

Asymptotic Theory for the Domination Number of Random Class Cover Catch Digraphs

John C. Wierman and Pengfei Xiang*
Department of Mathematical Sciences
The Johns Hopkins University

Abstract

We discuss the asymptotic behavior of the domination number of random class cover catch digraphs (CCCDs). The CCCD problem is motivated by its applications in statistical pattern classification. For uniformly distributed data in one dimension, Priebe, Marchette and Devinney found the exact distribution of the domination number of the random data-induced CCCD, and Devinney and Wierman proved the Strong Law of Large Numbers (SLLN). We will present progress toward the SLLN and the Central Limit Theorem (CLT) for general data distributions in one dimension. The ultimate goal is to establish SLLN and CLT results for higher dimensional CCCDs.

Keywords: Class Cover Catch Digraph, Domination Number, Strong Law of Large Numbers, Central Limit Theorem, Pattern Classification

1 Introduction

This article describes research in progress on the asymptotic behavior of the domination number of random data-induced class cover catch digraphs. Class cover catch digraphs (CCCDs) and the class cover problem (CCP) were introduced by Cowen and Cannon [1], motivated by statistical pattern classification [9]. In the remainder of this Introduction, we briefly describe the pattern classification problem, define the class cover problem, class cover catch digraphs, and domination number, and describe their relevance to pattern classification. Section 2 discusses previous results by Priebe, Marchette, and DeVinney [15] and DeVinney and Wierman [3] on one-dimensional CCCDs induced by uniformly distributed data. In section 3, we describe and sketch the proof of a general Strong Law of Large Numbers (SLLN) for the domination number in CCCDs induced by non-uniform data distributions. Asymptotic results for the variance of the domination number of a one-dimensional CCCD are described in section 4. Section 5 provides comments about future research goals, which focus on asymptotic results for higher-dimensional distributions.

*The research of both authors is supported in part by the Acheson J. Duncan Fund for the Advancement of Statistics.

1.1 Pattern Classification

Pattern classification is “the assignment of a physical object or event to one of several pre-specified categories” (See [5, page 2]). It is widely applied to problems such as automated speech recognition, DNA sequence identification, and fingerprint identification. For a comprehensive discussion of pattern classification, see the classic books [4] and [6], for example.

A mathematical model of the pattern classification problem is formulated as follows [9]. For simplicity, but without loss of generality, suppose we have two classes of objects of interest, which we will call the *target class*, denoted by \mathcal{X} , and the *non-target class*, denoted by \mathcal{Y} . Objects of both classes belong to a common *dissimilarity space* (Ω, d) , which consists of a set Ω and a *dissimilarity function* d , i.e. a function $d : \Omega \times \Omega \rightarrow \mathbf{R}$ such that $d(\alpha, \beta) = d(\beta, \alpha) \geq d(\alpha, \alpha) = 0$ for all $\alpha, \beta \in \Omega$. (A dissimilarity function is similar to a metric, but need not satisfy the triangle inequality.)

The uncertainty about class membership of the objects is modelled by *prior probabilities* $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ for these two classes ($P_{\mathcal{X}} + P_{\mathcal{Y}} = 1$). Given the class \mathcal{X} or \mathcal{Y} , the objects of that class are distributed according to the *class-conditional* distribution functions $F_{\mathcal{X}}(x)$ or $F_{\mathcal{Y}}(y)$, respectively. Thus, a random pair $(c(\Psi), \Psi)$ is generated by a two-step process: first choose the random class label $c(\Psi) \in \{\mathcal{X}, \mathcal{Y}\}$ according to the prior probabilities; then based on the chosen class, select Ψ according to the corresponding class-conditional distribution function.

In the classification problem, for an observation $(c(\psi), \psi)$ generated as above, only the data ψ is given, while the class label $c(\psi)$ is unknown. The goal of a *classifier* is to correctly guess whether $c(\psi)$ is \mathcal{X} or \mathcal{Y} . In *supervised classification*, we are given a training sample of size k with known classification:

$$D_k = \left\{ (c(\psi_1), \psi_1), \dots, (c(\psi_k), \psi_k) \right\}.$$

A *classifier* is a function $\hat{c}(\psi) = \hat{c}_k(\psi, D_k)$, which, based on the training data D_k , assigns a class label \mathcal{X} or \mathcal{Y} to any input point $\psi \in \Omega$. The performance of a classifier \hat{c} can be measured by the *misclassification rate*, given by

$$E[P(\hat{c}(\Psi) \neq c(\Psi) \mid D_k)].$$

1.2 Classification and the Class Cover Problem

The class cover problem arises in the context of a nonparametric approach to supervised pattern classification. Suppose that the training data consists of n data points $\{X_1, X_2, \dots, X_n\}$ in the target class \mathcal{X} and m data points $\{Y_1, Y_2, \dots, Y_m\}$ in the non-target class \mathcal{Y} , which are generated by independent and identically distributed random variables with distribution functions $F_{\mathcal{X}}$ and $F_{\mathcal{Y}}$, respectively. To construct a classifier, we wish to use the training data to identify a region of Ω that is representative of the class \mathcal{X} .

For each X_i , define its *covering ball* as the open ball

$$B(X_i) = \{\omega \in \Omega : d(\omega, X_i) < \min_j d(Y_j, X_i)\}$$

in the dissimilarity measure. The covering ball of X_i is the largest open ball centered at X_i which does not contain any of the Y_j 's. A *class cover* is a set of covering balls whose union contains all $X_i, i = 1, \dots, n$. Clearly, the set consisting of all covering balls is a class cover. The *class cover problem* (CCP) is to find a minimum

cardinality class cover. In the context of classification, the goal is to find a small set of data points to represent the class, in order to make the classifier less complex while keeping most of the relevant information.

The CCP has been actively studied recently, since its solution can be directly used to generate classifiers competitive with the other methods currently in use. A simple classifier can be constructed as follows: by switching the roles of \mathcal{X} and \mathcal{Y} , we obtain a pair of dual CCP's, resulting in two solutions such as $\mathcal{B}_{\mathcal{X}} = \{B(X_i) : i \in I, I \subset \{1, \dots, n\}\}$ and $\mathcal{B}_{\mathcal{Y}} = \{B(Y_j) : j \in J, J \subset \{1, \dots, m\}\}$, respectively. Define $\mathcal{C}_{\mathcal{X}} = \{\omega \in \Omega : \omega \in B(X_i) \text{ s.t. } B(X_i) \in \mathcal{B}_{\mathcal{X}}\}$, $\mathcal{C}_{\mathcal{Y}} = \{\omega \in \Omega : \omega \in B(Y_i) \text{ s.t. } B(Y_i) \in \mathcal{B}_{\mathcal{Y}}\}$. We can incorporate these two solutions into a classifier $\hat{c}(\psi) : \Omega \rightarrow \{\mathcal{X}, \mathcal{Y}\}$ as follows:

$$\hat{c}(\psi) = \begin{cases} \mathcal{X} & \psi \in \mathcal{C}_{\mathcal{X}} \cap \mathcal{C}_{\mathcal{Y}}^c, \\ \mathcal{Y} & \psi \in \mathcal{C}_{\mathcal{Y}} \cap \mathcal{C}_{\mathcal{X}}^c, \\ \text{undetermined} & \text{otherwise.} \end{cases}$$

Of course, more elaborate classifiers, which eliminate the undetermined region, may also be constructed. More details about the CCP's application to classification are presented in [2, 16, 17].

The CCP can be converted into a purely graph-theoretic problem. First, define the *class cover catch digraph (CCCD)* induced by a CCP as the digraph $D = (V, A)$ with the vertex set $V = \{X_i : i = 1, \dots, n\}$ and the edge set A such that there is a directed edge (X_i, X_j) if and only if $X_j \in B(X_i)$. A set $S \subseteq V$ is a *dominating set* of a digraph $D = (V, A)$ if and only if for all $v \in V$, either $v \in S$ or $(s, v) \in A$ for some $s \in S$. It is an easy consequence of these definitions that the CCP is actually equivalent to the problem of finding a minimum cardinality dominating set of the corresponding CCCD. Cowen and Cannon [1] proved that the dominating set problem is essentially a special case of the CCP, and since the dominating set problem is NP-Hard, it follows that the CCP is also NP-Hard.

The principal quantity of study in this article is the domination number. The *domination number* of a CCCD is the minimum cardinality of a dominating set of the CCCD. Due to the equivalence of the CCP and the CCCD dominating set problem, the domination number serves as a measure of the separation of the set of \mathcal{X} points from the set of \mathcal{Y} points. If they are well separated, the domination number is low, while if they are highly intermixed, the domination number is high. Thus, the domination number may serve as a basis for testing for equality of distributions of the target class and non-target class data.

The term "domination number" was first used in graph theory by Ore [13] in 1962. Due to many applications in such fields as computer networks, social sciences and computational complexity, there has been increasing interest in this topic. Haynes, Hedetniemi and Slater [7] provide a comprehensive discussion of domination in graphs, with more advanced topics covered in [8].

In the remainder of this article, we denote the domination number of the CCCD based on training data consisting of n \mathcal{X} points and m \mathcal{Y} points by $\Gamma_{n,m}(F_{\mathcal{X}}, F_{\mathcal{Y}})$, or simply by $\Gamma_{n,m}$. Clearly $\Gamma_{n,m}$ is a random variable whose distribution is determined by $n, m, F_{\mathcal{X}}$ and $F_{\mathcal{Y}}$. The article reports on progress in determining the asymptotic behavior of this random variable.

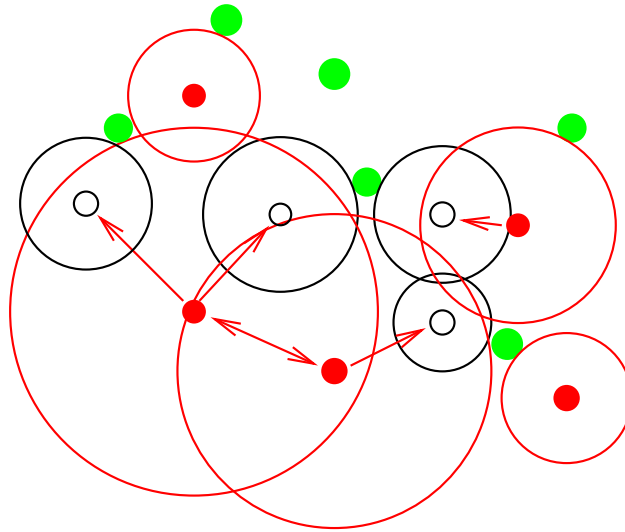


Figure 1: An example of a class cover catch digraph: Green disks indicate class \mathcal{Y} points. Red disks and small black circles indicate class \mathcal{X} points. Larger circles are the covering balls corresponding to the \mathcal{X} points at their centers. The five red arrows are the directed edges in the class cover catch digraph. Red indicates those disks and covering balls corresponding to a minimum cardinality dominating set. The domination number of this class cover catch digraph is 5.

2 One Dimensional CCCD's from Uniform Data

In this section we describe research of Priebe, DeVinney, and Marchette [15] and DeVinney and Wierman [3] for uniformly distributed data in one dimension. Both articles consider $\Gamma_{n,m}$ in the case where $\Omega = \mathbf{R}$, both $F_{\mathcal{X}}$ and $F_{\mathcal{Y}}$ are $U[0,1]$, the uniform distribution on the interval $[0,1]$, and the dissimilarity measure d is the Euclidean distance.

In this one dimensional situation, the \mathcal{Y} data points partition the interval $[0,1]$ into subintervals which may be considered separately. Let $Y_{(j)}$ denote the j -th order statistic of $Y_0 = 0, Y_1, \dots, Y_m, Y_{m+1} = 1$, and let the random variable $\alpha_{j,m}$ be the minimum number of covering balls needed to cover the $N_{j,m}$ \mathcal{X} -class points located between $Y_{(j)}$ and $Y_{(j+1)}$. Note that

$$\Gamma_{n,m} = \sum_{j=0}^m \alpha_{j,m},$$

so we are able to decompose the problem into $m + 1$ sub-problems of finding the domination number $\alpha_{j,m}$ in the interval $[Y_{(j)}, Y_{(j+1)}]$. Since they have different distributions, we will refer to $\alpha_{j,m}$ ($j = 0, m$) as *external components*, and to $\alpha_{j,m}$ ($j = 1, \dots, m - 1$) as *internal components*.

Certain aspects of the components' distributions can be determined relatively easily. Clearly $\alpha_{j,m} = 0$ if and only if $N_j = 0$. For the external components, $\alpha_{j,m} = 1$ if and only if $N_{j,m} > 0$, by taking the smallest \mathcal{X} point for $j = 0$ and the largest \mathcal{X} point for $j = m$ as the dominating sets. For the internal components, it is not difficult to see that $\alpha_{j,m}$ can be at most 2, because all X_i 's in $[Y_{(j)}, Y_{(j+1)}]$

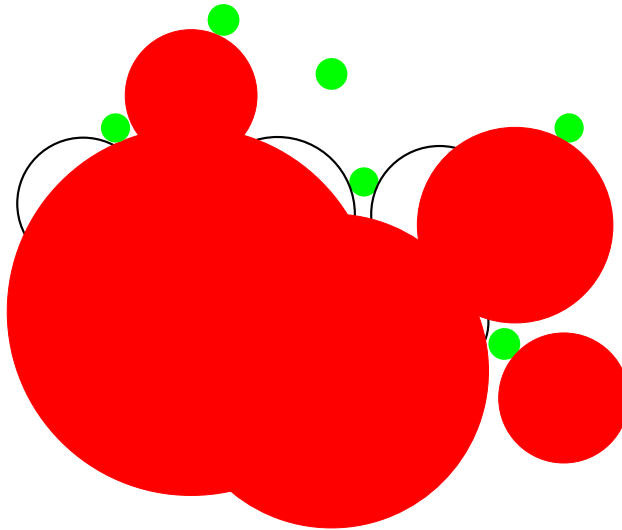


Figure 2: The red region is the union of the covering balls of the minimum cardinality dominating set for the class cover catch digraph in Figure 1. The black arcs correspond to parts of covering balls of points of \mathcal{X} which are not in the dominating set.

are contained in the covering balls of the two \mathcal{X} points that are closest to midpoint of this interval on the right and left. Thus, since the domination number is a non-negative integer, $\alpha_{j,m}$ can only be 0, 1 or 2.

Priebe, DeVinney, and Marchette [15] found the exact conditional probability that $\alpha_{j,m} = 2$ given $N_{j,m} \geq 2$ for $U[0,1]$ data. Note that given the value of $N_{j,m}$, this number of \mathcal{X} points are conditionally independent and uniformly distributed in the interval between $Y_{(j)}$ and $Y_{(j+1)}$. Since rescaling the length of the interval does not change the class cover catch digraph, we may without loss of generality consider each internal component to correspond to the case of $N_{j,m}$ \mathcal{X} points in the interval $[0,1]$ with \mathcal{Y} points at 0 and 1. Simple geometric considerations show that if there is an \mathcal{X} point in the interval $(\max\{X_i\}/2, (1 + \min\{X_i\})/2)$, its covering ball will cover all the \mathcal{X} points in the interval. (See Figure 4.) Using the joint distribution of the maximum and minimum of the uniform order statistics, the probabilities that the dominating set is of size one or two may be computed exactly.

The Priebe, DeVinney, and Marchette [15] exact conditional distribution results may be summarized in the following theorem.

Theorem 2.1. *If $\Omega = \mathbf{R}$, d is Euclidean distance, and $F_{\mathcal{X}} = F_{\mathcal{Y}} = U[0,1]$, then:*

- For $j \in \{0, 1, \dots, m\}$, if $N_{j,m} = 0$ then $\alpha_{j,m} = 0$.
- For $j \in \{0, m\}$, if $N_{j,m} > 0$ then $\alpha_{j,m} = 1$.
- For $j \in \{1, 2, \dots, m-1\}$, if $N_{j,m} = n_j > 0$ then

$$\begin{aligned} P(\alpha_{j,m} = 1 \mid N_{j,m} = n_j) &= 1 - P(\alpha_{j,m} = 2 \mid N_{j,m} = n_j) \\ &= \frac{5}{9} + \frac{4}{9} \frac{1}{4^{n_j-1}}. \end{aligned}$$

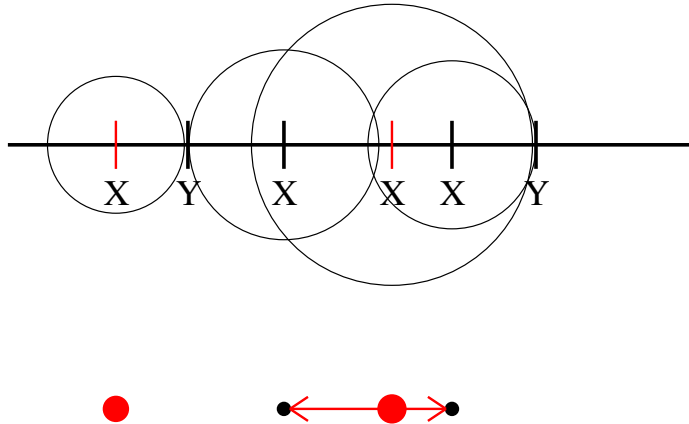


Figure 3: An example of a one-dimensional class cover problem and class cover catch digraph. In the upper diagram, vertical bars denote the positions of class \mathcal{X} and \mathcal{Y} points on the line, while red indicates the class \mathcal{X} points in the dominating set. (For ease of visualization, we have illustrated the covering balls with circles rather than intervals.) The lower graph shows the corresponding class cover catch digraph. Red indicates the dominating set of vertices and the edges of the CCCD.

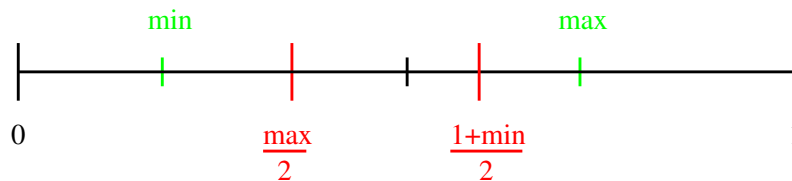


Figure 4: An \mathcal{X} point in the interval between the two red values will cover all \mathcal{X} points.

The theorem's precise formula for the conditional probability in the third part plays a crucial role in the proof of the strong law of large numbers for the domination number and the calculation of the asymptotic expression for the variance given in later sections of this article.

From the theorem, we see that for $j \in \{1, 2, \dots, m-1\}$, given $N_j = n_j > 0$, the conditional probability of $\alpha_{j,m} = 2$ is an increasing function of n_j , so that $\alpha_{j,m}$ tends to become larger as the number of \mathcal{X} points increases. However, it is interesting to note that this conditional probability does not increase to one, but tends to $\frac{4}{9}$ instead.

Under the same assumptions as Theorem 1.1, Devinney and Wierman [3] proved a strong law of large numbers for $\Gamma_{n,m}$:

Theorem 2.2. *If $\Omega = \mathbf{R}$, d is Euclidean distance, $F_{\mathcal{X}} = F_{\mathcal{Y}} = U[0, 1]$, and $m = \lfloor rn \rfloor, r \in (0, \infty)$, then*

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

where

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

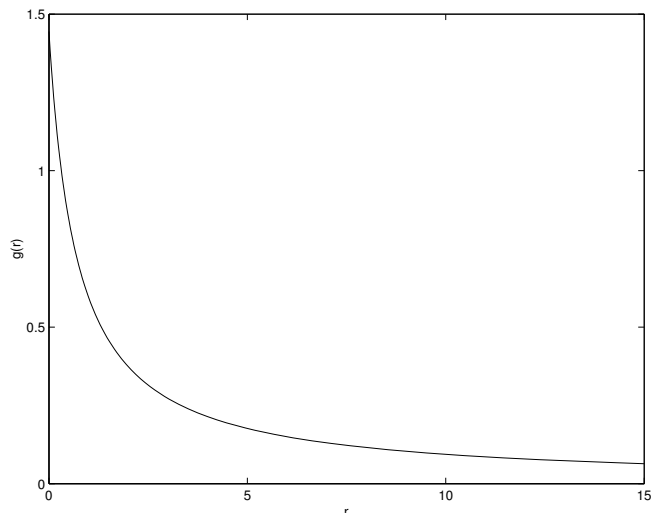


Figure 5: A graph of the limit function $g(r)$, plotted using MATLAB.

Figure 5 assists in interpreting the result of the Theorem 2.2. It suggests that $g(r) \rightarrow 0$ as $r \rightarrow \infty$, which is true since asymptotically the interval between $Y_{(j)}$ and $Y_{(j+1)}$ contains no \mathcal{X} point almost surely. Moreover, $g(r) \rightarrow \frac{13}{9}$ as $r \rightarrow 0$. To see this, consider the situation where each interval $(Y_{(j)}, Y_{(j+1)})$ contains a very large number of \mathcal{X} points. According to Theorem 2.1, the probability of $\alpha_{j,m} = 1$ is approximately $\frac{5}{9}$, while the probability of $\alpha_{j,m} = 2$ is approximately $\frac{4}{9}$. Thus, the expectation of $\alpha_{j,m}$ tends to $\frac{5}{9} \cdot 1 + \frac{4}{9} \cdot 2 = \frac{13}{9}$.

While the intuition in the previous paragraph identifies the limiting value of the expectation, the domination number is a sum of dependent random variables to which the standard Strong Law of Large Numbers does not apply. Using a ‘‘Poissonization’’ technique, DeVinney and Wierman [3] first proved the Strong Law of Large Numbers for the special case of $r = 1$. They constructed two independent Poisson processes A and B , with common rate $\lambda \in (0, \infty)$. Points of A play the role of \mathcal{X} data points, and points of B play the role of \mathcal{Y} data points. Due to the independence of interarrival times and the lack of memory property, the classical SLLN can be applied to the CCCD induced from these A and B points. Using the conditional uniformity property of Poisson processes, and conditioning on the $(n + 1)$ -st arrival of the B process, the result is transferred back to the original setting, but with a random number of \mathcal{X} points which is near n . A procedure of inserting or deleting \mathcal{X} points, while showing that the effects are negligible in the limit, produces the SLLN for the original situation. In order to transfer almost sure results between the two settings, it was necessary to prove a stronger form of convergence, namely complete convergence. For the $r \neq 1$ case, the proof is easily extended by letting process A having rate $r\lambda$ and process B having rate λ .

Remark: We have found an alternative proof to Theorem 2.2 using an existing SLLN for *quadrant dependent* random variables which is due to Matula [11]. The

concept of quadrant dependence was first introduced by E.L. Lehmann [10], and the limit theory for quadrant dependent random variables is comprehensively discussed in Newman [12].

3 Strong Law of Large Numbers (SLLN) for Non-uniform Data

In Theorem 2.2, it was assumed that the classes \mathcal{X} and \mathcal{Y} both have the same uniform distribution. But in real applications, they have different non-uniform distributions. Indeed, this is the reason for trying to assign objects to different classes. We have proved an extension to Theorem 2.2 for more general non-identical distribution functions in the one dimensional case:

Theorem 3.1. *Suppose $\Omega = \mathbf{R}$ and d is the Euclidean distance. Assume the densities $f_{\mathcal{X}}(x)$ and $f_{\mathcal{Y}}(y)$ are bounded functions with a finite number of discontinuities. If $m/n \rightarrow r$, then*

$$\lim_{n \rightarrow +\infty} \frac{\Gamma_{n,m}}{m} = \int g\left(r \frac{f_{\mathcal{Y}}(u)}{f_{\mathcal{X}}(u)}\right) f_{\mathcal{Y}}(u) du \quad a.s. \quad (1)$$

Proof Sketch. The proof consists of two phases: We first consider piece-wise constant densities $f_{\mathcal{X}}$ and $f_{\mathcal{Y}}$, i.e.

$$f_{\mathcal{X}}(x) = \sum_{l=1}^k a_l I_{[c_{l-1}, c_l]}(x),$$

$$f_{\mathcal{Y}}(y) = \sum_{l=1}^k b_l I_{[c_{l-1}, c_l]}(y)$$

where $c_0 < c_1 < \dots < c_k$. To prove (1) for this type of density function, we divide the CCP into sub-CCP's with conditional uniform distributions for the intervals $[c_{l-1}, c_l]$. In each interval, the ratio between the number of \mathcal{Y} points and \mathcal{X} points is asymptotically $r \frac{f_{\mathcal{Y}}(u)}{f_{\mathcal{X}}(u)}$. So from Theorem 2.2, we know that

$$\frac{\text{domination number in } [c_{l-1}, c_l]}{\text{number of } \mathcal{Y} \text{ points in } [c_{l-1}, c_l]}$$

is asymptotically $g\left(r \frac{f_{\mathcal{Y}}(u)}{f_{\mathcal{X}}(u)}\right)$, $u \in [c_{l-1}, c_l]$. By summing the domination numbers for all the intervals, we get an approximation Γ' to $\Gamma_{n,m}$, since we add external components for each endpoint c_l rather than the actual domination number of the interval containing c_l . We can prove that equation (1) holds if $\frac{\Gamma_{n,m}}{m}$ is replaced by $\frac{\Gamma'}{m}$. Since the difference between Γ' and $\Gamma_{n,m}$ is bounded by $2k$ where k is fixed, we conclude that equation (1) is also true.

For the general case, we construct a sequence of piece-wise constant density functions $f_{\mathcal{X},k}$ and $f_{\mathcal{Y},k}$ converging to $f_{\mathcal{X}}$ and $f_{\mathcal{Y}}$, respectively. Based on $\{X_i\}$ and $\{Y_j\}$, we define two new sequences of random variables $\{X_{i,k}\}$ and $\{Y_{j,k}\}$, which are respectively distributed according to $f_{\mathcal{X},k}$ and $f_{\mathcal{Y},k}$. From the first phase of the proof, we know that the SLLN is true for the domination number of the CCCD induced by the newly defined points $X_{i,k}$ and $Y_{j,k}$. By using the relation between X_i and $X_{i,k}$, and between Y_j and $Y_{j,k}$, we can argue that the SLLN still holds for the original densities $f_{\mathcal{X}}$ and $f_{\mathcal{Y}}$. \square

Intuitively, when class \mathcal{X} and class \mathcal{Y} both have the same distribution pattern, it means that their objects tend to be mixed together. Hence a larger dominating set is needed to distinguish \mathcal{X} from \mathcal{Y} than when they have different distributions. In the following theorem, we have proved that equal $f_{\mathcal{X}}$ and $f_{\mathcal{Y}}$ give the maximum limit in the SLLN.

Theorem 3.2. *Under the same assumptions as in Theorem 3.1,*

$$\int g\left(r \frac{f_{\mathcal{Y}}(u)}{f_{\mathcal{X}}(u)}\right) f_{\mathcal{Y}}(u) du \leq g(r)$$

where the equality holds if and only if $f_{\mathcal{X}} = f_{\mathcal{Y}}$ except possibly at the finite number of discontinuity points.

Proof. Note that $g^*(r) = g(\frac{1}{r})$ is a concave function. Expressing g in terms of g^* , and applying Jensen's inequality, we obtain

$$\begin{aligned} \int g\left(r \frac{f_{\mathcal{Y}}(u)}{f_{\mathcal{X}}(u)}\right) f_{\mathcal{Y}}(u) du &= \int g^*\left(\frac{1}{r} \frac{f_{\mathcal{X}}(u)}{f_{\mathcal{Y}}(u)}\right) f_{\mathcal{Y}}(u) du \\ &\leq g^*\left(\int \frac{1}{r} \frac{f_{\mathcal{X}}(u)}{f_{\mathcal{Y}}(u)} f_{\mathcal{Y}}(u) du\right) \\ &= g^*\left(\frac{1}{r}\right) = g(r). \quad \square \end{aligned}$$

Note that the maximum limit is the achieved for all cases of equal densities with a finite number of discontinuities. Thus, Theorem 3.2 could be used to construct asymptotically distribution-free statistical tests of a null hypothesis of equal densities versus an alternative hypothesis of different densities. Of course, limiting distributions would be needed to determine critical values for such tests.

4 Variance for CCCD's for Uniform Data

Our ultimate goal is to prove the CLT for $\Gamma_{n,m}$. To achieve this, an important first step is to calculate the limiting variance:

Theorem 4.1. *Suppose $\Omega = \mathbf{R}$, d is Euclidean distance, and $F_{\mathcal{X}} = F_{\mathcal{Y}} = U[0, 1]$. If $m/n \rightarrow r$, then*

$$\lim_{n \rightarrow \infty} \frac{\text{Var}(\Gamma_{n,m})}{m} = v(r) \quad (2)$$

where

$$v(r) \equiv \frac{1536r^5 + 6848r^4 + 11536r^3 + 8836r^2 + 2793r + 180}{9(r+1)^3(4r+3)^4}$$

Proof Sketch. By decomposing $\Gamma_{n,m}$ into internal and external components, we can write its variance as follows:

$$\text{Var}(\Gamma_{n,m}) = \text{Var}(\alpha_{0,m} + \sum_{i=0}^{m-1} \alpha_{j,m} + \alpha_{m,m}),$$

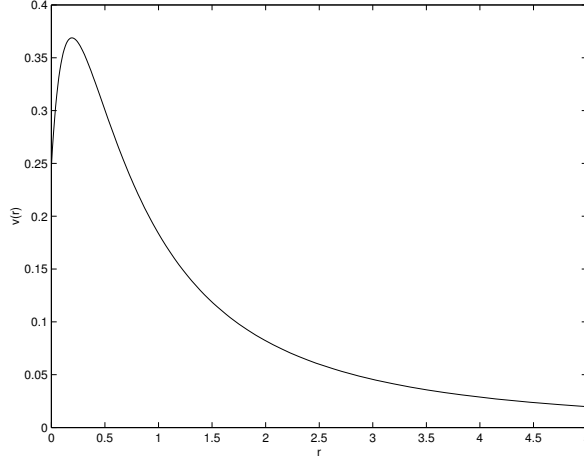


Figure 6: A graph of the limit function $v(r)$, plotted using MATLAB.

which can be expressed as a sum of variances and covariances of the components. Thus, in order to obtain (2), we need to calculate the limiting values of $Var(\alpha_{j,m})$ and $Cov(\alpha_{j_1,m}, \alpha_{j_2,m})$.

We begin by calculating $Var(\alpha_{j,m}) = E(\alpha_{j,m}^2) - (E(\alpha_{j,m}))^2$. Since $E(\alpha_{j,m}^k) = E[E(\alpha_{j,m}^k | N_{j,m})]$, $k = 1, 2$, we will proceed by computing the conditional expectation $E(\alpha_{j,m}^k | N_{j,m})$. Due to the form of the exact conditional probabilities given in Theorem 1.1, these conditional expectations can be expressed in terms of $4^{-N_{j,m}}$.

For convenience, we calculate $E(4^{-N_{j,m}} I_{\{N_{j,m} > 0\}})$ by conditioning on the spacing $L_{j,m} = Y_{(j+1)} - Y_{(j)}$ and computing $E[E(4^{-N_{j,m}} I_{\{N_{j,m} > 0\}} | L_{j,m})]$. This iterated expectation can be calculated by using the fact that the conditional distribution of $N_{j,m}$ given $L_{j,m}$ is Binomial($n, L_{j,m}$):

$$E[4^{-N_{j,m}} I_{\{N_{j,m} > 0\}} | L_{j,m}] = \sum_{1 \leq q \leq n} \left(\frac{1}{4}\right)^q \binom{n}{q} L_{j,m}^q (1 - L_{j,m})^{n-q}$$

We can compute the distribution of $L_{j,m}$ from the joint distribution of the order statistics $Y_{j,m}$, and after integration obtain:

$$E[4^{-N_{j,m}} I_{\{N_{j,m} > 0\}}] = \frac{m}{m+n} \left[\sum_{0 \leq q \leq n} \left(\frac{1}{4}\right)^q \frac{(m+n-q-1)!n!}{(n-q)!(m+n-1)!} - 1 \right]$$

After an intricate evaluation of the limit, using a form of the dominated convergence theorem, the following asymptotic expression is obtained.

$$Var(\alpha_{j,m}) = \frac{144r^3 + 360r^2 + 237r + 20}{9(r+1)^2(4r+3)^2} + o(1)$$

for $j \in \{1, \dots, m-1\}$.

For $Cov(\alpha_{j_1,m}, \alpha_{j_2,m})$, similarly we wish to compute the conditional expectation $E(\alpha_{j_1,m} \alpha_{j_2,m} | N_{j_1,m}, N_{j_2,m})$, by calculating

$$E[4^{-(N_{j_1,m} + N_{j_2,m})} I_{\{N_{j_1,m} > 0, N_{j_2,m} > 0\}}]$$

to obtain exactly

$$m(m-1) \left[\sum_{0 \leq q \leq n} \left(\frac{1}{4}\right)^q \frac{(m+n-q-2)!n!}{(n-q)!(m+n)!} (q-1) + \frac{1}{(m+n)(m+n-1)} \right].$$

This expression must be combined with the square of the expression for the mean, which produces a quantity which tends to zero. To obtain the order of convergence to zero, a more intricate limiting analysis is needed than for the expectation, and finally leads to

$$\begin{aligned} & Cov(\alpha_{j_1, m}, \alpha_{j_2, m}) \\ &= -\frac{1}{m} \frac{r^2(2304r^4 + 9984r^3 + 16096r^2 + 11440r + 3025)}{9(r+1)^3(4r+3)^4} + O\left(\frac{1}{m^2}\right) \end{aligned}$$

for $j_1, j_2 \in \{1, \dots, m-1\}$.

The means and covariances involving the external components are more easily calculated. The conclusion of the theorem is obtained by summing the expressions above for the variances and covariances of the components and taking the limit. \square

The calculation in the proof of Theorem 4.1 shows that the $\alpha_{j, m}$'s are rather weakly dependent, in the sense that the covariances tend to 0 in the order of $O(\frac{1}{m})$. Intuitively, this supports our belief in a Central Limit Theorem for the domination number.

5 Future Research Directions

We plan to continue investigating the limiting behavior of the domination number, namely, the strong law of large numbers and central limit theorem for $\Gamma_{n, m}$. This research will continue in two directions: one is to prove the SLLN for the higher dimensional spaces $\Omega = \mathbf{R}^d$, $d \geq 2$; the other is to prove the CLT for $\Gamma_{n, m}$ for the case of $\Omega = \mathbf{R}$ and $F_X = F_Y = U[0, 1]$, then extend it to more general distribution functions, and finally to higher dimensional spaces. In this effort, we are investigating the use of subadditive process methods for the SLLN, and characteristic function methods, linear quadrant dependence, Stein's method, and methods of Penrose and Yukich [14] for the CLT.

It should be noted that the one dimensional problem is mainly a testing ground for identifying approaches that might be useful in higher dimensions. The real goals are the SLLN and CLT in higher dimensional CCCD problems. One difficulty we encounter in higher dimension situations is dividing the sample space into regions, as we divided $[0, 1]$ into intervals $(Y_{(j)}, Y_{(j+1)})$ in the one dimensional case, since most likely we will not have such a simple identity as $\Gamma_{n, m} = \sum_{j=1}^m \alpha_{j, m}$.

In addition to the CLT and SLLN for $\Gamma_{n, m}$, we would like to apply the developing methods to other similar functions of the CCCD besides the domination numbers. One example is the size of greedy algorithm approximation to the minimum dominating set.

References

- [1] Adam Cannon and Lenore Cowen (2000) Approximation algorithms for the class cover problem, *6th International Symposium on Artificial Intelligence and Mathematics*.

- [2] Jason G. DeVinney (2003) *The Class Cover Problem and Its Applications in Pattern Recognition*, Doctoral dissertation, Johns Hopkins University Department of Mathematical Sciences.
- [3] Jason DeVinney and John C. Wierman (2002) A SLLN for a one-dimensional class cover problem, *Statistics and Probability Letters* 59, 425-435.
- [4] Luc Devroye, Laszlo Györfi, and Gabor Lugosi (1996) *A Probabilistic Theory of Pattern Recognition*, Springer.
- [5] Richard O. Duda and Peter E. Hart (1973) *Pattern Classification and Scene Analysis*, Wiley-Interscience.
- [6] Richard O. Duda, Peter E. Hart, and David G. Stork (2001) *Pattern Classification*, Wiley-Interscience.
- [7] Teresa W. Haynes, Stephen T. Hedetniemi, and Peter J. Slater (1998) *Domination in Graphs, Fundamentals*, Marcel Dekker, Inc., New York.
- [8] Teresa W. Haynes, Stephen T. Hedetniemi, and Peter J. Slater (1998) *Domination in Graphs, Advanced Topics*, Marcel Dekker, Inc., New York.
- [9] S.R. Kulkarni, G. Lugosi, and S.S. Venkatesh (1998) Learning pattern classification – A survey, *IEEE Transactions on Information Theory* 44, 2178-2206.
- [10] E.L. Lehmann (1966) Some concepts of dependence, *Annals of Mathematical Statistics* 37, 1087-1437.
- [11] Przemyslaw Matula (1992) A note on the almost sure convergence of sums of negatively dependent random variables, *Statistics and Probability Letters* 15, 209-213.
- [12] Charles M. Newman (1984) Asymptotic independence and limit theorems for positively and negatively dependent random variables, *Inequalities in Statistics and Probability*, IMS Lecture Notes-Monograph Series, Volume 5, 127-140.
- [13] O. Ore (1962) *Theory of Graphs*, Amer. Math. Soc. Publ., Providence, RI.
- [14] Mathew D. Penrose and J.E. Yukich (2001) Central limit theorem for some graphs in computational geometry, *Annals of Applied Probability* 11, 1005-1041.
- [15] Carey Priebe, Jason DeVinney, and David Marchette, (2001) On the distribution of the domination number for random class cover catch digraphs, *Statistics and Probability Letters* 55, 239-246.
- [16] Carey Priebe and David Marchette (2003) Characterizing the scale dimension of a high-dimensional classification problem, *Pattern Recognition* 36, 45-60.
- [17] Carey Priebe, David Marchette, Jason DeVinney, and Diego Socolinsky (2003) Classification using class cover catch digraphs, *Journal of Classification*, to appear.