons of diagnostic markers*, J. Am. Statist. Assoc. 91 (1996), pp. 689–698.
[14] M. Csörgő and P. Révész, *Strong Approximations in Probability and Statistics*, Academic press, New York, 1981.
[15] E. Slud, *Entropy and maximal spacings for random partitions*, Z. Wahrsch. Verw. Gebiete 41 (1978), pp. 341–352.

Appendix A. The following theorems are used

THEOREM A [Implied by Komlós *et al.* [7], Theorem 4]: *Let $X_1, X_2, \ldots \sim$ i.i.d. F, where F is any arbitrary continuous distribution function and $\mathbb{J}_n(x)$ be the empirical process. Then there exists a Kiefer process $\{K(y, t); 0 \le y \le 1, t \ge 0\}$, such that* $\sup_x |\mathbb{J}_n(x) - n^{-1/2} K(F(x), n)| =_{a.s.} O(n^{-1/2} \log^2 n)$.

THEOREM B [Csörgő and Révész [8], Theorem 6]: *Let $X_1, X_2, \ldots \sim$ i.i.d. F, satisfying Condition A. Then the quantile process $\mathbb{Q}_n(x)$ of X can be approximated by a Kiefer process $\{K(y,t); 0 \le y \le 1, t \ge 0\}$, as* $\sup_{\delta_n \le y \le 1-\delta_n} \mid f(F^{-1}(y))\mathbb{Q}_n(y) - n^{-1/2}K(y,n) \mid =_{a.s.} O(l_n)$.

THEOREM C [Lo [9], Lemma 6.3]: *There exists a Kiefer process $\{K(s,t); 0 \le s \le 1, t \ge 0\}$ independent of X such that* $\sup_x \mid \mathbb{J}_n^{\#}(x) - n^{-1/2}K(\mathbb{F}_n(x), n) \mid =_{a.s.} O(l_n)$.

THEOREM D [Csörgő and Révész [8], Theorem 3]: *Let $X_1, X_2, \ldots$ be i.i.d. F satisfying Condition A. Then* $\limsup_{n\to\infty} \sqrt{n}/(\log\log n) \sup_{\delta_n \le y \le 1-\delta_n} \mid f(F^{-1}(y))\mathbb{Q}_n(y) - \mathbb{W}_n(y) \mid \le 40\gamma 10^{\gamma}$ *a.s.* (γ *is defined in Condition A*)

THEOREM E [Csörgő and Révész [14], Lemma 4.5.1]: *Let $U_1, U_2, \ldots$ be i.i.d. $U(0,1)$ random variables, and let Kiefer process $\{K(y,t); 0 \le y \le 1, 0 \le t\}$ defined on the same probability space. Then* $\sup_{1\le k \le n} n^{-1/2}|K(U_{k:n}, n) - K(k/n, n)| =_{a.s.} O(l_n)$.

THEOREM F [Csörgő and Révész [14], Theorem 1.15.2]: *Let h_n be a sequence of positive numbers for which* $\lim_{n\to\infty} \log h_n^{-1}/(\log\log n) = \infty$, $\gamma_n = (2nh_n \log h_n^{-1})^{-1/2}$. *Then*

$$\lim_{n\to\infty} \sup_{0\le t\le 1-h_n} \sup_{0\le s\le h_n} \gamma_n |K(t+s,n) - K(t,n)| =_{a.s} 1.$$

THEOREM G *Let $K(s,n)$ be a Kiefer process,* $\sup_{0<s<1} |(K(s,n-1) - K(s,n))/(\sqrt{n-1})| =_{a.s.} O(l_n)$.

The result is contained in the proof of Lemma 6.3 in Lo [9].

THEOREM H [Law of Iterated Logarithm (LIL)]: *Let $K(y,n)$ be Kiefer process,*

$$\limsup_{n\to\infty} \sup_{0\le y\le 1} \frac{\mid K(y,n) \mid}{(2n\log\log n)^{1/2}} =_{a.s.} \frac{1}{2}.$$

THEOREM I [Special case of Lo [9], Lemma 6.2 when $F = U(0,1)$]: *Let $U_1, \ldots, U_m \sim$ i.i.d. $U(0,1)$, $\mathbb{U}_m(s)$ is the empirical function of U_is. For any Kiefer process K independent of $U_1, \ldots, U_m$, then* $\sup_{0<s<1} |(K(\mathbb{U}_m(s), m) - K(s,m))/\sqrt{m}| =_{a.s.} O(l_m)$.

THEOREM J [Csörgő and Révész [8], Theorem D]: *For $\mathbb{H}_n(x)$, the uniform empirical process and $\epsilon_n = 0.236(\log\log n)/n$, we have* $\limsup_{n\to\infty} \sup_{\epsilon_n \le x \le 1-\epsilon_n} (x(1-x)\log\log n)^{-1/2} |\mathbb{H}_n(x)| =_{a.s.} 2$.

THEOREM K [Csörgő and Révész [8], Theorem 2]: *Let $\mathbb{W}_n(y)$ be the uniform quantile process, then* $\limsup_{n\to\infty} \sup_{\delta_n \le y \le 1-\delta_n} (y(1-y)\log\log n)^{-1/2} |\mathbb{W}_n(y)| =_{a.s.} 4$.

THEOREM L [Csörgő and Révész [8], Lemma 3.2]: *Under Condition A,*

$$\frac{f(F^{-1}(y_1))}{f(F^{-1}(y_2))} \le \left\{ \frac{y_1 \vee y_2}{y_1 \wedge y_2} \cdot \frac{1 - y_1 \wedge y_2}{1 - y_1 \vee y_2} \right\}^{\gamma} \text{ for any pair } y_1, y_2 \in (0,1).$$

THEOREM M [Slud [15]]: *Let $U_1, \ldots, U_n \sim$ i.i.d. $U(0,1)$ and* $M_n = \max_{0\le k\le n}(U_{k+1:n} - U_{k:n})$. *Then,* $nM_n/\log n \to_{a.s} 1$.

THEOREM N *Under Condition A, for any* $y \in [\delta_n^\#, 1 - \delta_n^\#]$, ξ *lies between* y *and* $\tilde{\mathbb{U}}_{n-1}(y)$,

$$\sup_{\delta_n^\# \le y \le 1-\delta_n^\#} \frac{y(1-y)}{\xi(1-\xi)} \le 5, \quad \sup_{\delta_n^\# \le y \le 1-\delta_n^\#} \frac{\xi}{\tilde{\mathbb{U}}_{n-1}(y)} \le 5,$$

$$\sup_{\delta_n^\# \le y \le 1-\delta_n^\#} \frac{1-\xi}{1-\tilde{\mathbb{U}}_{n-1}(y)} \le 5, \quad \sup_{\delta_n^\# \le y \le 1-\delta_n^\#} f(F^{-1}(\xi))/f(F^{-1}(\tilde{\mathbb{U}}_{n-1}(y))) \le 10^\gamma.$$

The result is contained in the proof of Theorem 3 of Csörgő and Révész [8].