A SLLN for a One-dimensional Class Cover Problem

Jason DeVinney* John C. Wierman†

Mathematical Sciences Department
Johns Hopkins University

July 31, 2002

Abstract

Class cover catch digraphs arise in classification problems in statistical pattern recognition. We prove a strong law of large numbers for the domination number in a random one-dimensional model of class cover catch digraphs. The proof avoids complicated computations due to the dependence of random variables by considering a related Poisson process problem where we may apply classical strong law results and Chernoff exponential probability bounds. Complete convergence in the Poisson representation establishes the desired result for the original problem.

Keywords - class cover problem, catch digraphs, domination, Poisson process, complete convergence, strong law of large numbers, classification, pattern recognition

1 Introduction

Consider a vector space Ω, a dissimilarity d, and two finite, non-empty sets, $X, Y \subset \Omega$. Recall that a dissimilarity d on Ω is a function $d : \Omega \times \Omega \to \Re$ such that $d(i,j) = d(j,i) \geq d(i,i) = 0$ [CF94]. We refer to $X = \{x_1, \ldots, x_n\}$ as the *target* class and $Y = \{y_1, \ldots, y_m\}$ as the *non-target* class. The general class cover problem (CCP) is to find a minimum cardinality set of balls whose union contains all of the set X and no points in Y. For each $x_i \in X$ we define a *covering ball* $B_i = \{z \in \Omega : d(z, x_i) < r_i\}$ where $r_i = \min\{d(y, x_i) : y \in Y\}$. We define a *cover* of the target class, X, as a set of covering balls C such that $\forall x \in X$, $\exists B \in C$ such that $x \in B$. The CCP we consider is to find the minimum cardinality cover of X or more formally,

$$\min \left\{ |J| : J \subset [n], X \subset \bigcup_{j \in J} B_j \right\} \tag{1}$$

*devinney@mts.jhu.edu - Supported in part by Office of Naval Research Grant N00014-01-1-0011
†Supported by a Navy-American Society for Engineering Education sabbatical fellowship

where the notation $[n]$ stands for the set $\{1, 2, \ldots, n\}$. Note that by this definition each covering ball is centered at an element of X and the cover cannot contain any elements of Y.

A directed graph $D = (V, A)$ consists of a set V of vertices and a set A of arcs which are ordered pairs of vertices. The *catch digraph* induced by a collection of sets $\mathcal{S} = \{S_1, S_2, \ldots, S_n\}$ and corresponding base points $\mathcal{T} = \{T_1, T_2, \ldots, T_n\}$ is the digraph with vertex set $V = \{v_1, v_2, \ldots, v_n\}$ and an arc from v_i to v_j if and only if $T_j \in S_i$ (see [MM99]). We call the catch digraph induced by the collection of B_i and their centers x_i the *class cover catch digraph* (CCCD) *induced by* (Ω, X, Y, d).

A dominating set of a directed graph $D = (V, A)$ is a set of vertices $S \subset V$ such that for any $v \in V$, either $v \in S$ or $\exists w \in S : (w, v) \in A$. We denote the size of a minimum cardinality dominating set of a digraph D as $\gamma(D)$. Let J be a collection of indices such that $\{B_j : j \in J\}$ is a solution to some CCP. Then the set $\{v_j : j \in J\}$ is a minimum cardinality dominating set in the CCCD induced by that CCP and vice versa. The CCP on (Ω, X, Y, d) is therefore equivalent to finding a minimum cardinality dominating set in the CCCD induced by (Ω, X, Y, d). Determining the size of a minimum cardinality dominating set in a general graph (or digraph) is NP-Hard [HHS98]. This does not immediately imply that the CCP is NP-Hard since we have not characterized which digraphs are CCCD's. This topic is more thoroughly covered in [CEHS02] and [DP02].

If X and Y are sets of independent identically distributed observations drawn from the class conditional distributions F_X and F_Y respectively, then we have a randomized version of the CCP. We define the random variable $\Gamma_{n,m}(F_X, F_Y)$ as the size of a minimum cardinality dominating set in a random CCCD induced by n observations from F_X and m observations from F_Y, all stochastically independent. We are interested in the properties of the probability distribution of $\Gamma_{n,m}(F_X, F_Y)$.

The CCP is motivated by supervised pattern classification (see [KLV98]). The cover can be used to provide a simple estimate of the discriminant region for the target class. Priebe, DeVinney, Marchette and Socolinsky [PMDS02] give the details of how this estimate can be achieved using the CCP. By switching the role of target class between the classes of observations, two different instances of the CCP can be solved, resulting in two covers C_X and C_Y. A simple classifier $g : \Omega \to \{0, 1, 2\}$ obeys the following rule:

$$g(z) = \begin{cases} 1 & : \quad z \in C_X \cap C_Y^c \\ 2 & : \quad z \in C_Y \cap C_X^c \\ 0 & : \quad \text{otherwise} \end{cases}$$

where $g(x) = 0$ indicates no decision. More elaborate methods of applying the CCP to classification are presented in [PMDS02].

The CCP was introduced by Cannon and Cowen [CC00]. They study a variation of the CCP in which the radii of the covering balls must all be the same. Cannon and Cowen construct a polynomial time approximation algorithm for this CCP. Priebe, DeVinney and Marchette [PDM02] consider the one-dimensional CCP ($\Omega = \Re$ and d is the euclidean metric) and find the exact distribution of $\Gamma_{n,m}(F_X, F_Y)$ when $F_X = F_Y = U[0, 1]$, where $U[0, 1]$ is the uniform distribution on the interval $[0, 1]$.

2 One Dimensional CCP

We consider the special case of the CCP where $\Omega = \Re$, d is the euclidean metric and $F_X = F_Y = U[0,1]$. Since we will only consider the case where $F_X = F_Y = U[0,1]$, we may simplify notation and denote $\Gamma_{n,m}(F_X, F_Y)$ as $\Gamma_{n,m}$. It will be convenient to write $\Gamma_{n,m}$ as the sum of $m+1$ random variables which correspond to intervals between succesive non-target class points. Let the random variable n_i, be the number of X points located between $Y_{(i)}$ and $Y_{(i+1)}$ (where $Y_{(j)}$ is the jth order statistic of the points in Y, with $Y_{(0)} = 0$ and $Y_{(n+1)} = 1$). Let the random variable α_i be the minimum number of covering balls needed to cover the n_i points of X located between $Y_{(i)}$ and $Y_{(i+1)}$. Then we have

$$\sum_{i=0}^{m} \alpha_i = \Gamma_{n,m}. \tag{2}$$

It will be useful to distinguish α_i for $i = 1, 2, \ldots, m-1$ as the *internal* components and α_0 and α_m as *external* components. Priebe, DeVinney and Marchette [PDM02] calculate the exact distribution of the α_i. This result is summarized in Lemma 1.

Lemma 1. *If $F_X = F_Y = U[0,1]$ then the following are true.*
(i) For $i \in \{0, 1, \ldots, m\}$, if $n_i = 0$ then $\alpha_i = 0$.
(ii) For $i \in \{0, m\}$, if $n_i > 0$ then $\alpha_i = 1$.
(iii) For $i \in \{1, 2, \ldots, m-1\}$, if $n_i = k$, $k > 0$ then

$$P[\alpha_i = 1 | n_i = k] = 1 - P[\alpha_i = 2 | n_i = k] = \frac{5}{9} + \frac{4}{9}\frac{1}{4^{k-1}}.$$

We see that $\alpha_i \in \{0, 1, 2\}$ for $i = 0, 1, \ldots, m$. [1]

3 Main Result

This article extends the results of Priebe, DeVinney and Marchette for $\Gamma_{n,m}$. We prove the following Strong Law of Large Numbers:

Theorem 1. *For $a \in (0, \infty)$,*

$$\lim_{n \to \infty} \frac{\Gamma_{\lfloor an \rfloor, n}}{n} = \frac{a(13a+12)}{3(a+1)(3a+4)} \quad a.s., \tag{3}$$

where $\lfloor an \rfloor$ is the greatest integer less than or equal to an.

The theorem is stated in terms of the average of the internal components, which corresponds most closely to the usual strong law of large numbers. Note that the limiting expression is an increasing function of a, the ratio of numbers of target class and non-target class points. As $a \to 0$, the limiting expression converges to zero, reflecting the fact that most intervals between sucessive Y points will contain no X point. More interestingly, as $a \to \infty$, the limiting expression converges to $\frac{13}{9}$. This corresponds to each interval between sucessive Y points containing a large number of

[1] The fact that $\alpha_i \in \{0, 1, 2\}$ for $i = 0, 1, \ldots, m$ is actually a property of the one-dimensional CCP and holds for all distributions F_X and F_Y.

X points. By Lemma 1, the probability that one ball covers all points in this interval is near $\frac{5}{9}$ and the probability that two balls are needed is near $\frac{4}{9}$, resulting in an expected value of $\frac{13}{9}$.

Alternatively, one may consider normalizing by the number of X points. In this case we see $lim_{n\to\infty}\frac{\Gamma_{\lfloor an\rfloor,n}}{\lfloor an\rfloor} = \frac{13a+12}{3(a+1)(3a+4)}$ $a.s.$ The limiting expression now converges to zero as $a \to \infty$ and to one as $a \to 0$. The quantity $\frac{\Gamma_{\lfloor an\rfloor,n}}{\lfloor an\rfloor}$ gives a measure of the reduction in complexity resulting from using the dominating set as a representation for the entire target class.

4 Sketch of Proof

For clarity of presentation we will prove the special case of the main theorem where $a = 1$, showing that $\frac{\Gamma_{n,n}}{n} \to \frac{25}{42}$ a.s. The proof is outlined in this section with details presented in Section 5. A description of the modifications necessary to prove the general case, $m = an$, is presented in Section 6.

Note that there are only two external components and that their value is bounded above by one, so the asymptotic behavior of $\Gamma_{n,n}$ is determined by the internal components. The n_i depend on the lengths $(Y_{(i)}, Y_{(i+1)})$ which are identically distributed implying that the n_i are identically distributed. This fact and Lemma 1 imply that the internal components are identically distributed. However, the α_i are not independent random variables. Due to this dependence we cannot apply the standard strong law of large numbers. Attempts to compute higher moments have resulted in complicated expressions which have not been useful in establishing convergence. We circumvent this problem with an approach that establishes the strong law of large numbers for the cardinality of a solution of a CCP in a Poisson process setting. We then transfer the result back to the original setting.

For a Poisson process W we let W_i denote the time of the ith arrival and $W(t)$ as the number of arrivals in W before time t. Consider two one-dimensional Poisson processes, A and B, with common rate λ with $0 < \lambda < \infty$. Points of A will play the role of target class points and points in B will play the role of non-target class. We let X_i be the number of A points in (B_i, B_{i+1}) and ρ_i be the minimum number of covering balls needed to cover the X_i points of A in (B_i, B_{i+1}).

We consider Γ'_n, defined as the solution to the CCP on the points of A and B in $(0, B_{n+1})$. This CCP has exactly n non-target class points and a random number, $N_n = A(B_{n+1})$, of target class points. It has an advantage over our original CCP; the ρ_i (analogous to the internal and external components) are independent random variables, allowing the application of the standard Strong Law of Large Numbers with fourth moment assumptions. A simple calculation in this Poisson process setting evaluates the limit as $\frac{25}{42}$.

Using the conditional uniformity property of Poisson processes, and a standard density transformation result, we see that the N_n points of A and the n points of B have the same distribution as the order statistics of $N_n + n$ observations of a uniform distribution on $(0, B_{n+1})$. Rescaling the interval $(0, B_{n+1})$ to $(0, 1)$ does not change the value of Γ'_n. For each n, we correct the number of target class points as follows: If $N_n < n$, we add $n - N_n$ points $A'_1, A'_2, \ldots, A'_{n-N_n}$ which are uniformly distributed on $(0, 1)$, mutally independent and independent of the Poisson processes. If $N_n > n$, we

choose a random subset of $N_n - n$ points from $\{A_1, A_2, \ldots, A_{N_n}\}$ with all subsets of size $N_n - n$ equally likely, to remove from consideration. We then calculate a revised CCP solution with cardinality Γ_n on this corrected set of points. The random variable Γ_n in the Poisson process setting has an identical distribution with $\Gamma_{n,n}$ in the original setting.

Adding or removing a target class point affects the solution of the one-dimensional CCP by at most one. The number of points to be added or removed, $N_n - n$, form a random walk that arises naturally from the Poisson processes A and B. The fluctuations in the random walk are sufficiently small that the effect of adding or removing points is negligible in the limit. Combining these ideas, we show that the almost sure limits of $\frac{\Gamma'_n}{n}$ and $\frac{\Gamma_n}{n}$ are identical.

To transfer the result from Γ_n in the modified Poisson process setting to $\Gamma_{n,n}$ in the original setting, there is one additional complication to overcome. While the marginal distributions of Γ_n and $\Gamma_{n,n}$ are identical for each n, the joint distributions of $\{\Gamma_n;\ n = 1, 2, \ldots\}$ and $\{\Gamma_{n,n};\ n = 1, 2, \ldots\}$ are not due to the adding or removing of different sets of points for each n. However, if care is taken to demonstrate complete covergence for Γ_n, we obtain complete covergence (and therefore almost sure covergence) for $\Gamma_{n,n}$.

5 Proof

In this section we provide the details to complete the proof sketch in the previous section.

5.1 Poisson Representation

For proving a limiting result, we find it useful to convert the model from one in which uniformly distributed points are added to a fixed interval, into a model where the limit corresponds to increasing the length of the interval. As mentioned above, we use a correspondence between uniformly distributed points and a Poisson process.

We rely upon two standard distributional results. First, from an undergraduate-level density transformation exercise, if $X_1, X_2, \ldots, X_{n+1}$ are independent identically-distributed random variables with an Exponential distribution with parameter λ, then

$$\left(\frac{X_1}{\sum_{i=1}^{n+1} X_i}, \frac{X_1 + X_2}{\sum_{i=1}^{n+1} X_i}, \frac{X_1 + X_2 + X_3}{\sum_{i=1}^{n+1} X_i}, ..., \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n+1} X_i} \right)$$

has the same joint distribution as the order statistics of n independent Uniform[0,1] random variables. Secondly, recall the "conditional uniformity" property of Poisson processes: If $W(t) = n$, then the n Poisson points on $[0, t]$ conditionally have the same distribution as the order statistics of n independent Uniform$[0, t]$ random variables.

We consider the B process on $(0, B_{n+1})$. By the first property above, the first n points in the B process may be considered to be uniformly distributed on $(0, B_{n+1})$. At time B_{n+1}, there is a random number of A points, N_n. If we further condition on $N_n = m$, then both A and B points are uniformly distributed on $(0, B_{n+1})$. By rescaling the interval, the class cover problem on $A_1, A_2, \ldots, A_m$ and $B_1, B_2, \ldots, B_n$ is equivalent to the CCP on $\frac{A_1}{B_{n+1}}, \frac{A_2}{B_{n+1}}, \ldots, \frac{A_m}{B_{n+1}}$ and $\frac{B_1}{B_{n+1}}, \frac{B_2}{B_{n+1}}, \ldots, \frac{B_n}{B_{n+1}}$. Therefore, in the case where we stop the B process at B_{n+1} and condition on $N = m$, the size of the solution to the CCP on the A and B points has the same distribution as $\Gamma_{m,n}$.

5.2 Expected Value of Internal Components

For now, we return to observing the Poisson processes in $(0, B_{n+1})$. Recall that we let Γ'_n represent the size of a solution to the CCP on $A_1, \ldots, A_{N_n}$ and $B_1, \ldots, B_n$. Also X_i is the number of A points in (B_i, B_{i+1}) and ρ_i is the minimum number of covering balls needed to cover the X_i points of A in (B_i, B_{i+1}). We proceed by showing $E[\rho_i] = E[\rho_1] = \frac{25}{42}$ $\forall i \in \{1, \ldots n-1\}$. By the lack of memory property of the exponential distribution, $X_i = Z - 1$ where Z is a geometric random variable with parameter $p = \frac{1}{2}$. By conditional uniformity, and Lemma 1 we see that the ρ_i depend only on the value of X_i. Therefore since the X_i are identically distributed [Dav81], ρ_i for $i \in \{1, 2, \ldots, n-1\}$ are also identically distributed. We now calculate $E[\rho_1]$.

$$P[\rho_1 = 0] = P[X_1 = 0] = \frac{1}{2}$$

and

$$\begin{aligned} P[\rho_1 = 1] &= \sum_{k=1}^{\infty} P[\rho_1 = 1 | X_1 = k] P[X_1 = k] \\ &= \sum_{k=1}^{\infty} \left[\frac{5}{9} + \left(\frac{4}{9}\right) 4^{1-k} \right] \frac{1}{2^{k+1}} \\ &= \frac{5}{9} \sum_{k=1}^{\infty} 2^{-(k-1)} + \frac{8}{9} \sum_{k=1}^{\infty} 8^{-k} \\ &= \frac{5}{18} + \frac{8}{63} \\ &= \frac{17}{42}, \end{aligned} \tag{4}$$

and by subtraction,

$$P[\rho_1 = 2] = \frac{2}{21},$$

from which we obtain

$$E[\rho_1] = \frac{25}{42}.$$

5.3 Complete Convergence of Γ'_n

Next we will show complete covergence of $\frac{\Gamma'_n}{n}$ to $\frac{25}{42}$.

$$\begin{aligned} \sum_{n=1}^{\infty} P\left[\left| \frac{\Gamma'_n}{n} - \frac{25}{42} \right| \geq \epsilon \right] &= \sum_{n=1}^{\infty} P\left[\left| \sum_{i=0}^{n} \rho_i - \frac{25}{42} n \right| \geq n\epsilon \right] \\ &= \sum_{n=1}^{\infty} P\left[\left| \rho_0 + \rho_n - \frac{25}{42} + \sum_{i=1}^{n-1} \left(\rho_i - \frac{25}{42} \right) \right| \geq n\epsilon \right] \\ &\leq \sum_{n=1}^{\infty} P\left[\left| \rho_0 + \rho_n - \frac{25}{42} \right| + \left| \sum_{i=1}^{n-1} \left(\rho_i - \frac{25}{42} \right) \right| \geq n\epsilon \right] \end{aligned} \tag{5}$$

which, using the fact that $0 \leq \rho_i \leq 1$ $i \in \{0, n\}$,

$$\leq \sum_{n=1}^{\infty} P\left[\left|\sum_{i=1}^{n-1}\left(\rho_i - \frac{25}{42}\right)\right| \geq n\epsilon - 2\right].$$

Now we use a fourth moment version of Markov's inequality, and then expand the fourth power on the sum and use the fact that $E[\rho_i - \frac{25}{42}] = 0$ for $i = 1, \ldots, n-1$.

$$\begin{aligned}\sum_{n=1}^{\infty} P\left[\left|\sum_{i=1}^{n-1}\left(\rho_i - \frac{25}{42}\right)\right| \geq n\epsilon - 2\right] &\leq \sum_{n=1}^{\infty} \frac{E[|\sum_{i=1}^{n-1}(\rho_i - \frac{25}{42})|^4]}{(n\epsilon - 2)^4} \\ &\leq \sum_{n=1}^{\infty} \frac{Cn^2}{(n\epsilon - 2)^4} \\ &< \infty\end{aligned} \tag{6}$$

And thus we have shown the complete covergence of $\frac{\Gamma'_n}{n}$ to $\frac{25}{42}$. This is similar to our desired result, but we need to correct the number of A points in such a way that the corrected set has the correct distribution.

5.4 Adding and Deleting Points

We would like to prove convergence results about $\Gamma_{n,n}$ working with our current result about Γ'_n. To make this connection, we will add or remove the necessary number of A points (exactly $|N_n - n|$) in a uniformly random way and then show that $|N_n - n|$ is not asymptotically large enough to change the limit. Note that once we condition on N_n to determine the number of points to be added or deleted, the A points will be uniformly distributed on (B_0, B_{n+1}). We then add or delete (as appropriate) $|N_n - n|$ A points in a uniform way. The remaining n points of A are therefore uniformly distributed. Let the random variable Γ_n represent the size of a solution to the class cover problem on this new corrected set of points. Note that $\Gamma_{n,n}$ has the same distribution as Γ_n.

To study the fluctuations of $|N_n - n|$ we construct a random walk on the real line based on the two Poisson processes. The random walk will take one step up at each A point and one step down for each B point. As before, let X_i denote the number of A points in $[B_i, B_{i+1}]$. Note that, as mentioned in the calculation of $E[\rho_1]$, by the lack of memory property of the exponential distribution, $X_i = Z - 1$ where Z has geometric distribution with parameter $\frac{1}{2}$. Let $Y_i = X_i - 1$ and define a random walk G_k by

$$G_k = \sum_{i=0}^{k-1} Y_i \tag{7}$$

for $k \geq 1$. G_k represents the difference between the number of A and B points up to the arrival time of the kth B point. Therefore $N_n - n = G_{n+1} + 1$ (a one is added since the $n+1$st B point is not considered). If $G_{n+1} + 1 < 0$, we must add $G_{n+1} + 1$ points of A, while if $G_{n+1} + 1 > 0$, we must delete $G_{n+1} + 1$ points of A. We will use Chernoff's theorem [Bil95] to obtain exponential probability bounds on the number of points added or removed. For $0 \leq \epsilon \leq 1$,

$$\begin{aligned}P[|G_{n+1} + 1| \geq n\epsilon] &= P[G_n \geq n\epsilon - 1] + P[G_n \leq -n\epsilon - 1] \\ &\leq C_1 e^{-\alpha_1(n\epsilon - 1)} + C_2 e^{-\alpha_2(n\epsilon + 1)}\end{aligned} \tag{8}$$

for all $n \geq 1$, where $\alpha_1, \alpha_2 > 0$ and C_1, C_2 are constants.

5.5 The Effect of Adding or Deleting Points

We also observe that the addition or deletion of one target class point changes the cardinality of the solution to the one dimensional CCP by at most one.

Lemma 2. *Let X, Y be finite subsets of $\Re$ and consider X to be the target class. Let D be the CCCD induced by X, Y and D^- be the CCCD formed from $X - \{x\}, Y$ where x is some element of X. Then $|\gamma(D) - \gamma(D^-)| \leq 1$*

Proof: Let X, Y be finite subsets of $\Re$ with $|X| = n$ and $|Y| = m$. We use the same notation as in Section 4 for α_i and n_i.
Case 1. Suppose $x < Y_{(1)}$ or $x > Y_{(m)}$. Then only α_0 or α_m respectively will be affected by the removal of x. Also since α_0 and α_m must be either zero or 1, then it must be that $|\gamma(D) - \gamma(D^-)| \leq 1$.
Case 2. Suppose $x \in (Y_{(i)}, Y_{(i+1)})$ for $i \in \{1, 2, \ldots, m-1\}$. Again, only α_i will be affected by the removal of x. If $n_i > 1$ then α_i may be either zero, one or two. We must rule out the case that α_i changes from two to zero because of the removal of x. (We don't have to consider the case that α_i switches from zero to two since $x \in (Y_{(i)}, Y_{(i+1)}) \Rightarrow n_i > 0 \Rightarrow \alpha_i > 0$.) If $\alpha_i = 2$ before x is removed then it must be the case that $n_i > 1$ and therefore $n_i \geq 1$ after x is removed. Therefore α_i must be at least one after x is removed. If $n_i = 1$ then $\alpha_i = 1$ and after the removal of x, $\alpha_i = 0$. Therefore $|\gamma(D) - \gamma(D^-)| \leq 1$. □

5.6 Complete Covergence of $\frac{\Gamma_{n,n}}{n}$

If we let $D_n = \Gamma_n - \Gamma'_n$, then Lemma 2 implies that $|D_n| \leq G_{n+1} + 1$. We obtain our result as follows.

$$\begin{aligned}
\sum_{n=1}^{\infty} P\left[\left|\frac{\Gamma_{n,n}}{n} - \frac{25}{42}\right| \geq \epsilon\right] &= \sum_{n=1}^{\infty} P\left[\left|\frac{\Gamma_n}{n} - \frac{25}{42}\right| \geq \epsilon\right] \qquad (9)\\
&= \sum_{n=1}^{\infty} P\left[\left|\frac{\Gamma'_n}{n} - \frac{25}{42} + \frac{D_n}{n}\right| \geq \epsilon\right]\\
&\leq \sum_{n=1}^{\infty} P\left[\left|\frac{\Gamma'_n}{n} - \frac{25}{42}\right| + \left|\frac{G_{n+1}+1}{n}\right| \geq \epsilon\right]\\
&\leq \sum_{n=1}^{\infty} P\left[\left|\frac{\Gamma'_n}{n} - \frac{25}{42}\right| \geq \frac{\epsilon}{2}\right] + \sum_{n=1}^{\infty} P\left[\left|\frac{G_{n+1}+1}{n}\right| \geq \frac{\epsilon}{2}\right]\\
&< \infty
\end{aligned}$$

We have thus shown complete convergence, and therefore almost sure convergence, of $\frac{\Gamma_{n,n}}{n}$ to $\frac{25}{42}$.

6 Extension to $m = an$

The proof is easily extended to the case of unequal numbers of target and non-target points. Let A and B be Poisson processes with rates $a\lambda$ and λ respectively. We view the B process until B_{n+1}, giving n points from B and a random number, $N_n = A(B_{n+1})$, of A points in $(0, B_{n+1})$.

We can use a complete convergence version of the Strong Law of Large Numbers as before to show that the limit of $\frac{\Gamma'_n}{n}$ converges to the mean of ρ_i. To compute this mean, we use Lemma 1 and the fact that the distribution of the number of A points in (B_i, B_{i+1}) has a shifted geometric distribution with parameter $\frac{a}{a+1}$. We calculate the following probabilities.

$$P[\rho_1 = 0] = P[X_i = 0] = \frac{1}{a+1}.$$

$$\begin{aligned} P[\rho_1 = 1] &= \sum_{k=1}^{\infty} P[\rho_1 = 1 | X_1 = k] P[X_1 = k] \qquad (10) \\ &= \sum_{k=1}^{\infty} \left[\frac{5}{9} + \frac{4}{9} 4^{1-k} \right] \frac{a^k}{(a+1)^{k+1}} \\ &= \frac{5a^2 + 12a}{3(a+1)(3a+4)}. \end{aligned}$$

By subtraction,

$$P[\rho_1 = 2] = 1 - P[\rho_1 = 0] - P[\rho_1 = 1] = \frac{12a^2}{3(a+1)(3a+4)},$$

therefore

$$E[\rho_1] = P[\rho_1 = 1] + 2P[\rho_1 = 2] = \frac{a(13a+12)}{3(a+1)(3a+4)}.$$

We must adjust the number of A points to obtain the desired number, $\lfloor an \rfloor$. We again consider a random walk based on two Poisson processes, which takes a step of size one up at each point in A and a step of size a down at each point in B. Using the notation of Section 5.4, let $X_i = Z_i - 1$, where Z_i has a geometric distribution with parameter $\frac{a}{a+1}$, then $Y_i = X_i - a$ are the steps in the random walk. As in the previous case, the Chernoff bounds show that the insertion or deletion of A points does not cause a difference between the asymptotics of Γ_n and Γ'_n.

7 Discussion

Our result provides a strong law of large numbers for a class cover problem in which the data are uniformly distributed on $(0, 1)$. Using the language of [PDM02], the CCP we consider is the constrained heterogeneous CCP. In this CCP, covering balls must be centered at data points and may vary in radius. The proof method uses an associated Poisson process viewpoint to consider independent summands rather than the dependent summands of the CCP.

Our ongoing research on the randomized CCP is focused in two primary directions. One goal is to establish a central limit theorem and the related rates of convergence for the CCP. The second focus is on applications in statistical pattern recognition and machine learning, where there is considerable interest in CCCD's arising from high-dimensional data. We continue to investigate the multi-dimensional setting of the CCP where the problem is significantly more challenging.

References

[Bil95] Billingsly. *Probability and Measure*. Wiley, third edition, 1995.

[CC00] A. Cannon and L. Cowen. Approximations algorithms for the class cover problem. In *6th International Symposium on Artificial Intelligence and Mathematics, 2000*, 2000.

[CEHS02] A. Cannon, J.M. Ettinger, D. Hush, and C. Scovel. Machine learning with data dependent hypothesis classes. *Journal of Machine Learning Research*, 2:335–358, 2002.

[CF94] F. Critchley and B. Fichet. *Lecture Notes in Statistics: Classification and Dissimilarity Analysis*, volume 93, chapter 2. Springer-Verlag, 1994.

[Dav81] H. David. *Order Statistics*. Wiley, 1981.

[DP02] J. DeVinney and C. Priebe. Class cover catch digraphs. 2002. Submitted for Publication. Available as Technical Report No. 633, Department of Mathematical Sciences, Johns Hopkins University, Baltimore, MD 21218-2682.

[HHS98] T. Haynes, S. Hedetniemi, and P Slater. *Fundamentals of Domination in Graphs*. Marcel Dekker, Inc., 1998.

[KLV98] S. Kulkarni, G. Lugosi, and S. Venkatesh. Learning pattern classification - a survey. *IEEE Transactions on Information Theory*, 44(6):2178–2206, October 1998.

[MM99] T McKee and F. McMorris. *Topics in Intersection Graph Theory*. SIAM, 1999.

[PDM02] C. Priebe, J. DeVinney, and D. Marchette. On the distribution of the domination number for random class cover catch digraphs. *Stat. Probab. Lett.*, (55), 2002.

[PMDS02] C. Priebe, D. Marchette, J. DeVinney, and D. Socolinsky. Classification using class cover catch digraphs. 2002. Submitted for publication. Available as Technical Report No. 628, Department of Mathematical Sciences, Johns Hopkins University, Baltimore, MD 21218-2682.