SPATIALLY-CORRELATED SENSOR DISCRIMINANT ANALYSIS

Kush R. Varshney

Business Analytics and Mathematical Sciences Department, IBM Thomas J. Watson Research Center
1101 Kitchawan Rd., Route 134, Yorktown Heights, NY 10598, USA

ABSTRACT

A study of generalization error in signal detection by multiple spatially-distributed and -correlated sensors is provided when the detection rule is learned from a finite number of training samples via the classical linear discriminant analysis formulation. Spatial correlation among sensors is modeled by a Gauss–Markov random field defined on a nearest neighbor graph according to inter-sensor spatial distance, where sensors are placed randomly on a growing bounded region of the plane. A fairly simple approximate expression for generalization error is derived involving few parameters. It is shown that generalization error is minimized not when there are an infinite number of sensors, but a number of sensors equal to half the number of samples in the training set. The minimum generalization error is related to a single parameter of the sensor spatial location distribution, derived based on weak laws of large numbers in geometric probability. The finite number of training samples acts like a budgeting variable, similar to a total communication power constraint.

Index Terms— signal detection, distributed sensors, linear discriminant analysis, generalization error, geometric probability

1. INTRODUCTION

We have embarked on an age in which inexpensive sensors are everywhere and in everything, producing data that enables a smarter planet. Signal detection is one important use of the measurements produced by collections of spatially-distributed sensors. Sensor measurements in natural settings tend to exhibit correlation [1, 2], but often, the design and analysis of detection rules has focused on the case of statistically independent measurements (conditioned on the hypothesis). Much of the focus has also been on the case when the likelihood functions of the measurements are known *a priori*, but this is seldom true in practice. In this paper, the focus is on the *supervised learning* of detection rules for spatially-correlated sensor measurements, in particular via linear discriminant analysis.

Common wisdom dictates that when monitoring a field, the more sensors, the better the detection performance; and it is only because of material costs that one limits the number of sensors. However, this work considers supervised binary classification with a finite training set and shows that increasing the number of sensors beyond a certain amount results in the degradation of detection performance, independent of any power, communication, or network considerations. This behavior is the manifestation of the fundamental phenomenon of overfitting. In some sense, power or communication budgets are replaced by a budget on the cardinality of the labeled training set in the learning setting.

The analysis in [3] of detection using spatially-correlated sensor measurements corrupted by noise during communication to a fusion center shows that for stochastic signals subject to transmission power constraints, a finite rather than infinite sensor density in space is optimal. Although the analysis herein is for a constant sensor density, the conclusion is similar in the sense that with degraded and constrained information (here due to a finite training set), it is better to use a small-dimensional measurement vector rather than an infinite-dimensional one.

While the majority of prior work on detection with multiple sensors deals with known likelihood functions, there has also been work on the supervised learning of detection rules, such as [4, 5]. These works address questions of what can be done when communication-limited sensors have a labeled training set. Here an even more fundamental question is addressed: even without limitation on communication, what is the generalization behavior of detection rules learned from spatially-correlated sensor measurements.

A method for supervised classification with dimensionality reduction and information fusion in tree-structured sensor networks is developed in [6]. In that work, it is also apparent that for a fixed number of training samples, adding more sensors or dimensions of measurements beyond a certain point results in a degradation of detection performance.

In the remainder of the paper, the detection rule learning paradigm considered is the classic plug-in formulation of linear discriminant analysis. Generalization error is analyzed based on expressions given in [7]. The spatially-correlated sensor system is modeled as a Gauss–Markov random field with nearest neighbor dependency among randomly placed sensors, quite similar to the model employed in [8]. Like in [8], the asymptotic analysis of [9] is employed for simplification purposes.

2. SYSTEM MODEL

Consider the system model with p sensors randomly deployed on the plane. The location of sensor i, denoted $\mathsf{v}_i \in \mathbb{R}^2$, is drawn according to the distribution $f_{\mathsf{v}}(\mathbf{v})$ which is supported on a square with area p. Each sensor measures a scalar random variable x_i, $i = 1, \dots, p$. The overall measurement $\mathsf{x} \in \mathbb{R}^p$ is related to two hypotheses $y \in \{0, 1\}$ by the Gaussian likelihood functions $f_{\mathsf{x}|y}(\mathbf{x}|y = 0) \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $f_{\mathsf{x}|y}(\mathbf{x}|y = 1) \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$. The particulars of the covariance matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ come from the spatial locations of the sensors; correlation decays as a function of distance between sensors. The task is to determine the hypothesis based on the measurement vector.

The probability density $f_{\mathsf{x},y}(\mathbf{x}, y)$ governing the sensor measurements for a particular sensor location realization is not given to the system *a priori*. Only a set of n i.i.d. training samples $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ drawn from the distribution are given. If the distribution were given, then the likelihood ratio test detection rule based on it would minimize error. However, since it is not given, a detection rule must be learned from the training set.

The detection rule studied here is the linear discriminant analysis

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

rule, a simple, classical, often-used technique. The detection rule is:

$$\hat{y}(\mathbf{x}) = \mathrm{step}(\boldsymbol{\theta}^T \mathbf{x} + \theta_0), \tag{1}$$

where $\boldsymbol{\theta} = \left(\hat{\boldsymbol{\Sigma}}_0 + \hat{\boldsymbol{\Sigma}}_1\right)^{-1} (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$, $\theta_0 = -\frac{1}{2}\boldsymbol{\theta}^T (\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1)$, and $\hat{\boldsymbol{\mu}}_0$, $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\Sigma}}_0$, and $\hat{\boldsymbol{\Sigma}}_1$ are the conditional sample means and covariances of the n training samples. Once the detection rule $\hat{y}(\cdot) : \mathbb{R}^p \to \{0, 1\}$ is learned, it is applied to new unseen and unlabeled measurements $\mathbf{x}$.

Three assumptions on $f_{\mathsf{x},y}(\mathbf{x}, y)$ are made for simplicity of exposition. Let the prior probabilities of the hypotheses be equal, i.e. $\Pr[y = 0] = \Pr[y = 1] = 1/2$. Let $\boldsymbol{\mu}_0 = \mathbf{0}$ (vector of all zeroes) and let $\boldsymbol{\mu}_1 = \mathbf{1}$ (vector of all ones). Let $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$; note that the assumptions are not known by the deployed system. The covariance structure is based on the Euclidean nearest neighbor graph of the sensors. The (undirected) nearest neighbor graph contains an edge between sensor i and sensor j if sensor i is the nearest neighbor of sensor j *or* if sensor j is the nearest neighbor of sensor i. The set of edges in the nearest neighbor graph is denoted $\mathcal{E}$.

It is most convenient to specify the p^2 entries of the covariance matrix in three parts. First, the diagonal elements of $\boldsymbol{\Sigma}_1$ are all equal to the constant σ^2. Second, the elements of $\boldsymbol{\Sigma}_1$ corresponding to edges in the nearest neighbor graph are:

$$\{\boldsymbol{\Sigma}_1\}_{ij} = \sigma^2 g(d(\mathbf{v}_i, \mathbf{v}_j)), \quad (i, j) \in \mathcal{E}, \tag{2}$$

where $g(\cdot) : \mathbb{R}^+ \to (0, 1)$ is a monotonically decreasing function that encodes the correlation decay with distance. The inverse covariance or information matrix, denoted $\mathbf{J}_1 = \boldsymbol{\Sigma}_1^{-1}$, is used to specify the remaining elements. The off-diagonal elements of $\mathbf{J}_1$ corresponding to sensor pairs (i, j) that do not have an edge in the nearest neighbor graph are zero, i.e.

$$\{\mathbf{J}_1\}_{ij} = 0, \quad i \neq j,\ (i, j) \notin \mathcal{E}. \tag{3}$$

3. GENERALIZATION ERROR

As the linear discriminant analysis detection rule is learned from training samples but applied to new test samples, the performance metric of interest is the generalization error $\Pr[\hat{y}(\mathbf{x}) \neq y]$, which is always greater than or equal to the Bayes optimal detection error achieved by the optimal likelihood ratio test detection rule. Generalization error for the sensor system model of Sec. 2 is first studied for a given realization of sensor locations and then as an average across realizations.

3.1. Linear Discriminant Analysis Error Approximation

Despite extensive study by many researchers, an exact closed form expression for the generalization error of linear discriminant analysis has not yet been found, but several highly accurate approximations exist [7]. One of the best ones when the true likelihood functions are Gaussian and the true prior probabilities are equal, applicable for a wide range of p and n values, is the following [7]:

$$\Pr[\hat{y}(\mathbf{x}) \neq y] \approx \Phi\left(-\frac{\delta}{2}\left[\left(1 + \frac{4p}{n\delta^2}\right)\frac{n}{n-p}\right]^{-1/2}\right), \tag{4}$$

where $\Phi(\cdot)$ is the Gaussian cumulative distribution function and

$$\delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \left(\frac{\mathbf{J}_0 + \mathbf{J}_1}{2}\right)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \tag{5}$$

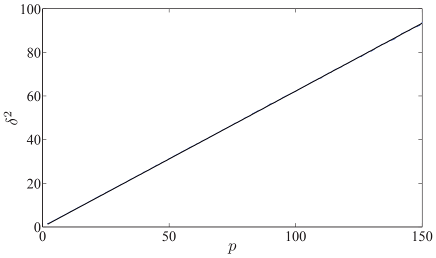

Fig. 1. Exact (blue line with marker dots) and approximate (black line) squared Mahalanobis distance as a function of the number of sensors. The two lines are nearly indistinguishable.

is a squared Mahalanobis distance.

With the assumptions $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{\mu}_1 = \mathbf{1}$, and $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$, the squared Mahalanobis distance simplifies to:

$$\delta^2 = \sum_{i=1}^{p}\sum_{j=1}^{p}\{\mathbf{J}_1\}_{ij}. \tag{6}$$

Off-diagonal entries of the information matrix with $(i, j) \in \mathcal{E}$ take the value [8]:

$$\{\mathbf{J}_1\}_{ij} = \frac{1}{\sigma^2} \cdot \frac{-g(d(\mathbf{v}_i, \mathbf{v}_j))}{1 - g(d(\mathbf{v}_i, \mathbf{v}_j))^2}. \tag{7}$$

Therefore the sum of the off-diagonal elements is:

$$\sum_{i=1}^{p}\sum_{j\neq i}\{\mathbf{J}_1\}_{ij} = \frac{2}{\sigma^2}\sum_{(i,j)\in\mathcal{E}} \frac{-g(d(\mathbf{v}_i, \mathbf{v}_j))}{1 - g(d(\mathbf{v}_i, \mathbf{v}_j))^2}. \tag{8}$$

The diagonal entries of the information matrix take the value [8]:

$$\{\mathbf{J}_1\}_{ii} = \frac{1}{\sigma^2}\left(1 + \sum_{\{j|(i,j)\in\mathcal{E}\}} \frac{g(d(\mathbf{v}_i, \mathbf{v}_j))^2}{1 - g(d(\mathbf{v}_i, \mathbf{v}_j))^2}\right). \tag{9}$$

Consequently, the sum of the diagonal elements is:

$$\sum_{i=1}^{p}\{\mathbf{J}_1\}_{ii} = \frac{p}{\sigma^2} + \frac{2}{\sigma^2}\sum_{(i,j)\in\mathcal{E}} \frac{g(d(\mathbf{v}_i, \mathbf{v}_j))^2}{1 - g(d(\mathbf{v}_i, \mathbf{v}_j))^2}. \tag{10}$$

Combining (8) and (10),

$$\delta^2 = \frac{p}{\sigma^2} - \frac{2}{\sigma^2}\sum_{(i,j)\in\mathcal{E}} \frac{g(d(\mathbf{v}_i, \mathbf{v}_j))}{1 + g(d(\mathbf{v}_i, \mathbf{v}_j))}. \tag{11}$$

3.2. Mahalanobis Distance Approximation

It may be noticed that in (11), the Mahalanobis distance depends on the Euclidean distances $d(\mathbf{v}_i, \mathbf{v}_j)$ which in turn depend on the particular realization of the random deployment of sensor locations. For analysis purposes, it is useful to characterize the average behavior of δ across realizations of $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$. Formulas obtained in [9] are used in developing this characterization.

Average behavior of functionals of the nearest neighbor graph can be described using average behavior of homogenous Poisson

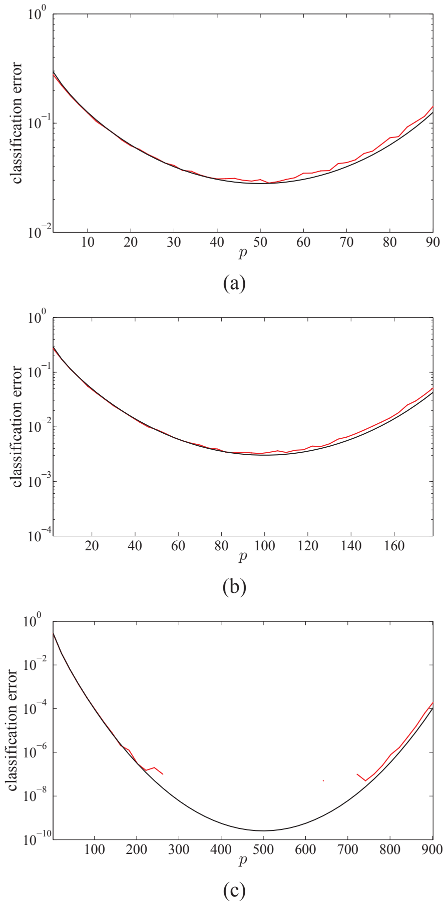

Fig. 2. Test error (red line with marker dots) and generalization error approximation (black line) as a function of the number of sensors for (a) $n = 100$, (b) $n = 200$, and (c) $n = 1000$.

point processes [9]. One specific result is that as p grows,

$$\frac{1}{p}\sum_{(i,j)\in\mathcal{E}}\phi(d(\mathbf{v}_i,\mathbf{v}_j)) \to \frac{1}{2}\int \mathrm{E}\left[\sum_{(0,a)\in\mathcal{F}}\phi\left(\frac{d(\mathbf{0},\mathbf{w}_a)}{\sqrt{f_{\mathbf{v}}(\mathbf{v})}}\right)\right] f_{\mathbf{v}}(\mathbf{v})d\mathbf{v}, \qquad (12)$$

where $\phi(\cdot)$ is an arbitrary function, $\mathbf{w}_k$ are spatial locations drawn according to the Poisson point process with unit rate over a unit square centered at the origin, and $\mathcal{F}$ is the set of edges of the nearest neighbor graph constructed from the origin point $\mathbf{0}$ and those points $\mathbf{w}_k$.

Consequently,

$$\delta^2 \approx \frac{p}{\sigma^2}\left(1-\zeta\right), \qquad (13)$$

where

$$\zeta = \int \mathrm{E}\left[\sum_{(0,a)\in\mathcal{F}}\phi\left(\frac{d(\mathbf{0},\mathbf{w}_a)}{\sqrt{f_{\mathbf{v}}(\mathbf{v})}}\right)\right] f_{\mathbf{v}}(\mathbf{v})d\mathbf{v}, \qquad (14)$$

and $\phi(\cdot) = g(\cdot)/(1+g(\cdot))$. It may be noted that the squared Mahalanobis distance is approximately a linear function of the number of sensors p. The approximation is in fact quite tight. Sec. 4 provides plots of quantities derived in this section as a function of p for a few examples, illustrating the quality of the approximations and illustrating the fact that more sensors is not necessarily better from an error rate perspective.

3.3. Optimal Number of Sensors

The optimal value of the number of sensors p may be found based on the Mahalanobis distance approximation (13), and the generalization error approximation (4). Combining approximations, an overall approximation for the generalization error is:

$$\Pr[\hat{y}(\mathbf{x}) \neq y] \approx \Phi\left(-\sqrt{\frac{p(1-\zeta)}{4\sigma^2}}\left[\left(1+\frac{4\sigma^2}{n(1-\zeta)}\right)\frac{n}{n-p}\right]^{-1/2}\right). \qquad (15)$$

The number of sensors that minimizes this generalization error approximation is $p^* = n/2$, irrespective of σ^2 and ζ. This optimal value is found by differentiating (15) with respect to p, setting it equal to zero, and solving for p. Of course, this achieved minimum approximate generalization error does depend on σ^2 and ζ:

$$\Pr[\hat{y}(\mathbf{x}) \neq y]^* = \Phi\left(-\frac{1}{4}\sqrt{\frac{n^2(1-\zeta)^2}{n\sigma^2(1-\zeta)+4\sigma^4}}\right). \qquad (16)$$

The expression (16) unsurprisingly reveals that the minimum generalization error is a monotonically increasing function of σ^2 bounded in the range zero to one half, and a monotonically decreasing function of n, the number of training samples. Furthermore, the minimum approximate generalization error is a monotonically increasing function of $\zeta \in [0, 1/2]$, implying that the sensor placement distribution $f_{\mathbf{v}}(\mathbf{v})$ should be chosen to minimize ζ in order to achieve the best system performance.

4. SIMULATIONS

Simulation examples are presented in this section showing test error, as well as values of the approximate Mahalanobis distance and approximate generalization error derived in Sec. 3. The particular correlation decay function, also known as the semivariogram, that is considered is $g(d) = \frac{1}{2}\exp(-\frac{d}{2})$; such exponential models of correlation decay in spatial signals often appear in geostatistics.

The sensor location distribution $f_{\mathbf{v}}(\mathbf{v})$ with support over the square with area p that is considered is an appropriately scaled and shifted version of the beta distribution independent and identically distributed in both components of $\mathbf{v}$. Both parameters of the beta distribution are taken to be equal to β. When $\beta = 1$, the sensors are placed uniformly over the square; they are concentrated in the middle of the square for $\beta > 1$ and concentrated at the edges of the square for $\beta < 1$.

4.1. Mahalanobis Distance Approximation

The squared Mahalanobis distance approximation (13) is first compared to the true value (11). The exact squared Mahalanobis distance with $\sigma^2 = 1$ is calculated for 100 realizations of $\mathbf{v}$ with the uniform distribution, i.e. $\beta = 1$, for different numbers of sensors. In Fig. 1, the blue curve with marker dots is the average of the realizations as

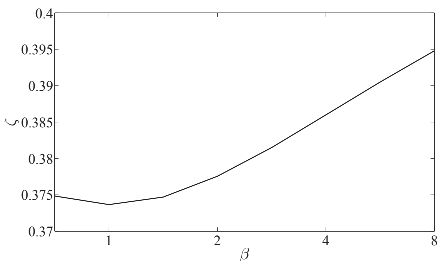

Fig. 3. Poisson point process expectation ζ as a function of the beta distribution parameter β for sensor spatial location distribution.

a function of p. The figure also includes a plot of the approximation to squared Mahalanobis distance using Poisson point processes as a black line. Even for small p, the approximation is so good that the blue curve and the black line are nearly indistinguishable.

Whereas finding the exact squared Mahalanobis distance involves constructing the $p \times p$ information matrix for each realization of $\mathbf{v}$ and each p, the approximation only involves finding ζ once, and it is applicable for all p, even as p approaches infinity.

4.2. Test Error and Generalization Error Approximation

Having empirically shown the high quality of the Mahalanobis distance approximation for all p (which converges in the limit as p goes to infinity), the generalization error approximation (15) for spatially-correlated sensors is now examined. Specifically examined is the test error of linear discriminant analysis as a function of p averaged over twenty realizations of $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$, ten realizations of the training set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ per sensor location realization, and 10^5 test samples per training set. This average test error is compared to the generalization error approximation, including the Mahalanobis distance approximation.

The red line with marker dots in Fig. 2(a) is the test error for $n = 100$; the black line shows the generalization error approximation. The first thing to notice is that as expected, error is minimized when $p = 50 = n/2$ both for the test error and the approximate generalization error expression. Detection performance suffers if too many sensors are thrown down. The second thing to notice is that the approximation is quite good for all p. Therefore, the approximate generalization error expression, which is simple to compute and has a nice analytic form, may be used in further analysis of the sensor system. Fig. 2(b) and Fig. 2(c) show error with larger numbers of samples in the training set: $n = 200$ and $n = 1000$. The excellent agreement gets even better for larger n. The $n = 1000$ case represents an instance in which it is computationally intractable to obtain the tiny error probabilities of the detection rule in a reasonable amount of time through test samples; the approximation is valid in this regime and may be used readily.

4.3. Different Sensor Placement Distributions

The simulations thus far have focused on the uniform distribution for $\mathbf{v}$. Now consider different distributions and consequently different values of ζ. Fig. 3 shows the value of the Poisson point process expectation ζ as a function of β. The uniform distribution with $\beta = 1$ minimizes this expectation. As discussed in Sec. 3.3, small ζ implies small generalization error. Therefore, among this family of distributions, the uniform distribution optimizes detection performance. The overall guideline is then that when using linear discriminant analysis detection given a budget n on the number of training samples, $n/2$ sensors placed uniformly should be used.

5. CONCLUSION

It is a fundamental truth that if nothing else, time is a limited resource, and limits a system to finite sets of training samples. It is shown in this work that when learning a linear discriminant analysis detection rule for spatially-correlated sensor measurements with local Gauss–Markov dependency and constant-density random sensor placement, it is optimal to use precisely half the number of sensors as training sample instances. This result that a finite rather than infinite number of sensors is optimal follows from the phenomenon of overfitting. Less is more.

In developing this result, generalization error has been approximated using an expression by Raudys that involves Mahalanobis distance. Mahalanobis distance has been exactly stated for Gauss–Markov sensor measurements, and has also been approximated using weak laws of large numbers. The approximations are found to be quite tight in comparison with empirically computed true values in all regimes of the number of sensors. It has been seen that within the particular family of sensor placement distributions considered in Sec. 4, the uniform distribution minimizes a Poisson point process expectation parameter and thus consequently minimizes generalization error.

6. REFERENCES

[1] M. C. Vuran, Ö. B. Akan, and I. F. Akyildiz, "Spatio-temporal correlation: Theory and applications for wireless sensor networks," *Comput. Netw.*, vol. 45, no. 3, pp. 245–259, Jun. 2004.

[2] A. Jindal and K. Psounis, "Modeling spatially correlated data in sensor networks," *ACM Trans. Sensor Netw.*, vol. 2, no. 4, pp. 466–499, Nov. 2006.

[3] J.-F. Chamberland and V. V. Veeravalli, "How dense should a sensor network be for detection with correlated observations?" *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5099–5106, Nov. 2006.

[4] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Nonparametric decentralized detection using kernel methods," *IEEE Trans. Signal Process.*, vol. 53, no. 11, pp. 4053–4066, Nov. 2005.

[5] J. B. Predd, S. R. Kulkarni, and H. V. Poor, "Distributed learning in wireless sensor networks," *IEEE Signal Process. Mag.*, vol. 23, no. 4, pp. 56–69, Jul. 2006.

[6] K. R. Varshney and A. S. Willsky, "Learning dimensionality-reduced classifiers for information fusion," in *Proc. Int. Conf. Inf. Fusion*, Seattle, WA, Jul. 2009, pp. 1881–1888.

[7] Š. Raudys and D. M. Young, "Results in statistical discriminant analysis: A review of the former Soviet Union literature," *J. Multivariate Anal.*, vol. 89, no. 1, pp. 1–35, Apr. 2004.

[8] A. Anandkumar, L. Tong, and A. Swami, "Detection of Gauss–Markov random fields with nearest-neighbor dependency," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 816–827, Feb. 2009.

[9] M. D. Penrose and J. E. Yukich, "Weak laws of large numbers in geometric probability," *Ann. Appl. Prob.*, vol. 13, no. 1, pp. 277–303, Jan. 2003.