RUTCOR RESEARCH REPORTS, 2011

RUTCOR – Rutgers Center for Operations Research
Busch Campus, Rutgers University
640 Bartholomew Road, Piscataway, NJ 08854-8003

RRR 1-2011 **GENERALIZATION ERROR BOUNDS FOR LOGICAL ANALYSIS OF DATA**, Martin Anthony

This report analyses the predictive performance of standard techniques for the 'logical analysis of data' (LAD), within a probabilistic framework. Improving and extending earlier results, we bound the generalization error of classifiers produced by standard LAD methods in terms of their complexity and how well they fit the training data. We also obtain bounds on the predictive accuracy which depend on the extent to which the underlying LAD discriminant function achieves a large separation (a 'large margin') between (most of) the positive and negative observations.

RRR 2-2011 **CONDORCET DOMAINS OF TILING TYPE**, Vladimir I. Danilov, Alexander Karzanov, Gleb Koshevoy

A Condorcet domain is a subset of the set of linear orders over a finite set of candidates defined by the following property: if voters' preferences are linear orders from such a subset then the simple majority rule does not yield cycles. We propose a method to construct "large" Condorcet domains by making use of the so-called rhombus tilings. We show that this method unifies several previously known constructions of Condorcet domains. Finally, we show that tree conjectures on the size of such domains are in fact equivalent and provide a counterexample.

Keywords: majority rule, rhombus tiling, weak Bruhat order, pseudo-line arrangement, alternating scheme, Fishburn's conjecture

RRR 3-2011 **A METHOD TO SCHEDULE BOTH TRANSPORTATION AND PRODUCTION AT THE SAME TIME IN A SPECIAL FMS**, Navid Hashemian, Béla Vizvári

There are many different types of FMSs. One of the simplest one is called Single-Stage Multi-machine System (SSMS). It consists of identical parallel work centers, a load/unload station, a material handling system and central tool magazine with a dedicated tool transportation system. SSMS has been investigated since the early 1990s. If the tool system is not a bottleneck then the main problem in scheduling an SSMS is that the schedules of the production and material handling systems must be harmonized, i.e. the best schedule of the production system must be determined such that it still can be served by the material handling system. The schedule of the SSMS has been addressed by some previous papers. But they scheduled either the production regardless the transportation system, or scheduled the transportation system only if the production had been already scheduled. The main contribution of the present paper is that a new single method is developed to schedule both the production and the material handling system simultaneously. All parts arrive to the assigned machine in time and the makespan is minimized. The two scheduling activities are not separated but they are done in the frame of the same algorithm.

RRR 4-2011 **MAXIMAL CONDORCET DOMAINS**, Danilov V.I., Koshevoy G.A.

A Condorcet domain is a subset of the set of linear orders on a finite set of candidates (alternatives to vote), such that if voters preferences are linear orders belonging to such a subset, then the simple majority rule does not yield cycles. It is well-known that the set of linear orders $\mathcal{LO}$ is the Bruhat lattice. We prove that a maximal Condorcet domain is a distributive sublattice in the Bruhat lattice. We give an explicit lattice formula for the simple majority rule. We introduce the notion of symmetric Condorcet domains and give a description of symmetric Condorcet domains of maximal size.

Keywords: Bruhat order, distributive lattice, binary plane tree, simple majority rule

RRR 5-2011 **MULTIVARIATE ARRIVAL RATE ESTIMATION BY SUM-OF-SQUARES POLYNOMIAL SPLINES AND SEMIDEFINITE PROGRAMMING**, Dávid Papp, Farid Alizadeh

An efficient method for the smooth estimation of the arrival rate of non-homogeneous, multi-dimensional Poisson processes from inexact arrivals is presented. The method provides a piecewise polynomial spline estimator using sum of squares polynomial optimization. It is easily parallelized exploiting the sparsity of the neighborhood structure of the underlying spline space; as a result, it is very efficient and scalable. Numerical illustration is included.

RRR 6-2011 **TO LAY OUT OR NOT TO LAY OUT?**, Sadegh Niroomand, Szabolcs Takács, Béla Vizvári

This paper seeks to raise awareness of some negative phenomena in science. It happens more and more frequently that somebody carries out research in one field but finds that the results are not strong enough for publication and then submits them to be published in a related but not identical field as an application. The Quadratic Assignment Problem (QAP) is known as one of the most difficult problems in integer programming, and it has a nice combinatorial formulation. Therefore, it is a suitable experimental field for many algorithmic ideas including artificial intelligence methods. On the other hand, these methods must compete with the special methods of QAP. The latter are far better in many cases. Thus, the experimental methods cannot be published in their own right. Their authors try to convert them to layout problems because QAP is well known as a basic model in that application area. However, it is easy to show by data analysis methods that the problems solved by some layout authors are not truly layout problems.

RRR 7-2011 **CHARACTERIZATION OF (QUASI)-ULTRAMETRIC FINITE SPACES IN TERMS OF (DIRECTED) GRAPHS**, Vladimir Gurvich, Michael Vyalyi

Given a complete directed graph (digraph) $D = (V, A)$ and a positive real weight function $d : A \to \{d_1, \ldots, d_k\} \subseteq \mathbf{R}_+$ such that $0 < d_1 < \ldots < d_k$, then, for any $i \in [k] = \{1, \ldots, k\}$ let us set $A_i = \{a = (u, w) \in A \mid d(a) \leq d_i\}$ and assume that every subgraph $D_i = (V, A_i), i \in [k]$, in the obtained nested family is transitive: $(u, w) \in A_i$ when $(u, v), (v, w) \in A_i$ for some $v \in V$ and for all $u, w \in V$ and $i \in [k]$.
It is not difficult to verify that the considered weighted digraph (D, d) defines a quasi-ultrametric finite space (QUMFS), that is,
$d(u, w) \geq 0$, $d(u, w) = 0$ iff $u = w$, and $d(u, w) \leq \max(d(u, v), d(v, w))$ $\forall\, u, v, w \in V$.

Moreover, each QUMFS is uniquely (up to an isometry) realized in such a way. This result implies that each QUMFS is realized by a multi-pole flow network.

In the symmetric case, $d(u,w) = d(w,u)$ for all $u, w \in V$, we obtain the following canonical representation of an ultrametric finite space (UMFS). Let $T = (V,E)$ be a rooted tree in which $L \subseteq V$ is the set of leaves and $v_0 \in N = V \setminus L$ is the root. For any leaf $v \in L$, there is a unique path $p(v)$ from v_0 to v. Furthermore, let $d : N \to \{d_1, \ldots, d_k\} \subseteq \mathbf{R}_+$ be a positive weight function ($0 < d_1 < \ldots < d_k$) whose weights strictly monotone decrease along each path $p(v)$, $v \in L$. Then, for every two distinct leaves $u, w \in L$, let us set $d(u,w) = d(v(u,w))$, where $v(u,w) \in N$ is the lowest common ancestor of u and w, or in other words, the last common vertex of the paths $p(u)$ and $p(w)$; standardly we set $d(u,w) = 0$ iff $u = w$. Again, it is easy to see that $d(u,w) = d(w,u) \geq 0$ and $d(u,w) \leq \max(d(u,v), d(v,w))$ for all $u, v, w \in L$. Thus, $(T = (V,E), d)$ forms an UMFS. Moreover, every UMFS is uniquely (up to an isometry) realized in this way if we additionally assume that $deg(v_0) \geq 2$ and $deg(v) \geq 3$ for any other $v \in N$.

By this result, somewhat surprisingly, an UMFS can be viewed as a k-person positional game of players $\{1, \ldots, k\}$ such that in every play $p(v)$ from v_0 to v the corresponding players move in a monotone strictly decreasing order.

Keywords: distance, ultrametric, spanning tree, minimum cut, maximum flow, Gomory-Hu tree, widest bottleneck path, decomposing n-graphs, positional game

RRR 8-2011 **RECONSTRUCTION OF WORLD BANK'S CLASSIFICATION OF COUNTRIES**, Nima Mirzaei, Béla Vizvári

The main objective of this paper is to analyze if the classification of countries provided by World Bank (WB) can be reconstructed with a linear and/or integer programming model called Multi-Group Hierarchical Discrimination if only data published by WB are used. WB has a public database of countries containing economical-financial and political factors. The parameters of the model have been determined on a collection of 44 countries. The model has been verified on other 39 countries. Only four out of 39 countries were misclassified which shows the power the elaborated model. Logical Analysis of Data (LAD) also has analyzed the problem. The attempt of the reconstruction of the classification uses 19 indicators. An important by-product of the reconstruction is that the methods select the most important indicators. Interestingly, the result proves that the World Bank classification is not based only on GNI per capita. In addition, more criteria that are important have main role in classification of countries.

RRR 9-2011 **ESTIMATING THE AVERAGE EFFECT SIZE AND THE PROPORTION OF MARKERS WITHOUT EFFECT IN GENOMEWIDE ASSOCIATION STUDIES**, József Bukszár, Edwin J. C. G. van den Oord

It has recently become possible to screen hundreds of thousands of genetic markers for their association with diseases. Knowledge of the proportion of markers without effect, p_0, and the effect sizes in these massive data sets has an intrinsic value and is required for a wide variety of applications. While numerous algorithms have been developed to estimate p_0, hardly any method is available to estimate effect sizes. We propose a maximum likelihood (ML) and a quasi-maximum likelihood (QML) approach for the simultaneous estimation of p_0 and the average effect size. The point estimate of any p_0 estimator can also be used in a 2-step procedure to estimate the average effect size through these (Q)ML methods. To avoid arbitrary choices of the fine-tuning parameter, needed for some p_0

estimators, we also developed a novel p_0 estimator where an (optimal) fine-tuning parameter is determined automatically through an iterative procedure. All estimators are illustrated for case-control studies for which we first derive an accurate approximation for the distribution of Pearson's statistic that depends on a single effect size parameter only. The two-step method appeared more accurate than the simultaneous estimation of both parameters. In this twostep procedure ML outperformed QML. ML combined with the Meinshausen-Rice estimator with fine-tuning parameter $\alpha = 0.5$ appeared to produce the best results in this genetic application. Our novel estimator was most precise among all studied p_0 estimators that did not require the pre-specification of a fine running parameter.

RRR 10-2011 **COMPUTING BOUNDS FOR THE PROBABILITY OF THE UNION OF EVENTS BY DIFFERENT METHODS**, József Bukszár, Gergely Mádi-Nagy, Tamás Szántai

Let $A_1, \ldots, A_n$ be arbitrary events. The underlying problem is to give lower and upper bounds on the probability $P(A_1 \cup \cdots \cup A_n)$ based on $P(A_{i_1} \cap \cdots \cap A_{i_k}), 1 \leq i_1 < \cdots < i_k \leq n$, where $k = 1, \ldots, d$, and $d \leq n$ (usually $d \ll n$) is a certain integer, called the order of the problem or the bound. Most bounding techniques fall in one of the following two main categories: those that use (hyper)graph structures and the ones based on binomial moment problems. In this paper we compare bounds from the two categories with each other, in particular the bounds yielded by univariate and multivariate moment problems are compared with Bukszár's hypermultitree bounds. In the comparison we considered several numerical examples, most of which have important practical applications, e.g., the approximation of the values of multivariate cumulative distribution functions or the calculation of network reliability. We compare the bounds based on how close they are to the real value and the time required to compute them, however, the problems arising in the implementations of the methods as well as the limitations of the usability of the bounds are also illustrated.

Keywords: Discrete moment problem, Linear programming, Hypergraphs, Expectation bounds, Probability bounds, Multivariate distribution function, Network reliability

RRR 11-2011 **TOTAL TIGHTNESS IMPLIES NASH-SOLVABILITY FOR THREE-PERSON GAME FORMS**, Endre Boros, Ondřej Čepek, Vladimir Gurvich

It is long known that every totally tight 2-person game form is Nash-solvable, that is, it has a Nash-equilibrium for any set of player preferences. Furthermore, it is also known that every totally tight 2-person game form is acyclic and dominance solvable. In this short paper we generalize the first result to 3-person game forms, leaving the general n-person case open. On the other hand, we show that the second result does not carry over to more than two players. We exhibit an example of a three-person game form which is totally tight but neither acyclic nor dominance solvable.

RRR 12-2011 **FURTHER GENERALIZATIONS OF THE WYTHOFF GAME AND MINIMUM EXCLUDANT**, Vladimir Gurvich

For any non-negative integers a and b, we consider the following game $WYT(a, b)$. Given two piles that consist of x and y matches, two players alternate turns; a single move consists of a player choosing x' matches from one pile and y' from the other, such that

$0 \le x' \le x,\ 0 \le y' \le y,\ 0 < x' + y'$, and $[\min(x', y') < b$ or $|x' - y'| < a]$.

The player who takes the last match is the winner in the normal version of the game and (s)he is the loser in its misere version.

It is easy to verify that the cases $(a = 0, b = 1), (a = b = 1)$, and $(b = 1, \forall a)$ correspond to the two-pile NIM, the Wythoff, and Fraenkel games, respectively. The concept of the minimum excludant mex is known to be instrumental in solving the last two games. We generalize this concept by introducing a function mex_b (such that $mex = mex_1$) to solve the normal and misere versions of the game $WYT(a, b)$.

Keywords: combinatorial games, impartial games, NIM, Wythoff game, Fraenkel game, minimal excludant, normal and misere versions, Sprague-Grundy function

RRR 13-2011 **POLYNOMIALLY COMPUTABLE BOUNDS FOR THE PROBABILITY OF THE UNION OF EVENTS**, E. Boros, A. Scozzari, F. Tardella, P. Veneziani

We consider the problem of finding upper and lower bounds for the probability of the union of events when the probabilities of the single events and the probabilities of the intersections of up to m events are given.

It is known that the best possible bounds can be obtained by solving linear programming problems with a number of variables that is exponential in the number of events. Due to their size and structure, these large linear programs are known to be very hard to solve. In the literature simpler, polynomially sized aggregations are considered and numerous closed form or polynomially computable bounds are derived from those.

We present here a new approach which introduces additional constraints to the dual linear programming problems in such a way that those become polynomially solvable. By using different sets of additional constraints, we introduce three new classes of polynomially computable upper and lower bounds. We show that they dominate almost all efficiently computable bounds known in the literature. Furthermore, by characterizing the vertices of two new classes of polyhedra, we can show that in two cases our bounds coincide with classical bounds, proving new extremal properties for those well-known bounds. Finally, we provide extensive numerical results comparing the average tightness of the various bounds on large number of instances.

RRR 14-2011 **SEMIDEFINITE CHARACTERIZATION OF SUM-OF-SQUARES CONES IN ALGEBRAS**, Dávid Papp, Farid Alizadeh

We extend Nesterov's semidefinite programming characterization of squared functional systems to cones of sum-of-squares elements in general abstract algebras. Using algebraic techniques such as isomorphism, linear isomorphism, tensor products, sums and direct sums, we show that many concrete cones are in fact sum-of-squares cones with respect to some algebra, and thus representable by the cone of positive semidefinite matrices. We also consider nonnegativity with respect to a proper cone K, and show that in some cases K-nonnegative cones are either sum-of-squares, or are semidefinite representable. For example we show that some well-known Chebyshev systems, when extended to Euclidean Jordan algebras, induce cones that are either Sum-of-Squares cones or are semidefinite representable. Finally we will discuss some concrete examples and applications, including minimum ellipsoid enclosing given space curves, minimization of eigenvalues of polynomial matrix pencils, approximation of functions by shape-constrained functions, and approximation of combinatorial optimization problems by polynomial programming.

RRR 15-2011 **GAMES SEKI AND SEKI-I**, Vladimir Gurvich

Let $A : I \times J \to \mathbb{Z}_+$, be a non-negative integer $m \times n$ matrix each row and column of which contain a strictly positive entry. The game *Seki* is defined as follows. Two players R and C take turns and it is specified who begins; this player is called the *first*, while the opponent *second*.

By one move a player can either reduce a strictly positive entry of A by 1 or pass. If both players pass then the game results in a draw. Player R (respectively, C) wins if a row (respectively, column) appears every entry of which is 0.

After a move, such a row and column may appear simultaneously. In game Seki we assume that the player who made this last move is the winner. Yet, we also study another version of the game, Seki-I, in which the above case is defined as a draw. If neither R nor C wins, even being first, A is called a *seki matrix (SM or SM-I)*. Furthermore, A is called a *complete seki matrix (CSM or CSM-I)* if A is a seki matrix and, moreover, each player must pass, that is, if (s)he makes an active move, the opponent wins.

Both Seki and Seki-I are difficult games. We cannot solve them and present only some partial results and conjectures mostly related to CSMs and CSM-Is. The latter family looks simpler, yet, game Seki, unlike Seki-I, is closely connected with the so-called seki (shared life) positions in the classical game of Go. Both Seki and Seki-I are of independent interest as combinatorial games.

Keywords: combinatorial games, games with positive incentive, Go, seki, shared life, complete seki, draw, integer matrix, integer doubly stochastic matrix.

RRR 16-2011 **MATHEMATICALLY-BASED INTEGRATION OF HETEROGENEOUS DATA**, József Bukszár, Edwin J. C. G. van den Oord

We present the exact formula, and based on that an estimate of the posterior probability that a hypothesis is true alternative in a multiple-hypothesis set-up utilizing information from several external data sets. Each external data set is a ranked list of a subset of the hypotheses, where for the ranks we merely assume that a hypothesis that is alternative in the external data set is more likely to have smaller rank than a hypothesis that is null in the external data set. An alternative hypothesis may be null in some or all of the external data sets, and also a null hypothesis may be alternative in some or all of the external data sets. The work is motivated by the problem of identifying biomarkers in a genetic data set that are associated with a complex disease utilizing several heterogeneous data sets.

RRR 17-2011 **THE PERFORMANCE OF A NEW HYBRID CLASSIFIER BASED ON BOXES AND NEAREST NEIGHBORS**, Martin Anthony, Joel Ratsaby

In this paper we present a new type of binary classifier defined on the unit cube. This classifier combines some of the aspects of the standard methods that have been used in the logical analysis of data (LAD) and geometric classifiers, with a nearest-neighbor paradigm. We assess the predictive performance of the new classifier in learning from a sample, obtaining generalization error bounds that improve as the 'sample width' of the classifier increases.

RRR 18-2011 **MISERABLE AND STRONGLY MISERABLE IMPARTIAL GAMES**, Vladimir Gurvich

An impartial game is called (strongly) miserable if its normal and
misère versions differ just slightly; more precisely, if some (all) zeros and ones of the Sprague-Grundy (SG) functions swap, while all larger values remain the same.
We obtain necessary and sufficient conditions for (strong) miserability and show that NIM, Wythoff's NIM, and game Euclid are miserable but not strongly miserable, while all subtraction games and the Fraenkel extension NIM(a) of Wythoff's NIM(1) for $a > 1$ are strongly miserable.
We also show that the sum of (strongly) miserable games is miserable but not necessarily strongly miserable.
SG theory enables us to solve efficiently the sum of games with known SG functions. If all summands are miserable, one can replace some of them by the corresponding misère versions, compute the modified SG functions making use of the above criteria, and solve the sum of the obtained games.

Keywords: combinatorial games, impartial games, standard and misère versions, Sprague-Grundy function, NIM, Wythoff's NIM, Fraenkel's NIM, game Euclid, game Mark, subtraction games, swap positions.

RRR 19-2011 **A POLYNOMIAL ALGORITHM FOR A TWO PARAMETER EXTENSION OF THE WYTHOFF NIM BASED ON THE PERRON-FROBENIUS THEORY**, Endre Boros, Vladimir Gurvich, Vladimir Oudalov

For any positive integer parameters a and b, the second author recently introduced a generalization mex_b of the standard minimum excludant $\mathrm{mex} = \mathrm{mex}_1$, along with a game NIM$(a, b)$ that extends further Fraenkel's NIM = NIM$(a, 1)$, which in its turn is a generalization of the classical Wythoff NIM = NIM$(1, 1)$. It was shown that P-positions (the kernel) of NIM(a, b) are given by the following typical recursion:

$$x_n = \mathrm{mex}_b\{x_i, y_i \mid 0 \le i < n\}, \;\; y_n = x_n + an; \;\; n \ge 0,$$

and conjectured that for all a, b the limits $\ell(a, b) = x_n(a, b)/n$ exist and are irrational algebraic numbers. In this paper we prove this conjecture showing that $\ell(a, b) = \frac{a}{r-1}$, where $r > 1$ is the Perron root of the polynomial

$$P(z) = z^{b+1} - z - 1 - \sum_{i=1}^{a-1} z^{\lceil ib/a \rceil},$$

whenever a and b are coprime; furthermore, it is known that $\ell(ka, kb) = k\ell(a, b)$.

In particular, $\ell(a, 1) = \alpha_a = \frac{1}{2}(2 - a + \sqrt{a^2 + 4})$. In 1982, Fraenkel (1982) introduced the game NIM(a) = NIM$(a, 1)$, obtained the above recursion and solved it explicitly getting $x_n = \lfloor \alpha_a n \rfloor$, $y_n = x_n + an = \lfloor (\alpha_a + a)n \rfloor$. Here we provide a polynomial time algorithm based on the Perron-Frobenius theory solving game NIM(a, b), although we have no explicit formula for its kernel.

Keywords: impartial games, NIM, Wythoff's NIM, Fraenkel's NIM, minimum excludant, algebraic number, asymptotic