Available online at www.sciencedirect.com
ScienceDirect
Statistics and Probability Letters 78 (2008) 1110–1118
STATISTICS & PROBABILITY LETTERS
www.elsevier.com/locate/stapro

A general SLLN for the one-dimensional class cover problem

John C. Wierman[a,*], Pengfei Xiang[b]

[a] *Department of Applied Mathematics and Statistics, 302 Whitehead Hall, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD, 21218, United States*
[b] *Wachovia Bank, United States*

Received 18 July 2007; received in revised form 20 November 2007; accepted 20 November 2007
Available online 23 November 2007

Abstract

The class cover catch digraph (CCCD) is motivated by applications in statistical pattern classification. For the special case of uniformly distributed data in one dimension, Priebe et al. [Priebe, C.E., DeVinney, J.G., Marchette, D.J., 2001. On the distribution of the domination number for random class cover catch digraphs. Statist. Probab. Lett. 55, 239–246] found the exact distribution of the domination number of the random data-induced CCCD, and DeVinney and Wierman [DeVinney, J.G., Wierman, J.C., 2002. A SLLN for a one-dimensional class cover problem. Statist. Probab. Lett. 59, 425–435] proved the Strong Law of Large Numbers (SLLN). This paper proves the generalized SLLN for the domination number of CCCDs for non-uniform data in one dimension.

1. Introduction

The class cover problem (CCP) originates from statistical pattern classification. (See Kulkarni et al. (1998) for a survey of pattern classification.) The CCP is defined in the following way in Priebe et al. (2003): Suppose $\mathcal{X} \equiv \{X_i : i = 1, \ldots, n\}$ and $\mathcal{Y} \equiv \{Y_j : j = 1, \ldots, m\}$ are two independent classes of i.i.d. random variables taking values in a sample space Ω, with class-conditional distribution functions F_X and F_Y, respectively. Consider a dissimilarity function $d : \Omega \times \Omega \to \mathbf{R}$ such that $d(\alpha, \beta) = d(\beta, \alpha) \geq d(\alpha, \alpha) = 0$ for any $\alpha, \beta \in \Omega$. For each $X_i \in \mathcal{X}$, define the covering ball $B(X_i) = \{\omega \in \Omega : d(\omega, X_i) < \min_{1 \leq j \leq m} d(Y_j, X_i)\}$. To make the covering ball definition valid, it is assumed that F_X and F_Y are continuous, so all $X_i \in \mathcal{X}$ and $Y_j \in \mathcal{Y}$ are distinct with probability one. A *class cover of* $\mathcal{X}$ is a subset of covering balls whose union contains all $X_i \in \mathcal{X}$. The CCP is to find a minimum cardinality class cover.

The *class cover catch digraph* (CCCD) induced by a CCP is defined as the digraph $D = (V, A)$, with the vertex set $V = \{X_i : i = 1, \cdots, n\}$ and the arc set $A = \{(X_i, X_j) : X_j \in B(X_i)\}$. By resorting to the terminology of "domination", the CCP can be converted to a graph theory problem on the induced CCCD. For a general digraph $D = (V, A)$, the set $S \subset V$ is a *dominating set* of D if and only if for all $v \in V$, either $v \in S$ or $(s, v) \in A$ for some $s \in S$. Hence, the CCP is equivalent to finding a minimum cardinality dominating set of the induced CCCD. Haynes et al. (1998) provided a comprehensive discussion of domination in graphs.

* Corresponding author. Tel.: +1 410 516 7211; fax: +1 410 516 7459.
E-mail address: wierman@jhu.edu (J.C. Wierman).

0167-7152/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2007.11.005

The *domination number* of a CCCD is the cardinality of the CCCD's minimum dominating set. In the CCCD setting, the domination number is a function of the sets $\mathcal{X}$ and $\mathcal{Y}$, so it is natural to denote it by $\Gamma_{n,m}(\mathcal{X}, \mathcal{Y})$. When the context is clear, we abbreviate it by $\Gamma_{n,m}$.

Marchette (2004, Ch. 4) gives an overview of CCCDs and their applications to classification. Minimum class covers for CCCDs have been used to construct classifiers that are competitive with other approaches in Ceyhan and Priebe (2006), DeVinney and Priebe (2006), Eveland et al. (2005) and Priebe et al. (2003). In particular, Priebe et al. (2003) have used CCCD classifiers for latent class discovery in gene expression monitoring by DNA microarrays. CCCDs' applications to high-dimensional classification problems are discussed in Marchette and Priebe (2003) and Solka et al. (2002). Related classification research based on proximity catch digraphs appears in Ceyhan and Priebe (2005), Ceyhan et al. (2007), and Ceyhan et al. (2006).

The size of the minimum class cover, the domination number, is an important measurement of the complexity of the CCCD classifiers. Two important previous results have been established regarding the probabilistic behavior of the domination number. In particular, Priebe et al. (2001) found the exact distribution of the domination number of CCCDs for uniformly distributed data in one dimension. Based on this distribution, DeVinney and Wierman (2002) proved the following SLLN for the special case of one-dimensional uniformly distributed data.

Theorem 1. *If* $\Omega = \mathbf{R}$, $F_X = F_Y = U[0, 1]$ *and* $m = \lfloor rn \rfloor$, $r \in (0, \infty)$, *then*

$$\lim_{n\to\infty} \frac{\Gamma_{n,m}}{n} = g(r) \quad a.s.,$$

where $g(r) \equiv \frac{r(12r+13)}{3(r+1)(4r+3)}$.

2. Results and proof sketch

We extend DeVinney and Wierman's SLLN to the general case for data with continuous densities in one dimension. Specifically, we prove the following theorem.

Theorem 2. *If* $\Omega = \mathbf{R}$, *the density functions* f_X *and* f_Y *are continuous and bounded on* $[a, b]$, $m \equiv m(n)$ *and* $m/n \to r$ *as* $n \to \infty$, *then*

$$\lim_{n\to\infty} \frac{\Gamma_{n,m}}{n} = \int_a^b g\left(r \cdot \frac{f_Y(u)}{f_X(u)}\right) f_X(u)\mathrm{d}u \quad a.s.,$$

where $g(r) = \frac{r(12r+13)}{3(r+1)(4r+3)}$ *is as in Theorem 1.*

Proof. We now outline the key steps in our proof, with the details deferred to Section 3. The basic approach is to first prove the SLLN for piece-wise constant densities, and then approximate general densities by piece-wise constant densities.

First, consider the case of piece-wise constant densities f_X and f_Y. Without loss of generality, the intervals of constancy for f_X and f_Y can be taken to be the same. Hence, we suppose that f_X and f_Y take respective constant values a_l and b_l in $[c_l, c_{l+1})$, $l = 0, \ldots, k-1$, where $c_0 = a$ and $c_k = b$. We decompose the original CCP into k sub-CCPs, each with uniformly distributed points in $[c_l, c_{l+1})$. We denote the random number of X-points in $[c_l, c_{l+1})$ by N_l, and the random number of Y-points in $[c_l, c_{l+1})$ by M_l. As shown in Lemma 2, for each l, $M_l/N_l \to r_l \equiv r \cdot \frac{b_l}{a_l}$ almost surely as $n \to \infty$. We wish to apply Theorem 1 to prove the following SLLN for the domination number $\Gamma^l_{n,m}$ for each sub-CCP in $[c_l, c_{l+1})$:

$$\frac{\Gamma^l_{n,m}}{N_l} \to g(r_l) \quad a.s.$$

However, a stronger form of Theorem 1 is needed, since Theorem 1 requires $m = \lfloor rn \rfloor$, while we need to accommodate the case in which $m \equiv m(n)$ and $m/n \to r$ as $n \to \infty$. Lemma 1 in Section 3 is obtained by slightly changing the proof by DeVinney and Wierman (2002). Summing up the convergence results for $\Gamma^l_{n,m}$ and adjusting for boundary effects near the interval endpoints yields an SLLN for the case of piece-wise constant densities:

$$\frac{\Gamma_{n,m}}{n} \to \sum_{l=0}^{k-1} g(r_l) b_l (c_{l+1} - c_l) = \int_a^b g\left(r \cdot \frac{f_Y(u)}{f_X(u)}\right) f_X(u)\mathrm{d}u \quad a.s.$$

Secondly, for any $\eta > 0$, for a carefully chosen $\epsilon_\eta \le \eta/3$ and $\delta \equiv \delta(\epsilon_\eta)$, we approximate the continuous densities f_X and f_Y by piece-wise constant densities $\hat{f}_X$ and $\hat{f}_Y$ respectively, where the intervals of constancy for $\hat{f}_X$ and $\hat{f}_Y$ have length δ that depends on η via ϵ_η. The ϵ_η is chosen to guarantee that

$$\left| \int_a^b g\left(r \cdot \frac{\hat{f}_Y(u)}{\hat{f}_X(u)}\right) \cdot \hat{f}_X(u)\mathrm{d}u - \int_a^b g\left(r \cdot \frac{f_Y(u)}{f_X(u)}\right) f_X(u)\mathrm{d}u \right| \le \eta/3,$$

due to the continuity of g.

We construct two classes of coupled random vectors, $\mathcal{X}$ vs. $\hat{\mathcal{X}}$, and $\mathcal{Y}$ vs. $\hat{\mathcal{Y}}$, in the following way. For any random vector (X_{1i}, X_{2i}) uniformly distributed over the region under the graph of f_X (i.e. the region bounded by the x-axis, the line $x = a$, the line $x = b$, and the graph of f_X), the random variable X_{1i} is distributed according to f_X; based on the random point (X_{1i}, X_{2i}), we construct a new random variable $\hat{X}_{1i}$, distributed according to $\hat{f}_X$. Denote $\mathcal{X} = \{X_{1i} : i = 1, \ldots, n\}$ and $\hat{\mathcal{X}} = \{\hat{X}_{1i} : i = 1, \ldots, n\}$, where the $X_{1i} \in \mathcal{X}$ are mutually independent. A similar procedure generates (Y_{1j}, Y_{2j}) uniformly distributed under the graph of f_Y, with Y_{1j} distributed according to f_Y, and a new random variable $\hat{Y}_{1j}$ distributed according to $\hat{f}_Y$. Similarly, denote $\mathcal{Y} = \{Y_{1j} : j = 1, \ldots, m\}$ and $\hat{\mathcal{Y}} = \{\hat{Y}_{1j} : j = 1, \ldots, m\}$, where all $Y_{1j} \in \mathcal{Y}$ are independent of each other. With this construction we show that with probability 1, for sufficiently large n,

$$\left| \frac{\Gamma_{n,m}(\mathcal{X}, \mathcal{Y}) - \Gamma_{n,m}(\hat{\mathcal{X}}, \hat{\mathcal{Y}})}{n} \right| \le \epsilon_\eta \le \eta/3.$$

By using the result of the second step, we show that with probability 1, for sufficiently large n, $\Gamma_{n,m}(\hat{\mathcal{X}}, \hat{\mathcal{Y}})$ satisfies

$$\left| \frac{\Gamma_{n,m}(\hat{\mathcal{X}}, \hat{\mathcal{Y}})}{n} - \int_a^b g\left(r \cdot \frac{\hat{f}_Y(u)}{\hat{f}_X(u)}\right) \hat{f}_X(u)\mathrm{d}u \right| \le \epsilon_\eta \le \eta/3.$$

Theorem 2 follows by combining the three inequalities above. □

Remark 1. Theorem 2 is still valid under more general conditions such as when densities have a finite number of discontinuities, have a finite number of vertical asymptotes (e.g. the arc-sine distribution), or are defined on unbounded intervals (e.g. the normal distribution). See Remark 3 for details on adjusting the proof.

In addition, we obtain an upper bound for the limiting value in Theorem 2.

Theorem 3. *Under the same conditions as Theorem 2,*

$$\int g\left(r \cdot \frac{f_Y(u)}{f_X(u)}\right) f_X(u)\mathrm{d}u \le g(r).$$

Proof. By elementary calculus, $g(r)$ is a concave continuous function. By Jensen's Inequality,

$$\int g\left(r \cdot \frac{f_Y(u)}{f_X(u)}\right) f_X(u)\mathrm{d}u \le g\left(\int r \cdot \frac{f_Y(u)}{f_X(u)} f_X(u)\mathrm{d}u\right) = g(r). \quad \square$$

Theorem 3 shows that when $f_X = f_Y$, the almost sure limit of $\frac{\Gamma_{n,m}}{n}$ is maximized. So roughly speaking, the domination number $\Gamma_{n,m}$ is maximized when X-points and Y-points have the same density, among all possible combinations of continuous and bounded densities. An intuitive explanation is that the domination number of CCCDs measures the degree of discrepancy between X-points and Y-points. Specifically, when there is no discrepancy (i.e. $f_X = f_Y$), more X-covering-balls are needed to segregate the class X from the class Y, thus the domination number is bigger than when a difference exists (i.e. $f_X \ne f_Y$).

Remark 2. As the almost sure limit of $\frac{\Gamma_{n,m}}{n}$ has the maximal value $g(r)$ when $f_X = f_Y$, Theorem 3 suggests a hypothesis test for the equality of densities which compares $\frac{\Gamma_{n,m}}{n}$ with $g(r)$.

3. Detailed proof

3.1. Extension of Theorem 1

First, we weaken the condition of $m = \lfloor rn \rfloor$ in Theorem 1 to $m/n \to r$.

Lemma 1. *If $F_X = F_Y = U[0,1]$, $m \equiv m(n)$ and $m/n \to r$, $r \in (0,\infty)$, then*

$$\lim_{n\to+\infty} \frac{\Gamma_{n,m}}{n} = g(r) \quad a.s.,$$

where $g(r)$ is as in Theorem 1.

Proof Sketch. Since the proof closely follows the arguments by DeVinney and Wierman (2002, pp. 431–433), we only address the differences from their argument. Consider $r = 1$. Construct two independent Poisson processes A and B, with common rate $\lambda \in (0,\infty)$. A-points play the role of X-points, and B-points play the role of Y-points. By conditioning on the $(m+1)$th arrival of the B process and re-scaling to the interval [0,1], we transfer the result back to the original setting, but with a random number $N_m = m + G_m$ of X-points. The classical SLLN can be applied to a CCP induced from these N_m X-points and m Y-points. Since $m/n \to r = 1$, for any $\epsilon > 0$, when n is sufficiently large, $\frac{|m-n|}{n} \le \frac{\epsilon}{2}$. Hence Chernoff's Theorem gives

$$P\left(\frac{|N_m - n|}{n} \ge \epsilon\right) \le P\left(\frac{|G_m|}{n} \ge \frac{\epsilon}{2}\right) \le C_1 e^{-\alpha_1(n\epsilon/2-1)} + C_2 e^{-\alpha_2(n\epsilon/2-1)}, \tag{1}$$

where $\alpha_1, \alpha_2 > 0$ and C_1 and C_2 are constants. This shows that the difference between N_m and n is negligible in the limit. As shown in DeVinney and Wierman (2002, page 431), due to the exponential probability bound in Inequality (1) and by the Borel–Cantelli Lemma, the SLLN still holds for the original setting with n X-points.

As in DeVinney and Wierman (2002), when $r \neq 1$, the proof can be easily extended by letting process A have rate $r\lambda$ and process B have rate λ. □

3.2. Piece-wise constant densities

Next, consider the case where f_X and f_Y are piece-wise constant densities. Recall that the intervals of constancy for f_X and f_Y can be taken to be the same without loss of generality, so

$$f_X(x) = \sum_{l=0}^{k-1} a_l I_{[c_l,c_{l+1})}(x) \quad \text{and} \quad f_Y(y) = \sum_{l=0}^{k-1} b_l I_{[c_l,c_{l+1})}(y),$$

where $a = c_0 < c_1 < \cdots < c_k = b$. Define the following random variables:

$$N_l = |\{X_i : X_i \in [c_l, c_{l+1})\}| \quad \text{and} \quad M_l = |\{Y_j : Y_j \in [c_l, c_{l+1})\}|.$$

Lemma 2. *If $m/n \to r$, $r \in (0,\infty)$, then for each interval $[c_l, c_{l+1})$, $l = 0,\ldots,k-1$, as $n \to \infty$,*

$$\frac{M_l}{m} \to b_l(c_{l+1} - c_l) \quad a.s. \quad \text{and} \quad \frac{N_l}{n} \to a_l(c_{l+1} - c_l) \quad a.s.,$$

and if $a_l \neq 0$, then $\frac{M_l}{N_l} \to r_l$ a.s., where $r_l \equiv r \cdot \frac{f_Y(u)}{f_X(u)} = r\frac{b_l}{a_l}$ for all $u \in [c_l, c_{l+1})$.

Proof. Since Y_j, $j = 1,\ldots,m$ are i.i.d., the indicator random variables $I_{\{Y_j \in [c_l,c_{l+1})\}}$ are also i.i.d., with

$$E\left(I_{\{Y_j \in [c_l,c_{l+1})\}}\right) = P\left(Y_j \in [c_l, c_{l+1})\right) = b_l(c_{l+1} - c_l).$$

Therefore, the standard SLLN yields

$$\frac{M_l}{m} = \frac{|\{Y_j : Y_j \in [c_l, c_{l+1})\}|}{m} = \frac{\sum_{j=1}^{n} I_{\{Y_j \in [c_l,c_{l+1})\}}}{m} \to E\left(I_{\{Y_j \in [c_l,c_{l+1})\}}\right) = b_l(c_{l+1} - c_l) \quad a.s.$$

Similarly,

$$\frac{N_l}{n} \to a_l(c_{l+1} - c_l) \quad a.s.$$

Hence, provided that $a_l \neq 0$,

$$\frac{M_l}{N_l} = \frac{m \cdot \frac{M_l}{m}}{n \cdot \frac{N_l}{n}} \to r \cdot \frac{b_l(c_{l+1} - c_l)}{a_l(c_{l+1} - c_l)} = r_l \quad a.s. \quad \square$$

Dividing the original CCP into k sub-CCPs, each induced by $\mathcal{X}^l = \{X_i : X_i \in [c_l, c_{l+1})\}$ and $\mathcal{Y}^l = \{Y_i : Y_i \in [c_l, c_{l+1})\}$, $l = 0, \ldots, k-1$, we denote the cardinality of a minimum class cover of the lth CCP by $\Gamma_{n,m}(\mathcal{X}^l, \mathcal{Y}^l)$. Since Lemma 2 shows that $M_l/N_l \to r_l$; from Theorem 1, it follows that

$$\frac{\Gamma_{n,m}(\mathcal{X}^l, \mathcal{Y}^l)}{N_l} \to g(r_l) \quad a.s.$$

The points $c_l, l = 1, \ldots, k-1$, are referred to as "filter" points in that for each $l \in \{1, \ldots, k\}$, only X-points and Y-points in $[c_{l-1}, c_l)$ determine $\Gamma_{n,m}(\mathcal{X}^l, \mathcal{Y}^l)$. Recall that the domination number in one dimension is additive over intervals between Y-points. Specifically, we have $\Gamma_{n,m}(\mathcal{X}, \mathcal{Y}) = \sum_{j=0}^{m} \alpha_{j,m}$, where each component $\alpha_{j,m}$ is determined by the X-points contained in $[Y_{(j)}, Y_{(j+1)})$. For any interval $[Y_{(j)}, Y_{(j+1)})$ containing no filter point, $\alpha_{j,m}$ must be a component of $\Gamma_{n,m}(\mathcal{X}^l, \mathcal{Y}^l)$ for the l such that $[Y_{(j)}, Y_{(j+1)}) \subset [c_{l-1}, c_l)$. However, if $[Y_{(j)}, Y_{(j+1)})$ contains one "filter" point c_l, then $\alpha_{j,m}$ is decomposed into the right external component of $\Gamma_{n,m}(\mathcal{X}^l, \mathcal{Y}^l)$ plus the left external component of $\Gamma_{n,m}(\mathcal{X}^{l+1}, \mathcal{Y}^{l+1})$. Finally, if $[Y_{(j)}, Y_{(j+1)})$ contains two or more "filter" points: $c_{l_1}, \ldots, c_{l_{T_j}} (T_j \geq 2)$, then $\alpha_{j,m}$ is divided into the following $T_j + 1$ components: the right external component of $\Gamma_{n,m}(\mathcal{X}^{l_1}, \mathcal{Y}^{l_1})$, plus $\Gamma_{n,m}(\mathcal{X}^{l_2}, \mathcal{Y}^{l_2}), \ldots, \Gamma_{n,m}(\mathcal{X}^{l_{T_j}}, \mathcal{Y}^{l_{T_j}})$, plus the left external component of $\Gamma_{n,m}(\mathcal{X}^{l_{T_j+1}}, \mathcal{Y}^{l_{T_j+1}})$. In summary, for any interval $[Y_{(j)}, Y_{(j+1)})$ containing no filter point, the corresponding component $\alpha_{j,m}$ of $\Gamma_{n,m}(\mathcal{X}, \mathcal{Y})$ is also a component of $\sum_{l=1}^{k} \Gamma_{n,m}(\mathcal{X}^l, \mathcal{Y}^l)$; for any interval $[Y_{(j)}, Y_{(j+1)})$ containing T_j filter points, the corresponding component $\alpha_{j,m}$ of $\Gamma_{n,m}(\mathcal{X}, \mathcal{Y})$ is decomposed into $T_j + 1$ components of $\sum_{l=1}^{k} \Gamma_{n,m}(\mathcal{X}^l, \mathcal{Y}^l)$. Furthermore, Lemma 3 of Priebe et al. (2001) shows that any component mentioned above could only be 0, 1 or 2, so the T_j "filter" points contained in a given interval $[Y_{(j)}, Y_{(j+1)})$ could contribute to the difference $\Gamma_{n,m}(\mathcal{X}, \mathcal{Y}) - \sum_{l=1}^{k} \Gamma_{n,m}(\mathcal{X}^l, \mathcal{Y}^l)$ by at least $0 - 2 * (T_j + 1) = -2T_j - 2$ and at most $2 - 0 * (T_j + 1) = 2$. Supposing that the set J consists of all j such that $[Y_{(j)}, Y_{(j+1)})$ contains at least one "filter" point, we have

$$\sum_{j \in J} (-2T_j - 2) \leq \Gamma_{n,m}(\mathcal{X}, \mathcal{Y}) - \sum_{l=1}^{k} \Gamma_{n,m}(\mathcal{X}^l, \mathcal{Y}^l) \leq \sum_{j \in J} 2.$$

Since there are $k-1$ "filter" points, there are at most $k-1$ such intervals $[Y_{(j)}, Y_{(j+1)})$ that contain one or more "filter" points, so $|J| \leq k-1$. Therefore, from the inequality above we obtain

$$-2 \sum_{j \in J} T_j - 2(k-1) \leq \Gamma_{n,m}(\mathcal{X}, \mathcal{Y}) - \sum_{l=1}^{k} \Gamma_{n,m}(\mathcal{X}^l, \mathcal{Y}^l) \leq 2(k-1).$$

By considering $\sum_{j \in J} T_j = k-1$, the inequality above becomes

$$-4(k-1) \leq \Gamma_{n,m}(\mathcal{X}, \mathcal{Y}) - \sum_{l=1}^{k} \Gamma_{n,m}(\mathcal{X}^l, \mathcal{Y}^l) \leq 2(k-1).$$

Since k is fixed,

$$\begin{aligned} \lim_{n \to \infty} \frac{\Gamma_{n,m}(\mathcal{X}, \mathcal{Y})}{n} &= \lim_{n \to \infty} \sum_{l=0}^{k-1} \frac{\Gamma_{n,m}(\mathcal{X}^l, \mathcal{Y}^l)}{n} \\ &= \sum_{l=0}^{k-1} \lim_{n \to \infty} \frac{\Gamma_{n,m}(\mathcal{X}^l, \mathcal{Y}^l)}{N_l} \cdot \frac{N_l}{n}. \end{aligned}$$

If $a_l \neq 0$, then by Lemma 1, $\frac{\Gamma_{n,m}(\mathcal{X}^l,\mathcal{Y}^l)}{N_l} \to g(r_l)$ $a.s.$, and by Lemma 2, $\frac{N_l}{n} \to a_l(c_{l+1}-c_l)$ $a.s.$ Hence $\lim_{n\to\infty} \frac{\Gamma_{n,m}(\mathcal{X}^l,\mathcal{Y}^l)}{N_l} \cdot \frac{N_l}{n} = g(r_l)a_l(c_{l+1}-c_l)$ $a.s.$ If instead $a_l = 0$, then, almost surely, there are no X-points in $[c_l, c_{l+1})$, so $\Gamma_{n,m}(\mathcal{X}^l, \mathcal{Y}^l) = 0$ $a.s.$ Thus we still have $\lim_{n\to\infty} \frac{\Gamma_{n,m}(\mathcal{X}^l,\mathcal{Y}^l)}{n} = 0 = g(r_l)a_l(c_{l+1}-c_l)$ $a.s.$ where $r_l = \infty$ and $g(\infty) \equiv \lim_{r\to\infty} g(r) = 0$. Therefore we obtain

$$\lim_{n\to\infty} \frac{\Gamma_{n,m}(\mathcal{X},\mathcal{Y})}{n} = \sum_{l=0}^{k-1} g(r_l)a_l(c_{l+1}-c_l) \quad a.s.$$

Rewriting the expressions in the sum in the form of integrals generates

$$\begin{aligned}\lim_{n\to\infty} \frac{\Gamma_{n,m}(\mathcal{X},\mathcal{Y})}{n} &= \sum_{l=0}^{k-1} \int_{c_l}^{c_{l+1}} g\left(r \cdot \frac{f_Y(u)}{f_X(u)}\right) f_X(u)\mathrm{d}u \\ &= \int_a^b g\left(r \cdot \frac{f_Y(u)}{f_X(u)}\right) f_X(u)\mathrm{d}u \quad a.s.\end{aligned}$$

3.3. Continuous densities

The formula obtained in the previous section is also valid when the densities f_X and f_Y are bounded and continuous:

Theorem 4. *If $\Omega = \mathbf{R}$, the density functions f_X and f_Y are bounded and continuous on $[a, b]$, and $m/n \to r, r \in (0, \infty)$, then*

$$\lim_{n\to\infty} \frac{\Gamma_{n,m}(\mathcal{X},\mathcal{Y})}{n} = \int_a^b g\left(r \cdot \frac{f_Y(u)}{f_X(u)}\right) \cdot f_X(u)\mathrm{d}u \quad a.s.,$$

where $g(r)$ is as in Theorem 1.

Proof. Since the density functions f_X and f_Y are bounded and continuous on $[a, b]$, f_X and f_Y are uniformly continuous. Thus for any $\epsilon > 0$, there exists a $\delta \equiv \delta(\epsilon) > 0$ such that for all x and y with $|x - y| < \delta$, $|f_X(x) - f_X(y)| \leq \frac{\epsilon}{4(b-a)}$ and $|f_Y(x) - f_Y(y)| \leq \frac{\epsilon}{4r(b-a)}$. Let $\Delta_l = [a + (l-1)\delta, a + l\delta) \cap [a, b]$ for $l \geq 1$. Define piece-wise constant functions that approximate f_X and f_Y by

$$\begin{aligned}\bar{f}_X(x) &= \min\{f_X(u) : u \in \Delta_l\} \quad \text{for } x \in \Delta_l, \\ \bar{f}_Y(y) &= \min\{f_Y(u) : u \in \Delta_l\} \quad \text{for } y \in \Delta_l.\end{aligned}$$

Note that $\bar{f}_X$ and $\bar{f}_Y$ both depend on ϵ via δ; hence all functions and random variables derived from $\bar{f}_X$ and $\bar{f}_Y$ are also ϵ-dependent, but for simplicity we drop an explicit reference to ϵ throughout the proof.

Since $\bar{f}_X \leq f_X$, $\bar{f}_Y \leq f_Y$, it follows that $\int_a^b \bar{f}_X \leq 1$ and $\int_a^b \bar{f}_Y \leq 1$. Re-scaling $\bar{f}_X$ and $\bar{f}_Y$ gives density functions $\hat{f}_X$ and $\hat{f}_Y$, which approximate f_X and f_Y, respectively. Our next step is to construct two classes of coupled random vectors: $\mathcal{X}$ vs. $\hat{\mathcal{X}}$, and $\mathcal{Y}$ vs. $\hat{\mathcal{Y}}$. Every component of the random vector $\mathcal{X}$ has density function f_X, whereas every component of $\hat{\mathcal{X}}$ has density function $\hat{f}_X$; and a similar property holds for $\mathcal{Y}$ and $\hat{\mathcal{Y}}$ as well. Now that we have introduced all the key notations, we first describe the overall structure of the proof before getting into the details. Recall that the ultimate goal is to prove that $\forall \eta > 0$, with probability one, there exists an $N_\eta > 0$ such that, when $n > N_\eta$,

$$\left|\frac{\Gamma_{n,m}(\mathcal{X},\mathcal{Y})}{n} - \int_a^b g\left(r \cdot \frac{f_Y(u)}{f_X(u)}\right) f_X(u)\mathrm{d}u\right| \leq \eta. \tag{2}$$

Hence it suffices to prove that when $n > N_\eta$,

$$\left|\frac{\Gamma_{n,m}(\mathcal{X},\mathcal{Y})}{n} - \frac{\Gamma_{n,m}(\hat{\mathcal{X}},\hat{\mathcal{Y}})}{n}\right| \leq \eta/3 \tag{3}$$

$$\left|\frac{\Gamma_{n,m}(\hat{\mathcal{X}},\hat{\mathcal{Y}})}{n}-\int_a^b g\left(r\cdot\frac{\hat{f}_Y(u)}{\hat{f}_X(u)}\right)\hat{f}_X(u)\mathrm{d}u\right|\le\eta/3 \tag{4}$$

$$\left|\int_a^b g\left(r\cdot\frac{\hat{f}_Y(u)}{\hat{f}_X(u)}\right)\hat{f}_X(u)\mathrm{d}u-\int_a^b g\left(r\cdot\frac{f_Y(u)}{f_X(u)}\right)f_X(u)\mathrm{d}u\right|\le\eta/3. \tag{5}$$

We first consider Inequality (5). Note that the expressions inside the above integral are polynomials in the density functions $\hat{f}_X$ and $\hat{f}_Y$. Since as $\epsilon\to 0$, $\hat{f}_X(u)\to f_X(u)$ and $\hat{f}_Y(u)\to f_Y(u)$ for any $u\in[a,b]$, the *Dominated Convergence Theorem* gives

$$\int_a^b g\left(r\cdot\frac{\hat{f}_Y(u)}{\hat{f}_X(u)}\right)\cdot\hat{f}_X(u)\mathrm{d}u\to\int_a^b g\left(r\cdot\frac{f_Y(u)}{f_X(u)}\right)f_X(u)\mathrm{d}u\quad\text{as }\epsilon\to 0.$$

Thus, for any given η, there must exist an $\epsilon_\eta\le\eta/3$ such that

$$\left|\int_a^b g\left(r\cdot\frac{\hat{f}_Y(u)}{\hat{f}_X(u)}\right)\cdot\hat{f}_X(u)\mathrm{d}u-\int_a^b g\left(r\cdot\frac{f_Y(u)}{f_X(u)}\right)f_X(u)\mathrm{d}u\right|\le\eta/3,$$

where $\hat{f}_X$ and $\hat{f}_Y$ are constructed as described in the very beginning of the proof by choosing $\epsilon=\epsilon_\eta$. In the rest of this proof, we show that for $\epsilon=\epsilon_\eta$, Inequalities (3) and (4) hold when n is sufficiently large, with probability 1.

We continue by describing the construction procedure of $\hat{\mathcal{X}}$ and $\hat{\mathcal{Y}}$. First, consider i.i.d. random points (X_{1i},X_{2i}), $1\le i\le n$, distributed uniformly over the region bounded by the x-axis, the line $x=a$, the line $x=b$, and the graph of f_X. Then,

$$P(s\le X_{1i}\le t)=\int_s^t f_X(u)\mathrm{d}u\quad\text{for all } a\le s\le t\le b,$$

so the marginal distribution of X_{1i} is f_X. Similarly, construct i.i.d. random points (Y_{1j},Y_{2j}), $1\le j\le m$, with Y_{1j}'s marginal distribution being f_Y. Denote $\mathcal{X}=\{X_{1i}:i=1,\dots,n\}$ and $\mathcal{Y}=\{Y_{1j}:j=1,\dots,m\}$.

Next, let $(\bar{X}_{1i},\bar{X}_{2i})$ and $(\bar{Y}_{1j},\bar{Y}_{2j})$ be i.i.d. random points uniformly distributed over the regions under the graph of $\bar{f}_X$ and $\bar{f}_Y$ respectively. Denote $\bar{R}_X$ as the region between the graphs of f_X and $\bar{f}_X$, and $\bar{R}_Y$ as the region between the graphs of f_Y and $\bar{f}_Y$.

Finally, define

$$\left(\hat{X}_{1i},\hat{X}_{2i}\right)=\left(X_{1i}I_{\{(X_{1i},X_{2i})\notin\bar{R}_X\}}+\bar{X}_{1i}I_{\{(X_{1i},X_{2i})\in\bar{R}_X\}},\,X_{2i}I_{\{(X_{1i},X_{2i})\notin\bar{R}_X\}}+\bar{X}_{2i}I_{\{(X_{1i},X_{2i})\in\bar{R}_X\}}\right)$$

and

$$\left(\hat{Y}_{1j},\hat{Y}_{2j}\right)=\left(Y_{1j}I_{\{(Y_{1j},Y_{2j})\notin\bar{R}_Y\}}+\bar{Y}_{1j}I_{\{(Y_{1j},Y_{2j})\in\bar{R}_Y\}},\,Y_{2j}I_{\{(Y_{1j},Y_{2j})\notin\bar{R}_Y\}}+\bar{Y}_{2j}I_{\{(Y_{1j},Y_{2j})\in\bar{R}_Y\}}\right).$$

Here the idea is to set $(\hat{X}_{1i},\hat{X}_{2i})=(X_{1i},X_{2i})$ if $(X_{1i},X_{2i})\notin\bar{R}_X$, and $(\hat{X}_{1i},\hat{X}_{2i})=(\bar{X}_{1i},\bar{X}_{2i})$ if $(X_{1i},X_{2i})\notin\bar{R}_X$. The same idea applies for Y-points. Denote $\hat{\mathcal{X}}=\{\hat{X}_{1i}:i=1,\dots,n\}$ and $\hat{\mathcal{Y}}=\{\hat{Y}_{1j}:j=1,\dots,m\}$.

Lemma 3. *$\hat{X}_{1i}$ and $\hat{Y}_{1j}$ have piece-wise constant density functions $\hat{f}_X$ and $\hat{f}_Y$, respectively.*

Proof. First, consider the simple case when the interval $[s,t]\subseteq\Delta_l$ for some l. Denote $\hat{f}_X(\Delta_l)\equiv\hat{f}_X(x)$ for all $x\in\Delta_l$, and $I_{X_i}\equiv I_{\{(X_{1i},X_{2i})\in\bar{R}_X\}}$. Then,

$$\begin{aligned}
P(s\le\hat{X}_{1i}\le t) &= P(s\le\hat{X}_{1i}\le t\mid I_{X_i}=1)P(I_{X_i}=1)+P(s\le\hat{X}_{1i}\le t\mid I_{X_i}=0)P(I_{X_i}=0)\\
&= P(s\le\hat{X}_{1i}\le t\mid(X_{1i},X_{2i})\in\bar{R}_X)P(I_{X_i}=1)\\
&\quad +P(s\le X_{1i}\le t\mid(X_{1i},X_{2i})\notin\bar{R}_X)P(I_{X_i}=0)\\
&= (t-s)\hat{f}_X(\Delta_l)P(I_{X_i}=1)+(t-s)\hat{f}_X(\Delta_l)P(I_{X_i}=0)\\
&= (t-s)\hat{f}_X(\Delta_l).
\end{aligned}$$

If the interval $[s,t] \not\subseteq \Delta_l$ for any l, then $[s,t]$ can be written as a union $\cup_{k=0}^{m}[s_k,t_k]$, where each $[s_k,t_k] \subseteq \Delta_k$ for distinct k. Similarly,

$$P(s \le \hat{X}_{1i} \le t) = \sum_{k=0}^{m}(t_k - s_k)\hat{f}_X(\Delta_k),$$

where $\hat{f}_X(\Delta_k) \equiv \hat{f}_X(x)$ for all $x \in \Delta_k$. A similar result for $\hat{Y}_{1j}$ can be obtained by the same argument. □

Recall that the random variable $\Gamma_{n,m}(\mathcal{X},\mathcal{Y})$ represents the size of a minimum class cover of $\mathcal{X} \equiv \{X_{1i}, i = 1,\ldots,n\}$ with respect to $\mathcal{Y} \equiv \{Y_{1j}, j = 1,\ldots,m\}$, and $\Gamma_{n,m}(\hat{\mathcal{X}},\hat{\mathcal{Y}})$ represents the size of a minimum class cover of $\hat{\mathcal{X}} \equiv \{\hat{X}_{1i}, i = 1,\ldots,n\}$ with respect to $\hat{\mathcal{Y}} \equiv \{\hat{Y}_{1j}, j = 1,\ldots,m\}$. For any point $(X_{1i}, X_{2i}) \in \bar{R}_X$, we have the set $\hat{X}_{1i} = \bar{X}_{1i}$, which is equivalent to replacing X_{1i} by $\bar{X}_{1i}$ so that the original domination number $\Gamma_{n,m}(\mathcal{X},\mathcal{Y})$ changes to the new domination number $\Gamma_{n,m}(\hat{\mathcal{X}},\hat{\mathcal{Y}})$. Note that deleting any X_{1i} can decrease the original domination number $\Gamma_{n,m}(\mathcal{X},\mathcal{Y})$ by at most 1, while adding any $\bar{X}_{1i}$ can further decrease $\Gamma_{n,m}(\mathcal{X},\mathcal{Y})$ by at most 1. Therefore, replacing X_{1i} by $\bar{X}_{1i}$ can contribute to the difference $|\Gamma_{n,m}(\mathcal{X},\mathcal{Y}) - \Gamma_{n,m}(\hat{\mathcal{X}},\hat{\mathcal{Y}})|$ by at most 2. Similarly, (Y_{1i}, Y_{2i}) in $\bar{R}_Y$ can also change the difference between $\Gamma_{n,m}(\mathcal{X},\mathcal{Y})$ and $\Gamma_{n,m}(\hat{\mathcal{X}},\hat{\mathcal{Y}})$ by at most 2. Thus,

$$|\Gamma_{n,m}(\mathcal{X},\mathcal{Y}) - \Gamma_{n,m}(\hat{\mathcal{X}},\hat{\mathcal{Y}})| \le 2\left(\sum_{i=1}^{n} I_{\{(X_{1i},X_{2i})\in\bar{R}_X\}} + \sum_{i=1}^{m} I_{\{(Y_{1i},Y_{2i})\in\bar{R}_Y\}}\right).$$

Since $I_{\{(X_{1i},X_{2i})\in\bar{R}_X\}}, i = 1,\ldots,n$ are i.i.d. random variables, applying the SLLN yields

$$\begin{aligned}&\frac{\sum_{i=1}^{n} I_{\{(X_{1i},X_{2i})\in\bar{R}_X\}}}{n} \xrightarrow{n\to\infty} E(I_{\{(X_{1i},X_{2i})\in\bar{R}_X\}})\\ &= P\left((X_{1i},X_{2i}) \in \bar{R}_X\right) \le (b-a)\cdot\frac{\epsilon_\eta}{4(b-a)} = \frac{\epsilon_\eta}{4} \quad a.s.,\end{aligned} \tag{6}$$

and

$$\begin{aligned}&\frac{\sum_{i=1}^{m} I_{\{(Y_{1i},Y_{2i})\in\bar{R}_Y\}}}{n} = \frac{m}{n}\cdot\frac{\sum_{i=1}^{m} I_{\{(Y_{1i},Y_{2i})\in\bar{R}_Y\}}}{m} \xrightarrow{n\to\infty} r\cdot E(I_{\{(Y_{1i},Y_{2i})\in\bar{R}_Y\}})\\ &= r\cdot P((Y_{1i},Y_{2i}) \in \bar{R}_Y) \le r\cdot(b-a)\cdot\frac{\epsilon_\eta}{4r(b-a)} = \frac{\epsilon_\eta}{4} \quad a.s.\end{aligned} \tag{7}$$

Recall that $\epsilon_\eta \le \eta/3$. Consequently, with probability 1, there exists an $N'(\epsilon_\eta) > 0$ such that when $n > N'(\epsilon_\eta)$,

$$\left|\frac{\Gamma_{n,m}(\mathcal{X},\mathcal{Y}) - \Gamma_{n,m}(\hat{\mathcal{X}},\hat{\mathcal{Y}})}{n}\right| \le 2\cdot\left(\frac{\epsilon_\eta}{4} + \frac{\epsilon_\eta}{4}\right) = \epsilon_\eta \le \eta/3,$$

which is exactly Inequality (3).

By the SLLN for piece-wise densities, with probability 1, there exists an $N''(\epsilon_\eta) > 0$ such that when $n > N''(\epsilon_\eta)$,

$$\left|\frac{\Gamma_{n,m}(\hat{\mathcal{X}},\hat{\mathcal{Y}})}{n} - \int_a^b g\left(r\cdot\frac{\hat{f}_Y(u)}{\hat{f}_X(u)}\right)\hat{f}_X(u)\mathrm{d}u\right| \le \epsilon_\eta \le \eta/3,$$

which is exactly Inequality (4).

Therefore, with probability 1, both Inequality (3) and Inequality (4) hold when $n > N_\eta \equiv \max\{N'(\epsilon_\eta), N''(\epsilon_\eta)\}$. Hence, Inequality (2) immediately follows as we have showed that Inequality (5) holds when choosing $\epsilon = \epsilon_\eta$. Since $\eta > 0$ is arbitrary, we conclude that

$$\lim_{n\to\infty}\frac{\Gamma_{n,m}(\mathcal{X},\mathcal{Y})}{n} = \int_a^b g\left(r\cdot\frac{f_Y(u)}{f_X(u)}\right)f_X(u)\mathrm{d}u \quad a.s. \quad \square$$

Remark 3. For simplicity, we assumed that the density functions f_X and f_Y are continuous and bounded. However, for some more general cases (e.g., when the densities are bounded but with only a finite number of discontinuities, when the densities have a finite number of vertical asymptotes, or when the densities are defined on unbounded intervals), our proof can apply as well by a slight modification. The key is to find appropriate piece-wise constant functions $\bar{f}_X$ (bounded above by f_X) and $\bar{f}_X$ (bounded above by f_Y), where $\bar{f}_X$ and $\bar{f}_Y$ sufficiently approximate f_X and f_Y, respectively, so that Inequality (5), $P\left((X_{1i}, X_{2i}) \in \bar{R}_X\right) \leq \frac{\epsilon_\eta}{4}$ as in Inequality (6), and $P\left((Y_{1i}, Y_{2i}) \in \bar{R}_Y\right) \leq \frac{\epsilon_\eta}{4}$ as in Inequality (7) still hold.

4. Summary

In this paper, we prove the generalized SLLN for the domination number of CCCDs in one dimension (Theorem 4). Our proof is based on the SLLN for uniform distributions proved by DeVinney and Wierman (2002). In addition, we give an upper bound for the almost sure limit of $\frac{\Gamma_{n,m}}{n}$. This upper bound has an interesting explanation from the point of view of pattern classification, and it may result in a statistical test for the identity of two densities.

For further research directions, we are interested in proving the SLLN in higher dimensions, where the CCCD problem is significantly more challenging because the exact distribution of $\Gamma_{n,m}$ is unknown. Moreover, we are also exploring ways to prove the Central Limit Theorem (CLT) for the domination number.

Acknowledgments

The authors gratefully acknowledge the financial support for this research from the Johns Hopkins University through the Acheson J. Duncan Fund for the Advancement of Research in Statistics. We thank Carey Priebe for the encouragement and helpful discussions regarding research on class cover catch digraphs and related statistical classification methods.

References

Ceyhan, E., Priebe, C.E., 2005. The use of domination number of a random proximity catch digraph for testing spatial patterns of segregation and association. Statist. Probab. Lett. 73, 37–50.

Ceyhan, E., Priebe, C.E., 2006. On the distribution of the domination number of a new family of parametrized random digraphs. Model Assist. Statist. Appl. 1, 231–255.

Ceyhan, E., Priebe, C.E., Marchette, D.J., 2007. A new family of graphs for testing spatial segregation. Canad. J. Statist. 35, 27–50.

Ceyhan, E., Priebe, C.E., Wierman, J.C., 2006. Relative density of the random r-factor proximity catch digraph for testing spatial patterns of segregation and association. Comput. Statist. Data Anal. 50, 1925–1964.

DeVinney, J.G., Priebe, C.E., 2006. A new family of proximity graphs: Class cover catch digraphs. Discrete Appl. Math. 154, 1975–1982.

DeVinney, J.G., Wierman, J.C., 2002. A SLLN for a one-dimensional class cover problem. Statist. Probab. Lett. 59, 425–435.

Eveland, C.K., Socolinsky, D.A., Priebe, C.E., Marchette, D.J., 2005. A hierarchical methodology for one-class problems with skewed priors. J. Classification 22, 17–48.

Haynes, T.W., Hedetniemi, S.T., Slater, P.J., 1998. Domination in Graphs: Fundamentals. Marcel Dekker, New York.

Kulkarni, S.R., Lugosi, G., Venkatesh, S.S., 1998. Learning pattern classification — a survey. IEEE Trans. Inform. Theory 44, 2178–2206.

Marchette, D.J., 2004. Random Graphs for Statistical Pattern Recognition. Wiley, New York.

Marchette, D.J., Priebe, C.E., 2003. Characterizing the scale dimension of a high-dimensional classification problem. Pattern Recognit. 36, 45–60.

Priebe, C.E., DeVinney, J.G., Marchette, D.J., 2001. On the distribution of the domination number for random class cover catch digraphs. Statist. Probab. Lett. 55, 239–246.

Priebe, C.E., Marchette, D.J., DeVinney, J.G., Socolinsky, D.A., 2003. Classification using class cover catch digraphs. J. Classification 20, 3–23.

Priebe, C.E., Solka, J.L., Marchette, D.J., Clark, B.T., 2003. Class cover catch digraphs for latent class discovery in gene expression monitoring by DNA microarrays. Comput. Statist. Data Anal. 43, 621–632.

Solka, J.L., Clark, B.C., Priebe, C.E., 2002. A visualization framework for the analysis of hyperdimensional data. Internat. J. Image Graphics 2, 145–161.