Dissertation Defense

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

Pengfei Xiang

`xiang@jhu.edu`

Department of Applied Mathematics and Statistics

The Johns Hopkins University

Background: Pattern Classification

- Abstract mathematical model:
  - ◇ (Ω, X, Y).
  - ◇ Random data: $\left(c(\Psi), \Psi\right)$ with the class label part $c(\Psi) \in \{X, Y\}$ and the data part $\Psi \in \Omega$.
  - ◇ Prior probabilities: P_X, P_Y.
    Class-conditional distribution functions: F_X, F_Y.
- Classifier:
  - ◇ For an observation $\left(c(\psi), \psi\right)$, given the data part ψ, guess the unknown class label part $c(\psi)$.

Class Cover Problem

Consider two sequences of i.i.d. random variables:

$$X_i \sim F_X, i = 1, \cdots, n,$$

$$Y_j \sim F_Y, j = 1, \cdots, m.$$

- Covering ball: For X_i, define its covering ball as $$B(X_i) \equiv \left\{\omega \in \Omega : d(X_i, \omega) < \min_{j \in \{1, \cdots, m\}} d(X_i, Y_j)\right\}.$$
- Class cover: A subset of covering balls whose union contains all X_i's.
- Class cover problem: Find a minimum cardinality class cover.

Class Cover Catch Digraph

- Definition: The 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 $(X_i, X_j) \in A$ iff $X_j \in B(X_i)$.
- Dominating set: The set $S \subset V$ is a dominating set of a digraph $D = (V, A)$ iff for all $v \in V$, either $v \in S$, or $(s, v) \in A$ for some $s \in S$.
- The CCP is equivalent to finding a minimum dominating set of the induced CCCD.
- CCCD and CCP in high dimensions are NP-Hard.

Construction of a CCCD

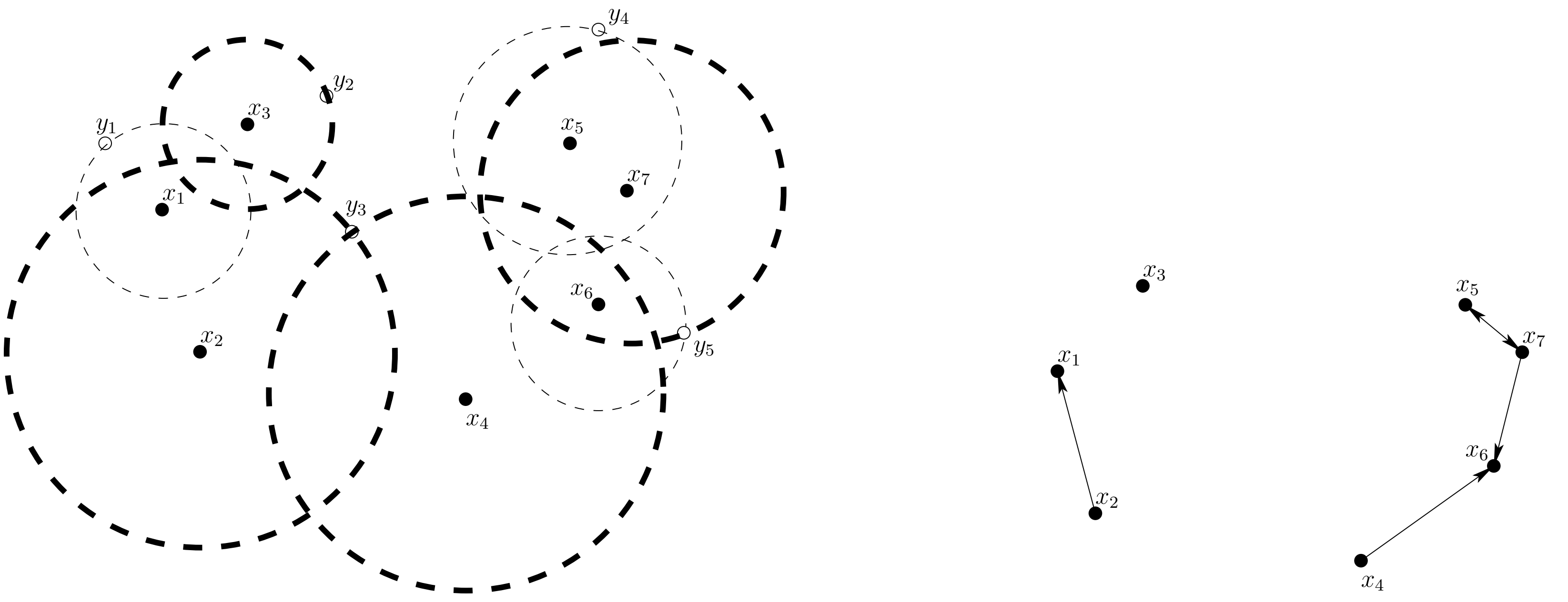

Figure 1: An illustration of the construction of a CCCD

Domination Number

- Definition: The domination number of a CCCD is the cardinality of the CCCD's minimum dominating set.
- Notation: letting $\mathcal{X} \equiv \{X_1, \cdots, X_n\}$ and $\mathcal{Y} \equiv \{Y_1, \cdots, Y_m\}$, we denote the domination number by $\Gamma_{n,m}(\mathcal{X}, \mathcal{Y})$, or simply by $\Gamma_{n,m}$.
- Research direction: The probabilistic limiting behavior of $\Gamma_{n,m}$.

Previous Results (1)

For the special case of $\Omega = \mathbf{R}$ and $F_X = F_Y = U[0,1]$,

- Denote $Y_{(j)}$ as the jth order statistic of $Y_1, \cdots, Y_m$, and define $Y_{(0)} \equiv 0, Y_{(m+1)} \equiv 1$.
- Let random variable $N_{j,m}$ be the number of X-points between $Y_{(j)}$ and $Y_{(j+1)}$, and $\alpha_{j,m}$ be the minimum number of covering balls needed to cover these $N_{j,m}$ X-points.
- $\Gamma_{n,m} = \sum_{j=0}^{m} \alpha_{j,m}$.

Previous Results (2)

Under the above assumptions, Priebe, Devinney and Marchette find the conditional distribution of $\alpha_{j,m}$ given $N_{j,m}$. Furthermore, Devinney and Wierman prove the following strong law of large numbers (SLLN) for $\Gamma_{n,m}$:

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

$$\lim_{n \to +\infty} \frac{\Gamma_{n,m}}{n} = g(r) \equiv \frac{r(12r + 13)}{3(r+1)(4r+3)} \quad a.s.$$

SLLN in One Dimension with General Densities

In this dissertation, we have proved the SLLN in one dimension for the more general case:

Theorem 2. *If* $\Omega = \mathbf{R}$, f_X *and* f_Y *are bounded and continuous density functions, and* $m/n \to r$, $r \in (0, \infty)$, *then*

$$\lim_{n\to\infty} \frac{\Gamma_{n,m}}{n} = \int g\left(r \cdot \frac{f_Y(u)}{f_X(u)}\right) \cdot f_X(u) du \qquad a.s.$$

where $g(r) \equiv \frac{r(12r+13)}{3(r+1)(4r+3)}$ *(same as in the SLLN for uniform densities).*

Proof of the SLLN(1)

Proof sketch:

- Extend the result for uniform density functions to piece-wise constant densities.
- Construct piece-wise constant approximation to the bounded continuous function case.

Proof of the SLLN(2)

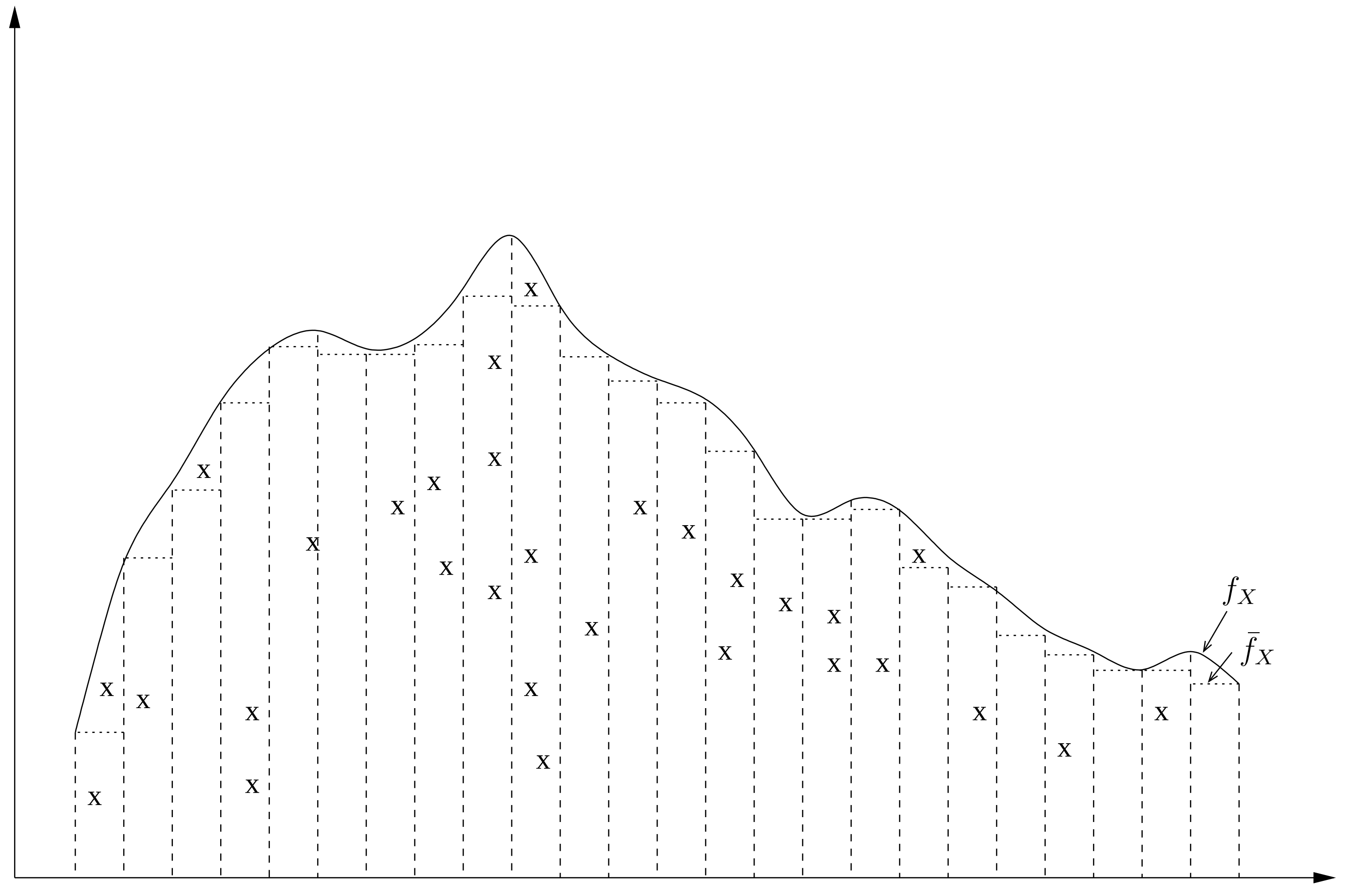

Figure 2: Illustration of the proof of the SLLN

One Corollary of the SLLN

Corollary 1. *Under the same conditions as in the SLLN, we have*

$$\int g\left(r \cdot \frac{f_Y(u)}{f_X(u)}\right) \cdot f_X(u)du \leq g(r)$$

with equality holding iff $f_X = f_Y \quad a.e.$

Applications Build some statistical test for equality of the distributions.

Variance of the Domination Number in One Dimension

Since $\Gamma_{n,m} = \sum \alpha_{j,m}$, we only need to calculate the variances and covariances of the components:

Theorem 3. *If* $\Omega = \mathbf{R}$, $F_X = F_Y = U[0,1]$ *and* $m/n \to r$, $r \in (0,\infty)$, *then*

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

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

Hence,

$$\frac{Var(\Gamma_{n,m})}{m} \to v(r) \equiv \frac{1536r^5+6848r^4+11536r^3+8836r^2+2793r+180}{9(r+1)^3(4r+3)^4}.$$

Calculation of the Variance

The calculation is very technical (taking about 40 pages in the dissertation). It's essentially done in two steps:

- first, we get the conditional expectations $E(\alpha_{j,m}^k \mid N_{j,m}), k = 1, 2,$ using the conditional probability of $\alpha_{j,m}$ given $N_{j,m}$;
- then we compute $E(\alpha_{j,m}^k), k = 1, 2,$ using $N_{j,m}$'s distribution. Note that given $L_{j,m} = l_{j,m}, j = 0, \cdots, m,$ the random vector $\{N_{j,m} : j = 0, \cdots, m\}$ is multinomially distributed with parameters $\{n, l_{j,m} : j = 0, \cdots, m\}$, where the distribution of $L_{j,m}$ can be easily calculated.

Verification of the Limiting Variance Formula

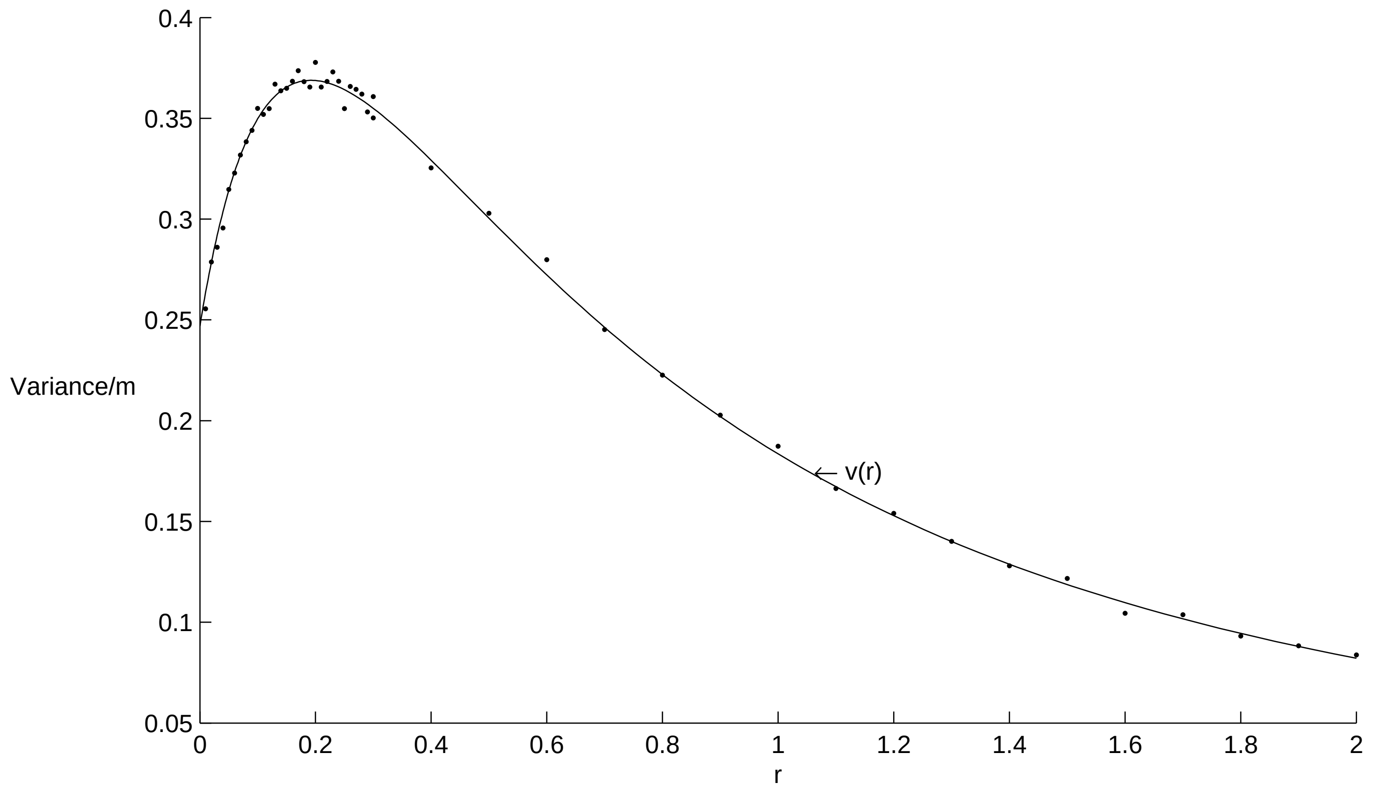

Figure 3: Verification of $\lim_{n\to\infty} \frac{Var(\Gamma_{n,m})}{m} = v(r)$

Central Limit Theorem (CLT) in One Dimension

Theorem 4. *If* $\Omega = \mathbf{R}, F_X = F_Y = U[0,1]$*, and* $m/n \to r$*,* $r \in (0,\infty)$*, then*

$$\frac{1}{m^{1/2}}\big(\Gamma_{n,m} - E[\Gamma_{n,m}]\big) \xrightarrow{\mathcal{L}} N(0,\sigma^2)$$

where $\sigma^2 = \lim_{m\to\infty} \frac{Var[\Gamma n,m]}{m}$*.*

Proof of the CLT (1)

- Issue: Recall $\Gamma_{n,m} = \sum_{j=0}^{m} \alpha_{j,m}$. Note that $\alpha_{j,m}$ solely depends on $N_{j,m}$, but $N_{j,m}$'s are dependent on each other. In fact, $N_{j,m}$'s are *negatively associated*.
- Solution: Project $\Gamma_{n,m}$ onto a conditional probability space where all the components $\alpha_{j,m}$'s become independent of each other, then apply the SLLN and CLT for negatively associated random variables.

next slide $\longrightarrow$

Proof of the CLT (2)

Define $\mathcal{F}_m$ as the σ-field generated by $N_{j,m}, j = 0, \cdots, m$. Let $Z_{j,m} = \frac{1}{m^{1/2}}\big(\alpha_{j,m} - E[\alpha_{j,m}]\big)$. Then define the conditional characteristic function $f_m(t)$ as follows:

$$\begin{aligned} f_m(t) &\equiv E\left[e^{it\sum_{j=0}^{m} Z_{j,m}} \mid \mathcal{F}_m\right] \\ &= \prod_{j=0}^{m} E\left[e^{itZ_{j,m}} \mid \mathcal{F}_m\right], \end{aligned}$$

where the last step holds because $Z_{j,m}$'s are independent given $\mathcal{F}_m$.

Proof of the CLT (3)

Applying the Taylor expansion yields

$$f_m(t) \approx \prod_{j=0}^{m} \Big(1 + itE[Z_{j,m} \mid N_{j,m}] - \tfrac{t^2}{2} E[Z_{j,m}^2 \mid N_{j,m}]\Big),$$

hence

$$log\big(f_m(t)\big) \approx it \sum_{j=0}^{m} E[Z_{j,m} \mid N_{j,m}] - \frac{t^2}{2} \sum_{j=0}^{m} Var[Z_{j,m} \mid N_{j,m}],$$

thus

$$\begin{aligned} E\left[e^{it\sum_{j=0}^{m} Z_{j,m}}\right] &= E\big[f_m(t)\big] \\ &\approx E\left[e^{it\sum_{j=0}^{m} E[Z_{j,m}|N_{j,m}]}\right] \cdot E\left[e^{-\frac{t^2}{2}\sum_{j=0}^{m} Var[Z_{j,m}|N_{j,m}]}\right] \\ &\rightarrow e^{-\frac{t^2\sigma_1^2}{2}} \cdot e^{-\frac{t^2\sigma_2^2}{2}} = e^{-\frac{t^2\sigma^2}{2}}. \end{aligned}$$

Weak Law of Large Numbers (WLLN) in 2 Dimensions

The CCCD problem becomes much more challenging in higher dimensions. Applying the SLLN for subadditive processes, we have proved the following WLLN in 2 dimensions.

Theorem 5. *If the densities f_X and f_Y are positive, bounded and continuous on $[0,1]^2$, and $m/n \to r, r \in (0,\infty)$, then*

$$\lim_{n\to\infty} \frac{\Gamma'_{n,m}}{n} = \iint\limits_{[0,1]^2} g\left(r \cdot \frac{f_Y(u,v)}{f_X(u,v)}\right) \cdot f_X(u,v)\,du\,dv \quad \textit{in probability}.$$

$\boxed{\leftarrow \textit{previous slide}}$

Proof Sketch of the WLLN in 2 Dimensions

The proof is done in three steps:

1. apply the SLLN for subadditive processes to prove the SLLN for the domination number in the Poisson case;
2. use the result in the Poisson case to prove the WLLN for the domination number in $[0, 1]^2$ with uniform densities;
3. extend the result above to the case with general densities.

Definition of 2-dimensional Subadditive Processes

Let $\{X_{s,t} : 0 \leq s < t, s, t \in \mathbf{R}^2\}$ be a collection of random variables. Then $\{X_{s,t}\}$ is called a *2-dimensional subadditive process* if it satisfies

- Subadditivity:
  For disjoint squares $I_i = \{u : a_i \leq u < b_i, a_i, b_i \in \mathbf{R}^2\}$, if $I = \cup_{i=1}^n I_i$ is also a square, then $X_I \leq \sum_{i=1}^n X_{I_i}$.
- Stationarity:
  The joint distributions of $\{X_{I_1+u}, \cdots, X_{I_n+u}\}$ is the same as that of $\{X_{I_1}, \cdots, X_{I_n}\}$, where $u \in \mathbf{R}^2$.
- Expectation Condition:
  $\gamma(X) \equiv \inf_I \left\{\frac{E[X_I]}{|I|} : I = [a_i, b_i), a_i, b_i \in \mathbf{R}^2\right\} > -\infty$.

Illustration of 2-dimensional Subadditive Processes

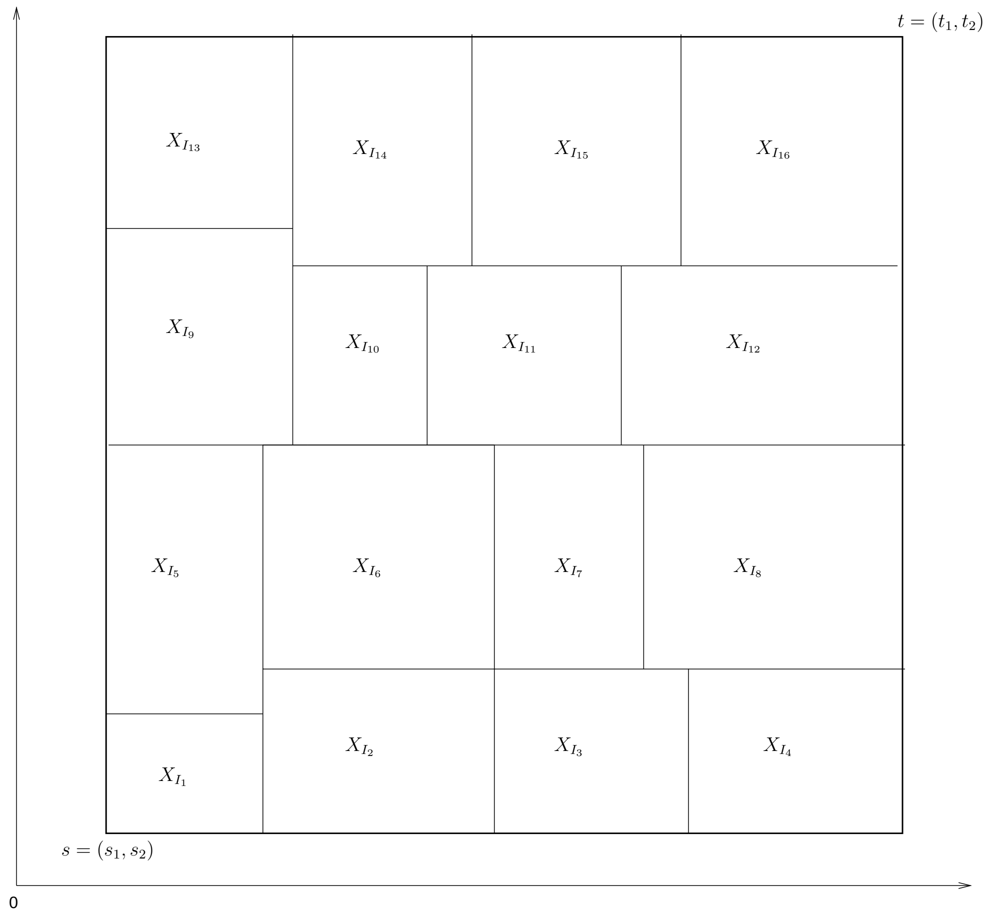

Figure 4: Subadditivity: $X_{\cup_{i=1}^{n} I_i} \leq \sum_{i=1}^{n} X_{I_i}$

SLLN for Multidimensional Subadditive Processes

The above definition can be easily generalized to the multidimensional case. Akcoglu and Krengel proved that

Theorem 6. *If* $\{X_{s,t}\}$ *is a multidimensional subadditive process, then*

$$\lim_{n\to\infty} \frac{X_{J_n}}{|J_n|} = \zeta \quad a.s.$$

and $E[\zeta] = \gamma(X)$*, where* $J_n = [\vec{0}, n\vec{e})$ *with* $\vec{0} = (0, \cdots, 0)$ *and* $\vec{e} = (1, \cdots, 1)$*.*

Note: if $\{X_{s,t}\}$ is independent, then $\zeta = \gamma(X)\ a.s.$

Subadditivity of the Domination Number in 2 dimensions (1)

Suppose X and Y are Poisson process points in $\mathbb{R}^2$. Let Γ_I denote the domination number generated by these X and Y points in any rectangles $I \subset \mathbb{R}^2$.

- Issue: $\{\Gamma_I\}$ is *not* a subadditive process.

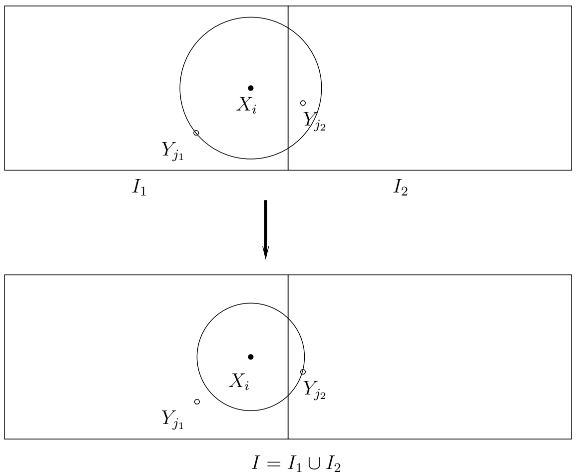

Figure 5: Non-subadditivity of $\{\Gamma_I\}$

Subadditivity of the Domination Number in 2 dimensions (2)

- Idea: Find a subadditive process that approximates $\{\Gamma_I\}$.
- Solution: Restrain the covering balls in I, and refer to corresponding domination number as *constrained domination number*, denoted by $\bar{\Gamma}_I$. Then $\{\bar{\Gamma}_I\}$ is subadditive.

SLLN in the Poisson Case

Since $\{\bar{\Gamma}_I\}$ is a multidimensional subadditive process, we have

$$\lim_{n\to\infty} \frac{\bar{\Gamma}_{J_n}}{|J_n|} = \zeta \quad a.s. \quad \text{with } E[\zeta] = \gamma(\Gamma).$$

Then we generalize this result to the SLLN for the original domination number Γ_{J_n}.

WLLN in $[0,1]^2$ *with Uniform Densities*

Next, we transfer the result in the Poisson case to $[0,1]^2$.

- Conditioning on the $(n+1)$th arrival of X-points, suppose there are n X-points and m_n Y-points uniformly distributed in $J_{t(n)}$.
- But we need m Y-points for the desired result in $[0,1]^2$.
- So we uniformly add $m - m_n$ or delete $m_n - m$ Y-points.
- We argue that the effect of adding or deleting $|m - m(n)|$ Y points is negligible, so the WLLN holds in $[0,1]^2$ with uniform densities.

WLLN in $[0,1]^2$ *with General Densities*

- We basically follow the same idea as in one dimension to extend the WLLN with uniform densities to general densities.
- But the detailed proof is much more complicated, since adding or deleting a X or Y point no longer only changes the domination number by at most 2 as in one dimension.

Simulation

We have used Monte Carlo simulations to check the limit theorems obtained in this dissertation, and also empirically verified some limit theorems that are not proved but are likely to be true, such as the CLT in two dimensions.

Future Research Directions

- CLT in 2 or higher dimensions.
- Other properties of CCCDs, such as the edge density.