Volume 15, No. 3 (2006), pp. Allerton Press, Inc.

MATHEMATICAL METHODS OF STATISTICS

THE CHANGE SET PROBLEM AND LOCAL COVERING NUMBERS

E. Khmaladze[1,3], R. Mnatsakanov[2,3], and N. Toronjadze[4]

[1]School of Math., Statist. and Computer Science, Victoria Univ. Wellington
P.O. Box 600, Wellington, New Zealand
E-mail: Estate.Khmaladze@mcs.vuw.ac.nz

[2]Dept. Statist., West Viginia Univ.
P.O. Box 6330, Morgantown, WV 26506, USA
E-mail: rmnatsak@stat.wvu.edu

[3]Dept. Probab. Theory and Math. Statist., A. Razmadze Math. Inst.
1 M. Alexidze St., 0193, Tbilisi, Georgia

[4]Financial Monitoring Service of Georgia, NBG
3/5 Leonidze St., 0105, Tbilisi, Georgia

For a pair (X, Y), suppose that $\mathbb{P}\{Y \in \cdot \mid X\} = P_1\{\cdot\}$ if X is outside certain region $G \subset \mathbb{R}^d$ and $\mathbb{P}\{Y \in \cdot \mid X\} = P_2\{\cdot\}$ if $X \in G$. We call this region a change-set and it could also be called an "image". Consider a Maximum estimator (M-estimator) $\hat{G}$ of G based on n independent pairs $(X_i, Y_i)_{i=1}^n$ under assumption that G belongs to a totally bounded class $\mathcal{C}$ of measurable subsets of $\mathbb{R}^d$ with the distance $d(G, G') = F(G \triangle G')$ induced by the distribution F of the X_i's. The classical characteristic of complexity of $\mathcal{C}$ is its covering number. However this characteristic is often not enough and one needs a more delicate characteristic of "local complexity". This is the local covering number, which is considered in Section 2 of the paper. Using it we derive an inequality for $\mathbb{P}\{d(\hat{G}, G) > \varepsilon\}$ and obtain the rate of convergence ε_n of $d(\hat{G}, G)$. Then we show that under broad conditions the deviations of $d(\hat{G}, G)$ from ε_n are of order $1/n$ regardless of what the rate ε_n is. We also study local covering numbers in the important case where $\mathcal{C}$ is formed by subgraphs of non-decreasing functions on $[0, 1]$. The results obtained for fixed P_1 and P_2 are carried over to the case when the "change" from P_1 to P_2 becomes asymptotically small as $n \to \infty$.

Key words: change-point problem, change-set, exponential inequality, local covering number.

2000 Mathematics Subject Classification: Primary 62G05, 62G20; secondary 62H35.

1. Introduction

There are several possible formulations of a spatial change-point problem, which we prefer to call a *change-set problem.* We choose here the one which seems to us

*To whom correspondence should be addressed, e-mail: Estate.Khmaladze@mcs.vuw.ac.nz

the most basic and transparent. Namely, following, e.g., the pattern of Mammen and Tsybakov ([1], Section 3), we consider a sequence $\{(X_i, Y_i)\}_1^n$ of independent pairs of random variables, where the X_i's take values in the d-dimensional Euclidean space $\mathbb{R}^d$, $d \geq 1$, and are commonly referred to as "locations", while the Y_i's take values in some measurable space $\{\mathbb{E}, \mathcal{E}\}$ and are called the corresponding "marks". In other words, $\{(X_i, Y_i)\}_1^n$ is a marked point process in $\mathbb{R}^d$. Concerning the X_i's we assume that they are identically distributed with some distribution F. Strictly speaking, the requirements that locations are random or i.i.d. are not necessary for what follows, but will make the presentation more transparent. As to the marks, they can be of very diverse nature.

For instance, suppose at each location X_i we may observe only whether or not pollution is present, in which case Y_i will be simply a $\{0, 1\}$-random variable (this case was studied in Mammen and Tsybakov [1]). In other cases at each location we may record the wind speed or the concentration of a chemical, or we can measure the energy of an earthquake at hypocentre X_i. In all of these cases, the Y_i's are presumably continuous random variables. It also may be that at each location we measure the concentration of several chemicals or measure these concentrations as functions of depth in a drill bore, in which case Y_i is a random vector or a vector-valued random function (of depth), and $\{\mathbb{E}, \mathcal{E}\}$ should be a properly selected space of trajectories of this random function. In the example with earthquakes Y_i can be an energy spectrum of an earthquake, which is a random function of a relatively complex behavior and in which case $\{\mathbb{E}, \mathcal{E}\}$ must be again a functional space, and so on.

In the present context, however, we do not need to know much about $\{\mathbb{E}, \mathcal{E}\}$ assuming only that there are two different distributions P_1 and P_2 on $\mathcal{E}$ and a measurable set $G \subset \mathbb{R}^d$ such that the conditional distribution of Y_i given X_i is

$$\mathbb{P}\{\cdot \mid X_i\} = P_1\{\cdot\} I_{\{X_i \notin G\}} + P_2\{\cdot\} I_{\{X_i \in G\}} = (P_1\{\cdot\})^{I_{\{X_i \notin G\}}} (P_2\{\cdot\})^{I_{\{X_i \in G\}}}. \tag{1.1}$$

In other words, we assume in (1.1) that there is a set G such that for all X_i outside G the corresponding mark has some "grey level" distribution P_1, while it has a different distribution P_2 if X_i is in G. The existence of a singular component of P_2 with respect to P_1 and vice versa will only simplify the statistical inference concerning G, and we assume that P_1 and P_2 are equivalent (mutually absolutely continuous). The set G in (1.1) will be called the *change-set* and this set is our parameter of interest. Note that G can also be called an *image* and the change-set problem can also be viewed as an *image reconstruction* problem. We will consider M-estimators $\hat{G}$ of this set, see (1.2), and will determine the rate of convergence of $\hat{G}$ to G including the constants.

The likelihood of the pair (X_i, Y_i) with respect to the reference measure $F \times P_1$ is $[dP_2/dP_1(Y_i)]^{I_{\{X_i \in G\}}}$ and the log-likelihood of $\{(X_i, Y_i)\}_1^n$ is

$$L_n(G) = \sum_1^n I_{\{X_i \in G\}} \log \frac{dP_2}{dP_1}(Y_i).$$

Therefore the logarithm of the likelihood ratio is

$$L_n(G) - L_n(G_0) = \sum_1^n \left[I_{\{X_i \in G\}} - I_{\{X_i \in G_0\}} \right] \log \frac{dP_2}{dP_1}(Y_i),$$

where G_0 stands for the true change-set. This suggests a slightly more general form of the processes in G which we will use here: choose some "score function" $\xi(y)$ on $\{\mathbb{E}, \mathcal{E}\}$ and consider

$$L_n(G, G_0) := \sum_1^n \big[I_{\{X_i \in G\}} - I_{\{X_i \in G_0\}}\big]\xi(Y_i).$$

We define an estimator $\widehat{G}_n$ of G_0 as

$$\widehat{G}_n = \widehat{G}_n(\delta) := \arg\max_{G \in \mathcal{N}_\delta} L_n(G, G_0), \tag{1.2}$$

with $\delta = \delta_n \to 0$ as $n \to \infty$. Then we study asymptotics of $d(\widehat{G}_n, G_0)$ as $n \to \infty$, where $d(G, G') = F(G \triangle G')$.

To say something fruitful about the rate of convergence one needs to assume apriori that G belongs to a certain relatively poor class of sets. Namely, let $\mathcal{B}$ be the Borel σ-algebra of subsets of $\mathbb{R}^d$ endowed with the pseudo-distance $d(G, G')$ and let $\mathcal{C}$ be a totally bounded subclass of the metric space $\{\mathcal{B}, d\}$. Denote by $\mathcal{N}_\delta$ the minimal δ-net of $\mathcal{C}$. Its cardinality, $N_\delta = \operatorname{card}\{\mathcal{N}_\delta\}$, is called the covering number of $\mathcal{C}$ and the function $H_\delta = \log N_\delta$ is called the metric entropy of $\mathcal{C}$. From now on we will assume that our unknown change-set G is an element of $\mathcal{C}$.

Although there are several relatively early papers devoted to statistical problems of set estimation, like, for example, Ripley and Rasson [2] or Moore [3], these problems attracted more interest in the last 10–15 years. Some papers, for example, Carlstein and Krishnamoorthy [4], Ferger [5], Müller and Song [6], Rudemo and Stryhn [7], and others, treat estimation of a set as the estimation problem of its boundary. Then the smoothness assumptions or other structural assumptions on the boundary, as in Korostelev *et al.* [8] or Puri and Ryumgaart [9], though very natural and clear in the context, are equivalent to the total boundedness assumption on the class $\mathcal{C}$. Many papers, like, e.g., already mentioned Mammen and Tsybakov [1] and Ferger [5], use the notion of the covering number (but not of the local covering number) explicitly.

Although N_δ as a function of δ is a very important characteristic of richness and complexity of the class $\mathcal{C}$, we will realize below that to determine the true rate of convergence in some practically important cases and to obtain more refined statements, see, e.g., Theorems 2.4 and 2.5 below, N_δ is not enough and we need a more delicate characteristic of the "local" richness of the class, which is the covering number of a neighborhood of a given element of the class. Namely, for each $G \in \mathcal{C}$ let $\mathcal{O}(t, G)$ be the neighborhood of G in $\mathcal{B}$ of radius t and let

$$N_\delta(t, G) = \operatorname{card} \mathcal{N}_\delta \cap \mathcal{O}(t, G). \tag{1.3}$$

Then we need to study the *local covering number* $N_\delta(t, G)$ as t and δ tend to 0 simultaneously. This allows us to obtain the correct rate of convergence, in many cases unattainable otherwise.

The local covering number was introduced and studied, in connection with the change-set problem, in Khmaladze *et al.* [10]. However this concept was considered and used earlier: Le Cam [11] considered $N_\delta(t)$ for the neighborhood of a point in $\mathbb{R}^d$ when $t = \text{const}\cdot\delta$ (see also reference in Section 3), while Birgé [12, 13] considered $\log N_\delta(t)/\log(t/\delta)$ for the neighborhood of a function also for $t = \text{const}\cdot\delta$. For further references and material on the now well established method to study the rate of convergence one can refer, e.g., to the recent fundamental paper of Birgé [14], as well as the papers by van de Geer [15], Shen and Wong [16], and Yang and Barron [17], and to the monograph van de Geer [18]. A concise presentation is available in Section 3 of van der Vaart and Wellner [19].

However, this method involves relatively complicated chaining technique and uses conditions, which cannot be met by some practically useful classes. In particular, Birgé's condition (Birgé [12] and also Condition 4 of Yang and Barron [17], cf. also p. 290 of van der Vaart and Wellner [19]), requires that the *supremum* of $[\log N_\delta(t)/\log(t/\delta)]$ in a neighborhood of δ be a positive bounded function $U(\delta)$ of δ with $n\delta^2 \geq U(\delta)$. Similar conditions are proposed in van de Geer [18] for different models. Namely, the function $U(\delta)$ is defined there as

$$U(\delta) = \int_{\delta^2/c_1}^{\delta} \big[\log N_u(\delta)\big]^{1/2}\, du,$$

and the corresponding rate has to satisfy the condition $\sqrt{n}\delta^2 \geq cU(\delta)$. Although these conditions proved to be useful in many cases, they cannot be met in the change-set problem by any Vapnik–Červonenkis class (VČ-class), where δ of interest is of order $1/n$ (see Khmaladze *et al.* [20]). It is also not satisfied for some Dudley classes, like, for example, the class of sub-graphs of bounded non-decreasing functions on $[0, 1]$, see Section 3 below.

We believe the approach of this paper is simpler. At the root of it lies the fact that we estimate G by an element of a finite approximating class. Indeed, we cannot think of any situation where one would *not* estimate the unknown set G by a representative of one or another approximating class. This allows us to stay with only relatively simple inequality (2.3), which we modify then to the form (2.8). If, for fixed n, we were obliged to consider $\delta \to 0$, these inequalities would become useless, because the number of summands would increase unboundedly, and we would be obliged to use the chaining argument. However, this is not necessary: for a given n, there exists a finite "resolution level" δ_n, see Theorem 2.2, and it is unreasonable to use δ smaller than δ_n. This leaves us with one geometric object to study, the distribution function (2.6), and thus provides a tool uniformly applicable to all classes $\mathcal{C}$.

In (2.11) we introduce the upper bound ε_n on the rate of convergence using inequality (2.8). We compare it with the sequence z_n, see (2.12), which is natural to consider as an upper bound on the rate of $d(\widehat{G}_n, G_0)$ if one does not use the local covering number. For some classes, ε_n is $o(z_n)$, but for other classes they may be of the same order of magnitude. However, under natural conditions, see Theorem 2.4, the sequence z_n is worse then ε_n in the sense that

$$\mathbb{P}\big\{d(\widehat{G}_n, G_0) > z_n\big\} \to 0.$$

And finally, still in Section 2, we show that, no matter what the rate of ε_n is, the deviations of $d(\widehat{G}_n, G_0)$ from ε_n are "typically" on the scale of $1/n$ in the sense that $\mathbb{P}\{d(\widehat{G}_n, G_0) > \varepsilon_n + L/n\}$ can be kept smaller than any given $p \in (0,1)$ for sufficiently large constant L (see Theorem 2.5). Consequently, if $1/n = o(\varepsilon_n)$, then ε_n has also the correct constant.

In Section 3 we consider two examples, one of which has been considered in Puri and Ryumgaart [9] and is of independent interest. While in Sections 2 and 3 we assume that P_1 and P_2 are fixed, in Section 4 we will see that most statements can be carried over to the case of converging P_1 and P_2 with sample size n replaced by the "effective" sample size.

The Local Covering Number and Inequalities for $\mathbb{P}\{d(\widehat{G}_n, G_0) > \varepsilon\}$

Throughout the paper we denote $d(G, G_0) = F(G \triangle G_0)$. Let us introduce some notation for the first two moments of $\xi(Y_i)$ and $L_n(G, G_0)$. As a score function ξ one can choose any bounded function such that

$$\alpha_2 := \int \xi(y)\, dP_2(y) > \int \xi(y)\, dP_1(y) := \alpha_1.$$

The reader will notice below that, although the larger $\alpha_2 - \alpha_1$ the better constants we will have, the rates as such will not be affected. To simplify the notation, we assume that ξ is shifted by $(\alpha_2 + \alpha_1)/2$ and hence

$$\alpha = \int \left\{\xi(y) - \frac{\alpha_2 + \alpha_1}{2}\right\} dP_2(y) > 0 > \int \left\{\xi(y) - \frac{\alpha_2 + \alpha_1}{2}\right\} dP_1(y) = -\alpha.$$

Then one obtains

$$\mu(G, G_0) := E\big[(I_{\{X_i \in G\}} - I_{\{X_i \in G_0\}})\xi(Y_i)\big] = -\alpha\, d(G, G_0). \tag{2.1}$$

For the variance, with $\beta_i = \int \xi^2(y)\, dP_i(y)$, $i = 1, 2$, and $\beta = \max(\beta_1, \beta_2)$, one obtains

$$\beta(X) := E[\xi^2(Y) \mid X] = \beta_2 I_{\{X \in G_0\}} + \beta_1 I_{\{X \notin G_0\}}$$

and

$$\sigma^2(G, G_0) := \mathrm{Var}([I_{\{X_i \in G\}} - I_{\{X_i \in G_0\}}]\xi(Y_i)) \le \int_{G \triangle G_0} \beta(x)\, dF(x),$$

so that

$$\sigma^2(G, G_0) \le \beta\, d(G, G_0). \tag{2.2}$$

Let $\delta < \varepsilon$. Denote $\mathcal{G}' = \{G \in \mathcal{N}_\delta \colon d(G, G_0) \ge \varepsilon\}$ and let $G'' \in \mathcal{N}_\delta$ be such that $d(G'', G_0) \le \delta$. We have

$$\mathbb{P}\{d(\widehat{G}_n, G_0) \ge \varepsilon\} \le \sum_{G' \in \mathcal{G}'} \mathbb{P}\{L_n(G', G_0) > L_n(G'', G_0)\}. \tag{2.3}$$

We will estimate each probability in the sum using Bennett's exponential inequality (see, e.g., Shorack and Wellner [21], p. 852, (d)) and this will lead us to an inequality for $\mathbb{P}\{d(\widehat{G}_n, G_0) \ge \varepsilon\}$ which we propose and study in this section.

Denote

$$\sup_y |\xi(y)| = b \qquad \text{and} \qquad \lambda = \frac{\alpha^2}{\beta}.$$

Lemma 2.1. (i) *If* $\varepsilon > \frac{3}{2}\delta$, *then*

$$\mathbb{P}\{L_n(G', G_0) > L_n(G'', G_0)\} \le \exp[-n\lambda c\{d(G', G_0) - \delta\}], \tag{2.4}$$

where $c = 0.1\psi(\gamma)$ *with* $\gamma = 0.2b\alpha/\beta$ *and*

$$\psi(\gamma) = \frac{2}{\gamma^2}\int_0^\gamma \log(1+y)\,dy.$$

(ii) *If* $\varepsilon > \delta + \bar{c}/n$ *and* $n\delta \to \infty$, *then*

$$\mathbb{P}\{L_n(G', G_0) > L_n(G'', G_0)\} \le \exp\Big[-\lambda\frac{\bar{c}}{4\delta}\{d(G', G_0) - \delta\}(1 + o(1))\Big]. \tag{2.5}$$

Remark 2.1. Using Hoeffding's inequality (see, e.g., Shorack and Wellner [21], p. 855) one could obtain the following inequality

$$\mathbb{P}\{L_n(G', G_0) > L_n(G'', G_0)\} \le \exp\Big[-\frac{n\alpha}{2b}\{d(G', G_0) - \delta\}^2\Big].$$

Since $d(G, G_0) \le 1$, this inequality gives in our situation a much less accurate bound.

Proof of Lemma 2.1. Let us abbreviate

$$L'_{n0} = L_n(G', G_0) - EL_n(G', G_0) \qquad \text{and} \qquad \mu' = \mu(G', G_0)$$

and define L''_{n0} and μ'' likewise. Then

$$\mathbb{P}\{L_n(G', G_0) > L_n(G'', G_0)\} = \mathbb{P}\{L'_{n0} - L''_{n0} > -n(\mu' - \mu'')\}.$$

Apply Bennett's inequality to this probability:

$$\mathbb{P}\big\{L'_{n0} - L''_{n0} > -n(\mu' - \mu'')\big\} \le \exp\bigg\{-\frac{n(\mu' - \mu'')^2}{2\sigma^2}\psi\Big(\frac{b|\mu' - \mu''|}{\sigma^2}\Big)\bigg\},$$

where σ^2 denotes the variance of one summand of the sum $L'_{n0} - L''_{n0}$. Now use (2.1) and (2.2) to bound the exponent from above. We have $\mu' - \mu'' \ge \alpha\{d(G', G_0) - \delta\}$, $\sigma^2 \le \beta\{d(G', G_0) + \delta\}$. Since $x\psi(x)$ is an increasing function, we can substitute these bounds in the previous inequality, which gives

$$\mathbb{P}\big\{L_n(G', G_0) > L_n(G'', G_0)\big\} \le \exp\Big[-\frac{n\lambda}{2}\frac{\{d(G', G_0) - \delta\}^2}{d(G', G_0) + \delta}\psi(\gamma')\Big],$$

where

$$\gamma' = \frac{b\alpha}{\beta}\,\frac{\{d(G',G_0)-\delta\}}{\{d(G',G_0)+\delta\}}.$$

Since $(z-\delta)/(z+\delta)$ is also an increasing function, we can simplify the exponent further: for $\varepsilon \geq (3/2)\delta$ we have

$$c \leq \frac{d(G',G_0)-\delta}{2\{d(G',G_0)+\delta\}}\psi(\gamma'),$$

which, after substitution into the previous inequality, gives (2.4), while for $\varepsilon \geq \delta + \bar{c}/n$ we have

$$\frac{\bar{c}}{2n\delta+\bar{c}}\psi\left(\frac{b\alpha}{\beta}\,\frac{\bar{c}}{2n\delta+\bar{c}}\right) \leq \frac{d(G',G_0)-\delta}{2\{d(G',G_0)+\delta\}}\psi(\gamma').$$

As $n\delta \to \infty$ the left-hand side becomes $\bar{c}/2n\delta(1+o(1))$, which leads to inequality (2.5). □

Using the local covering number (1.3), let us introduce now

$$V_\delta(t,G_0) = V_\delta(t) = \frac{N_\delta(t,G_0)}{N_\delta}. \tag{2.6}$$

Clearly V_δ is a discrete distribution function with a finite number of jumps, and this number increases as $\delta \to 0$. As a result of (2.3) and (2.4) we obtain that, for $\varepsilon > \frac{3}{2}\delta > 0$,

$$\mathbb{P}\big\{d(\widehat{G}_n,G_0) > \varepsilon\big\} \leq N_\delta \int_\varepsilon^1 e^{-n\lambda c(t-\delta)}\,V_\delta(dt). \tag{2.7}$$

In certain cases $n_1 = n\lambda$ becomes a natural quantity (see Section 4). For the present, however, it is better to keep n. Besides, denote $c_1 = \lambda c$.

The probability $\mathbb{P}\{d(\widehat{G}_n,G_0) > \varepsilon\}$ and its upper bound depend on δ, and it is natural to make this upper bound as small in δ as we can for every ε. One can argue that unlike N_δ the distribution function V_δ is "stable" in δ. With this in mind we summarize the construction in the following statement.

Theorem 2.2. *For* $\varepsilon > \frac{3}{2}\delta > 0$

$$\min_\delta \mathbb{P}\{d(\widehat{G}_n,G_0) > \varepsilon\} \leq N_{\delta_n} e^{nc_1\delta_n} \int_\varepsilon^1 e^{-nc_1 t}\,V_{\delta_n}(dt), \tag{2.8}$$

where

$$\delta_n := \arg\min_\delta N_\delta e^{nc_1\delta}. \tag{2.9}$$

The *proof* follows from (2.3) and (2.4) and the definition of δ_n. The choice of δ_n as (2.9) could be interpreted as a (quasi-) optimal resolution level. It is uniform in ε, which is quite convenient. The choice of δ as a solution of the equation $n\delta = \log N_\delta$ is very closely related to (2.9) and was systematically used, e.g., in Yang and Barron [17]. We now state some asymptotic properties of δ_n, $n \geq 1$.

Lemma 2.3. (i) *$\delta_n \to 0$ and $N_{\delta_n} e^{nc_1\delta_n} = o(e^{nc_1\Delta})$ for any $\Delta > 0$, $n \to \infty$;*

(ii) *if $N_\delta \to \infty$ as $\delta \to 0$, then $N_{\delta_n} e^{nc_1\delta_n} \to \infty$ as $n \to \infty$;*

(iii) *$n\delta_n \to \eta$, $0 < \eta < \infty$, as $n \to \infty$ iff the metric entropy $H_\delta = \log N_\delta$ satisfies the condition: there is a constant μ, which may depend on H, but not on δ, such that*

$$H_\delta - \frac{\mu}{\delta} x \leq H_{\delta+x}, \qquad -\delta \leq x \leq 1-\delta. \tag{2.10}$$

Proof. (i) Let δ'_n be such that

$$N_{\delta'_n} = e^{nc_1\delta'_n}.$$

Then $\delta'_n \to 0$ as $n \to \infty$, because if for some subsequence $\delta'_{n'} \to \Delta > 0$, then $\exp(nc_1\delta'_{n'}) \to \infty$, while $N_{\delta'_{n'}} \to N_\Delta < \infty$, which contradicts the definition of δ'_n. Then for δ_n we obtain

$$N_{\delta_n} e^{nc_1\delta_n} \leq N_{\delta'_n} e^{nc_1\delta'_n} = e^{2nc_1\delta'_n} = o(e^{nc_1\Delta})$$

for any $\Delta > 0$ and $\delta_n \to 0$.

(ii) Follows from the fact that $\delta_n \to 0$ and the condition that $N_\delta \to \infty$ as $\delta \to 0$.

(iii) Suppose (2.10) is satisfied. Take $\delta_n = \mu(nc_1)^{-1}$. Then

$$H_{\delta_n} - nc_1(\delta - \delta_n) \leq H_\delta, \qquad 0 \leq \delta \leq 1,$$

or

$$H_{\delta_n} + nc_1\delta_n \leq H_\delta + nc_1\delta, \qquad 0 \leq \delta \leq 1,$$

which is equivalent to (2.9). Now suppose the last inequality is satisfied and $n\delta_n \to \eta$. Then

$$H_{\delta_n} - nc_1\delta_n \frac{\delta - \delta_n}{\delta_n} \leq H_\delta \qquad \text{or} \qquad H_{\delta_n} - \frac{\eta c_1}{\delta_n} x \leq H_{\delta_n + x}. \qquad \square$$

We introduce now two sequences, which will be systematically used in this paper. Let $\varepsilon_n(p)$, $n \geq 1$, be a sequence such that

$$\lim_{n\to\infty} N_{\delta_n} \int_{\varepsilon_n(p)}^1 e^{-nc_1(t-\delta_n)} V_{\delta_n}(dt) = p, \qquad 0 < p \leq 1, \tag{2.11}$$

and let $z_n(p)$, $n \geq 1$, be a sequence such that

$$\lim_{n\to\infty} N_{\delta_n} e^{-nc_1\{z_n(p)-\delta_n\}} = p, \qquad p \leq 1. \tag{2.12}$$

A frequently used bound for the sum in (2.3) would be

$$N_\delta \max_{G'} \mathbb{P}\big\{L_n(G', G_0) > L_n(G'', G_0)\big\},$$

which would lead to the inequality

$$\mathbb{P}\big\{d(\widehat{G}_n, G_0) > \varepsilon\big\} \le N_\delta e^{-n\lambda c(\varepsilon-\delta)}. \tag{2.13}$$

Therefore, as follows from (2.13), $z_n(p)$, $n \ge 1$, would provide the upper bound for the rate of convergence of $d(\widehat{G}_n, G_0)$ to 0, if we do not exploit the local covering number. If we do, the upper bound will be given by $\varepsilon_n(p)$.

From Lemma 2.3 (i), (ii) one can deduce that if $N_\delta \to \infty$ as $\delta \to 0$, then

$$z_n(p) \to 0 \qquad \text{but} \qquad n z_n(p) \to \infty \tag{2.14}$$

for any $p > 0$. From (2.14) we see that in no case can $z_n(p)$ be of order $1/n$. We will find later that in some cases $\varepsilon_n(p) = o(z_n(p))$, $n \to \infty$. However, more interesting is that even if $\varepsilon_n(p)$ and $z_n(p)$ are of the same order of magnitude, the inequalities (2.8) and (2.13) lead to entirely different bounds.

Theorem 2.4. *Assume $\delta_n \to 0$. If either $V_{\delta_n}(z) \to 0$ for $z \to 0$ or $V_{\delta_n}(z + T/n) - V_{\delta_n}(z) \to 0$ for any $T > 0$ and all sufficiently small z, then for any $0 < p \le 1$*

$$\mathbb{P}\{d(\widehat{G}_n, G_0) > z_n(p)\} \to 0.$$

Remark 2.3. The condition of Theorem 2.4 requires that either $V_{\delta_n}(t)$ does not concentrate around G_0 or the increment of $V_{\delta_n}(t)$ is small on $1/n$ scale. In Example 3.1 one can see that this is still true even in the extreme case of classes $\mathcal{C}$ having only one limit point.

Proof of Theorem 2.4. Since $z_n(p) \ge z_n(1)$, it is sufficient to consider $z_n(1)$. First use integration by parts for the integral in the right-hand side of (2.8). With $\varepsilon = z_n(1)$ we obtain

$$e^{-nc_1}[1 - V_{\delta_n}(z_n(1))] + nc_1 \int_{z_n(1)}^{1} e^{-nc_1 t}\left[V_{\delta_n}(t) - V_{\delta_n}(z_n(1))\right] dt.$$

According to (2.14) the first summand here is $o(e^{-nc_1 z_n(1)})$. Moreover,

$$nc_1 \int_{z_n(1)+T/nc_1}^{1} e^{-nc_1 t}\left[V_{\delta_n}(t) - V_{\delta_n}(z_n(1))\right] dt \le e^{-nc_1 z_n(1) - T}.$$

At the same time

$$\begin{aligned} nc_1 &\int_{z_n(1)}^{z_n(1)+T/nc_1} e^{-nc_1 t}\left[V_{\delta_n}(t) - V_{\delta_n}(z_n(1))\right] dt \\ &\qquad \le \left[V_{\delta_n}\left(z_n(1) + \frac{T}{n}\right) - V_{\delta_n}(z_n(1))\right] e^{-nc_1 z_n(1)} = o(e^{-nc_1 z_n(1)}). \end{aligned}$$

Since $e^{-nc_1 z_n(1)} = (N_{\delta_n} e^{nc_1\delta_n})^{-1}$, this completes the proof. □

Under the conditions of this theorem the behavior of the δ-net beyond shrinking the $z_n(1)$-neighborhood of G_0 has no influence on $d(\widehat{G}_n, G_0)$.

In the next theorem we consider how far can $\varepsilon_n(p)$ lie from $\varepsilon(1)$.

Theorem 2.5. *With δ_n defined in* (2.9)*, let $\varepsilon_n(1)$, $n \geq 1$, be a sequence defined in* (2.11)*. If the sequence of distributions*

$$d\widetilde{V}_n(\tau) = \frac{e^{-\tau} dV_{\delta_n}\big(\frac{\tau}{nc_1} + \varepsilon_n(1)\big)}{\int_0^\infty e^{-\tau} dV_{\delta_n}\big(\frac{\tau}{nc_1} + \varepsilon_n(1)\big)}, \qquad \tau \geq 0, \tag{2.15}$$

is weakly compact, then for any $p \in (0,1)$ there is a constant $L = L(p)$ such that

$$\mathbb{P}\bigg\{ d(\widehat{G}_n, G_0) > \varepsilon_n(1) + \frac{L(p)}{nc_1} \bigg\} \leq p.$$

In particular, if $\varepsilon_n(1) \geq \mathrm{const}/nc_1$, then $d(\widehat{G}_n, G_0) = O_P(\varepsilon_n(1))$.

Proof. According to the definition of $\varepsilon_n(1) = \varepsilon_n$, $n \geq 1$,

$$N_{\delta_n} e^{nc_1(\delta_n - \varepsilon_n)} \omega_n(\varepsilon_n) \to 1, \qquad n \to \infty,$$

where

$$\omega_n(\varepsilon_n) = \int_{\varepsilon_n}^1 e^{-nc_1(t-\varepsilon_n)}\, dV_{\delta_n}(t).$$

The weak compactness condition of $\widetilde{V}_n$, $n \geq 1$, implies that for $\varepsilon'_n = \varepsilon_n + L/nc_1$

$$\begin{aligned}\mathbb{P}\{d(\widehat{G}_n, G_0) > \varepsilon'_n\} &\leq N_{\delta_n} e^{nc_1(\delta_n - \varepsilon'_n)} \omega_n(\varepsilon'_n) \sim e^{-nc_1(\varepsilon'_n - \varepsilon_n)} \frac{\omega_n(\varepsilon'_n)}{\omega_n(\varepsilon_n)} \\ &= \int_{\varepsilon'_n}^1 \frac{e^{-nc_1(t-\varepsilon_n)} V_n(dt)}{\omega_n(\varepsilon_n)} = \int_L^\infty \frac{e^{-\tau} dV_n\big(\frac{\tau}{nc_1} + \varepsilon_n\big)}{\omega_n(\varepsilon_n)}\end{aligned}$$

and the right-hand side can be made arbitrarily small. □

Observe that the sequence $\varepsilon_n(p)$, $n \geq 1$, required by definition (2.11) does not always exist as well as the weak compactness condition for the sequence of distributions (2.15) is not always satisfied as the following lemma shows. However, the situations when this occurs are rather exceptional.

Lemma 2.6. *If G_0 is an isolated element of $\mathcal{C}$, that is, if*

$$\inf_{G \neq G_0} d(G, G_0) = t_0 > 0,$$

then

$$\mathbb{P}\{d(\widehat{G}_n, G_0) > 0\} \leq N_{\delta_n} e^{nc_1(\delta_n - t_0)} = o(e^{-nt}), \qquad n \to \infty, \tag{2.16}$$

for any $t < t_0$.

Proof. For any ε_n such that $t_0 > \varepsilon_n > (3/2)\delta_n$ the inequality

$$\mathbb{P}\{d(\widehat{G}_n, G_0) > \varepsilon_n\} \leq N_{\delta_n} e^{nc_1(\delta_n - t_0)}, \qquad n \to \infty,$$

follows from (2.7). Its right-hand side is $o(e^{-nt})$ as follows from Lemma 2.3 (i). However, since G_0 is an isolated point, it is clear that $\mathbb{P}\{d(\widehat{G}_n, G_0) > \varepsilon_n\} = \mathbb{P}\{d(\widehat{G}_n, G_0) > 0\}$. □

Theorem 2.5 shows that under a mild assumption $d(\widehat{G}_n, G_0)$ can exceed $\varepsilon_n(1)$ only by a quantity of order $1/nc_1$. However, it is very interesting to learn how far can $\varepsilon_n(1)$ itself lie from δ_n. The next theorem describes conditions for $\varepsilon_n(1)$ also to be not further than const $/nc_1$ from $(3/2)\delta_n$. In practice one usually obtains upper bounds for $N_\delta(t)$, and therefore the conditions of the next theorem are given in terms of $N_\delta(t)$ rather than of $N_\delta(dt)$.

Theorem 2.7. *Let* $b_n := z_n(1) - \delta_n = (nc_1)^{-1} \log N_{\delta_n}$. *If*

$$\limsup_{n\to\infty} nc_1 \int_0^{b_n} e^{-nc_1 t}\big[N_{\delta_n}(t + q_n) - N_{\delta_n}(q_n)\big]\, dt \leq \varphi(L), \tag{2.17}$$

$$q_n = \frac{3}{2}\delta_n + \frac{L}{nc_1},$$

for every $L > 0$ *and* $\varphi(L)e^{-L} \to 0$ *as* $L \to \infty$, *then for any* $p \in (0, 1]$ *there exists* $L = L(p)$ *such that*

$$\mathbb{P}\{d(\widehat{G}_n, G_0) > \delta_n + L(p)/nc_1\} \leq p\{1 + o(1)\}, \qquad n \to \infty,$$

and if there exists $\varepsilon_n(p)$, $n \geq 1$, *satisfying* (2.11), *then*

$$\varepsilon_n(p) \leq \max\big(\delta_n + L(p)/nc_1\, (\tfrac{3}{2})\delta_n\big).$$

Conversely, if there exists a constant L *such that* $\varepsilon_n(p) \leq \delta_n + L/nc_1$, *then* $\varphi(L) < \infty$ *for this* L.

Remark 2.4. Examples show that if $\varphi(L)$ exists, then the requirement

$$e^{-L}\varphi(L) \to 0 \qquad \text{as} \quad L \to \infty$$

is not strong one. In many cases $\varphi(L)$ remains simply bounded. However, in Khmaladze et al. [20] one can see that the upper limit in (2.17) can be infinity (see, e.g., Corollary 2.1 (ii) and Example 3 in that paper) and the difference $\varepsilon_n(1) - \delta_n$ is indeed larger than $1/n$.

Proof. Remark first that for the given choice of b_n,

$$nc_1 \int_{b_n}^1 e^{-nc_1 t}\big[N_{\delta_n}(t + c\delta_n + L/nc_1) - N_{\delta_n}(c\delta_n + L/nc_1)\big]\, dt$$
$$\leq N_{\delta_n} nc_1 \int_{b_n}^1 e^{-nc_1 t}\, dt \leq N_{\delta_n} e^{-nc_1 b_n} = 1.$$

In inequality (2.7) put $\varepsilon = q_n$ and choose $L = L(p)$ such that $e^{-L}\{\varphi(L)+1\} = p \leq 1$. Then integration by parts yields

$$\begin{aligned}
&N_\delta e^{-nc_1(q_n-\delta_n)} \int_{q_n}^1 e^{-nc_1(t-q_n)} V_{\delta_n}(dt) \\
&\quad = e^{-L} nc_1 \int_0^{1-q_n} e^{-nc_1 t}\bigl[N_{\delta_n}(t+q_n) - N_{\delta_n}(q_n)\bigr]\, dt \\
&\quad \leq e^{-L}\{\varphi(L)+1\} + o(1) = p + o(1), \qquad n \to \infty.
\end{aligned}$$

Since the right-hand side of (2.7) is a decreasing function of ε, then if there exists a sequence $\varepsilon_n(p)$, $n \geq 1$, satisfying (2.11), it must be such that $\varepsilon_n(p) \leq \max\bigl(\delta_n + L(p)/nc_1\, 1.5\delta_n\bigr)$ for all sufficiently large n. Now suppose the last requirement on $\varepsilon_n = \varepsilon_n(p)$, $n \geq 1$, is true. Then

$$\begin{aligned}
&e^{-L} nc_1 \int_0^{b_n} e^{-nc_1 t}\bigl[N_{\delta_n}(t+q_n) - N_{\delta_n}(q_n)\bigr]\, dt \\
&\quad \leq e^{-L} nc_1 \int_0^{1-q_n} e^{-nc_1 t}\bigl[N_{\delta_n}(t+q_n) - N_{\delta_n}(q_n)\bigr]\, dt \\
&\quad = N_{\delta_n} e^{-nc_1(q_n-\delta_n)} \int_{q_n}^1 e^{-nc_1(t-q_n)} V_{\delta_n}(dt) \\
&\quad \leq N_{\delta_n} e^{-nc_1(\varepsilon_n-\delta_n)} \int_{\varepsilon_n}^1 e^{-nc_1(t-\varepsilon_n)} V_{\delta_n}(dt)
\end{aligned}$$

because of monotonicity in ε. The last expression converges to p by definition of $\varepsilon_n(p)$, $n \geq 1$. Hence

$$nc_1 \int_0^{b_n} e^{-nc_1 t}\bigl[N_{\delta_n}(t+q_n) - N_{\delta_n}(q_n)\bigr]\, dt \leq e^L + o(1), \qquad n \to \infty. \qquad \square$$

3. Two Examples

We start with an example, which may look artificial and which indeed carries no practical importance. However, it illustrates a point of some theoretic value.

One can think that the difference between the application of local covering number and of the covering number will be unimportant at least for the classes which are "highly concentrated" around few elements. One can argue that in such classes everything is already so very much "local" that the use of local covering number will hardly bring anything better. We will see, however, that this is not generally true.

Example 3.1. Consider the situation when for arbitrarily small but fixed t

$$N_\delta(t)/N_\delta \to 1 \qquad \text{as} \quad \delta \to 0.$$

Namely, suppose $\mathcal{C}$ is a closed monotone sequence, $\mathcal{C} = \{G_k,\, k \geq 1,\, G_0\}$, and either $G_1 \supset G_2 \supset \dots$, $G_0 = \bigcap_{k=1}^\infty G_k$, or $G_1 \subset G_2 \subset \dots$, $G_0 = \bigcup_{k=1}^\infty G_k$. Denote $x_k = d(G_k, G_0)$. Then the problem reduces to estimation of N_δ and $N_\delta(t)$ for a positive sequence $x_k \to 0$. Denote $y_k = x_k - x_{k+1}$ and to avoid unnecessary complications suppose that y_k form a monotone sequence. For any $\delta > 0$ let

$$k(\delta) = \inf\{k\colon y_i \le \delta \ \text{ for } \ \text{all } i \ge k\}$$

and take

$$N_\delta = k(\delta) + \left[\frac{x_{k(\delta)}}{2\delta}\right] + 1,$$

where $[z]$ stands for the integer part of z. This N_δ corresponds to the δ-net constructed as follows: include in $\mathcal{N}_\delta$ all elements with $x_i \ge x_{k(\delta)}$ and for the rest of the sequence, starting from $x_{k(\delta)}$, take the δ-net of uniformly spaced G's, not necessarily in $\mathcal{C}$, located at distance δ from each other. There will be no more than $[x_{k(\delta)}/2\delta] + 1$ of such G's. Below we neglect the difference between $[x_{k(\delta)}/2\delta] + 1$ and $x_{k(\delta)}/2\delta$ for simplicity of notation.

Denote $x^{-1}(t) = \inf\{k\colon x_k \le t\}$. Then for $N_\delta(t)$ we obtain

$$N_\delta(t) = \begin{cases} k(\delta) - x^{-1}(t) + x_{k(\delta)}/2\delta, & t \ge x_{k(\delta)}, \\ t/2\delta, & t \le x_{k(\delta)}. \end{cases}$$

It is more interesting to consider “quickly” converging sequences. Let $x_k = a^k$, $0 < a < 1$, form a geometrically converging sequence. Then, $y_k = (1-a)a^k = \delta$ leads to $k(\delta) = \log(\delta/(1-a))/\log a$ and $x_{k(\delta)} = \delta/(1-a)$. Hence

$$N_\delta = \frac{\log\delta - \log(1-a)}{\log a} + \frac{1}{2(1-a)},$$

so that it increases quite slowly with $\delta \to 0$. The optimal δ_n of (2.9), the upper bound $z_n(1)$, and the bound b_n of Theorem 2.7 are

$$\delta_n = \frac{1}{nc_1\log nc_1} + O\Big(\frac{\log\log nc_1}{nc_1\log^2 nc_1}\Big), \quad z_n(1) \sim b_n = \frac{\log nc_1}{nc_1} + O\Big(\frac{\log\log nc_1}{nc_1}\Big).$$

We have $q_n = \delta_n + L/nc_1 > x_{k_{\delta_n}}$, while $x^{-1}(t) = \log t/\log a$ and the integrability condition (2.17) of

$$N_{\delta_n}(t + q_n) - N_{\delta_n}(q_n) = \frac{1}{\log a}\log\Big[1 + \frac{t}{\delta_n + L/nc_1}\Big]$$

for all $L > 0$ becomes apparent:

$$\lim_{n\to\infty} nc_1 \int_0^{b_n} e^{-nc_1 t}\log\Big[1 + \frac{t}{\delta_n + L/nc_1}\Big]\,dt \le L\int_0^\infty e^{-\tau L}\log(1+\tau)\,d\tau = \varphi(L),$$

which actually is a decreasing function in L. Consequently if $e^{-L}\{\varphi(L) + 1\} = 1$, then

$$\mathbb{P}\Big\{d(\widehat{G}_n, G_0) > \frac{3}{2n\log n} + \frac{L'}{nc_1}\Big\} \le e^{-(L'-L)}\{1 + o(1)\}, \qquad n \to \infty.$$

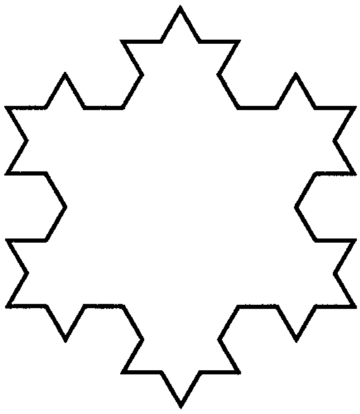

FIGURE 1. Depicts G_3 of the sequence.

One possible example of $\{G_k,\, k \geq 1,\, G_0\}$ forming a geometrically converging sequence is illustrated by the Serpinski star (Figure 1).

Here $a = 4/9$ and for $k \geq 2$ we have $d(G_k, G_{k+1}) = (3\sqrt{3}/4)4^{k+1}(a/3^{k+1})^2$. The figure clearly shows that the "regularity" of the boundary of the change set per se is irrelevant to our problem.

Many formulations of the classical change-point problem are connected with change-sets forming VČ-classes. We consider this situation in more detail in Khmaladze *et al.* [20]. As our next example in this section consider the class of subgraphs of bounded monotone functions on a compact set. The change-set problem for this class was studied earlier by Puri and Ryumgaart [9]. The covering number of this class, shown in Lemma 3.1, is essentially larger than any power of δ (cf. (3.2)).

Example 3.2. Let $\mathcal{C}' = \{f\colon [0,1] \to [0,1],\, \nearrow\}$ and let

$$\mathcal{C} = \{f_{\text{sub}},\, f \in \mathcal{C}'\} \quad \text{with} \quad f_{\text{sub}} = \{(x,y) \in [0,1]^2 \colon f(x) \geq y\}.$$

Let Λ_2 denote the Lebesgue measure on $[0,1]^2$ and take

$$d(f_{\text{sub}}, g_{\text{sub}}) = \Lambda_2(f_{\text{sub}} \triangle g_{\text{sub}}) = \int |f(x) - g(x)|\, dx. \tag{3.1}$$

Hence $\mathcal{C}'$ with L_1-distance is isometric to $\mathcal{C}$ with the distance $\Lambda_2(f_{\text{sub}} \triangle g_{\text{sub}})$.

Below we present asymptotics for the covering number and local covering numbers at two different elements of $\mathcal{C}$. As a corollary, this will show that the conditions of Theorem 2.4 are satisfied for this case. It will also reveal (Theorem 3.2) that the behavior of local covering numbers is uneven in f: different f have $N_\delta(t, f)$ of different rate in t and δ.

First consider a δ-net of $\mathcal{C}$. Assume $m = 1/\delta$ an integer number for simplicity of notation and let $x_j = j/m$, $y_k = k/m$, $j = 0, \ldots, m$, $k = 0, \ldots, m$. Let

$$\mathcal{N}_\delta = \{f_\delta \colon f_\delta \in \{y_k\}_0^m,\ f_\delta \text{ is constant on each } [x_j, x_{j+1}),\ \nearrow\}.$$

Lemma 3.1. (i) $\mathcal{N}_\delta$ *is a δ-net for $\mathcal{C}$.*
(ii) *With $m = 1/\delta$*

$$N_\delta = \frac{(2m)!}{(m!)^2} \sim 2^{2m}\frac{1}{\sqrt{\pi m}}, \qquad m \to \infty, \tag{3.2}$$

while

$$\delta_n \sim \frac{\sqrt{2\log 2}}{\sqrt{nc_1}} \qquad \text{and} \qquad z_n(1) \sim 2\frac{\sqrt{2\log 2}}{\sqrt{nc_1}}, \quad n \to \infty.$$

The proof of (i) is left to reader, while the proof of (ii) can be obtained in a way similar to the proof of (i) in Theorem 3.2 below.

We see that the rate of convergence for this class is at least $1/\sqrt{nc_1}$. As far as we understand it, the rate of convergence shown in Puri and Ryumgaart [9] depended on the way the bounding function was estimated and was slower than $1/\sqrt{nc_1}$.

We also see that the difference between δ_n and $z_n(1) \sim 2\delta_n$ is what can be called "practically unimportant". However, there is certain refinement of the "rate of convergence" statement if we realize that actually $\mathbb{P}\{d(\widehat{G}_n, G_0) > z_n(1)\} \to 0$.

To show this we need to consider $N_\delta(t)$. First consider the sup-metric on $\mathcal{C}$ instead of L_1-metric, and denote $N_{\delta,u}(t) := N_{\delta,u}(t,f)$ the number of elements of $\mathcal{N}_\delta$ satisfying the inequality

$$\sup_{0\le x\le 1} |f_\delta(x) - f(x)| \le t.$$

Denote

$$\varphi_k(l) = \frac{(l+1)\cdots(l+k)}{k!}, \qquad l = 0, 1, \dots, m, \tag{3.3}$$

and let $\varphi_0 = \mathbf{1}$ be the $m+1$-dimensional vector with all coordinates equal to 1.

Theorem 3.2. *Let $t = L\delta$ and assume L is an integer.*
(i) *Let $f_1(x) = \text{const}$, with $t < \text{const} < 1 - t$. Then*

$$N_{\delta,u}(t, f_1) = \frac{(2L+m)!}{(2L)!m!}$$

and for $L = O(\sqrt{m})$

$$N_{\delta,u}(t, f_1) \sim \frac{m^{2L}e^{2L^2/m}}{(2L)!}, \qquad m \to \infty.$$

(ii) *Let $f_2(x) = x$, $0 \le x \le 1$. Then*

$$\varphi_{m+1}(2L) + \sum_{j=1}^{2L}\binom{m}{j}\varphi_{m-j}(2L) - \varphi_L(L)$$
$$\le N_{\delta,u}(t, f_2) \le \sum_{j=0}^{2L}\binom{m}{j}\varphi_{m-j+1}(2L) + \varphi_L(L)$$

and for $L = O(\sqrt{m})$

$$N_{\delta,u}(t,f_2) \sim \binom{m}{2L}\varphi_{m-2L}(2L) \sim \left(\frac{m^{2L}}{(2L)!}\right)^2.$$

Proof. (i) Direct counting shows that

$$N_{\delta,u}(t,f_1) = \sum_{i_m=0}^{2L} \cdots \sum_{i_1=0}^{i_2} \sum_{i_0=0}^{i_1} 1.$$

This can be rewritten as

$$N_{\delta,u}(t,f_1) = \langle \mathbf{1}_L, S_L^{m-1}\mathbf{1}_L\rangle,$$

where $\mathbf{1}_L = (1,\dots,1)^T \in \mathbb{R}^{2L+1}$ and the operator S_L has $(2L+1)\times(2L+1)$ matrix of the form

$$S_L = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ \multicolumn{5}{c}{\dotfill} \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix}.$$

For the factorial moments φ_k defined by (3.3) and restricted to $1, 2, \dots, 2L+1$ we have (cf., e.g., Gelfond [23], p. 31) that $S_L\varphi_k = \varphi_{k+1}$. Therefore

$$N_{\delta,u}(t,f_1) = \langle \mathbf{1}_L, \varphi_{m-1}\rangle = \varphi_m(2L) = \frac{(2L+1)\cdots(2L+m)}{m!}.$$

The asymptotics of $N_{\delta,u}(t,f_1)$ can now be obtained by the Stirling formula.

(ii) It can also be seen that the number $N'_{\delta,u}(t,f_2)$ of step-functions f_δ in the uniform $L\delta$-neighborhood of f_2 which are allowed to start at $x=0$ from a value $\geq -L\delta$ and finish at $x=1$ at a value $1+L\delta$ differs from $N_{\delta,u}(t,f_2)$, for fixed L, only by a quantity depending on L but not on m:

$$0 < N'_{\delta,u}(t,f_2) - N_{\delta,u}(t,f_2) < \frac{1}{2}\varphi_L(L),$$

while $N'_{\delta,u}(t,f_2)$ itself is equal to

$$N'_{\delta,u}(t,f_2) = \sum_{j_m=0}^{2L} \cdots \sum_{j_1=0}^{j_2\wedge 2L} \sum_{j_0=0}^{j_1\wedge 2L} 1.$$

The expression on the right-hand side can be rewritten as

$$N'_{\delta,u}(t,f_2) = \langle \mathbf{1}_L, M^{m-1}\mathbf{1}_L\rangle, \qquad \mathbf{1}_L = (1,\dots,1)^T \in \mathbb{R}^{L+1},$$

where M is the operator with the $(2L+1)\times(2L+1)$-matrix

$$M = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ \multicolumn{5}{c}{\dots\dots\dots\dots\dots} \\ 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix}.$$

Let us represent M as the sum $M = S_L + N$, where the operator N is, obviousely, nilpotent: $N^{2L+1} = 0$ (see, e.g., Glazman and Ljubič [24], p. 123, or Hirsch and Smale [25], p. 116). As a consequence, we have

$$\langle \mathbf{1}_L, M^{m-1}\mathbf{1}_L\rangle = \sum_{j=0}^{2L} \binom{m-1}{j} \langle \mathbf{1}_L, S_L^{m-j-1} N^j \mathbf{1}_L\rangle.$$

Since $N^j\mathbf{1}_L = (1,\dots,1,0,\dots,0)^T$ is a vector with the last j coordinates equal to zero, we obtain $e_1 = (1,0,\dots,0)^T \le N^j\mathbf{1}_L \le \mathbf{1}_L$. Therefore

$$S_L^{m-j-1}e_1 \le S_L^{m-j-1}N^j\mathbf{1}_L \le S_L^{m-j-1}\mathbf{1}_L.$$

Note also that $S_L^{m-j-1}e_1 = S_L^{m-j-2}\mathbf{1}_L$ and $\langle \mathbf{1}_L, S_L^{m-j-1}\mathbf{1}_L\rangle = \varphi_{m-j}(2L)$. Thus

$$\varphi_{m-j}(2L) \le \langle \mathbf{1}_L, S_L^{m-j}N^j\mathbf{1}_L\rangle \le \varphi_{m-j+1}(2L).$$

Therefore

$$\varphi_m(2L) + \sum_{j=1}^{2L}\binom{m}{j}\varphi_{m-j-1}(2L) - \varphi_L(L)$$
$$\le N_{\delta,u}(t,f_2) \le \sum_{j=0}^{2L}\binom{m}{j}\varphi_{m-j}(2L).$$

Using the asymptotic relation

$$\varphi_k(2L) \sim \frac{k^{2L}}{(2L)!}e^{\frac{4L^2}{k}}, \qquad k\to\infty,$$

we find that the summands with $j = 2L$ on both sides of the inequalities are the leading terms and are both of the same order. Hence

$$N_{\delta,u}(t,f_2) \sim \binom{m}{2L}\varphi_{m-2L}(2L) \sim \left(\frac{m^{2L}}{(2L)!}\right)^2. \qquad \square$$

Hence we see that due to the geometry of our compact set $\mathcal{C}$ the neighborhood of the same width $t = L\delta$ of the increasing function f_2 is much richer than that of the constant function f_1. However, both neighborhoods are just VČ-classes, while the whole compact set is a Dudley class.

Let us turn back to the metric (3.1) and consider $N_\delta(t, f)$. Since

$$\sup_{0\le x\le 1} |f(x) - g(x)| \le t \Rightarrow \int |f(x) - g(x)|\, dx \le t$$
$$\Rightarrow \sup_{0\le x\le 1} |f(x) - g(x)| \le \sqrt{2t},$$

it follows that

$$N_{\delta,u}(t, f) \le N_\delta(t, f) \le N_{\delta,u}(\sqrt{2t}, f).$$

If we choose now $t = L'\delta$ with constant L', we get $\sqrt{2t} = \sqrt{L'\delta}$ and the asymptotic expressions of Theorem 3.2 can be used. Therefore we immediately obtain that

$$N_\delta(t, f)/N_\delta \to 0 \qquad \text{for both} \quad f = f_1 \quad \text{and} \quad f = f_2$$

and the condition of Theorem 2.4 is satisfied.

4. On Asymptotically Small Changes

In this section we consider what happens if the possible change of distribution of marks on G_0 is getting smaller as the sample size n increases, that is, if P_2 and P_1 converge to each other as $n \to \infty$. This question of clear practical as well as theoretical importance was in special situations considered earlier. For instance, the case of converging P_1 and P_2 in the change-point problem on the real line (with one change-point) was considered in Dümbgen [26].

The basic observation is that nothing essentially changes in the framework of previous sections apart from the fact that the sample size n should be replaced by smaller "effective" sample size $n_1 = n\lambda$ as soon as $n_1 \to \infty$. What we need to clarify is the asymptotic behavior of the constants α, β_j, $j = 1, 2$, involved in the basic inequality (2.8), which will now vary with n.

Suppose the distribution P_2 of the marks on the change-set converges to P_1:

$$\left[\frac{dP_2}{dP_1}(y)\right]^{1/2} = 1 + \frac{1}{2\sqrt{m}} h_m(y), \qquad \int h_m^2(y) dP_1(y) \to 1,$$
$$m = m(n) \to \infty \quad \text{as} \quad n \to \infty,$$

and suppose there is no "complete mismatch" between the score function ξ and the "direction" $h_m(y)$ along which P_2 tend to P_1:

$$\liminf \int \xi(y) h_m(y)\, dP_1(y) = \alpha_0 > 0.$$

Theorem 4.1. *If $n/m \to \infty$, then under the above conditions all previous statements remain valid with n replaced by $n_1 = n\lambda \sim n/m$.*

Remark 4.1. For $n_1 \to \infty$ we need that $m = o(n)$ rather than

$$m = o(n/(\log\log n)^2)$$

as can be found in the literature.

Proof. Under the assumptions above

$$\alpha = \frac{1}{2} \int \xi(y) \left[\left\{ 1 + \frac{1}{2\sqrt{m}} h_m(y) \right\}^2 - 1 \right] dP_1(y)$$
$$= \frac{1}{2\sqrt{m}} \int \xi(y)\, h_m(y)\, dP_1(y) + O\Big(\frac{1}{m}\Big),$$

while $\beta_1 = \int \xi^2(y) dP_1(y)$ remains constant. Since the score function ξ is bounded and P_2 converges to P_1, we get $\beta_2 \to \beta_1$. Therefore

$$\frac{1}{4} \geq \limsup m\lambda \geq \liminf m\lambda \geq \frac{\alpha_0^2}{4\beta_1},$$

while the parameter $\gamma \to 0$ and $c \to 0.1$. Inequality (2.5) is then still true. The rest of the proof follows from the formulations of the statements above since we everywhere indicated the rates in terms of nc_1 rather than just n. □

Example 4.1. Mammen and Tsybakov [1] consider the MLE and the score function ξ chosen as $\xi = \log(dP_2/dP_1)$, while the marks Y_i, $i = 1, \ldots, n$, are Bernoulli random variables with $P_2\{Y_i = 1\} = p_2$ and $P_1\{Y_i = 1\} = p_1$. The authors assume in addition that $p_1 = \frac{1}{2} - p$ and $p_2 = \frac{1}{2} + p$, $p < 1/2$, which leads to the equality $-\alpha_1 = \alpha_2$.

According to Theorem 4.1 this equality is asymptotically true, i.e., $-\alpha_1/\alpha_2 \to 1$, whenever $p_2, p_1 \to p_0$, $0 < p_0 < 1$. The "effective" sample size is of order $n|p_1 - p_2|^2$. In the other interesting case when $p_1 \to 0$ and $p_2 = \rho p_1$, $\rho =$ const, the limit for $-\alpha_1/\alpha_2$ is different from 1,

$$\alpha_1 \sim p_1(\log\rho + 1 - \rho), \qquad \alpha_2 \sim p_1(\rho\log\rho + 1 - \rho),$$

and α_1/α_2 can converge to any number depending on ρ. The "effective" sample size in this case is clearly of order $p_1 n$.

Acknowledgement.

The authors wish to thank the referee for very careful reading and a number of suggestions which lead to better presentation.

References

[1] E. Mammen and A. B. Tsybakov, *Asymptotic minimax recovery of sets with smooth boundaries*, Ann. Statist., 23 (1995), 502–524.

[2] B. D. Ripley and J. P. Rasson, *Finding the edge of a Poisson forest*, J. Appl. Probab., 14 (1977), 483–491.

[3] M. Moore, *On the estimation of a convex set*, Ann. Statist., 12 (1984), 1090–1099.

[4] E. Carlstein and C. Krishnamoorthy, *Boundary estimation*, J. Amer. Statist. Assoc., 87 (1992), 430–438.

[5] D. Ferger, *Boundary estimation based on set-indexed empirical processes*, J. Nonparam. Statist., 16 (2004), 245–260.

[6] H. G. Müller and K. S. Song, *A set-indexed process in a two-region image*, Stochastic Process. Appl., 62 (1996), 87–101.

[7] M. Rudemo and H. Stryhn, *Approximating the distribution of maximum likelihood contour estimators in two-region images*, Scand. J. Statist., 21 (1994), 41–55.

[8] A. P. Korostelev, L. Simar, and A. B. Tsybakov, *Efficient estimation of monotone boundaries*, Ann. Statist., 23 (1995), 476–489.

[9] M. L. Puri and F. H. Ruymgaart, *Change curves in the presence of dependent noise.* In: E. Carlstein *et al.*, eds., IMS Lecture Notes-Monograph Series, 23, pp. 242–254. Hayward, CA: Inst. Math. Statist., 1994.

[10] E. Khmaladze, R. Mnatsakanov, and N. Toronjadze, *The Change Set Problem.* Part 1: *Estimation of the Change Set and the Rate of Convergence*, Report S97-12, Dept. of Statist., UNSW, Sydney, 1997.

[11] L. Le Cam, *Convergence of estimates under dimensionality restrictions*, Ann. Statist., 1 (1973), 38–53.

[12] L. Birgé, *Approximation dans les espaces metriques et theorie de l'estimation*, Z. Wahrsch. verw. Gebiete, 65 (1983), 181–237.

[13] L. Birgé, *On estimating a density using Hellinger distance and some other strange facts*, Probab. Theory Rel. Fields, 68 (1986), 271–291.

[14] L. Birgé, *Model selection via testing: an alternative to (penalized) maximum likelihood estimators*, Ann. Inst. H. Poincaré, Probab. Statist., 42 (2006), 273–325.

[15] S. A. van de Geer, *Hellinger-consistency of certain nonparametric maximum likelihood estimators*, Ann. Statist., 21 (1993), 14–44.

[16] X. Shen and W. H. Wong, *Convergence rate of sieve estimates*, Ann. Statist., 22 (1994), 580–615.

[17] Y. Yang and A. Barron, *Information-theoretic determination of minimax rates of convergence*, Ann. Statist., 27 (1999), 1564–1599.

[18] S. A. van de Geer, *Empirical Processes in M-estimation*, Cambridge University Press, 2000.

[19] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*, Springer, New York, 1996.

[20] E. Khmaladze, R. Mnatsakanov, and N. Toronjadze, *The change-set problem for Vapnik–Červononkis classes*, Math. Methods Statist., 15 (2006), 224–231.

[21] G. R. Shorack and J. A. Wellner, *Empirical Processes with Applications to Statistics*, Wiley, New York, 1986.

[22] A. A. Borovkov and A. A. Mogulskii, *On probabilities of small deviations for processes*, Trudy Inst. Mat. (Novosibirsk), 13 (1989), Asympt. Analiz Raspred. Sluch. Process., 147–168.

[23] A. O. Gelfond, *Calculus of Finite Differences*, Hindustan Publ. Corp., Delhi, 1971.

[24] I. M. Glazman and Ju. I. Ljubič, *Finite-Dimensional Linear Analysis: A Systematic Presentation in Problem Forms*, MIT Press, Cambridge, Mass., 1974.

[25] M. W. Hirsh and S. Smale, *Differential Equations, Dynamical Systems, and Linear algebra*, Academic Press, New York, 1974.

[26] L. Dümbgen, *The asymptotic behavior of some nonparametric change-point estimators*, Ann. Statist., 19 (1991), 1471–1495.

[*Received April* 2004; *revised August* 2006]