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

Pengfei Xiang* John C. Wierman†

Abstract

Finding a minimum dominating set is one of the NP complete core problems. In this paper, we will discuss the limiting behavior of the domination number of random class cover catch diagraphs (CCCDs). The CCCD problem is motivated by its applications in pattern classification. For the special case of 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 CCCD.

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

1 Introduction

1.1 Class Cover Problem

The class cover problem (CCP) is motivated by its applications in statistical pattern classification [**?**]. It was first initiated by Cowen and Cannon [**?**], and has been actively studied recently, since the solution to it can be

*Department of Mathematical Sciences, The Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21218, USA. xiang@jhu.edu. Supported in part by Acheson J.Duncan Fund for the Advancement of Research in Statistics.

†Department of Mathematical Sciences, The Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21218, USA. wierman@jhu.edu. Supported in part by Acheson J.Duncan Fund for the Advancement of Research in Statistics.

directly used to generate classifiers[1] competitive with the other methods. Roughly speaking, in the setting of classification, the CCP is just a problem of selecting a small set of data points to be representative of a class.

Now we give a formal description of the CCP. 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 $\forall \alpha, \beta \in \Omega$. We suppose $\{X_i : i = 1, \cdots, n\}$ and $\{Y_j : j = 1, \cdots, m\}$ are two sets of i.i.d. random variables taking values in Ω, with class-conditional distribution functions $F_{\mathcal{X}}$ and $F_{\mathcal{Y}}$, respectively. We assume X_i's are independent of Y_j's.

Definition 1.1. *For each X_i, we define the covering ball $B(X_i) = \{\omega \in \Omega : d(\omega, X_i) < min_j d(Y_j, X_i)\}$.*

A *class cover of* $\mathcal{X}$ is a subset of covering balls whose union contains all X_i's. Obviously the set consisting of all covering balls is a class cover. However, we want to choose a class cover to represent class $\mathcal{X}$ that is as small as possible, to make the classifier less complex while keeping most of the relevant information. Therefore, the CCP we consider here is to find a minimum cardinality class cover.

1.2 Class Cover Catch Digraph

We can convert the CCP to a purely graph theory problem as follows:

Definition 1.2. *The class cover catch digraph (CCCD) induced by a CCP is the digraph $D = (V, A)$ with the vertex set $V = \{X_i : i = 1, \cdots, n\}$ and the edge set A such that there is a directed edge (X_i, X_j) if and only if $X_j \in B(X_i)$.*

Definition 1.3. *The set $S \subset V$ is a dominating set of a diagraph $D = (V, A)$ 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 easy to see that the CCP is actually equivalent to finding a minimum cardinality dominating set of the corresponding CCCD. Cowen and Cannon prove 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 [**?**].

1.3 Domination Number

Definition 1.4. *The domination number of a CCCD is the cardinality of the CCCD's minimum dominating set.*

[1] We will formally give the definition of classifier in Section 1.4.

In 1962, Ore first used the name "domination number" in his book [**?**]. For its many applications in such fields as computer networks, social sciences, computational complexity, etc, there has been a rising interest in this area, with a lot of results obtained very recently. The book by Haynes, Hedetniemi and Slater provides a comprehensive discussion of domination in graphs [**?**]. More advanced topics are covered in [**?**].

In the CCCD problem, the domination number is especially important. It is useful in approximating minimum dominating sets. Here we denote the domination number by $\Gamma_{n,m}(F_{\mathcal{X}}, F_{\mathcal{Y}})$, or simply by $\Gamma_{n,m}$. Obviously, $\Gamma_{n,m}$ is a random variable whose distribution depends on $n, m, F_{\mathcal{X}}$ and $F_{\mathcal{Y}}$.

1.4 Applications in Pattern Classification

Pattern classification is "the assignment of a physical object or event to one of several pre-specified categories" (See [**?**, page 2]). It is widely applied to real world problems such as automated speech recognition, DNA sequence identification, fingerprint identification, etc.

The abstract mathematical model of the pattern classification problem is formulated as follows [**?**]. For simplicity, but without loss of generality, suppose we have two classes of objects of interest, which we will call class $\mathcal{X}$ and class $\mathcal{Y}$, respectively. We assume that the objects of both classes belong to a common dissimilarity space Ω. To model the uncertainty about which class the objects we encounter belong, we assume that there are *prior* probabilities $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ for these two classes ($\sum_{c\in\{\mathcal{X},\mathcal{Y}\}} P_c = 1$). Furthermore, we assume that given the class $\mathcal{X}$ or $\mathcal{Y}$, the objects of that class are drawn according to the *class-conditional* distribution function $F_{\mathcal{X}}(x)$ or $F_{\mathcal{Y}}(y)$. We can generate a random pair $\big(c(\Psi), \Psi\big)$ in 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 a classification problem, for an observation pair $(c(\psi), \psi)$ generated as above, only the data part ψ is given while the class label part $c(\psi)$ is unknown, so the goal of a *classifier* is to guess whether $c(\psi)$ is $\mathcal{X}$ or $\mathcal{Y}$. In many situations, in addition to ψ, we are also given a training sample of size k with known classification:

$$D_k = \Big\{\big(c(\psi_1), \psi_1\big), \cdots, \big(c(\psi_k), \psi_k\big)\Big\}.$$

So generally, a classifier is a function $\hat{c}_k(\psi) = \hat{c}_k(\psi, D_k)$, 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 *probability of error*, or *misclassification rate*, given by

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

The CCP can be used to build classifiers. Just shown here as an example, a simple classifier can be constructed as follows: by switching the roles of $\mathcal{X}$ and $\mathcal{Y}$, we can get a pair of dual CCP's, resulting in two solutions such as $\mathcal{B}_{\mathcal{X}} = \{B(X_i) : i \in I, I \subset \{1, \cdots, n\}\}$ and $\mathcal{B}_{\mathcal{Y}} = \{B(Y_j) : j \in J, J \subset \{1, \cdots, m\}\}$, 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 $\hat{c}(\psi) : \Omega \to \{\mathcal{X}, \mathcal{Y}\}$ as follows:

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

For a thorough description of pattern classification, see the two classic books [**?**] and [**?**]. More details about the CCP's application to classification are presented in [**?**].

2 Previous Results

There have been several research results on the probabilistic properties of $\Gamma_{n,m}$ in the case of $\Omega = \mathbf{R}$. In this one dimensional situation, we denote $Y_{(j)}$ as the jth order statistic of $Y_0 = 0, Y_1, \cdots, Y_m, Y_{m+1} = 1$, 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 $Y_{(j+1)}$. We refer to $\alpha_{j,m}$ $(j = 0, m)$ as external components, and $\alpha_{j,m}$ $(j = 1, \cdots, m-1)$ as internal components. It should be noted that $\Gamma_{n,m} = \sum_{j=0}^{m} \alpha_{j,m}$. This way we are able to decompose the problem into $m+1$ sub-problems of finding the domination number $\alpha_{j,m}$ in the interval $[Y_{(j)}, Y_{(j+1)}]$. It is easy to see that $\alpha_{j,m}$ can be at most 2, because all X_i's in $[Y_{(j)}, Y_{(j+1)}]$ 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.

Priebe, Devinney and Marchette [**?**] find the conditional distribution of $\alpha_{j,m}$ given $N_{j,m}$ for the special case of $\Omega = \mathbf{R}$ and $F_{\mathcal{X}} = F_{\mathcal{Y}} = U[0,1]$, where $U[0,1]$ is the uniform distribution on the interval $[0,1]$:

Theorem 2.1. *Suppose $\Omega = \mathbf{R}$. If $F_{\mathcal{X}} = F_{\mathcal{Y}} = U[0,1]$, then the following are true:*

- *For $j \in \{0, 1, \cdots, m\}$, 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 $j \in \{1, 2, \cdots, m-1\}$, if $N_{j,m} = n_j > 0$ then*

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

From the above theorem, we know that $\alpha_{j,m} \in \{0, 1, 2\}$, and $\alpha_{j,m} = 0$ iff $N_j = 0$. Given $N_j = n_j > 0$, for $j \in \{1, 2, \cdots, m-1\}$, the conditional probability of $\alpha_{j,m} = 2$ is an increasing function of n_j, which just means that $\alpha_{j,m}$ tends to become larger as the number of $\mathcal{X}$ points increases.

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

Theorem 2.2. *Suppose* $\Omega = \mathbf{R}$. *For the special case of* $F_{\mathcal{X}} = F_{\mathcal{Y}} = U[0, 1]$, *we have*

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

where $g(r) \equiv \frac{12r+13}{3(r+1)(4r+3)}$, *and* $m = \lfloor rn \rfloor, r \in (0, \infty)$.

Using $MATLAB$, we draw a graph of $\lim_{n \to +\infty} \frac{\Gamma_{n,m}}{m} = g(r)$ as a function of r:

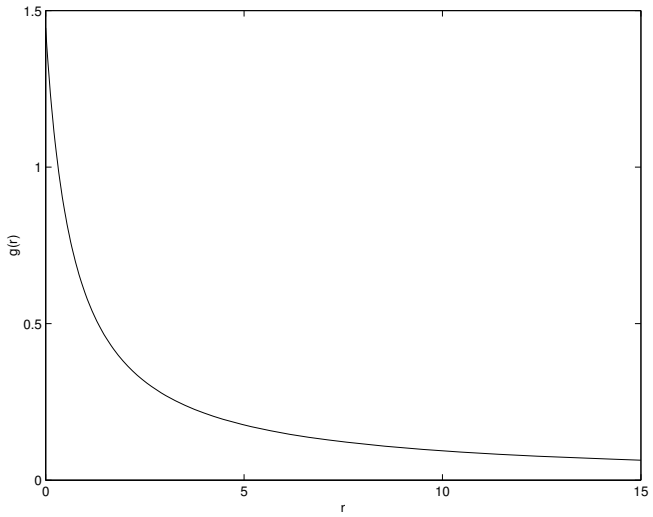

Figure 1: $g(r)$

As shown in Figure 1, we can see that when $r \to \infty, g(r) \to 0$, which is justified by the fact that asymptotically the interval between $Y_{(j)}$ and $Y_{(j+1)}$ contains no $\mathcal{X}$ point almost surely. Moreover, as $r \to 0, g(r) \to \frac{13}{9}$. This corresponds to the situation where each interval between $Y_{(j)}$ and $Y_{(j+1)}$ contains very large number of $\mathcal{X}$ points. According to Theorem 1.1, the probability of $\alpha_{j,m} = 1$ is approximately $\frac{5}{9}$, while the probability of

$\alpha_{j,m} = 2$ is approximately $\frac{4}{9}$, therefore $\frac{13}{9} = \frac{5}{9} \cdot 1 + \frac{4}{9} \cdot 2$ can be just viewed as an expectation value of $\alpha_{j,m}$.

In their proof in [**?**], DeVinney and Wierman first prove the special case of $r = 1$. They construct two related Poisson processes A and B, with common rate $\lambda \in (0, \infty)$. Points of A play the role of $\mathcal{X}$ points, and points of B play the role of $\mathcal{Y}$ points. Classical SLLN can be applied to a CCP induced from these A and B points, then the result is transferred back to the original setting.

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 also find an alternative proof to Theorem 2.2 by using an existing SLLN theorem for *quadrant dependent* random variables [**?**]. The concept of quadrant dependence was first introduced by E.L. Lehmann in [**?**], and the limiting theory for quadrant dependent random variables is comprehensively discussed in [**?**].

3 Strong Law of Large Numbers (SLLN)

In Theorem 2.2, we assume that classes $\mathcal{X}$ and $\mathcal{Y}$ both have uniform distribution. But in real world applications, they usually have different non-uniform distributions. Our research has proved an extension to Theorem 2.2 for more general distribution functions in the one dimensional case:

Theorem 3.1. *Suppose $\Omega = \mathbf{R}$. Assume the densities $f_{\mathcal{X}}(x)$ and $f_{\mathcal{Y}}(y)$ are bounded functions with a finite number of discontinuities. Then*

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

where $g(r) = \frac{12r+13}{3(r+1)(4r+3)}$ and $m/n \to r$.

Proof Sketch. Our proof is conducted in two phases:

We first consider piece-wise constant densities $f_{\mathcal{X}}$ and $f_{\mathcal{Y}}$, i.e.

$$\begin{aligned} f_{\mathcal{X}}(x) &= \sum_{l=1}^{k} a_l I_{[c_{l-1}, c_l)}(x), \\ f_{\mathcal{Y}}(y) &= \sum_{l=1}^{k} b_l I_{[c_{l-1}, c_l)}(y) \end{aligned}$$

where $a = c_0 < c_1 < \cdots < c_k = b$. To prove (1) for this type of density function, we divide the CCP into sub-CCP's with uniform distribution 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 $g\big(r\frac{f_{\mathcal{Y}}(u)}{f_{\mathcal{X}}(u)}\big), u \in [c_{l-1}, c_l]$. By adding up the domination numbers for all the intervals, we get an approximation Γ' to $\Gamma_{n,m}$. 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.

Then for the general continuous case, we construct a sequence of piecewise 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 step in our 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_i and $Y_{i,k}$, we can argue that the SLLN still holds for the original densities $F_{\mathcal{X}}$ and $F_{\mathcal{Y}}$. □

In addition, we prove that equal $f_{\mathcal{X}}$ and $f_{\mathcal{Y}}$ give the maximum limit in the SLLN, that is:

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

$$\int g\Big(r\frac{f_{\mathcal{Y}}(u)}{f_{\mathcal{X}}(u)}\Big) f_{\mathcal{Y}}(u)du \le g(r)$$

where the equality holds if and only if $f_{\mathcal{X}}(u) = f_{\mathcal{Y}}(u) \quad a.s.$

Proof. Note that $g(r)$ is a convex function, hence $g^*(r) = g(\frac{1}{r})$ is a concave function. Therefore by Jensen's inequality, we have

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

An intuitive explanation for the above inequality is that when class $\mathcal{X}$ and class $\mathcal{Y}$ both have the same distribution pattern, a larger dominating set is needed to distinguish $\mathcal{X}$ from $\mathcal{Y}$.

This result could be used to construct distribution-free statistical tests.

4 Asymptotic Variance

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* $\Omega = \mathbf{R}$, *and* $F_{\mathcal{X}} = F_{\mathcal{Y}} = U[0,1]$. *Then*

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

where $v(r) \equiv \frac{2304r^6+13056r^5+29792r^4+34512r^3+20697r^2+5586r+360}{18(r+1)^3(4r+3)^4}$, *and* $m/n \to r$.

Using $MATLAB$, we draw a graph of $\lim_{n\to+\infty} \frac{Var(\Gamma_{n,m})}{m} = v(r)$ as a function of r:

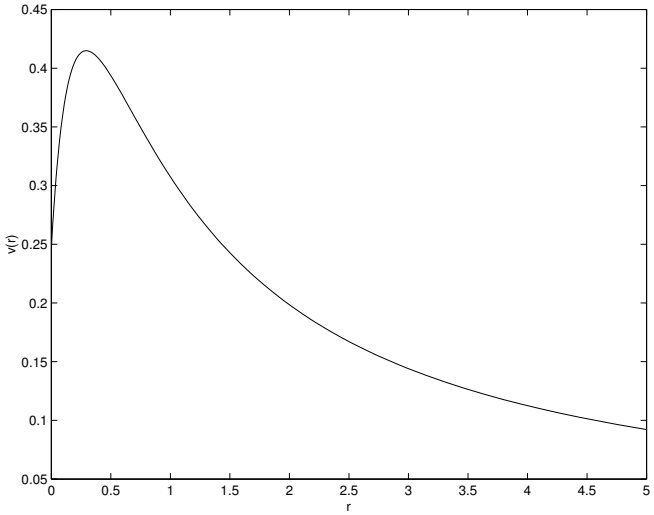

Figure 2: $v(r)$

As an intermediate step in the proof of Theorem 4.1, we obtain that for any j_1, j_2 such that $1 \le j_1, j_2 \le m-1$ and $j_1 \ne j_2$,

$$Cov(\alpha_{j_1,m}, \alpha_{j_2,m}) \quad = \quad \frac{-r^2(2304r^4+9984r^3+16096r^2+11440r+3025)}{9(r+1)^3(4r+3)^4} \cdot\frac{1}{m} + o(\frac{1}{m^2}).$$

This says that $\alpha_{j,m}$'s are weakly dependent in the sense that the covariances tend to 0 in the order of $O(m)$. This fact may be helpful in proving the Central Limit Theorem.

Proof Sketch. First we compute the conditional expectation $E(\alpha_{j,m}^k \mid N_{j,m})$, $k = 1, 2$ and $E(\alpha_{j_1,m}\alpha_{j_2,m} \mid N_{j_1,m}, N_{j_2,m})$ using those formulas in Theorem 1.1; then we need to calculate $E(4^{N_{j,m}})$ and $E(4^{N_{j_1,m}+N_{j_2,m}})$ to get the final result. □

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 space $\Omega = \mathbf{R}^d, d \geq 2$; the other is to prove the CLT for $\Gamma_{n,m}$ for the case of $\Omega = \mathbf{R}$ and $F_{\mathcal{X}} = F_{\mathcal{Y}} = U[0, 1]$, then extend it to more general distribution functions, and finally to higher dimensional spaces.

To directly prove the CLT using characteristic function methods, we might need to calculate the 3rd or even 4th moment. Considering that it has taken us a long time to get the limiting variance (i.e., 2nd moment), we certainly want to avoid this complicated computation by using or improving some existing CLT theorem. Some possible ways include Stein's method and quadrant dependence. Another useful reference we would like to look into is a series of results by J.E. Yukich and his collaborators [**?**].

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 how to divide the whole sample space into regions, as we divided the $[0, 1]$ into intervals $(Y_{(j)}, Y_{(j+1)})$ in the one dimensional case. Therefore most likely we will not have such a simple identity as $\Gamma_{n,m} = \sum_{j=1}^{m} \alpha_{j,m}$.

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