

Sequential Importance Sampling for Multiway Tables

Yuguo Chen, Duke University
Ian H. Dinwoodie, Duke University
Seth Sullivant, UC Berkeley

Abstract

We describe an algorithm for the sequential sampling of entries in multiway contingency tables. Properties of the sampling values at each step are related to properties of the associated toric ideal using computational commutative algebra. In particular, order of cell sampling and the set of sampling values at each step are related to properties of the initial terms. We apply the algorithm to examples of contingency tables which appear in the social and medical sciences.

AMS 1991 subject classifications. Primary: 62H17, 62F03. Secondary: 13P10.

Key words and phrases. Conditional inference, contingency table, counting problem, exact test, Monte Carlo, sequential importance sampling, toric ideal.

1 Introduction

Sampling from multiway contingency tables can be used to compute exact Monte Carlo p -values of goodness-of-fit and parameter significance for conditional methods of inference. This is desirable when the tables of interest are numerous but have entries that raise doubts about the validity of asymptotic methods. A classical application is testing for Hardy-Weinberg equilibrium with genotype with multiple alleles, where some alleles may be quite rare and result in sparse tables [Guo and Thompson (1992)]. Other applications are described in Besag and Clifford (1989), Diaconis and Efron (1985), and Chen, Diaconis, Holmes, and Liu (2003). A more general problem is sampling from nonnegative integer lattice points. This includes contingency tables, and further applications such as Monte Carlo EM algorithms with incomplete data [Vardi (1996)] and Bayesian computation of posterior distributions [Tebaldi and West (1998)].

This paper shows that sequential importance sampling (SIS) can be implemented efficiently for a class of multiway contingency table problems that have been studied mostly with Markov chains. We show how basic geometric features of the sampling can be deduced from algebraic conditions on a collection of Markov moves that may be smaller than the full Markov basis. The results of this paper extend the applicability of SIS from two-way tables [Chen *et al.* (2003)] to a wider range of multiway tables and allow further comparison with Markov chain methods.

For some loglinear models, the constraints from sufficient statistics on multiway tables make it difficult to design irreducible Markov chains. The work of Diaconis and Sturmfels (1998) gives a method to produce Markov moves in the space of tables that connect all tables, but in some practical cases, such as large logistic regression examples, the moves cannot be computed. It is sometimes possible to do computations with a smaller collection of moves by letting some entries in the space of tables go negative – this idea is used in Bunea and Besag (2000) and Chen, Dinwoodie, Dobra, and Huber (2004). The cost is a longer running time for the Markov chain. In general the running times of these Markov chains are very difficult to judge. Therefore, Markov chains have three disadvantages: they can be hard to design, they can take a long time to run to stationarity, and the time to run to stationarity may not be clear. Their advantages are memory efficiency, ease of programming, and wide applicability.

The idea behind SIS is to sample cell entries in the contingency table one after the other so that the final joint distribution (i.e., the

proposal distribution) is close to a target distribution. The target distribution on the collection of tables may be uniform, which is useful for counting and some Bayesian applications where a uniform prior on probabilities leads to equally likely tables, and for the volume test of independence [Diaconis and Efron (1985)]. The target distribution may be hypergeometric, which arises in conditional inference with multinomial sampling, or it may be another related distribution such as the one for Hardy-Weinberg proportions.

SIS does not have the same disadvantages as a Markov chain, because the method terminates at the last cell value and generates i.i.d. samples from the proposal distribution. However, SIS raises a new set of implementation issues. The main problems are approximating the support of the marginal distribution quickly, and then approximating the marginal distribution on the support set with a proposal distribution.

Ideally, one would like to sample a cell value from the marginal distribution of a cell entry, conditional on the entries that have already been sampled. However, these marginal distributions are quite difficult to compute explicitly except in very small examples. Sequential importance sampling uses a proposal distribution on the set (or a superset) of all possible marginal values. The algorithm works successfully if the proposal distribution and the true distribution are “close enough” in ways that show up in low variability of compensating weights at the end.

When the support of the marginal cell distribution is an interval $[l, u]$, which turns out to be quite common, one needs the values of the upper and lower bounds. We state a simple but useful result that gives conditions on when the linear programming (LP) procedure will give the exact integer bounds. This is valuable, because using an integer programming (IP) algorithm at each step in the procedure would be much slower than using an LP solver. A precise algebraic relationship between LP and IP is developed in Hosten and Sturmfels (2003), but this theory can be hard to apply for large multiway tables.

SIS can yield an approximate count of constrained tables very quickly when the target distribution is uniform. This application has been carried out thoroughly in Chen *et al.* (2003), where SIS was shown to be more efficient than Markov chains for counting and testing two-way tables. In our multiway examples, we found approximate counts of tables without difficulty. By comparison, the counting software Latte [DeLoera, Haws *et al.* (2003)] had difficulty with most of our examples.

The paper is organized as follows. In Section 2 we introduce es-

quential methods of SIS. The algebraic conditions for efficient sampling are formulated in Sections 3 and 4. Many of the algebraic ideas of Markov chains on lattice points are used here. But the theory here is different in that the full list of moves in a Markov basis is not needed for the computation. One only needs certain algebraic properties on a subset of moves, and the algebraic properties can be checked easily. Section 3 treats the ideal case where conditions are independent of the actual data and only depend on the model. Section 4 is more technical and develops stronger methods that can apply when the observed margins imply conditions of positivity on the tables consistent with the margin values. Section 5 has practical algebraic conditions that guarantee that integer programming and linear programming give the same answer. In practice, it is not essential that LP and IP be identical. Section 6 discusses sampling distributions for different target distributions. In Section 7 we give a range of examples to show how well SIS can work in real problems.

2 Elements of SIS

Let Ω denote the set of all contingency tables with given constraints. Assume Ω is nonempty. The p -value for conditional inference on contingency tables can often be written as

$$\mu = E_p f(T) = \sum_{T \in \Omega} f(T) p(T), \quad (1)$$

where $p(T)$ is the underlying distribution on Ω , which is usually uniform or hypergeometric and only known up to a normalizing constant, and $f(T)$ is a function of the test statistic. For example, if we let

$$f(T) = 1_{\{p(T) \leq p(T_0)\}}, \quad (2)$$

where T_0 is the observed table, formula (1) gives the p -value of the exact test [Guo and Thompson (1992)]. In many cases, sampling from $p(T)$ directly is difficult. The importance sampling approach is to simulate a table $T \in \Omega$ from a different distribution $q(\cdot)$, where $q(T) > 0$ for all $T \in \Omega$, and estimate μ by

$$\hat{\mu} = \frac{\sum_{i=1}^N f(T_i) \frac{p(T_i)}{q(T_i)}}{\sum_{i=1}^N \frac{p(T_i)}{q(T_i)}}, \quad (3)$$

where T_1, \dots, T_N are i.i.d. samples from $q(T)$. We can also estimate the total number of tables in Ω by

$$|\widehat{\Omega}| = \frac{1}{N} \sum_{i=1}^N \frac{1}{q(T_i)}, \quad (4)$$

because $|\Omega| = \sum_{T \in \Omega} \frac{1}{q(T)} q(T)$. The underlying distribution on Ω corresponding to this case is uniform.

In order to evaluate the efficiency of an importance sampling algorithm, we can look at the number of i.i.d. samples from the target distribution that are needed to give the same standard error for $\hat{\mu}$ as N importance samples. A rough approximation for this number is the *effective sample size* [Kong, Liu and Wong (1994)]

$$\text{ESS} = \frac{N}{1 + cv^2}, \quad (5)$$

where the *coefficient of variation* (cv) is defined as

$$cv^2 = \frac{\text{var}_q\{p(T)/q(T)\}}{E_q^2\{p(T)/q(T)\}}. \quad (6)$$

Accurate estimation generally requires a low cv^2 , i.e. $q(T)$ must be sufficiently close to $p(T)$. We will use cv^2 as a measure of efficiency for an importance sampling scheme. In practice the theoretical value of the cv^2 is unknown, so its sample counterpart is used to estimate cv^2 . The standard error of $\hat{\mu}$ or $|\widehat{\Omega}|$ can be simply estimated by further repeated sampling [Chen *et al.* (2003)].

A central problem in implementing an importance sampling algorithm is the construction of a good proposal distribution $q(\cdot)$. Since the target space Ω is usually rather complicated, e.g., multiway tables with certain fixed marginals, it is not immediately clear what proposal distribution $q(\cdot)$ can be employed. We found that a computationally efficient method is to sample a table sequentially, cell by cell, in such a way to guarantee that every table in Ω can be produced. More precisely, we stack all entries of the table into a long vector t , and start by sampling the first cell of the vector t conditional on the constraints imposed on the table. This is equivalent to choosing one of the possible values for this entry. Conditional on the realization of the first cell, we sample the second cell in a similar manner and then move forward recursively until all the cells are sampled. Denoting the configurations of the cells of t by t_1, \dots, t_d , we can write $q(\cdot)$ as:

$$q(t = (t_1, \dots, t_d)) = q(t_1)q(t_2|t_1)q(t_3|t_2, t_1) \cdots q(t_d|t_{d-1}, \dots, t_1).$$

Thus, SIS raises a whole new set of problems if it is to be employed effectively: 1) When is the support of the marginal distribution $t_i|(t_{i-1}, \dots, t_1)$ nice? 2) How can the support of the marginal distribution be quickly determined or approximated? 3) How to sample from the support of the marginal distribution so that the proposal distribution is close to the true underlying distribution? We address these questions one by one in the following.

3 Sequential Intervals

When we apply SIS to the problem of sampling two-way contingency tables with fixed marginal sums [Chen *et al.* (2003)], we notice that the support of the marginal distribution $t_i|(t_{i-1}, \dots, t_1)$ is an interval of integers $[l_i, u_i]$, for $i = 1, \dots, d$. Therefore we can sample a value from the interval at each step, and always produce a table in Ω , i.e., every table satisfies the constraints. This saves a lot of computing time comparing to rejection sampling. Another advantage of having this interval property is that we can come up with a good proposal distribution $q(t_i|t_{i-1}, \dots, t_1)$ relatively easily, comparing to the case that there are gaps in the interval.

SIS tends to perform better when the sequential interval property holds, but for the general polytope, it is not always true that one can fill in entries in sequence and expect the range of feasible values to be an interval of integers. Examples of where the sequential interval property does not hold are very sparse logistic regression [Chen *et al.* (2004)] and many 3-way tables with certain margin constraints [see DeLoera and Onn (2004) for the full range of difficulties with 3-way tables]. Typically, there may be a problem if the moves of a Markov basis involve changes in some entry that are of size ± 2 or larger. A precise condition is more complicated and may depend on the margin values and the order of the sequential sampling. In this section we give the basic theorems that do not use the actual values of the margin constraints. In the next section we strengthen the results.

Let A be a $p \times d$ matrix of nonnegative integers, denoted Z_+ . We assume that a sum of some subset of the rows of A is a strictly positive vector. A Markov basis M_A for A is subset of $\ker_Z(A)$ such that for each pair of vectors $\mathbf{u}, \mathbf{v} \in Z_+^d$ with $A\mathbf{u} = A\mathbf{v}$, there is a sequence of vectors $\mathbf{m}_i \in M_A, i = 1, \dots, l$ such that

$$\mathbf{u} = \mathbf{v} + \sum_{i=1}^l \mathbf{m}_i,$$

$$\mathbf{0} \leq \mathbf{v} + \sum_{i=1}^j \mathbf{m}_i, \quad j = 1, \dots, l.$$

That is, two nonnegative vectors with the same linear constraints can be connected with a sequence of increments from M_A while always maintaining the linear constraints and the nonnegativity. A Markov basis is generally larger than a lattice basis for the kernel of A , but always exists independently of the actual values of the linear constraints.

For $\mathbf{t} \in Z_+^p$, let

$$A^{-1}[\mathbf{t}] := \{\mathbf{n} \in Z_+^d : \mathbf{A}\mathbf{n} = \mathbf{t}\}.$$

This is a collection of tables or nonnegative integer vectors, with linear constraints, sometimes called a polytope. The linear constraint value \mathbf{t} will sometimes informally be called margin constraint. The value of \mathbf{t} will typically be the sufficient statistics for a loglinear model. Our primary goal is to sample from $A^{-1}[\mathbf{t}]$.

There are two fundamental algebraic ideas related to Markov bases. Define the polynomial ring $Q[x_1, \dots, x_d]$ in variables x_1, \dots, x_d , one for each cell. Define the toric ideal

$$I_A := \langle \mathbf{x}^{\mathbf{n}} - \mathbf{x}^{\mathbf{m}} : \mathbf{A}\mathbf{n} = \mathbf{A}\mathbf{m} \rangle,$$

where $\mathbf{x}^{\mathbf{n}} := x_1^{n_1} x_2^{n_2} \dots x_d^{n_d}$ is the usual monomial notation for a nonnegative integer vector of exponents $\mathbf{n} = (n_1, \dots, n_d)$. Diaconis and Sturmfels (1998) showed that a finite generating set of binomials $\{\mathbf{x}^{\mathbf{m}_i^+} - \mathbf{x}^{\mathbf{m}_i^-}\}$ for I_A defines Markov moves $\pm(\mathbf{m}_i^+ - \mathbf{m}_i^-)$ that are a Markov basis in that they connect all of $A^{-1}[\mathbf{t}]$ when chosen randomly as vector increments, regardless of the actual value of \mathbf{t} . Further more, a subcollection of moves will connect two tables \mathbf{n}, \mathbf{m} if $\mathbf{x}^{\mathbf{n}} - \mathbf{x}^{\mathbf{m}} \in I$, where I is the ideal generated by the subcollection. This idea can be used to show connectivity (irreducibility) for subcollections of the full Markov basis for particular values of \mathbf{t} . The book Kreuzer and Robbiano (2000) defines toric ideals and operations of saturation that we will use later.

Definition 3.1. Define the projection operator $\pi_1 : Z^d \rightarrow Z$ by $\pi_1(z_1, \dots, z_d) = z_1$.

Lemma 3.1. Suppose a Markov basis M_A satisfies $\pi_1(M_A) \subset \{-1, 0, +1\}$. Then $\pi_1(A^{-1}[\mathbf{t}])$ is an interval of integers $[l_1, u_1]$.

Proof. One can connect tables $\mathbf{m}, \mathbf{n} \in A^{-1}[\mathbf{t}]$ with values m_1 and n_1 in the first coordinate by changing the first coordinate only ± 1 at each step, so the gap between possible values cannot be greater than 1. ■

If the columns of A are $\mathbf{a}_1, \dots, \mathbf{a}_d$, let $A_i = (\mathbf{a}_i, \mathbf{a}_{i+1}, \dots, \mathbf{a}_d)$ be the matrix that deletes the first $i - 1$ columns and keeps the last $d - i + 1$ columns of A .

Definition 3.2. *The polytope $A^{-1}[\mathbf{t}]$ has the sequential interval property if $\pi_1(A^{-1}[\mathbf{t}])$ is an interval $[l_1, u_1]$, and for $i = 1, \dots, d - 1$: if $n_i \in \pi_1(A_i^{-1}[\mathbf{t} - n_1\mathbf{a}_1 - \dots - n_{i-1}\mathbf{a}_{i-1}])$, then $\pi_1(A_{i+1}^{-1}[\mathbf{t} - n_1\mathbf{a}_1 - \dots - n_{i-1}\mathbf{a}_{i-1} - n_i\mathbf{a}_i])$ is also an interval $[l_{i+1}, u_{i+1}]$.*

For $\mathbf{m} \in Z^d$, define $\mathbf{m}^+ = \max\{0, \mathbf{m}\}$, $\mathbf{m}^- = \max\{0, -\mathbf{m}\}$, so $\mathbf{m} = \mathbf{m}^+ - \mathbf{m}^-$.

Proposition 3.1. *Suppose a Markov basis $M_A = \{\pm\mathbf{m}_1, \dots, \pm\mathbf{m}_g\}$ has the property that $G := \{\mathbf{x}^{\mathbf{m}_i^+} - \mathbf{x}^{\mathbf{m}_i^-}, i = 1, \dots, g\}$ is a lex Gröbner basis with ordering $x_1 > x_2 > \dots > x_d$ on indeterminates and suppose the elements of $G \cap Q[x_i, \dots, x_d]$ are square-free in x_i for each i . Then $A^{-1}[\mathbf{t}]$ has the sequential interval property for all \mathbf{t} .*

Proof. By the elimination theorem, the lex basis G has the property that $G \cap Q[x_i, \dots, x_d]$ is a Gröbner basis for the ideal $I_{A_i} = \langle \mathbf{x}^{\mathbf{m}} - \mathbf{x}^{\mathbf{n}}, A_i\mathbf{m} - A_i\mathbf{n} \rangle$. Hence by results of Diaconis and Sturmfels (1998), the difference of the exponents (together with signs \pm) of elements in $G \cap Q[x_i, \dots, x_d]$ are a Markov basis with 0 in coordinates $1, 2, \dots, i - 1$. An application of Lemma 3.1 to the matrix A_i with first coordinate n_i finishes the proof. ■

When using this result, some orders on the cells may have the desired property and others may not, so it can be used to find good orderings on the cells. In fact, the converse to Proposition 3.1 is also true, in the sense that matrices A such that $A^{-1}[\mathbf{t}]$ has the sequential interval property regardless of \mathbf{t} are characterized by their lex Gröbner bases.

Proposition 3.2. *Let A be a matrix such that $A^{-1}[\mathbf{t}]$ has the sequential interval property for all \mathbf{t} . Then the reduced lex Gröbner basis G for I_A with ordering $x_1 > x_2 > \dots > x_d$ has $G \cap Q[x_i, \dots, x_d]$ square-free in x_i for all i .*

Proof. It suffices for prove the claim when $i = 1$, the rest following by induction. Let $G := \{\mathbf{x}^{\mathbf{m}_i^+} - \mathbf{x}^{\mathbf{m}_i^-}\}$ be the reduced lex Gröbner basis. In particular, none of the monomials $\mathbf{x}^{\mathbf{m}_i^+}$ are divisible by the leading monomial of any binomial in I_A . Suppose there is some $\mathbf{x}^{\mathbf{m}^+} - \mathbf{x}^{\mathbf{m}^-} \in G$ with $\pi_i(\mathbf{m}^+) = a > 1$. Let $\mathbf{t} = A\mathbf{m}^+$. Since $A^{-1}[\mathbf{t}]$ has the sequential interval property and $\pi_1(\mathbf{m}^-) = 0$, there exists $\mathbf{n} \in A^{-1}[\mathbf{t}]$ with $\pi_1(\mathbf{n}) = a - 1$. Then the binomial $x_1^{-a+1}(\mathbf{x}^{\mathbf{m}^+} - \mathbf{x}^{\mathbf{n}}) \in I_A$, is not equal to $\mathbf{x}^{\mathbf{m}^+} - \mathbf{x}^{\mathbf{m}^-}$, and has leading term $x_1^{-a+1}\mathbf{x}^{\mathbf{m}^+}$ which divides $\mathbf{x}^{\mathbf{m}^+}$. This is a contradiction and $\mathbf{x}^{\mathbf{m}^+} - \mathbf{x}^{\mathbf{m}^-}$ is not in the reduced Gröbner basis G . ■

4 Markov Subbases

In this section we give results that can be used when the full Markov basis does not have the required properties to guarantee sequential intervals. The results use the particular values of the margin constraints that may allow a smaller and simpler connecting set that we call a Markov subbasis.

A Markov subbasis $M_{A,\mathbf{t}}$ for $\mathbf{t} \in \mathbb{Z}_+^p$ and A is a finite subset of $\ker_{\mathbb{Z}}(A)$ such that for each pair of vectors $\mathbf{u}, \mathbf{v} \in A^{-1}[\mathbf{t}]$, there is a sequence of vectors $\mathbf{m}_i \in M_{A,\mathbf{t}}, i = 1, \dots, l$ such that

$$\begin{aligned} \mathbf{u} &= \mathbf{v} + \sum_{i=1}^l \mathbf{m}_i, \\ \mathbf{0} &\leq \mathbf{v} + \sum_{i=1}^j \mathbf{m}_i, \quad j = 1, \dots, l. \end{aligned}$$

The connectivity through nonnegative lattice points only is required to hold for this specific \mathbf{t} .

Lemma 4.1. *Suppose a Markov subbasis $M_{A,\mathbf{t}}$ satisfies $\pi_1(M_{A,\mathbf{t}}) \subset \{-1, 0, +1\}$. Then $\pi_1(A^{-1}[\mathbf{t}])$ is an interval of integers $[l_1, u_1]$.*

Proof. One can connect tables with feasible values n_1 and m_1 in the first coordinate by changing the first coordinate only ± 1 at each step, so the gap between possible values cannot be greater than 1. ■

The following proposition is used in Examples 7.3 and 7.4 where Proposition 3.1 cannot be used.

Proposition 4.1. *Suppose a Markov subbasis $M_{A,\mathbf{t}} = \{\pm \mathbf{m}_1, \dots, \pm \mathbf{m}_g\}$ and let $G := \{\mathbf{x}^{\mathbf{m}_i^+} - \mathbf{x}^{\mathbf{m}_i^-}, i = 1, \dots, g\}$. Suppose G has the following three properties: G is a lex Gröbner basis for the generated ideal $I_{M_{A,\mathbf{t}}}$ with order $x_1 > x_2 > \dots > x_d$ on indeterminates; and $G \cap Q[x_i, \dots, x_d]$ are square-free in x_i for each i ; and $I_{M_{A,\mathbf{t}}} : x_i^\infty \cap Q[x_{i+1}, \dots, x_d] \subset I_{M_{A,\mathbf{t}}}$ for each $i = 1, 2, \dots, d-1$. Then the polytope $A^{-1}[\mathbf{t}]$ has the sequential interval property.*

Proof. By Lemma 4.1, $\pi_1(A^{-1}[\mathbf{t}])$ is an interval. Suppose $n_1 \in \pi_1(A^{-1}[\mathbf{t}])$. We must show that two tables in $A^{-1}[\mathbf{t}]$ with a common entry in coordinate 1 can be connected with moves in $M_{A,\mathbf{t}}$ without touching coordinate 1. To see this, suppose tables $\mathbf{u}', \mathbf{v}' \in A^{-1}[\mathbf{t}]$ have common first coordinate $u_1 = v_1 = c$.

Let $\mathbf{u} = (0, u_2, u_3, \dots, u_d), \mathbf{v} = (0, v_2, v_3, \dots, v_d)$. We must show that $\mathbf{x}^{\mathbf{u}} - \mathbf{x}^{\mathbf{v}} \in \langle G \cap Q[x_2, \dots, x_d] \rangle$ to be able to connect them with moves in G that only involve changing the second coordinate (by only ± 1 at each step). Since G is a lex basis, $\langle G \cap Q[x_2, \dots, x_d] \rangle = I_{M_{A,\mathbf{t}}} \cap Q[x_2, \dots, x_d]$, and it is enough to show that $\mathbf{x}^{\mathbf{u}} - \mathbf{x}^{\mathbf{v}} \in I_{M_{A,\mathbf{t}}}$. We have that $\mathbf{x}^{\mathbf{u}} - \mathbf{x}^{\mathbf{v}} \in I_A$.

Since $x_1^c(\mathbf{x}^{\mathbf{u}} - \mathbf{x}^{\mathbf{v}}) = \mathbf{x}^{\mathbf{u}'} - \mathbf{x}^{\mathbf{v}'} \in I_{M_{A,\mathbf{t}}}$, the binomial $\mathbf{x}^{\mathbf{u}} - \mathbf{x}^{\mathbf{v}} \in I_{M_{A,\mathbf{t}}} : x_1^\infty \cap Q[x_2, \dots, x_d]$. Under the assumption $I_{M_{A,\mathbf{t}}} : x_1^\infty \cap Q[x_2, \dots, x_d] \subset I_{M_{A,\mathbf{t}}}$, the first step is proven.

Suppose now that two tables $\mathbf{u}', \mathbf{v}' \in A^{-1}[\mathbf{t}]$ have common first two coordinates $u_1 = v_1 = c_1, u_2 = v_2 = c_2$. Let $\mathbf{u} = (0, 0, u_3, u_4, \dots, u_d), \mathbf{v} = (0, 0, v_3, v_4, \dots, v_d)$. We must show that $\mathbf{x}^{\mathbf{u}} - \mathbf{x}^{\mathbf{v}} \in \langle G \cap Q[x_3, \dots, x_d] \rangle$ to be able to connect them with moves in G that only involve changing the third coordinate (by only ± 1 at each step). By the argument above, we have that $x_2^{c_2} \mathbf{x}^{\mathbf{u}} - x_2^{c_2} \mathbf{x}^{\mathbf{v}} \in I_{M_{A,\mathbf{t}}}$. Then by the saturation condition on x_2 , $\mathbf{x}^{\mathbf{u}} - \mathbf{x}^{\mathbf{v}} \in I_{M_{A,\mathbf{t}}} : x_2^\infty \cap Q[x_3, \dots, x_d] \subset I_{M_{A,\mathbf{t}}}$.

The argument continues likewise for each cell in the order $1, 2, \dots, d$.

■

Lemma 4.2 below can be used to identify useful Markov subbases that may be candidates for Proposition 4.1. It is used for Example 7.3. The saturation method is computationally more practical than primary decomposition for determining connected components.

Lemma 4.2. *Let $M \subset \ker_Z(A)$ be Markov moves with ideal I_M . Suppose each element $\mathbf{n} \in A^{-1}[\mathbf{t}]$ satisfies $n_s > 0$ for all $s \in S \subset \{1, \dots, d\}$, and suppose that $I_M : \prod_{s \in S} x_s = I_A$, the toric ideal. Then the moves in M connect all of $A^{-1}[\mathbf{t}]$ and are therefore a Markov subbasis.*

Proof. Let $\mathbf{u}, \mathbf{v} \in A^{-1}[\mathbf{t}]$, and let $\mathbf{u}' = \mathbf{u} - I_S, \mathbf{v}' = \mathbf{v} - I_S$, where I_S is the vector with 1 in the coordinates that are in the set S , and 0 elsewhere. Clearly $\mathbf{x}^{\mathbf{u}'} - \mathbf{x}^{\mathbf{v}'} \in I_A$, so by the saturation assumption, $(\mathbf{x}^{\mathbf{u}'} - \mathbf{x}^{\mathbf{v}'}) \prod_{s \in S} x_s \in I_M$. The fundamental result of Diaconis and Sturmfels says that the moves in M connect $\mathbf{u} = \mathbf{u}' + I_S$ with $\mathbf{v} = \mathbf{v}' + I_S$ through the nonnegative tables. ■

5 Bounds on Cell Entries

When the conditions for sequential intervals are met, linear programming in the rational numbers can be asked dynamically for bounds on the interval at each step in the sampling. Linear programming is much faster than integer programming and under conditions that hold in most examples, LP gives the same answer as IP, which gives the exact integer bounds. The conditions we formulate are concrete algebraic conditions that can be checked with a preliminary calculation.

Another way to approximate the intervals in addition to LP and IP is called the shuttle algorithm, described in Buzzigoli and Giusti (1999) and Dobra and Fienberg (2001). This is an iterative method that usually does not give exact integer programming results. It has two small advantages in special cases: it is easy to program, and it can be implemented without explicitly constructing a constraint matrix, a task which may be impossible for very large problems with millions of cells. In our numerical examples, LP works better than the shuttle algorithm, in some cases much better.

The difference between LP and IP is studied in Hosten and Sturmfels (2003) from a different point of view. They give the worst possible difference over all constraint values, whereas our results may use the particular constraint values to prove a smaller difference for the specific data set. Our results do not use the irreducible redundant decomposition of their work, which can be large and slow to compute.

The numerical implementation of LP to determine an interval $[l, u]$ must be done carefully. Linear programming tends to give wider intervals than the true interval because LP considers solutions in a larger space. Roundoff of numerical approximations that come from floating point operations or interior point methods can result in sampling a number out of the feasible range $[l, u]$ or into a strict subset of the feasible range which can lead to errors. The program that we embedded into the sampling code and that worked well is lpsolve [Berkelaar, Eikland, and Notebaert (2004)]. In very special cases, one can use known formulas for the interval, such as the Fréchet bounds. This

can work in two-way tables and some decomposable graphical models [Chen *et al.* (2003)]. Usually this is not possible.

Consider the IP and LP problems

$$\begin{aligned} u_j(\mathbf{b}) &:= \max\{n_j : A_j \mathbf{n} = \mathbf{b}, \mathbf{n} \in Z_+^d\} \\ l_j(\mathbf{b}) &:= \min\{n_j : A_j \mathbf{n} = \mathbf{b}, \mathbf{n} \in Z_+^d\} \\ U_j(\mathbf{b}) &:= \max\{q_j : A_j \mathbf{q} = \mathbf{b}, \mathbf{q} \in Q_+^d\} \\ L_j(\mathbf{b}) &:= \min\{q_j : A_j \mathbf{q} = \mathbf{b}, \mathbf{q} \in Q_+^d\} \end{aligned}$$

where Z_+, Q_+ are the nonnegative integers and nonnegative rational numbers respectively. We are interested in bounding the nonnegative quantities $U_j - u_j$ and $l_j - L_j$.

For the following proposition, let $A_Q^{-1}[\mathbf{t}] := \{\mathbf{q} \in Q_+^d : A\mathbf{q} = \mathbf{t}\}$, the set of nonnegative rational numbers with constraints \mathbf{t} .

Proposition 5.1. *Suppose a Markov subbasis $M_{A,\mathbf{t}} = \{\pm \mathbf{m}_1, \dots, \pm \mathbf{m}_g\}$ has the property that $G := \{\mathbf{x}^{\mathbf{m}_i^+} - \mathbf{x}^{\mathbf{m}_i^-}, i = 1, \dots, g\}$ is a lex Gröbner basis with ordering $x_1 > x_2 > \dots > x_d$ on indeterminates for the generated ideal $I_{M_{A,\mathbf{t}}}$. Also, suppose $I_{M_{A,\mathbf{t}}} : \prod_{s \in I_{S_Q}} x_s = I_A$ where S_Q is the collection of coordinates which are always positive for elements in $A_Q^{-1}[\mathbf{t}]$, and suppose $I_{M_{A,\mathbf{t}}} : x_i^\infty \cap Q[x_{i+1}, \dots, x_d] \subset I_{M_{A,\mathbf{t}}}$ for each $i = 1, 2, \dots, d-1$.*

If the coordinate values of \mathbf{m}_i^+ are in $\{0, 1\}$, then $l_j(\mathbf{t}_j) = L_j(\mathbf{t}_j)$ for all $j = 1, 2, \dots, d$ and all \mathbf{t}_j given by $\mathbf{t}_1 = \mathbf{t}, \mathbf{t}_j = \mathbf{t} - \mathbf{a}_1 n_1 - \mathbf{a}_2 n_2 - \dots - \mathbf{a}_{j-1} n_{j-1}, j = 2, \dots, d$.

Proof. We show first the result that $l_1 \leq L_1$. Let $\mathbf{m} \in A^{-1}[\mathbf{t}]$.

Use long division to compute the normal form of $\mathbf{x}^{\mathbf{m}}$ with respect to $I_{M_{A,\mathbf{t}}}$. Let the normal form be the monomial $\mathbf{x}^{\mathbf{n}^*}$. It is nearly immediate that $n_1^* \geq l_1$, since the first coordinate of the normal form when dividing by a Gröbner basis for the full ideal I_A is l_1 .

Let \mathbf{q}^* solve $L_1 = \min\{q_1 : A\mathbf{q} = \mathbf{t}, \mathbf{q} \in Q_+^d\}$. We show that $q_1^* \geq n_1^*$, which together with $n_1^* \geq l_1$ will prove the result $L_1 = l_1$.

Suppose by way of contradiction that $n_1^* > q_1^*$. Since \mathbf{q}^* is rational, an integer multiple, say $\lambda \mathbf{q}^*$ is integral. Then $A(\lambda \mathbf{q}^*) = A(\lambda \mathbf{n}^*)$ so $\mathbf{x}^{\lambda \mathbf{n}^*} - \mathbf{x}^{\lambda \mathbf{q}^*} \in I_A$. Furthermore, by the assumption of positivity of coordinates S_Q on elements in $A_Q^{-1}[\mathbf{t}]$, it follows that $\mathbf{q}_s^*, \mathbf{n}_s > 0$ for $s \in S_Q$. Then $\mathbf{x}^{\lambda \mathbf{n}^*} - \mathbf{x}^{\lambda \mathbf{q}^*} \in I_{M_{A,\mathbf{t}}}$, by the saturation assumption on I_A .

Since G is a Gröbner basis for this ideal, one of the lead terms of the basis must divide the lead monomial $\mathbf{x}^{\lambda \mathbf{n}^*}$. This means that

the indices of positive coordinates of the exponents \mathbf{m}_i^+ of the lead monomial must be included in the positive coordinates of \mathbf{n}^* . Since the corresponding coordinate values are 0 or 1, the divisor must also divide \mathbf{n}^* . This contradicts its construction above as the normal form without divisors. Hence it cannot be the case that $n_1^* > q_1^*$. This proves that $l_1 \leq n_1^* \leq q_1^* = L_1$.

We show next the result that $l_2 \leq L_2$. Let $\mathbf{m} \in A^{-1}[\mathbf{t}]$.

Use long division to compute the normal form of $\mathbf{x}^{\mathbf{m}}$ with respect to $G_2 := G \cap Q[x_2, x_3, \dots, x_d]$, the elements of the subbasis that only involve coordinates $2, 3, \dots, d$. Let the normal form be the monomial $\mathbf{x}^{\mathbf{n}^*}$, where $n_1^* = m_1$, which has not changed in the division. It is nearly immediate that $n_2^* \geq l_2$, since the first coordinate of the normal form when dividing $x_2^{m_2} \cdots x_d^{m_d}$ by a Gröbner basis for the full ideal I_{A_2} is l_2 .

Let \mathbf{q}^* solve $L_2 = \min\{q_2 : A\mathbf{q} = \mathbf{t}, q_1 = m_1, \mathbf{q} \in Q_+^d\}$. We show that $q_2^* \geq n_2^*$, which together with $n_2^* \geq l_2$ will prove the result $L_2 = l_2$.

Suppose by way of contradiction that $n_2^* > q_2^*$. Since \mathbf{q}^* is rational, an integer multiple, say $\lambda\mathbf{q}^*$ is integral. Then $A(\lambda\mathbf{q}^*) = A(\lambda\mathbf{n}^*)$ so $\mathbf{x}^{\lambda\mathbf{n}^*} - \mathbf{x}^{\lambda\mathbf{q}^*} \in I_A$. Furthermore, by the assumption of positivity of coordinates S_Q on elements in $A_Q^{-1}[\mathbf{t}]$, it follows that $\mathbf{q}_s^*, \mathbf{n}_s > 0$ for $s \in S_Q$. Then $\mathbf{x}^{\lambda\mathbf{n}^*} - \mathbf{x}^{\lambda\mathbf{q}^*} \in I_{M_{A,\mathbf{t}}}$, by the saturation assumption on I_A . Also, $\mathbf{x}^{\lambda(0, n_2^*, \dots, n_d^*)} - \mathbf{x}^{\lambda(0, q_2^*, \dots, q_d^*)} \in I_{M_{A,\mathbf{t}}} : x_1^\infty \cap Q[x_2, \dots, x_d]$. By the other saturation assumption, it follows that $\mathbf{x}^{\lambda(0, n_2^*, \dots, n_d^*)} - \mathbf{x}^{\lambda(0, q_2^*, \dots, q_d^*)} \in I_{M_{A,\mathbf{t}}}$.

Since G_2 is a lexicographic Gröbner basis for the ideal $I_{M_{A,\mathbf{t}}} \cap Q[x_2, \dots, x_d]$ by the elimination theorem, one of the lead terms of the basis G_2 must divide the lead monomial $\mathbf{x}^{\lambda(0, n_2^*, \dots, n_d^*)}$, since we have just shown that this is the lead monomial in a binomial that belongs to $I_{M_{A,\mathbf{t}}} \cap Q[x_2, \dots, x_d]$. This means that the indices of positive coordinates of the exponents \mathbf{m}_i^+ of the lead monomial must be included in the positive coordinates of \mathbf{n}^* . Since the corresponding coordinate values are 0 or 1, the divisor must also divide \mathbf{n}^* . This contradicts its construction above as the normal form without divisors. Hence it cannot be the case that $n_2^* > q_2^*$. This proves that $l_2 \leq n_2^* \leq q_2^* = L_2$.

The remaining coordinates are proved similarly. ■

We state without proof an analogous result for the upper bounds.

Proposition 5.2. *Suppose a Markov subbasis $M_{A,\mathbf{t}} = \{\pm\mathbf{m}_1, \dots, \pm\mathbf{m}_g\}$ has the property that $G := \{\mathbf{x}^{\mathbf{m}_i^+} - \mathbf{x}^{\mathbf{m}_i^-}, i = 1, \dots, g\}$ is a grevlex Gröbner basis with ordering $x_d > x_{d-1} > \dots > x_1$ on indeterminates for the generated ideal $I_{M_{A,\mathbf{t}}}$. Also, suppose $I_{M_{A,\mathbf{t}}} : \prod_{s \in I_{S_Q}} x_s = I_A$*

where S_Q is the collection of coordinates which are always positive for elements in $A_Q^{-1}[\mathbf{t}]$. If the coordinate values of \mathbf{m}_i^+ are in $\{0, 1\}$, then $u_1(\mathbf{t}) = U_1(\mathbf{t})$.

The corollary below applies to five of the examples. For the lower bounds, it requires finding or characterizing the exponents on the lead terms of a lex Gröbner basis for the full toric ideal. For the upper bounds, it is more work because one must look at grevlex bases, which are related to maximizing coordinates in integer programming, for each of the constraint matrices A_j on the remaining cells (j, \dots, d) .

Corollary 5.1. *If a lex Gröbner basis for I_A has square-free exponents on the lead monomials, then $l_j = L_j$ for all $j = 1, \dots, d$. If each grevlex Gröbner basis for I_{A_j} , $j = 1, \dots, d$ and indeterminate ordering $x_d > x_{d-1} > \dots > x_1$ has square-free exponents on the lead monomials, then $u_j = U_j$ for all $j = 1, \dots, d$.*

Proof. The saturation assumptions of Proposition 5.1 hold if $I_{M_{A,\mathbf{t}}} = I_A$, so the lower bounds from LP and IP are equal. For the upper bounds, the statement is a restatement of Proposition 5.2 for each step in the sequential sampling. ■

6 Sampling Distributions

Assume that the sequential interval property holds for a multiway table with given constraints and the intervals can be found by LP, the next question is how to sample from these intervals. Ideally, we want to sample a cell value from the true marginal distribution of a cell entry conditional on the entries that have already been sampled. However, these marginal distributions are quite difficult to compute explicitly except in very small examples. SIS skirts this difficulty by sampling from a simple proposal distribution on the set of all possible marginal values for each entry. The algorithm works well if the proposal distribution and the true distribution are close enough to each other.

For a target uniform distribution, which is useful for counting the total number of tables, we propose a uniform distribution on the available interval for each cell. We call this “uniform sampling method.” With the length of the proposed sampling interval, the SIS weights can be computed exactly for reweighting at the end. This strategy gives low cv^2 (≤ 5) and works very well on the examples in Section 7.

For a target hypergeometric distribution, which arises in conditional inference with multinomial sampling, we propose to sample a cell value from the hypergeometric distribution $p(x) = \binom{u}{x} \binom{u}{l+u-x} / \binom{2u}{l+u}$ on the interval of available integers $[l, u]$. We call this “hypergeometric sampling method,” which is usually (but not always, see Example 7.4.) better than the uniform sampling method when the target distribution is hypergeometric. The hypergeometric sampling method gives useful results for examples in Section 7, although the cv^2 is not consistently small. For sparse tables, approximating the marginal mass function of the count in a single cell can be difficult. This seems to be a problem that goes beyond algebra.

7 Examples

In all the real examples in the following, the property of sequential intervals holds and the LP approximation was very close to or exactly equal to the exact IP range. The starting point to verify the conditions of Sections 3, 4, and 5 for a particular example is to attempt to compute the toric ideal I_A . It is most efficient to use an algorithm based on saturation, such as the algorithm in the `toric_ideal(A, "hs")` command in Singular (2003) or the `groebner` command in 4ti2 (2003). The software 4ti2 was used to construct constraint matrices for several examples (but one must know that the order of the columns corresponds to lex order on indeterminates when the coordinate labels are in the reverse order). For more detailed analysis, one may need operations such as division and saturation to verify conditions of Sections 3, 4, 5. These operations were fast on the examples we worked on in this section. In some examples, the Markov basis computation may be unnecessary if one knows already a Markov subbasis that connects tables with the particular constraint values.

In the following examples, all results are based on 1000 random samples using either the uniform random sampling method or the hypergeometric sampling method. The code was written in R and the software `lpsolve` was called from R. The running time ranges from several seconds to a few minutes on a Pentium IV laptop for the following examples, where IP usually takes hours and sometimes cannot finish within a reasonable amount of time for large examples.

Example 7.1. Consider the 3-way case/control data in the 4x4x2 table below from the Ille-et-Verlaine cancer study of the age 35-44 group (Breslow and Day (1980), Appendix I). The factors are Alcohol level (A), Tobacco level (T) and Response R, where R=0 is a control

measurement, R=1 is a case.

		A				
		1	2	3	4	
R=0	T	1	60	35	11	1
		2	13	20	6	3
		3	7	13	2	2
		4	8	8	1	0
R=1	T	1	0	0	0	2
		2	1	3	0	0
		3	0	1	0	2
		4	0	0	0	0

With a retrospective odds-ratio model $p(a, t|1)/p(a, t|0) = e^{\alpha_a + \beta_t}$, the appropriate margins to fix for conditional inference (treating $p(a, t|0)$ as unknown nuisance parameters) are [A,T] (sum over case/control counts at each level), [A,R], and [T,R] (sums over other factor at each response level). The constraints imply that the Graver basis for the independence model on T and R is a Markov basis, and the Graver basis equivalent to the collection of square-free circuit moves on one level of the Response factor. Thus the results of Section 3 and Corollary 5.1 imply the property of sequential intervals and LP will give the exact integral interval bounds at each step.

The simulation with LP gave 100% good tables. When the underlying distribution is uniform, the uniform sampling method gave cv^2 of 0.24 and estimated the total number of tables to be 25. When the underlying distribution is hypergeometric, the hypergeometric sampling method gave cv^2 of 0.5, and the estimated p -value for the exact test (defined by (1) and (2)) is 0.04.

Example 7.2. Consider the 4-way abortion opinion data from Christensen (1990, p. 129).

Race	Sex	Opinion	18-25	26-35	36-45	46-55	56-65	66+
White	Male	Yes	96	138	117	75	72	83
		No	44	64	56	48	49	60
		Undec.	1	2	6	5	6	8
	Female	Yes	140	171	152	101	102	111
		No	43	65	58	51	58	67
		Undec.	1	4	9	9	10	16
Non White	Male	Yes	24	18	16	12	6	4
		No	5	7	7	6	8	10
		Undec.	2	1	3	4	3	4
	Female	Yes	21	25	20	17	14	13
		No	4	6	5	5	5	5
		Undec.	1	2	1	1	1	1

Christensen fits the model: [RSO], [RSA], [ROA], [SOA]. The lex basis of 165 elements is square-free in the lead monomials, so the

sequential interval property holds by Section 3 and the IP and LP lower bounds are identical. A more detailed calculation to verify the conditions of Corollary 5.1 requires computing a grevlex basis for each of the submatrices of A_i , defined in Section 3 as the matrix that has columns $i, i + 1, \dots, d$ from A . This can be done and the condition is verified, proving that LP and IP upper bounds are always the same.

The LP method for finding the interval bounds gave 100% good tables in practice. When the underlying distribution is uniform, the uniform sampling method gave cv^2 of 2.92 and estimated the total number of tables to be 9.1×10^7 . When the underlying distribution is hypergeometric, the value of cv^2 using the hypergeometric sampling method was around 102.9, and the estimated p -value for the exact test (defined by (1) and (2)) is 0.85. Some loglinear models for this data are worse in that more bad tables are generated and the sequential interval property may not hold.

Example 7.3. Consider the 6-way binary Czech autoworker data below. Implementing SIS for this example requires techniques beyond the basic methods of Section 3.

F	E	D	C	B		A	
				no	yes	no	yes
neg	< 3	< 140	no	44	40	112	67
			yes	129	145	12	23
		≥ 140	no	35	12	80	33
			yes	109	67	7	9
	≥ 3	< 140	no	23	32	70	66
			yes	50	80	(0) 7	13
		≥ 140	no	24	25	73	57
			yes	51	63	7	16
pos	< 3	< 140	no	5	7	21	9
			yes	(0) 9	17	(0) 1	(0) 4
		≥ 140	no	(0) 4	3	11	8
			yes	14	17	5	(0) 2
	≥ 3	< 140	no	7	(0) 3	14	14
			yes	9	16	(0) 2	(0) 3
		≥ 140	no	(0) 4	(0) 0	13	11
			yes	(0) 5	14	(0) 4	4

A model that fits well is given by: [ACDEF], [ABDEF], [ABCDE], [BCDF], [ABCF], [BCEF] [Dobra, Tebaldi, West (2003)]. In the above table, (0) means that cell can be 0 in the rationals with the constraints from the model above, the others are strictly positive. The lex basis for the toric ideal with lex order in indeterminates yields 20 elements, one of which p_1 has an exponent of 2 on the lead indeterminate x_{111111} .

Therefore Proposition 3.1 cannot be applied directly. However, the ideal generated by the other 19 polynomials saturates in one step with respect to the monomial $\prod_{s \in S} x_s$ where S is the set of 41 coordinates that must be positive. Hence by Lemma 4.2, these 19 moves are a Markov subbasis. They are a lex Gröbner basis for themselves, and they have the saturation property required in Proposition 4.1, so the sequential interval property holds. Furthermore Proposition 5.1 show that the IP and LP lower bounds are always the same. Corollary 5.1 does not apply to show that the LP and IP upper bounds are the same, because exponents of 2 appear in the grevlex bases. We can use Proposition 5.2 on successive cells. One can show that the 19 moves in the lex basis for the Markov subbasis remain a subbasis for the later cells after fixing the first cells, using the method of Proposition 4.1 (if we were using the full Markov basis it would be immediate). Then using Proposition 5.2 at each stage shows that LP and IP are the same beginning at cell 111211 in lexicographic order 111111, 111112, etc., but the condition of the square-free grevlex basis does not hold before that point.

If the cells are filled in across rows and then down, the order is 111111, 211111, 121111, etc. (the order from 4ti2). The sequential interval holds in this order as well.

For this model, using LP for interval bounds gave 100% good tables. The shuttle algorithm gave 99% good tables with one iteration, and 99.5% with two iterations. When the underlying distribution is uniform, the uniform sampling method gave cv^2 of 1.09 and estimated the total number of tables to be 841. The quantity cv^2 when targeting the hypergeometric distribution using hypergeometric sampling method was 50.7, and the estimated p -value for the exact test (defined by (1) and (2)) is 0.27.

For this data with the model of all 15 four element constraints like [A,B,C,D], LP gave 100% good tables whereas the shuttle algorithm gave only 2% good tables after 10 iterations. The Markov basis for this case is too difficult to compute, but SIS still works well with $cv^2 = 5.0$ when the target distribution is uniform.

Example 7.4. Logistic regression tables have the sequential interval property when column sums are positive. Positive column sums give a Markov subbasis that is a lex Gröbner basis with square-free exponent on the lead indeterminate, but not entirely square-free. Thus the LP /IP difference can be positive.

The following data on leukemia deaths at age level 10-19 years at exposure is from Sugiura and Otake (1974).

dose:	1	2	3	4	5	6
leukemia deaths	5	4	6	1	3	6
	5973	11811	2620	771	792	820

The logistic regression model shows a residual deviance of 7.13 on 4 degrees of freedom, which is inconclusive for the goodness-of-fit.

A Markov subbasis with positive column sums consists of the moves that are differences of adjacent minors. These are square-free in the lead indeterminate, and they are a lex Gröbner basis, and finally the saturation condition of Proposition 4.1 holds, so the sampling has sequential intervals. The IP /LP gap of 0 is not guaranteed by Corollary 5.1 because of exponents of 2 in some of the Markov moves, and a calculation shows that the exact maximal difference between LP and IP on the interval bounds over all cell entries given all previous values, for these particular constraint values, is 0.8. Thus, despite the difference between LP and IP, rounding LP results in excellent performance of the LP approximation at every stage. The LP method for finding the intervals at each step gave 100% good tables and the shuttle algorithm gave 2.5% good tables after 10 iterations.

When the underlying distribution is uniform, the uniform sampling method gave cv^2 of 1.05 and estimated the total number of tables to be 3060. Latte [DeLoera *et al.* (2003)] finds 3053 tables with the same sufficient statistics. When the underlying distribution is hypergeometric, the hypergeometric sampling method gave cv^2 of 32.7, whereas the uniform sampling method gave $cv^2 = 20.3$. The estimated p -value for the exact test (defined by (1) and (2)) is 0.09.

Example 7.5. Consider the $3 \times 3 \times 3$ example from Diaconis and Sturmfels (1998, p. 379).

9	16	41	8	8	46	11	14	38
85	52	105	35	29	54	47	35	115
77	30	38	37	15	22	25	21	42

Proposition 4.1 and Corollary 5.1 imply sequential intervals with an IP /LP gap of 0 at every step. In simulation, LP gave 100% good tables, and the shuttle algorithm also gave 100% good tables after one iteration.

When the underlying distribution is uniform, the uniform sampling method gave cv^2 of 2.08 and estimated the total number of tables to be 1.9×10^{12} . When targeting the hypergeometric distribution, the hypergeometric sampling method gave $cv^2 = 180.7$.

Example 7.6. Consider data of genotype pairs from Guo and Thompson (1992).

1236									
120	3								
18	0	0							
982	55	7	249						
32	1	0	12	0					
2582	132	20	1162	29	1312				
6	0	0	4	0	4	0			
2	0	0	0	0	0	0	0		
115	5	2	53	1	149	0	0	4	

The constraints are the 9 allele counts, which are 9 linear functions that count twice the diagonal entry, so the A matrix has entries 0, 1, 2. For sequential sampling, the order of cells given by the following table leads to sequential intervals by Proposition 3.1:

1									
10	2								
11	18	3							
12	19	25	4						
13	20	26	31	5					
14	21	27	32	36	6				
15	22	28	33	37	40	7			
16	23	29	34	38	41	43	8		
17	24	30	35	39	42	44	45	9	

In general for the genotype problem, sampling across rows will not give intervals. Furthermore, the lead monomials in a lex basis have exponents that are all 0 or 1, so LP gives the exact interval bounds by Corollary 5.1. The simulation with LP produced 100% good tables. See Huber *et al.* (2004) for a direct sampling strategy and some further discussion of this example.

Example 7.7. Consider a constraint matrix A of the form $A = (A_0 \mid I)$ with 0 or 1 entries. Here I is the $e \times e$ identity matrix and A_0 is size $e \times f$ with columns $\mathbf{a}_1, \dots, \mathbf{a}_f$. This occurs in a tomography problem introduced by Vardi (1996) where A is a routing matrix where routes between adjacent vertices use the connecting edge, and the edge counts are put last as slack variables. The integer data $\mathbf{y} = A\mathbf{x}$ where \mathbf{x} are traffic counts between ordered pairs of nodes on a graph and \mathbf{y} is the aggregate traffic across links.

The property of sequential intervals holds for the entries of \mathbf{x} under the constraint $A\mathbf{x} = \mathbf{y}$ in the order of the columns. With indeterminates w_1, \dots, w_f for the first f columns and z_1, \dots, z_e for the last e slack variables, a lex Gröbner basis in $Q[w_1, \dots, w_f, z_1, \dots, z_e]$ consists of the f binomials $w_i - \mathbf{z}^{\mathbf{a}_i}$.

Linear programming will give the exact interval bounds each step because of the square-free lead monomials. The shuttle algorithm will

also give the exact intervals in one step. The interval for the first cell is exactly $[0, \min_{a_{i,1} > 0} \{y_i\}]$ and the same type of problem recurs at each step $1, \dots, f$.

The sampling method of West and Tebaldi (1998) for Bayesian computation of the posterior distribution is closely related to sequential sampling. Dinwoodie (2000) shows how fast sampling can be used in a Monte Carlo EM algorithm for estimating traffic rates.

To conclude, the examples indicate that algebraic theory is useful for establishing the interval property of cell sampling in practical problems, and that linear programming can be used efficiently at each step to find the interval. Also, practical problems are often easier with new refined methods than the earlier margin-independent Markov bases would indicate.

What seems difficult and requires further work is to formulate a method to design the proposal distribution at each step. We have seen that the uniform sampling method works very well when the underlying distribution is uniform. However, when the underlying distribution is hypergeometric, the hypergeometric sampling method still has space for improvement.

In practice, one may try sequential sampling even if the sequential interval property does not hold or the algebraic conditions are not satisfied. In this case, SIS tends to be less efficient because it may generate many bad tables and the proposal distribution may be hard to design. If one can find rough bounds for each entry and design the proposal distribution carefully, so that the fraction of valid tables is high and the cv^2 is low, SIS can still give satisfactory results.

Acknowledgments. This work was supported by NSF DMS-0200888, NSF DMS-0203762, and by SAMSI under grant DMS-0112069. We have used the software 4ti2, lpsolve, R, and Singular for computations.

References

- Berkelaar, M., Eikland, K., and Notebaert, P. (2004). *lpsolve: Open Source (Mixed-Integer) Linear Programming System*. GNU LGPL (Lesser General Public Licence).
- Besag, J., and Clifford, P. (1989). Generalized Monte Carlo significance tests. *Biometrika*, **76**, 633-642.

- Breslow, N. E., and N. E. Day (1980). *Statistical Methods in Cancer Research, Volume 1*. International Agency for Research on Cancer, Lyon, France.
- Bunea, F., and Besag, J. (2000). MCMC in $I \times J \times K$ contingency tables. Fields Institute Communications, **26**, AMS, Providence Rhode Island.
- Buzzigoli, L., and Giusti, A. (1999). An algorithm to calculate the lower and upper bounds of the elements of an array given its marginals. In *Statistical Data Protection, Proceedings of the Conference (Lisbon, 25 to 27 March 1998)*, Eurostat, Luxembourg, 131-147.
- Chen, Y., Diaconis, P., Holmes, S., and Liu, J. S. (2003). Sequential Monte Carlo methods for statistical analysis of tables. *Journal of the American Statistical Association*, in press.
- Chen, Y., Dinwoodie, I. H., Dobra, A., and Huber, M. (2004). Lattice points, contingency tables and sampling. To appear in *Contemporary Mathematics*. A. Barvinok, M. Beck, C. Haase, B. Reznick and V. Welker eds. American Mathematical Society.
- Christensen, R. (1990). *Log-Linear Models*. Springer, New York.
- DeLoera, J. A., Haws, D., Hemmecke, R., Huggins, P., Tauzer, J., Yoshida, R. (2003). *A User's Guide for LattE v1.1*. Available at <http://www.math.ucdavis.edu/latte/>.
- DeLoera, J. A., and Onn, S. (2004). Universality of Markov bases of slim three-way tables. Manuscript.
- Diaconis, P., and Efron, B. (1985). Testing for independence in a two-way table: new interpretations of the chi-square statistic. *Annals of Statistics*, **13**, 845-874.
- Diaconis, P., and Sturmfels, B. (1998). Algebraic methods for sampling from conditional distributions. *Annals of Statistics*, **26**, 363-397.
- Dinwoodie, I. H. (2000). Conditional expectations in network traffic estimation. *Statistics and Probability Letters*, **47**, 99-103.
- Dobra, A., and Fienberg, S. (2001). Bounds for cell entries in contingency tables induced by fixed marginal totals with applications to disclosure limitation. *Statistical Journal of the United Nations ECE*, **18**, 363-371.
- Dobra, A., Tebaldi, C., and West, M. (2003). Bayesian inference in incomplete multi-way tables. *Journal of Statistical Planning and Inference*, to appear.

- Greuel, G.-M., Pfister, G., and Schoenemann, H. (2003). SINGULAR: A computer algebra system for polynomial computations. <http://www.singular.uni-kl.de>.
- Guo, S. W., and Thompson, E. A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics*, **48**, 361-372.
- Hemmecke, R., and Hemmecke, R. (2003). *4ti2 Version 1.1: Computation of Hilbert bases, Graver bases, toric Gröbner bases, and more*. <http://www.4ti2.de>.
- Hosten, S., and Sturmfels, B. (2003). Computing the integer programming gap. Manuscript.
- Huber, M., Chen, Y., Dinwoodie, I. H., Dobra, A., and Nicholas, M. (2004). Monte Carlo algorithms for Hardy-Weinberg proportions. Under revision for *Biometrics*.
- Kong, A., Liu, J. S., and Wong, W. H. (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, **89**, 278-288.
- Kreuzer, M., and Robbiano, L. (2000). *Computational Commutative Algebra*. Springer, New York.
- R Development Core Team (2004). *R: A Language and Environment for Statistical Computing*. <http://www.R-project.org>.
- Sugiura, N., and Otake, M. (1974). An extension of the Mantel-Haenszel procedure to $K \times c$ contingency tables and the relation to the logit model. *Communications in Statistics*, **A 3**, 829-842.
- Tebaldi, C., and West, M. (1998). Bayesian inference on network traffic using link count data. *Journal of the American Statistical Association*, **93**, 557-573.
- Vardi, Y. (1996). Network Tomography: Estimating source-destination traffic intensities from link data. *Journal of the American Statistical Association*, **91**, 365-377.

Institute of Statistics and Decision Sciences
 Box 90251
 Duke University
 Durham, NC 27708-0251