Efficient Importance Sampling for Binary Contingency Tables

By Jose Blanchet
Harvard University

October 29, 2007

Abstract

Importance sampling has been reported to produce algorithms with excellent empirical performance in counting problems. However, the theoretical support for its efficiency in these applications has been very limited. In this paper, we propose a methodology that can be used to design efficient importance sampling algorithms for counting and test their efficiency rigorously. We apply our techniques after transforming the problem into a rare-event simulation problem – thereby connecting complexity analysis of counting problems with efficiency in the context of rare-event simulation. As an illustration of our approach, we consider the problem of counting the number of binary tables with fixed column and row sums, c_j's and r_i's respectively, and total marginal sums $d = \sum_j c_j$. Assuming that $\max_j c_j = o\left(d^{1/2}\right)$, $\sum c_j^2 = O\left(d\right)$ and the r_j's are bounded we show that a suitable importance sampling algorithm (proposed by Chen, Diaconis, Holmes and Liu (2005)) requires $O(d^2\varepsilon^{-2}\delta^{-1})$ operations to produce an estimate that has ε-relative error with probability $1-\delta$. In addition, if $\max_j c_j = o\left(d^{1/4-\delta_0}\right)$ for some $\delta_0 > 0$, the same coverage can be guaranteed with $O\left(d^2\varepsilon^{-2}\log\left(\delta^{-1}\right)\right)$ operations.

1 Introduction

We are interested in the complexity analysis of sequential or state-dependent importance sampling algorithms (SIS) for counting problems. The development of algorithms for approximate counting in polynomial time has been a topic of great interest in theoretical computer science (see Valiant (1979)). Successful techniques have been developed for efficient approximate counting based on the Markov Chain

[0]AMS 2000 subject classifications. Primary: 68W20, 60J20; secondary: 05A16, 05C30, 62Q05.
Key words and phrases. approximate counting, bipartate graphs, binary tables, importance sampling, Markov processes, Doob h-transform, changes-of-measure, rare-event simulation.

Monte Carlo (MCMC) method (see the texts by Sinclair (1993) and Jerrum (2003) for detailed information on these techniques). A different class of randomized algorithms for approximate counting, based on importance sampling, has received substantial attention recently (basic notions on importance sampling are discussed in Section 2; for additional background on importance sampling see Asmussen and Glynn (2007) and Liu (2001)). Chen, Diaconis, Holmes and Liu (2005) proposed an algorithm based on importance sampling for counting the number of bipartite graphs with a given degree sequence. They tested their algorithm empirically and observed that it achieved excellent performance. Recently, Blitzstein and Diaconis (2006) have also used importance sampling algorithms for approximately counting the number of acyclic and undirected graphs with a given degree sequence. In addition, Rubinstein (2006) has applied adaptive importance sampling algorithms to a variety of combinatorial problems, including counting and optimization. Although many of these algorithms based on importance sampling seem to have excellent practical performance, the theoretical framework to carry through a rigorous analysis of their performance is still under development.

Our purpose is to illustrate a framework that can be used to design efficient importance sampling algorithms for counting and provide a rigorous analysis of their computational complexity. Our method provides a direct connection between asymptotic approximations and efficient importance sampling and we believe that the principle underlying this connection can be applied in substantial generality. In order to illustrate our proposed techniques, we shall consider the problem of counting the number of 0-1 matrices with specified column and row sums – these types of matrices are also called binary contingency tables in statistical applications. In the context of graph theory, this problem is equivalent to that of counting the number of bipartite graphs with a given degree sequence.

Returning to the problem that we consider here we mention that statistical analysis of binary contingency tables is a problem that has been motivated by several application domains, including some in Biology as explained in Chen et al (2005). Our goal is to provide rigorous support for the observed experimental efficiency of a class of SIS algorithms proposed by Chen et al (2005) for counting binary contingency tables. Formally, the problem consists in developing fast computational algorithms for counting the number of solutions $\{x_{ij} : 1 \leq i \leq m, 1 \leq j \leq n\}$ to

$$\begin{aligned} \sum_{j=1}^{n} x_{ij} &= r_i, \quad i \in \{1, 2, ..., m\}, && (1) \\ \sum_{i=1}^{m} x_{ij} &= c_j, \quad j \in \{1, 2, ..., n\}, && (2) \\ x_{ij} &\in \{0, 1\}. \end{aligned}$$

Let us define $d = \sum_{i=1}^{m} r_i = \sum_{j=1}^{n} c_j$. Our complexity analysis is performed by sending $d \nearrow \infty$ (and n, $m \nearrow \infty$ as well) in the context of sparse matrices under regularity

conditions. In particular, we assume that $\max_{j\leq n} c_j = o\left(d^{1/2}\right)$, $\sum_{j=1}^{n} c_j^2 = O\left(d\right)$ and that the r_i's are bounded. We shall construct and analyze a SIS-based estimator that is ε-close (in relative terms) to the total number of solutions to (1) - (2) with coverage probability of at least $1-\delta$ and that requires $O\left(d^2\varepsilon^2\delta^{-1}\right)$ operations as $d \nearrow \infty$ and $\varepsilon, \delta \searrow 0$ for its construction. Moreover, by imposing an additional growth condition of the form $\max_{i\leq n} c_j = o\left(d^{1/4-\delta_0}\right)$ for some $\delta_0 > 0$ as $d \nearrow \infty$ we obtain that SIS yields a "fully polynomial randomized approximation scheme" in the sense that $O((md+n)\varepsilon^{-2}\log(\delta^{-1})) = O(d^2\varepsilon^{-2}\log(\delta^{-1}))$ operations are required to produce an estimator that has ε-relative precision with probability $1-\delta$.

The proposed strategy for constructing and designing SIS algorithms proceeds as follows. The first step is to transform the counting problem into a so-called rare-event estimation problem, that is, the problem of estimating the probability of a rare event. Such probability is given by the ratio of the number of required solutions (x_{ij}'s) satisfying both (1) and (2) divided by the number of solutions satisfying only the set of constraints (2) (i.e. there is no restriction on the row sums), which can be easily computed. The second step is to recognize that such probability can be characterized by a system of linear equations. This is achieved by noting that the required probability can be computed in terms of a suitably defined m-dimensional random walk and then applying first transition analysis (i.e. conditioning on the first increment of the random walk). The solution to this system of equations provides the means of constructing the optimal importance sampling distribution, which is state-dependent. Such optimal importance sampling distribution corresponds to the so-called Doob's h-transform which arises in the context of positive harmonic functions and exponential martingales. The next step is to use results developed by McKay (1984) (see also Greenhill, McKay and Wang (2006) and Bekessy, Bekessy and Komlos (1972)) that approximate the number of solutions satisfying constraints (1) and (2) in the context of large and sparse matrices (that we have adopted here). We then use these approximations to construct an importance sampling distribution that mimics the behavior of the optimal importance sampler (thus, the better the approximation the closer the importance sampler to the optimal one).

It turns out that the importance sampling algorithm suggested by the previous strategy coincides with one of the algorithms studied in Chen et al (2005). Our results imply that in the context of large and sparse matrices satisfying the assumptions indicated above, the variance of the estimator obtained by the procedure has the best possible performance. That is, the coefficient of variation of the estimator (the ratio of the standard deviation to the probability of interest) remains bounded as $d \nearrow \infty$. In the context of rare-event simulation, an estimator that has a bounded coefficient of variation is said to be *strongly efficient* (a notion that will be reviewed in Section 2). Moreover, we show that if $\max_{i\leq n} c_j = o\left(d^{1/4-\delta_0}\right)$ then the proposed estimator is *exponentially efficient*, a concept that is introduced in Section 2 and in particular implies strong efficiency. Exponential efficiency allows to conclude that under our assumptions the proposed SIS estimator is ε-close in relative terms with coverage

probability $1-\delta$ and has complexity $O(d^2\varepsilon^{-2}\log(\delta^{-1}))$ as indicated above.

A recent algorithm for counting binary tables developed by Bezakova, Bhatnagar and Vigoda (2006), based on MCMC techniques (simulated annealing), has been shown to have complexity of (roughly) $O(d^3(nm)^2\max_{i,j}(c_i,r_j))$ operations. However, it is important to note that the procedure proposed by Bezakova et al (2006) works in complete generality (i.e. it does not require the sparsity assumptions imposed here). Other algorithms based on MCMC have been devised for counting binary contingency tables with certain regularity conditions on the degree sequence (such as ours). For instance, Kim and Vu (2003) assumed that $\max(r_i,c_j)=o\left(d^{1/3}\right)$ and proposed an algorithm that allows to generate an almost uniform bipartite graph (with given degree sequence) in time $O(d^3)$. Kannan, Tetali, and Vempala (1997) also study the problem of uniform generation of bipartite graphs with given degrees.

In a recent paper, Bayati, Kim and Saberi (2007) used ideas based on SIS to construct an algorithm for generation of simple graphs with a given degree sequence (a slightly different problem than the one that we study here). Under regularity conditions (similar to those imposed here) they proved that their proposed algorithm has excellent performance for asymptotically uniform generation, basically linear complexity, which makes the algorithm optimal in the sense that no faster complexity rate is possible. Their methods seem completely different from those developed here. In particular, we do not require the explicit use of concentration inequalities but instead the application of bounds related to Lyapunov inequalities for Markov chains. In addition, our methods suggest a natural way to develop efficient SIS in a variety of settings – basically if the optimal importance sampling distribution is described by a Markov chain and there are asymptotic approximations for the quantity of interest.

It is important to emphasize that although our complexity analysis of SIS suggests excellent performance (which is validated by the computer experiments performed by Chen et al (2005)), such performance can only be guaranteed under certain regularity conditions. This has been noted by Bezakova et al (2007) who constructed a counterexample showing that a SIS related to the one presented here (also proposed by Chen et al (2005)) can take exponential complexity if the degree sequence is allowed to grow arbitrarily. Nevertheless, one of the main points that we intend to communicate is that the general method outlined here could be adapted to specific contexts in which the problem at hand seems to have certain regularity properties that allow to develop approximations. The strategy would be then to enhance the approximations by means of efficient computational algorithms that can be shown to have desirable complexity properties in an asymptotic regime related to the developed approximations.

The basic principles behind the design and analysis of the SIS algorithms discussed here can be applied more broadly. For instance, Blanchet and Glynn (2007) apply these principles in the context of first-passage time probabilities with an emphasis on one-dimensional random walk problems with general heavy-tailed increments (which are of particular interest in insurance and queueing). Another example is given in

Blanchet and Liu (2007), which develops strongly efficient rare-event simulation algorithms for large deviation probabilities of regularly varying random walks. The analysis of SIS algorithms in rare-event simulation involves constructing so-called Lyapunov functions, which are solutions to certain inequalities that are used in stability analysis of Markov processes. The use of Lyapunov inequalities in the context of counting problems as the one considered here is particularly interesting because the dimension of the state-variables of the underlying Markov process (in our context m) is growing. As we shall discuss in Section 5 the construction of a suitable Lyapunov function often requires a good understanding of the local likelihood ratio obtained at each step of the simulation.

The rest of the paper is organized as follows. In Section 2, we provide a short discussion of importance sampling and performance analysis of rare-event simulation algorithms. Section 3 relates the problem of counting binary contingency tables to its rare-event simulation counterpart. The characterization of the optimal change-of-measure by means of a linear system of equations is also given in Section 3. In Section 4, we apply approximations for the number of tables into the design of an SIS algorithm. The complexity analysis of the counting algorithm is explained in Section 5.

2 Importance Sampling and Complexity in Rare-event

Let us briefly discuss basic concepts related to importance sampling and rare-event simulation methodology. For a more detailed discussion on importance sampling, see for instance Asmussen and Glynn (2007), Glynn and Iglehart (1989) and Liu (2001). Efficiency of rare-event simulation estimators is discussed in Asmussen and Glynn (2007), Bucklew (2004) and Juneja and Shahabuddin (2006).

Suppose that we want to estimate $P(Z \in A) > 0$, for a given random object taking values on a space $\mathcal{X}$ with a σ-field $\mathcal{B}$. Let us define the probability measure $F_Z(dz)$ on $(\mathcal{X}, \mathcal{B})$ via $F_Z(dz) = P(Z \in dz)$, so that

$$P(Z \in A) = \int_A F_Z(dz).$$

Let $G(dz)$ be any probability measure on $(\mathcal{X}, \mathcal{B})$ and assume that $F_Z(dz) I_A(z)$ is absolutely continuous with respect to $G(dz)$ (note that $F_Z(dz) I_A(z)$ is not necessarily a probability measure). In other words, assume that if $G(B) = 0$ then $P(A \cap B) = 0$, so that the Radon-Nikodym derivative $L(z) = I_A(z)(dF_Z/dG)(z)$ is well defined. Then,

$$P(Z \in A) = E^G L(Z) = \int L(z) G(dz). \tag{3}$$

We are using $E^G(\cdot)$ to denote an expectation that is computed under the probability measure $G(\cdot)$ (similarly, we will use $\mathrm{Var}_G(\cdot)$ for variances under $G(\cdot)$). The measure $G(\cdot)$ is an *admissible* choice for an *importance sampler* or *change-of-measure* that can be used to produce the *importance sampling estimator* L, which is clearly an unbiased estimator of $P(Z \in A)$. In some discussions on importance sampling the likelihood $W(z) \triangleq (dF_Z/dG)(z)$, when is well defined, is said to be the "importance sampling weight" (see, for instance, Liu (2001)). When $W(z)$ is well defined then one can write $L(z) = W(z) I_A(z)$.

The idea behind importance sampling is to take advantage of representation (3) in order to estimate $P(Z \in A)$. In particular, one can simulate k iid (independent and identically distributed) copies of Z using the distribution $G(\cdot)$ and output the estimator

$$W_{IS}^{(k)} = \frac{1}{k} \sum_{j=1}^{k} L(Z_j). \tag{4}$$

By the LLN's and identity (3), $W_{IS}^{(k)}$ is a consistent estimator of $P(Z \in A)$ as $k \nearrow \infty$. Note that importance sampling can in principle achieve zero variance. Indeed, if one chooses as change-of-measure

$$G^*(dz) = P(Z \in dz | Z \in A) = \frac{P(Z \in dz) I(z \in A)}{P(Z \in A)}$$

we obtain, say if $m = 1$ and $Z = z_1$,

$$\begin{aligned} W_{IS} = L(z_1) &= P(Z \in dz_1) \left[\frac{P(Z \in dz_1)}{P(Z \in A)} \right]^{-1} \\ &= P(Z \in A). \end{aligned}$$

So, our estimate of $P(Z \in A)$ is exact (in particular, it has zero variance). Obviously, such importance sampling estimator is not feasible to implement in practical cases because it requires knowledge of $P(Z \in A)$, which is the quantity of interest. However, the form of the zero-variance importance sampler indicates that a good change-of-measure should be similar to the conditional distribution of Z given that $Z \in A$.

Now, let us briefly discuss notions of efficiency that are helpful to calculate the computational cost (in terms of the number of replications) of estimating small probabilities via simulation using an estimator of the form (4). Let $\beta \triangleq P(Z \in A)$ and suppose that $\beta \approx 0$. In order to be precise, we shall introduce a parameter d such that $\beta_d \triangleq P(Z_d \in A) \longrightarrow 0$ as $d \nearrow \infty$ and perform our cost analysis under this asymptotic regime.

Our goal is to produce an estimator, $\widehat{\beta}_{d,k}$, with the property that that for given $\varepsilon, \delta \in (0,1)$, $\left| \widehat{\beta}_{d,k} - \beta_d \right| \leq \beta_d \varepsilon$ with probability $(1-\delta)$. If $\widehat{\beta}_{d,k}$ has this property, we say that $\widehat{\beta}_{d,k}$ *has* ε*-relative precision with* $1-\delta$ *confidence*. Here we use the subindex

k to denote the number of iid replications required to produce $\widehat{\beta}_{d,k}$. Suppose that $L_{d,j} = L_{d,j}(Z_{d,j})$ denotes the importance sampling estimator obtained in the j-th replication using a given change-of-measure G (a generic copy of such estimator will be denoted by $L_d = L_d(Z_d)$). We then consider the unbiased estimator

$$\widehat{\beta}_{d,k} = \frac{1}{k}\sum_{j=1}^{k} L_{d,j}(Z_{d,j}).$$

A standard way to measure the efficiency of the estimator $\widehat{\beta}_{d,k}$ in the rare-event simulation literature relates to its variance measured in relative terms. This approach gives rise to the notion of *strong efficiency*. More precisely, if $\sigma_d^2 = \mathrm{Var}_G(L_d) < \infty$, then the L_d's are said to be strongly efficient if the corresponding coefficient of variation, $cv_d \triangleq \sigma_d/\beta_d$, is uniformly bounded for $d \geq 0$. One often says that L_d is strongly efficient meaning that the family of L_d's is strongly efficient. In order to motivate strong efficiency in terms of the computational cost (measured by the number of iid replications) required to produce an estimator that has ε-relative precision with $1-\delta$ confidence one can use Chebyshev's inequality to obtain

$$P\left(\left|\widehat{\beta}_{d,k} - \beta_d\right| \geq \varepsilon\beta_d\right) \leq \frac{\sigma_d^2}{k\varepsilon^2\beta_d^2}.$$

Therefore $k \geq \varepsilon^{-2}\delta^{-1}(\sigma_d/\beta_d)^2$ replications are required to produce an estimator that achieves ε relative precision with $1-\delta$ confidence. Consequently, if L_d is strongly efficient, the number of replications required to obtain ε-relative precision with $1-\delta$ confidence is bounded as $\beta_d \longrightarrow 0$. Obviously, strong efficiency alone is not a useful concept for measuring computational complexity because nothing has been said about the computational cost attached to each replication.

When dealing with discrete structures, such as binary contingency tables, it makes sense to measure the cost per replication in terms of the amount of information (number of bits) required to encode the family of problems at hand (i.e. *the size of the problem*). In the context of binary contingency tables, statistical applications such as those described by Chen et al (2005), require estimating the whole distribution of statistics that depend on all the entries in the table in order to perform an hypothesis test. As a consequence, it makes sense to parameterize the size of the problem, say d, in terms of the number of bits required to encode a binary table, which can be taken to be the number of ones (or the number zeros, but if the table is sparse, it is obviously cheaper to encode it in terms of the number of ones).

The total complexity involves multiplying the number of replications, k, times the cost attached to the generation of each replication which we shall denote by $\kappa(d)$ (the cost per replication is measured by the total number of operations such as additions, multiplications and comparisons in terms of the size of the problem). Therefore, in the presence of strong efficiency, by setting the number of replications $k = O\left(\varepsilon^{-1}\delta^{-1}\right)$ we

see that $\widehat{\beta}_{d,k}$ requires $O\left(\kappa\left(d\right)\varepsilon^{-2}\delta^{-1}\right)$ operations as $d \nearrow \infty$ and $\varepsilon, \delta \searrow 0$ to achieve ε-relative precision with $1-\delta$ confidence.

The notions of efficiency discussed in the previous paragraph are related to standard notions found in randomized algorithms and approximate counting, such as that of *fully polynomial randomized approximation schemes* (FPRAS) (see, Mitzenmacher (2003) p. 254). In particular, an algorithm that outputs and estimator that has ε-relative precision with $1-\delta$ confidence in $O\left(\kappa\left(d\right)\varepsilon^{-k_1}\log(\delta^{-1})^{k_2}\right)$ operations, for some k_1, $k_2 > 0$, as $d \nearrow \infty$ and $\varepsilon, \delta \searrow 0$ is a FPRAS if $\kappa\left(d\right)$ grows polynomially in the size of the problem, say, d. In order to relate FPRAS to the types of efficiency notions encountered in rare-event simulation we need a stronger form of efficiency that we shall call *exponential efficiency*.

Definition We say that the family of estimators $\left(L_d\left(Z_d\right) : d \geq 1\right)$ is *exponentially efficient* for estimating β_d if there exists $\theta > 0$ such that

$$\psi\left(\theta\right) \triangleq \sup_{d\geq 1} \log E \exp\left(\theta L_d\left(Z_d\right)/\beta_d\right) < \infty.$$

The next lemma, which is a uniform version of Chernoff's bound, will be useful to relate an estimator of the form $\widehat{\beta}_{d,m}$ to a FPRAS.

Lemma 1 *Suppose that the family of estimators* $\left(L_d\left(Z_d\right) : d \geq 1\right)$ *is exponentially efficient for estimating* β_d*, then for* $\varepsilon > 0$ *we have*

$$P\left(\left|\widehat{\beta}_{d,k} - \beta_d\right| \geq \varepsilon\beta_d\right) \leq 2\exp\left(-k\min(I\left(\varepsilon\right), I\left(-\varepsilon\right)\right) \tag{5}$$

where $I\left(h\right) = \sup_\theta\left(\theta\left(1+h\right) - \psi\left(\theta\right)\right)$*. Moreover,* $I\left(\varepsilon\right), I\left(-\varepsilon\right) > 0$ *and* $I\left(h\right) \geq \rho h^2$ *for some* $\rho > 0$*.*

Proof. Just as in the proof of Chernoff's bound, (5) follows by an application of Chebyshev's inequality. Let $\psi_d\left(\theta\right) = \log E\exp\left(\theta L_d\left(Z_d\right)/\beta_d\right)$ we then obtain

$$\begin{aligned} P\left(\widehat{\beta}_{d,k} - \beta_d \geq \varepsilon\beta_d\right) &\leq \exp\left(-k\sup_{\theta\geq 0}\left(\theta\left(1+\varepsilon\right) - \psi_d\left(\theta\right)\right)\right) \\ &= \exp\left(-k\sup_{\theta}\left(\theta\left(1+\varepsilon\right) - \psi_d\left(\theta\right)\right)\right) \leq \exp\left(-kI\left(\varepsilon\right)\right). \end{aligned}$$

Similarly, one obtains

$$P\left(\beta_d - \widehat{\beta}_{d,k} \geq \varepsilon\beta_d\right) \leq \exp\left(-kI\left(-\varepsilon\right)\right).$$

Inequality (5) is obtained by adding up the left and right hand sides of the previous displays after simple manipulations. The last part of the lemma follows from the

convexity of $\psi(\cdot)$ (supremum of convex functions is convex) combined with the fact that $\psi_d(\theta) - \theta = cv_d^2\theta^2/2 + O(\theta^3)$ as $\theta \searrow 0$ uniformly over d (which holds by Taylor's theorem and exponential efficiency) and the bound $\sup_{d\geq 1} cv_d^2 < \infty$ (which once again follows from exponential efficiency). ■

An immediate consequence of the previous results is that if L_d is exponentially efficient (which is to say that the family $(L_d : d \geq 1)$ is exponentially efficient) and $\kappa(d)$ operations are required to generate a single replication of L_d. Then, $\widehat{\beta}_{d,k}$ requires $O\left(\kappa(d)\,\varepsilon^{-2}\log\left(\delta^{-1}\right)\right)$ operations as $d \nearrow \infty$ and $\varepsilon, \delta \searrow 0$ to achieve ε-relative precision with $1-\delta$ confidence. If $\kappa(d)$ grows polynomially in the size of the problem d, then $\widehat{\beta}_{d,k}$ is the output of a FPRAS. One way to verify exponential efficiency is by showing that L_d is bounded above by some deterministic constant, say c_d^*, such that $c_d^* = O(\beta_d)$ as $d \nearrow \infty$.

3 Efficient Counting via IS and its Connection to Rare-event Simulation

A 0-1 table with specified marginals is a binary array (0-1 elements) of dimensions $m \times n$ such that the sum of the elements in the i-th row equals r_i ($i \in \{1, ..., m\}$) and the corresponding sum over the j-th column equals c_j, $j \in \{1, ..., n\}$.

Notational convention: Throughout the rest of the paper we shall use the notation $\mathbf{c} = (c_1, ..., c_n)$, $\mathbf{r} = (r_1, ..., r_m)$ and $\sum_{j=1}^n c_j = d = \sum_{i=1}^m r_i$. In addition, we shall reserve the use of boldface letters to denote vectors or high dimensional objects. Random variables are denoted using capital letters and the use of lower case is restricted to deterministic quantities (including specific realizations of random objects).

We are interested in developing an importance sampling algorithm that allows to efficiently count the number of such arrays, which we shall denote by $\mu(\mathbf{r}, \mathbf{c})$.

Note that the number of tables with m rows and given only column marginals $\mathbf{c}$ is

$$\eta(\mathbf{c}, m) = \binom{m}{c_1} ... \binom{m}{c_n}.$$

So, the number of tables with given column and row marginals, $\mathbf{c}$ and $\mathbf{r}$ respectively, can be evaluated via

$$\eta(\mathbf{c}, m) \cdot P(\mathbf{T}(\mathbf{Y}) = \mathbf{r}),$$

where $\mathbf{T}(\mathbf{Y}) \in \mathbb{R}^m$ is the row marginals of a table $\mathbf{Y}$ sampled uniformly over the space of tables with column marginals $\mathbf{c}$ (that is, $\mathbf{T}(\mathbf{Y})_i = \sum_{j=1}^n \mathbf{Y}_{i,j}$ for $1 \leq i \leq m$). The problem of counting the number of binary contingency tables is equivalent to that of estimating $P(\mathbf{T}(\mathbf{Y}) = \mathbf{r})$. We can see that efficient estimation of this probability with good relative precision is not straightforward because the probability in question

may become arbitrarily small as the size of the table increases. In other words, the event $\{\mathbf{T}(\mathbf{Y}) = \mathbf{r}\}$ would typically be rare.

We shall exploit this connection in order to design a counting algorithm that can be rigorously proved to be efficient in some asymptotic sense. In particular, we shall formulate the problem of estimating $P(\mathbf{T}(\mathbf{Y}) = \mathbf{r})$ as a sequential rare-event simulation problem involving a suitably defined random walk. Importance sampling will then arise as a natural technique for estimating this probability.

Define an m dimensional random walk (rw) $\mathbf{S} = (\mathbf{S}_k : 0 \leq k \leq n)$ as follows. Given vectors of non-negative integers $\mathbf{c}$ and $\mathbf{r}$, set $\mathbf{S}_0 = \mathbf{r}$ and (for $k \in \{1, ..., n\}$) define $\mathbf{S}_k = \mathbf{S}_{k-1} - \mathbf{X}_k$ where $\mathbf{X}_k$ is a 0-1 entry vector (of dimension m) with uniform distribution over the space of configurations $(x_{k,1}, ..., x_{k,m})$ such that $\sum_{j=1}^{m} x_{k,j} = c_k$ and (for $1 \leq j \leq m$) $x_{k,j} \in \{0, 1\}$. The vector $\mathbf{X}_k$ represent the k-th column of the table. The random vectors $(\mathbf{X}_k : 1 \leq k \leq n)$ are assumed to be independent. Finally, let us write $P_{\mathbf{r},\mathbf{c}}(\cdot)$ for the probability law generated by the rw $\mathbf{S}$ subject to $\mathbf{S}_0 = \mathbf{r}$ and $E_{\mathbf{r},\mathbf{c}}(\cdot)$ for the corresponding expectation operator. Note that $P_{\mathbf{r},\mathbf{c}}(\cdot)$ is defined via a time inhomogeneous Markov chain, this is because the distributions of the $\mathbf{X}_k$'s change in time according to $\mathbf{c}$.

Now, given an arbitrary vector $\boldsymbol{\rho}$, we shall use $size(\boldsymbol{\rho})$ to denote its dimension. So, for instance, $size(\mathbf{c}) = n$. Observe that

$$u(\mathbf{r}, \mathbf{c}) \triangleq P_{\mathbf{r},\mathbf{c}}(\mathbf{S}_n = 0) = P(\mathbf{T}(\mathbf{Y}) = \mathbf{r}).$$

Our discussion on importance sampling suggests that we use a probability measure that is as close as possible to the optimal choice (which achieves zero-variance). In this case, we can describe the optimal change of measure, $P^{Q^*}(\cdot)$, in terms of a suitably defined Markov process as follows. Note (using first transition analysis) that $u(\mathbf{r}, \mathbf{c})$ satisfies

$$u(\mathbf{r}, \mathbf{c}) = E_{\mathbf{r},\mathbf{c}}(u(\mathbf{S}_1, \boldsymbol{\rho}_1)),$$

where $\boldsymbol{\rho}_1 = (c_2, ..., c_n)$; note that $size(\boldsymbol{\rho}_1) = n - 1$. More generally, at time $0 \leq k \leq n - 1$

$$u(\mathbf{s}_k, \boldsymbol{\rho}_k) = E_{\mathbf{s}_k,\boldsymbol{\rho}_k}\left(u\left(\mathbf{S}_{k+1}, \boldsymbol{\rho}_{k+1}\right)\right), \tag{6}$$

where $\boldsymbol{\rho}_{k+1} = (c_{k+2}, ..., c_n)$, $size(\boldsymbol{\rho}_k) = n - k$. If we denote the empty vector by the symbol $*$, we must have that $u(\mathbf{0}, *) = 1$ and $u(\mathbf{r}, *) = 0$ for $\mathbf{r} \neq 0$.

Let us define a Markov kernel $Q^*_{\boldsymbol{\rho}_k}(\cdot)$ (for $0 \leq k \leq n$) via

$$Q^*_{\boldsymbol{\rho}_k}(\mathbf{s}_k, \mathbf{s}_{k+1}) = \binom{m}{c_{k+1}}^{-1} \frac{u\left(\mathbf{s}_{k+1}, \boldsymbol{\rho}_{k+1}\right)}{u(\mathbf{s}_k, \boldsymbol{\rho}_k)}.$$

Note that (6) guarantees that $Q^*_{\boldsymbol{\rho}_k}$ is a well defined Markov kernel (i.e. the probabilities $Q^*_{\boldsymbol{\rho}_k}(\mathbf{s}_k, \cdot)$ is a probability mass function). In order to simplify the notation in what follows, we shall drop the explicit dependent on the subindex $\boldsymbol{\rho}_k$. If one could use $P^{Q^*}(\cdot)$ as importance sampler for simulation (i.e. simulate the process

$(\mathbf{S}_k : 0 \leq k \leq n)$ according to transitions generated by $Q^*(\cdot)$) then our likelihood ratio estimator would be (given $\mathbf{S}_0 = \mathbf{r}$ and $\boldsymbol{\rho}_0 = \mathbf{c}$)

$$I(\mathbf{S}_n = 0) \prod_{k=0}^{n-1} \frac{u(\mathbf{S}_k, \boldsymbol{\rho}_k)}{u\left(\mathbf{S}_{k+1}, \boldsymbol{\rho}_{k+1}\right)} = \frac{u(\mathbf{S}_0, \boldsymbol{\rho}_0)}{u(\mathbf{0}, *)} = u(\mathbf{r}, \mathbf{c}),$$

which has zero variance. Therefore, $Q^*(\cdot)$ corresponds to the zero-variance importance sampling distribution. The kernel $Q^*(\cdot)$ is the so-called Doob's h-transform and describes the conditional distribution of the process $\mathbf{S}$ given that $\mathbf{S}_n = 0$, see Doob (1957).

The efficient design of importance sampling algorithms should take advantage of any available information about $u(\cdot)$. For instance, if we know that $u(\cdot)$ is in some sense close to some computable function $v(\cdot)$, then, given our previous discussion, it is natural to consider a transition kernel of the form

$$Q(\mathbf{s}_k, \mathbf{s}_{k+1}) = \binom{m}{c_{k+1}}^{-1} \frac{v\left(\mathbf{s}_{k+1}, \boldsymbol{\rho}_{k+1}\right)}{w(\mathbf{s}_k, \boldsymbol{\rho}_k)},$$

where $w(\mathbf{s}_k, \boldsymbol{\rho}_k)$ is the appropriate normalizing constant that makes $Q(\cdot)$ a well defined Markov transition kernel. Once again, for notational simplicity, we shall suppress the explicit dependence of $\boldsymbol{\rho}_k$ in $Q(\cdot)$ but keep in mind that $Q(\cdot)$ is a time inhomogeneous Markov kernel. This is the strategy that we shall pursue in the next section in order to describe an importance sampling scheme that can be rigorously shown to be efficient in a context of sparse tables. An early reference that explores the connection between h-transforms and importance sampling is Glynn and Iglehart (1989), see also Asmussen and Glynn (2007) and Juneja and Shahabuddin (2006) for more applications of this idea.

4 Approximating the Optimal Change-of-measure and Algorithm Design

In order to apply the strategy outlined at the end of the previous section we need to find a suitable approximation $v(\cdot)$ to $u(\cdot)$. Results from McKay (1984) and Greenhill, McKay and Wang (2006) will allow us to obtain valuable information on $u(\cdot)$ that we will exploit in order to design an efficient importance sampling algorithm. In order to develop the required approximations, it is useful to introduce some notation.

As we indicated at the beginning of the previous section, $\mu(\mathbf{r}, \mathbf{c})$ represents the number of tables with fixed column sums vector $\mathbf{c}$ and marginal row sums given by $\mathbf{r}$. Note that with this definition of $\mu(\mathbf{r}, \mathbf{c})$ we have $u(\mathbf{r}, \mathbf{c}) = \mu(\mathbf{r}, \mathbf{c}) / \eta(\mathbf{c}, m)$. We shall assume that the c_j's are ordered in a non-increasing way so that $c_1 \geq c_2 \geq ... \geq c_n$. Having the c_j's ordered in this way does not affect the asymptotic approximations

that we are about to describe but, as we shall see, the ordering is important for the good performance of SIS.

We now introduce some convenient notation as in Greenhill, McKay and Wang (2006) that will be useful throughout the rest of the paper. Given a number s and an integer $k \geq 0$ we define $[s]_k = s(s-1)...(s-k+1)$ and $[s]_0 = 1$. Given a vector $\mathbf{s} = (s_1, ..., s_{n_0})$ of dimension n_0 we set $[\mathbf{s}]_0 = 1$ and define, for any integer $k \geq 1$,

$$[\mathbf{s}]_k = \sum_{j=1}^{n_0} [s_j]_k, \quad \left[\mathbf{s}^k\right]_1 = \sum_{j=1}^{n_0} s_j^k$$

and also $\mathbf{s}! = s_1!s_2!...s_{n_0}!$.

Define

$$\varphi(\mathbf{r}, \mathbf{c}) = \frac{[\mathbf{c}]_1!}{\mathbf{r}!\mathbf{c}!} \text{ and } \alpha(\mathbf{r}, \mathbf{c}) = \frac{[\mathbf{c}]_2 [\mathbf{r}]_2}{2[\mathbf{c}]_1^2}.$$

We now are ready state the following result, the proof of which is given at the end of the section. The next theorem is basically an adaptation of results from McKay (1984) (see also Theorem 1.1 of Greenhill, McKay and Wang (2006)).

Theorem 1 *Assume that* $\max([\mathbf{c}^2]_1, [\mathbf{r}^2]_1) = O(d)$ *and* $\max_{j \leq n, i \leq m}(c_j, r_i) = o\left(d^{1/2}\right)$ *as* $d \nearrow \infty$. *Then,*

$$\mu(\mathbf{r}, \mathbf{c}) \sim \varphi(\mathbf{r}, \mathbf{c}) \exp(-\alpha(\mathbf{r}, \mathbf{c}))$$

as $d \nearrow \infty$.

The previous result is slightly different from that of McKay (1984) who required $\max_{i \leq m, j \leq n}\{r_i, c_j\} = o\left(d^{1/4}\right)$ but did not assume $\max([\mathbf{c}^2]_1, [\mathbf{r}^2]_1) = O(d)$. Further refinements have been given in Theorem 1.3 of Greenhill, McKay and Wang (2006) who introduce additional correction terms by assuming $\max_{i \leq m} r_i \times \max_{j \leq n} c_j = O\left(d^{2/3}\right)$.

Continuing in the spirit of our discussion at the end of the previous section. We are interested in proposing a function $v(\cdot)$ that mimics the behavior of $u(\cdot)$ in some sense in order to construct our importance sampling algorithm. Theorem 1 suggest using the approximation

$$v(\mathbf{r}, \mathbf{c}) \triangleq \frac{\varphi(\mathbf{r}, \mathbf{c}) \exp(-\alpha(\mathbf{r}, \mathbf{c}))}{\eta(\mathbf{c}, m)}.$$

Let us define $v(\mathbf{0}, *) = 1$ and $v(\mathbf{s}, \boldsymbol{\rho}) = 0$ if at least one component of $\mathbf{s}$ is negative. As we indicated at the end of last section, our discussion of the zero-variance change-of-measure, Q^*, suggests designing the importance sampling distribution via a Markov transition kernel of the form

$$Q(\mathbf{s}_k, \mathbf{s}_{k+1}) = \binom{m}{c_{k+1}}^{-1} \frac{v\left(\mathbf{s}_{k+1}, \boldsymbol{\rho}_{k+1}\right)}{w(\mathbf{s}_k, \boldsymbol{\rho}_k)},$$

where $\boldsymbol{\rho}_k = (c_{k+1}, ..., c_n)$ and

$$w(\mathbf{s}_k, \boldsymbol{\rho}_k) = \sum_{(\mathbf{s}_k, \boldsymbol{\rho}_k) \to \left(\mathbf{s}_{k+1}, \boldsymbol{\rho}_{k+1}\right)} \binom{m}{c_{k+1}}^{-1} \frac{\varphi\left(\mathbf{s}_{k+1}, \boldsymbol{\rho}_{k+1}\right)}{\eta\left(\boldsymbol{\rho}_{k+1}, m\right)}$$

is the normalizing constant that makes $Q(\cdot)$ a well defined Markov transition kernel. In the previous display and in the discussion that follows we use $(\mathbf{s}_k, \boldsymbol{\rho}_k) \to \left(\mathbf{s}_{k+1}, \boldsymbol{\rho}_{k+1}\right)$ to denote an admissible transition step (i.e. $\mathbf{s}_k - \mathbf{s}_{k+1}$ is an m-dimensional 0-1 whose components add up to c_{k+1} and $\boldsymbol{\rho}_{k+1} = (c_{k+2}, ..., c_n)$).

We shall mention how to simulate transitions under $Q(\cdot)$ right after the precise description of the proposed algorithm below. We will use $P^Q_{\mathbf{r},\mathbf{c}}(\cdot)$ to denote the probability measure induced by the random walk $\mathbf{S}$ under the transition kernel given that $\mathbf{S}_0 = \mathbf{r}$ and $E^Q_{\mathbf{r},\mathbf{c}}(\cdot)$ to denote the corresponding expectation operator associated to $P^Q_{\mathbf{r},\mathbf{c}}(\cdot)$.

Note that under the change-of-measure $P^Q_{\mathbf{r},\mathbf{c}}(\cdot)$ we may have $P^Q_{\mathbf{r},\mathbf{c}}(\mathbf{S}_n \neq 0) > 0$. Therefore, when running an importance sampling algorithm based on transitions according to $Q(\cdot)$ we may obtain realizations for which $\{\mathbf{S}_n \neq 0\}$. A sufficient condition that implies $\{\mathbf{S}_n \neq 0\}$ and which can be easily checked at a time $k < n$ is that the number of strictly positive components of $\mathbf{S}_k$ is less than c_{k+1}. So, the path generation under $P^Q_{\mathbf{r},\mathbf{c}}(\cdot)$ will be done sequentially according to the transition kernel $Q(\cdot)$ until time n (in which case we have that the event $\{\mathbf{S}_n = 0\}$ has occurred) or up to the first time k such that the number of strictly positive components of $\mathbf{S}_k$ is less than c_{k+1} (in which case we have that $\{\mathbf{S}_n \neq 0\}$)..

In order to explain this path generation scheme more formally let us define

$$\Phi(\mathbf{S}_k) = card\{j : \mathbf{S}_{k,j} > 0\}.$$

That is, $\Phi(\mathbf{S}_k)$ is the number of strictly positive components of $\mathbf{S}_k$. Put $c_{n+1} \triangleq 1$ and define a stopping time τ via

$$\tau = \inf\{0 \leq k \leq n : \Phi(\mathbf{S}_k) < c_{k+1}\}.$$

Observe that when $\{\tau < n\}$ occurs one of the components of the vector $\mathbf{S}_n$ must be negative and therefore $\{\mathbf{S}_n \neq 0\}$. On the other hand, if $\mathbf{S}_\tau = 0$ we must have that $\tau = n$ because the c_i's are strictly positive and $\sum_{j=1}^n c_j = d$. Therefore, we have that $\{\mathbf{S}_n = 0\} = \{\mathbf{S}_\tau = 0\}$ and consequently

$$u(\mathbf{r}, \mathbf{c}) = P_{\mathbf{r},\mathbf{c}}(\mathbf{S}_\tau = 0).$$

The path generation scheme that we described before under the measure $P^Q_{\mathbf{r},\mathbf{c}}(\cdot)$ will be done sequentially up to the stopping time τ. Note that the k-th column, namely $\mathbf{X}_k$, is generated under $P^Q_{\mathbf{r},\mathbf{c}}(\cdot)$ during the course of the path generation only if $\tau > k-1$. In turn, $\mathbf{X}_k$ is a binary vector such that the sum of its components equals c_k and

$P^{Q}_{\mathbf{r},\mathbf{c}}(\cdot)$ avoids assigning negative components to the random walk $\mathbf{S}$ and therefore generation of increments under $P^{Q}_{\mathbf{r},\mathbf{c}}(\cdot)$ can be performed up to time τ. If $\tau < n$, then the τ-th assignment under $P^{Q}_{\mathbf{r},\mathbf{c}}(\cdot)$ cannot be done and the estimator is just zero. If $\tau = n$, then the table is constructed satisfying the row and column sums. We then conclude that $P^{Q}_{\mathbf{r},\mathbf{c}}(\cdot)$ is admissible in the sense that it does not assign zero mass to outcomes that are possible under $P_{\mathbf{r},\mathbf{c}}(\cdot)$ and for which $\mathbf{S}_n = 0$. In fact, it turns out that the sequential importance sampling algorithm generated by $Q(\cdot)$ coincides with one of the procedures studied by Chen, Diaconis, Holmes and Liu (2005). In order to see this, note that

$$\begin{aligned} Q(\mathbf{s}_k, \mathbf{s}_{k+1}) &= \binom{m}{c_{k+1}}^{-1} \frac{v(\mathbf{s}_k, \boldsymbol{\rho}_k)}{w(\mathbf{s}_k, \boldsymbol{\rho}_k)} \frac{v\left(\mathbf{s}_{k+1}, \boldsymbol{\rho}_{k+1}\right)}{v(\mathbf{s}_k, \boldsymbol{\rho}_k)} \\ &\propto \frac{v\left(\mathbf{s}_{k+1}, \boldsymbol{\rho}_{k+1}\right)}{v(\mathbf{s}_k, \boldsymbol{\rho}_k)} = \frac{\varphi\left(\mathbf{s}_{k+1}, \boldsymbol{\rho}_{k+1}\right)\eta(\boldsymbol{\rho}_k, m)}{\varphi(\mathbf{s_k}, \boldsymbol{\rho_k})\eta\left(\boldsymbol{\rho}_{k+1}, m\right)} \\ &\propto \frac{\varphi\left(\mathbf{s}_{k+1}, \boldsymbol{\rho}_{k+1}\right)}{\varphi(\mathbf{s_k}, \boldsymbol{\rho_k})} \propto \frac{\mathbf{s}_k!}{\mathbf{s}_{k+1}!} \exp\left(\alpha(\mathbf{s}_k, \boldsymbol{\rho}_k) - \alpha\left(\mathbf{s}_{k+1}, \boldsymbol{\rho}_{k+1}\right)\right) \\ &\propto \prod_{j \in \{j : \mathbf{s}_{k+1,j} \neq \mathbf{s}_{k,j}\}} (\mathbf{s}_{k,j} \exp(2\gamma_k \mathbf{s}_{k,j})) \end{aligned}$$

where $\gamma_k = \sum_{j=2}^{n-k}(\boldsymbol{\rho}^2_{k,j} - \boldsymbol{\rho}_{k,j})/(2(d - \boldsymbol{\rho}_{k,1}))$ (note that $card\{j : \mathbf{s}_{k+1,j} \neq \mathbf{s}_{k,j}\} = \boldsymbol{\rho}_{k,1}$). The expression indicated above coincides with the description given in p. 112 of Chen et al (2005) (the complete details of the computation corresponding to the difference $\alpha(\mathbf{s}_k, \boldsymbol{\rho}_k) - \alpha\left(\mathbf{s}_{k+1}, \boldsymbol{\rho}_{k+1}\right)$ are given in Section 5, see equation (13)). Thus, adopting the terminology used by Chen et al (2005), one can sample from $Q(\mathbf{s}_k, \cdot)$ according to the so-called "drafting method" for Conditional-Poisson (CP) distributions (the precise form of a CP distribution is given in STEP 2 of Algorithm 1 given below, the drafting method was introduced by Chen, Dempster and Liu (1994)).

Chen et al (2005) also proposed a more refined importance sampling procedure which can be explained as follows. Note that we constructed our importance sampling transition kernel, $Q(\cdot)$, via a suitable approximation $v(\mathbf{r}, \mathbf{c})$ of $u(\mathbf{r}, \mathbf{c})$ that is valid as $d \nearrow \infty$. Furthermore, we introduced additional information into $v(\cdot)$ by defining $v(\mathbf{s}, \boldsymbol{\rho}) = 0$ if $\mathbf{s}$ contains at least one negative component. Intuitively, we could have done even better by setting $v(\mathbf{s}, \boldsymbol{\rho}) = 0$ whenever $u(\mathbf{s}, \boldsymbol{\rho}) = 0$, this is the idea behind the refinement proposed by Chen et al (2005). One immediate difficulty here is the question of how to easily test the pairs $(\mathbf{s}, \boldsymbol{\rho})$ for which $u(\mathbf{s}, \boldsymbol{\rho}) = 0$. This is achieved by making use of a characterization of so-called graphical sequences (i.e. degree sequences that can give rise to a bipartite graph) in terms of certain constraints that can be easily checked during the course of the simulation. Introducing these types of constraints on the support of Q together with asymptotic approximations may help produce efficient importance sampling estimators (in terms of the discussion given in Section 2). However, in our current context, the vanilla version of the importance

sampling procedure proposed in the previous paragraph will already be proved to be efficient.

Let us now provide an explicitly description of the importance sampling algorithm proposed by Chen et al (2005) in its basic form, excluding the refinement idea. This is also the algorithm generated by our description of $Q(\cdot)$ which we analyze later.

Algorithm 1

STEP 1 Order the c_i's so that $c_1 \geq ... \geq c_n$ and set $\mathbf{s} \longleftarrow \mathbf{r}$, $\boldsymbol{\rho} \longleftarrow \mathbf{c}$, $L \longleftarrow 1$ and $l \longleftarrow 0$.

STEP 2 Let $\mathcal{A} = \{i : s_i > 0\}$ and define $m_{\mathcal{A}} = card(\mathcal{A})$. Put $c_1 \longleftarrow \rho_1$, $\boldsymbol{\rho} \longleftarrow (\rho_2, ..., \rho_n)$ and $l \longleftarrow l + 1$. If $m_{\mathcal{A}} < c_1$ put $L = 0$ and GO TO STEP 3. Otherwise, if $l < n$ then evaluate

$$\gamma \longleftarrow [\boldsymbol{\rho}]_2 / 2 [\boldsymbol{\rho}]_1^2$$

else, if $n = 0$ set $\gamma = 0$. Sample $(Y_{i_1}, ..., Y_{i_{m_{\mathcal{A}}}})$ according to the distribution

$$P\left(Y_{i_1} = y_{i_1}, ..., Y_{i_{m_{\mathcal{A}}}} = y_{i_{m_{\mathcal{A}}}}\right) = \frac{1}{w'} \prod_{j=1}^{m_{\mathcal{A}}} (s_j \exp(-2\gamma s_j))^{y_{i_j}},$$

where $\sum_{j=1}^{m_{\mathcal{A}}} y_{i_j} = c_1$, $y_{i_j} \in \{0, 1\}$ and

$$w' = \sum_{\{(y_{i_1}, ..., y_{i_{m_{\mathcal{A}}}}) : y_{i_1} + ... + y_{i_j} = c_1\}} \prod_{j=1}^{m_{\mathcal{A}}} (s_j \exp(-2\gamma s_j))^{y_{i_j}}.$$

Then update

$$L \longleftarrow \frac{w'}{\prod_{j=1}^{m_{\mathcal{A}}} (s_j \exp(-2\gamma s_j))^{Y_{i_j}}} L$$

and for $j \in \mathcal{A}$ put $\mathbf{s}_j \longleftarrow \mathbf{s}_j - Y_j$.

STEP 3 If $L = 0$ or $n = 0$ output L and STOP, otherwise, GO TO STEP 2.

The sampling procedure required to generate the Y_{i_k}'s in STEP 2 can be done in several ways, as is explained in Chen, Dempster and Liu (1994) and Chen and Liu (1997). Chen et al (2005) use the drafting method of Chen, Dempster and Liu (1994). The detailed algorithm and complexity analysis of Chen, Dempster and Liu (1994) implies that the drafting method, which we suggest to apply in STEP 2, can be implemented in $O(mc_k)$ operations (this computational effort includes also the cost of evaluating w' and corresponds to the generation of $(Y_{i_1}, ..., Y_{i_{m_{\mathcal{A}}}})$). This analysis implies that the computational cost per replication of an importance sampling algorithm based on $Q(\cdot)$ is of order $O(md + n) = O(d^2)$ as $d \nearrow \infty$. The contribution of the term n corresponds to ordering the c_j's in STEP 1 and computing γ in STEP 2. Note that subsequent updates of γ can be done recursively so there is

no need to add an extra factor from the fact that the algorithm goes through STEP 2 n times. Hence, according to our discussion in the previous section, if we show that the importance sampling estimator induced by $Q(\cdot)$ achieves exponential efficiency we would obtain that the overall complexity of the importance sampling algorithm generated by $Q(\cdot)$ is $O(d^2)$ as $d \nearrow \infty$. We shall develop this result in the next section.

Proof of Theorem 1. We follow closely the steps in the proof of Lemma 2.2 and Theorem 1.3 in Greenhill, McKay and Wang (2006) (GMW). First, we introduce their counting model: We consider a set of d labeled points arranged on m cells, say $\varrho_1, ..., \varrho_m$. The cell ϱ_i contains r_i elements. Similarly, we consider another set of d labeled points arranged in n cells denoted by $\Xi_1, ..., \Xi_n$ and assume that the j-th cell Ξ_j contains c_j elements. We then have $2d$ labeled points in total. A partition of the $2d$ elements into d unordered pairs is called a *pairing*. Each pair is denoted by $e = (\rho, \xi)$ where $\rho \in \varrho_i$ for some $1 \leq i \leq m$ and $\xi \in \Xi_j$ for some $1 \leq j \leq n$. We also write $v(\rho)$ to denote the cell corresponding to the point ρ and similarly $v(\xi)$ to denote the cell corresponding to the point ξ. A *random pairing* is a pairing that is chosen uniformly at random out of the $d!$ possible pairings. Two pairs are called *parallel* if they involve the same cells. An *error* is an unordered set of two parallel pairs.

It follows easily that the probability of obtaining $l \geq 0$ given pairs occurring in a random pairing is $1/[d]_l$. Let p_d be the probability that no errors occur in a random pairing. As noted by GMW, we have

$$\mu(\mathbf{r}, \mathbf{d})\, \mathbf{r}!\mathbf{c}! = d! p_d$$

because (up to a permutation in the labels of the elements in each of the cells, each contingency table corresponds to a pairing that has no errors). Therefore, it suffices to estimate p_d which is done, once again following GMW, using inclusion-exclusion and Bonferrioni's inequalities. Note that $p_d = 1 - \overline{p}_d$, where $\overline{p}_d$ is the probability that at least one error occurs in a random pairing. In turn, $\overline{p}_d$ is less or equal to the total contribution corresponding to placements with one error, which we denote by $\overline{b}_d^{(1)}$. More generally, let us define $\overline{b}_d^{(k)}$ as the total contribution in the inclusion-exclusion development corresponding to pairings that contain k errors or more. We then have that for $k \geq 1$

$$\overline{b}_d^{(1)} - \overline{b}_d^{(2)} + ... + \overline{b}_d^{(2k-1)} - \overline{b}_d^{(2k)} \leq \overline{p}_d \leq \overline{b}_d^{(1)} - \overline{b}_d^{(2)} + ... + \overline{b}_d^{(2k-1)}. \tag{7}$$

Note that $\overline{b}_d^{(k)}$, $k \geq 2$, can be divided in two parts, namely, one that contains errors that do not have a pair in common, which we denote by $\beta_{d,0}^{(k)}$, and another part that contains errors that have pairs in common, which we denote by $\beta_{d,1}^{(k)}$. We define

$\beta_{d,0}^{(1)} = \overline{b}_d^{(1)}$ and note that

$$\overline{b}_d^{(1)} = \frac{1}{d\,(d-1)} \left(\sum_{i=1}^{n} c_i\,(c_i-1) \sum_{j=1}^{m} r_j\,(r_j-1) \right) /2 = \alpha\,(\mathbf{r},\mathbf{c}) + o\,(1)$$

as $d \nearrow \infty$. We claim that for each $k \geq 2$

$$\beta_{d,0}^{(k)} = \frac{\alpha\,(\mathbf{r},\mathbf{c})^k}{k!} + o\,(1) \tag{8}$$

as $d \nearrow \infty$. To see this let us first define $\mathcal{N}_k^0$ as the set of ordered $2k$-tuples of pairs $(e_1, e_2, ..., e_k, e_1', ..., e_k')$ (with $e_j = \left(\rho_j, \xi_j\right)$ and $e_j' = \left(\rho_j', \xi_j'\right)$) satisfying for each $l \in \{1, ..., k\}$, $i_l = i_l'$ and $j_l = j_l'$ where

$$\begin{aligned} v\,(\rho_l) &= i_l, \;\; v\,(\rho_l') = i_l, \\ v\,(\xi_l) &= j_l, \;\; v\,(\xi_l') = i_l \end{aligned}$$

with $i_l \neq i_s$ and $j_l \neq j_s$ is $l \neq s$.

Note that

$$\beta_{d,0}^{(k)} = \frac{|\mathcal{N}_k|}{2^k k!} \frac{1}{[d]_{2k}}.$$

We claim that

$$|\mathcal{N}_{k+1}| = |\mathcal{N}_k| \left(\left(\sum_{i=1}^{n} [c_i]_2 \sum_{j=1}^{m} [r_j]_2 \right) + o\left(d^2\right) \right). \tag{9}$$

To verify this claim let us pick an arbitrary element $(e_1, e_2, ..., e_k, e_1', ..., e_k') \in \mathcal{N}_k$. We obtain an element of $\mathcal{N}_{k+1}$ by adding two parallel pairs (e_{k+1}, e_{k+1}') so that we obtain $k+1$ errors that do not have pairs in common. This is achieved in

$$\left(\sum_{i=1}^{n} [c_i]_2 - \sum_{l=1}^{k} [c_{i_l}]_2 \right) \left(\sum_{j=1}^{m} [r_j]_2 - \sum_{l=1}^{k} [r_{j_l}]_2 \right)$$

many ways. Now, since $\max_{j \leq n} c_j = o\left(d^{1/2}\right)$ we have that

$$\frac{\sum_{l=1}^{k} [c_{i_l}]_2}{d} \leq O\left(\frac{\max_{j \leq n} c_j^2}{d} \right) \longrightarrow 0$$

as $d \nearrow \infty$ (a completely analogous estimate also applied to the sum involving the r_{j_l}'s). This implies (9) and, as a consequence, (8). To study $\beta_{d,1}^{(k)}$ it suffices to perform a very rough analysis. Indeed, note that

$$\beta_{d,1}^{(k)} = \sum_{l=1}^{\lfloor k/2 \rfloor} O\left(\frac{\sum_{i=1}^{n} [c_i]_{2+l} \sum_{j=1}^{m} [r_i]_{2+l}}{[d]_{2k-l}} \right),$$

where l-th term in the previous sum corresponds to the cases in which there are pairs that are common to l errors. Note that the sum goes up to $\lfloor k/2 \rfloor$ because, by the pigeon hole principle, there cannot be cases for which there are $\lfloor k/2 \rfloor + 1$ pairs that are common to all k errors (otherwise we would have more than k errors meaning that we would be counting at least one error twice). Now, we have that

$$\begin{aligned} & \frac{\sum_{i=1}^{n} c_i^{2+l} \sum_{j=1}^{m} r_i^{2+l}}{d^{2k-l}} \\ \leq & \left(\frac{\max_{i \leq n} c_i}{d^{(k-1)/l - 1/2}} \right)^l \times \left(\frac{\max_{j \leq n} r_j}{d^{(k-1)/l - 1/2}} \right)^l \frac{\sum_{i=1}^{n} c_i^2 \sum_{j=1}^{m} r_i^2}{d^2} \longrightarrow 0. \end{aligned}$$

as $d \nearrow \infty$ (and $m, n \nearrow \infty$) because of our assumptions since $(k-1) \geq \lfloor k/2 \rfloor \geq l$ for $k \geq 2$. We then conclude, combining the previous estimate together with (8), that

$$\overline{b}_d^{(k)} = \beta_{d,0}^{(k)} + \beta_{d,1}^{(k)} = \frac{\alpha(\mathbf{r}, \mathbf{c})^k}{k!} + o(1)$$

as $d \nearrow \infty$. In order to complete the argument recall that for each $c \in (0, \infty)$,

$$\lim_{k \longrightarrow 0} \sup_{0 \leq x \leq c} \left| \exp(x) \sum_{j=0}^{k} (-1)^j \frac{x^j}{j!} - 1 \right| = 0.$$

Under our current assumptions we have that $\alpha(\mathbf{r}, \mathbf{c}) = O(1)$, therefore the previous estimate for the exponential function together with (7) yields the conclusion of the result. ■

5 Complexity Analysis

This section is dedicated to the proof of the following theorem which is our main result.

Theorem 2 *Suppose that* $\max_{i \leq m} r_i = O(1)$, $\max_{j \leq n} c_j = o\left(d^{1/2}\right)$ *and that* $[\mathbf{c}^2]_1 = O(d)$ *as* $d \nearrow \infty$.

i) Then, the estimator L provided by Algorithm 1 is strongly efficient as $d \nearrow \infty$. Therefore, since each replication of L requires $O(d^2)$ operations, the computational complexity required to estimate $u(\mathbf{r}, \mathbf{c})$ with ε-relative precision and $(1-\delta)$ confidence is of order $O\left(\varepsilon^2 \delta^{-1} d^2\right)$ as $d \nearrow \infty$ and ε, $\delta \searrow 0$.

ii) Moreover, if in addition we have that $\max c_j = o\left(d^{1/4-\delta_0}\right)$ *for some* $\delta_0 > 0$ *as* $d \nearrow \infty$, *then the estimator L provided by Algorithm 1 is exponentially efficient as $d \nearrow \infty$. Consequently, $O\left(\varepsilon^{-2} \log(\delta^{-1}) d^2\right)$ operations are required to estimate $u(\mathbf{r}, \mathbf{c})$ with ε relative precision and $(1-\delta)$ confidence as $d \nearrow \infty$ and ε, $\delta \searrow 0$.*

The following basic result (whose proof is given at the end of this section) will be very useful in the analysis of the likelihood ratio produced by our importance sampler.

Lemma 2 *Let $\{x_j : j \geq 1\}$ be a sequence of positive integers and let us write $\{x_{i,n} : 1 \leq i \leq n\}$ to denote any non-increasing arrangement of the set $\{x_i : 1 \leq i \leq n\}$ so that*

$$x_{1,n} \geq x_{2,n} \geq ... \geq x_{n,n}.$$

Define $y_{k,n}^{(1)} = \sum_{j=k+1}^{n} x_{j,n}$ and $y_{k,n}^{(2)} = \sum_{j=k+1}^{n} x_{j,n}^2$ for $0 \leq k \leq n-1$. Then,

i)

$$\frac{y_{k+1,n}^{(2)}}{y_{k+1,n}^{(1)}} \leq \frac{y_{k,n}^{(2)}}{y_{k,n}^{(1)}} \leq \frac{y_{0,n}^{(2)}}{y_{0,n}^{(1)}}.$$

ii) If $y_{0,n}^{(2)}/y_{0,n}^{(1)} = O(1)$ as $n \nearrow \infty$, then there exists a constant $a > 0$ (independent of n and k) such that

$$y_{k,n}^{(2)} \leq a(n-k) \tag{10}$$

as $n \nearrow \infty$. Moreover, if $x_{1,n} = o\left(n^{\beta_0 - \delta_0}\right)$ for $0 \leq \delta_0 < \beta_0 \leq 1/2$ then we also have that

$$\frac{x_{j,n}}{y_{j-1,n}^{(1)}} \leq \frac{a^{1/2}}{1 + (n-j)^{1-\beta_0+\delta_0}}. \tag{11}$$

iii) Under the assumptions of part ii), if $\delta_0 > 0$, then

$$\sup_{n \geq 1} \sum_{j=1}^{n} \frac{x_{j,n}^{1/\beta_0}}{y_{j-1,n}^2} < \infty. \tag{12}$$

The previous result will be applied repeatedly to the sequence of c_k's which is assumed to be ordered in a non-increasing way, namely, $c_1 \geq c_2 \geq ... \geq c_n$. So, for instance, assuming that $[\mathbf{c}^2] = O(d)$, then given $\boldsymbol{\rho}_0 = \mathbf{c}$ and

$$\boldsymbol{\rho}_k = \left(\rho_{k,1}, ..., \rho_{k,size(\boldsymbol{\rho}_k)}\right) = (c_{k+1}, ..., c_n)$$

for $j \leq n-1$, (11) implies that there exists n_0 such that for all $k \leq n - n_0$ we have that then $\rho_{k,1}/[\boldsymbol{\rho}_k]_1 \leq 1/2$. Similar implications are immediate from Lemma 2 and will be invoked in our future discussion.

We now proceed with the development behind Theorem 2, we first start with part ii). By running Algorithm 1 we obtain the estimator

$$\begin{aligned} L &= L_d \triangleq \prod_{k=0}^{\tau-1} \frac{w\left(\mathbf{S}_k, \boldsymbol{\rho}_k\right)}{v\left(\mathbf{S}_{k+1}, \boldsymbol{\rho}_{k+1}\right)} I\left(\mathbf{S}_\tau = 0\right) \\ &= \prod_{k=0}^{n-1} \frac{w\left(\mathbf{S}_k, \boldsymbol{\rho}_k\right)}{v\left(\mathbf{S}_{k+1}, \boldsymbol{\rho}_{k+1}\right)} I\left(\mathbf{S}_n = 0\right) \\ &= \frac{v\left(\mathbf{r}, \mathbf{c}\right)}{v\left(\mathbf{0}, *\right)} \prod_{k=0}^{n-1} \frac{w\left(\mathbf{S}_k, \boldsymbol{\rho}_k\right)}{v\left(\mathbf{S}_k, \boldsymbol{\rho}_k\right)} I\left(\mathbf{S}_n = 0\right), \end{aligned}$$

where (as indicated before) $v\left(\mathbf{0}, *\right)$ is defined as 1. Recall that

$$v\left(\mathbf{r}, \mathbf{c}\right) \sim u\left(\mathbf{r}, \mathbf{c}\right)$$

as $d \nearrow \infty$. Therefore, in order to show exponential or strong efficiency, we must study the properties of R_d defined as

$$R_d\left(\left(\mathbf{S}_0, \boldsymbol{\rho}_0\right), ..., \left(\mathbf{S}_{n-1}, \boldsymbol{\rho}_{n-1}\right)\right) \triangleq \prod_{k=0}^{n-1} \frac{w\left(\mathbf{S}_k, \boldsymbol{\rho}_k\right)}{v\left(\mathbf{S}_k, \boldsymbol{\rho}_k\right)} I\left(\mathbf{S}_n = 0\right),$$

given $\mathbf{S}_0 = \mathbf{r}$ and $\boldsymbol{\rho}_0 = \mathbf{c}$. The analysis of R_d involves studying the ratio $w(\mathbf{s}_0, \boldsymbol{\rho}_0)/v(\mathbf{s}_0, \boldsymbol{\rho}_0)$

$$\frac{w\left(\mathbf{s}_0, \boldsymbol{\rho}_0\right)}{v\left(\mathbf{s}_0, \boldsymbol{\rho}_0\right)} = \binom{m}{\rho_{0,1}}^{-1} \sum_{(\mathbf{r},\mathbf{c}) \to (\mathbf{s},\boldsymbol{\rho})} \frac{v\left(\mathbf{s}_1, \boldsymbol{\rho}_1\right)}{v\left(\mathbf{s}_0, \boldsymbol{\rho}_0\right)}.$$

Note that

$$\frac{v\left(\mathbf{s}_1, \boldsymbol{\rho}_1\right)}{v\left(\mathbf{s}_0, \boldsymbol{\rho}_0\right)} = \frac{\varphi\left(\mathbf{s}_1, \boldsymbol{\rho}_1\right) \eta\left(\boldsymbol{\rho}_0, m\right)}{\varphi\left(\mathbf{s}_0, \boldsymbol{\rho}_0\right) \eta\left(\boldsymbol{\rho}_1, m\right)},$$

and (using the notation $\rho_{i,j}$ to denote the j-th component of the vector $\boldsymbol{\rho}_i$ for $i \in \{0, 1\}$ and recalling that $\rho_{1,i} = \rho_{0,i+1}$)

$$\frac{\eta\left(\boldsymbol{\rho}_0, m\right)}{\eta\left(\boldsymbol{\rho}_1, m\right)} = \binom{m}{\rho_{0,1}} ... \binom{m}{\rho_{0,n}} \left[\binom{m}{\rho_{0,2}} ... \binom{m}{\rho_{0,n}}\right]^{-1} = \binom{m}{\rho_{0,1}}.$$

Next, observe that $\mathbf{s}_1$ is obtained by selecting a set $\Gamma = \{i_1, ...i_{\rho_{0,1}}\}$ of (ordered) subindices and by picking the i-th component of the vector $\mathbf{s}_1$, namely $s_{1,i}$, via $\mathbf{s}_{1,i} = s_{0,i} - 1\left(i \in \Gamma\right)$. Consequently, we have

$$\frac{\varphi\left(\mathbf{s}_1, \boldsymbol{\rho}_1\right)}{\varphi\left(\mathbf{s}_0, \boldsymbol{\rho}_0\right)} = \binom{d_0}{\rho_{0,1}}^{-1} \left(\Pi_{i \in \Gamma} s_{0,i}\right) \exp\left(-\left(\alpha\left(\mathbf{s}_1, \boldsymbol{\rho}_1\right) - \alpha\left(\mathbf{s}_0, \boldsymbol{\rho}_0\right)\right)\right),$$

where $d_0 = \sum_{j=1}^{m} s_{0,j} = \sum_{j=1}^{n_0} \rho_{0,j}$ (with $n_0 = size(\boldsymbol{\rho}_0)$). Therefore

$$\frac{w(\mathbf{s}_0, \boldsymbol{\rho}_0)}{v(\mathbf{s}_0, \boldsymbol{\rho}_0)} = \binom{d_0}{\rho_{0,1}}^{-1} \sum_{(\mathbf{s}_0, \boldsymbol{\rho}_0) \to (\mathbf{s}_1, \boldsymbol{\rho}_1)} (\Pi_{i \in \Gamma} s_{0,i}) \exp\left(-(\alpha(\mathbf{s}_1, \boldsymbol{\rho}_1) - \alpha(\mathbf{s}_0, \boldsymbol{\rho}_0))\right).$$

Let us provide a more convenient expression for the previous ratio, first we write $\mathbf{s}_{0,\Gamma} = \left(s_{i_1}, ..., s_{i_{\rho_{0,1}}}\right)$ and define $\gamma = [\boldsymbol{\rho}_1]_2/(2[\boldsymbol{\rho}_1]_1^2)$. Then, (using $\mathbf{1}$ to denote the vector of ones) we have

$$\begin{aligned}
\alpha(\mathbf{s}_1, \boldsymbol{\rho}_1) &= \gamma[\mathbf{s}_1]_2 = \gamma\left([\mathbf{s}_0]_2 - [2(\mathbf{s}_{0,\Gamma} - \mathbf{1})]_1\right), \\
\alpha(\mathbf{s}_0, \boldsymbol{\rho}_0) &= \left(\gamma\left(1 - \frac{\rho_{0,1}}{[\boldsymbol{\rho}_0]_1}\right)^2 + \frac{\left[\rho_{0,1}\right]_2}{2[\boldsymbol{\rho}_0]_1^2}\right)[\mathbf{s}_0]_2 \\
&= \left(\gamma - \frac{2\gamma\rho_{0,1}}{[\boldsymbol{\rho}_0]_1} + \gamma\frac{\rho_{0,1}^2}{[\boldsymbol{\rho}_0]_1^2} + \frac{\left[\rho_{0,1}\right]_2}{2[\boldsymbol{\rho}_0]_1^2}\right)[\mathbf{s}_0]_2 \\
&= \gamma[\mathbf{s}_0]_2 - \frac{2\gamma\rho_{0,1}[\mathbf{s}_0]_2}{[\boldsymbol{\rho}_0]_1} + \frac{\gamma\rho_{0,1}^2[\mathbf{s}_0]_2}{[\boldsymbol{\rho}_0]_1^2} + \frac{\left[\rho_{0,1}\right]_2[\mathbf{s}_0]_2}{2[\boldsymbol{\rho}_0]_1^2}.
\end{aligned}$$

Therefore we have

$$\begin{aligned}
&\alpha(\mathbf{s}_1, \boldsymbol{\rho}_1) - \alpha(\mathbf{s}_0, \boldsymbol{\rho}_0) \\
= \; &-\gamma\left[2(\mathbf{s}_{0,\Gamma} - \mathbf{1})\right]_1 + \frac{2\gamma\rho_{0,1}[\mathbf{s}_0]_2}{[\boldsymbol{\rho}_0]_1} - \frac{\gamma\rho_{0,1}^2[\mathbf{s}_0]_2}{[\boldsymbol{\rho}_0]_1^2} - \frac{\left[\rho_{0,1}\right]_2[\mathbf{s}_0]_2}{2[\boldsymbol{\rho}_0]_1^2}. \qquad (13)
\end{aligned}$$

Define $\beta(\mathbf{s}_0, \boldsymbol{\rho}_0)$ via

$$\log\beta(\mathbf{s}_0, \boldsymbol{\rho}_0) = -\frac{2\gamma\rho_{0,1}[\mathbf{s}_0]_2}{[\boldsymbol{\rho}_0]_1} + \frac{\gamma\rho_{0,1}^2[\mathbf{s}_0]_2}{[\boldsymbol{\rho}_0]_1^2} + \frac{\left[\rho_{0,1}\right]_2[\mathbf{s}_0]_2}{2[\boldsymbol{\rho}_0]_1^2}$$

and

$$h(\mathbf{s}_{0,\Gamma}, s_{0,1}) = \Pi_{j=1}^{\rho_{0,1}}\left(s_{0,i_j}\exp\left(2\gamma\left(s_{0,i_j} - 1\right)\right)\right).$$

We now are ready to provide an estimate for the ratio $w(\mathbf{s}_0, \boldsymbol{\rho}_0)/v(\mathbf{s}_0, \boldsymbol{\rho}_0)$.

Lemma 3 *Assuming that* $\max_{i \leq m} r_i = O(d)$ *and that* $[\mathbf{c}^2]_1 = O(d)$ *as* $d \nearrow \infty$ *there exists a constant* $\lambda \in (0, \infty)$ *such that*

$$\frac{w(\mathbf{s}_0, \boldsymbol{\rho}_0)}{v(\mathbf{s}_0, \boldsymbol{\rho}_0)} \leq \exp\left(\lambda\frac{\left[\rho_{0,1}\right]_4}{[\boldsymbol{\rho}_0]_1^2}\right).$$

Proof. Let us define $\rho_{0,1}$ iid random variables J_1, J_2,..., $J_{\rho_{0,1}}$ with distribution

$$\widetilde{P}\left(J_1 = j\right) = \frac{\exp\left(2\gamma s_{0,j}\right) s_{0,j}}{\widetilde{w}},$$

where $\widetilde{w} = \sum_{j=1}^{m} \exp\left(2\gamma s_{0,j}\right) s_{0,j}$ and $m = size\left(\mathbf{s}_0\right)$. In addition, we shall use $\widetilde{E}\left(\cdot\right)$ to denote the expectation operator associated with $\widetilde{P}\left(\cdot\right)$ and define the event $A = \left\{J_i \neq J_j : i \neq j\right\}$ (i.e. all the J_i's are different). We have

$$\binom{[\mathbf{s}_0]_1}{\rho_{0,1}}^{-1} \sum_{\Gamma \subset \{1,...,m\}} h\left(\mathbf{s}_{0,\Gamma}, s_{0,1}\right) = \binom{[\mathbf{s}_0]_1}{\rho_{0,1}}^{-1} \widetilde{w}^{\rho_{0,1}} \exp\left(-2\gamma\rho_{0,1}\right) \widetilde{P}\left(A\right).$$

Let us first analyze $\widetilde{w}$. Note that under our assumptions

$$\begin{aligned} \widetilde{w} &= \sum_{j=1}^{m} s_{0,j} \left(1 + 2\gamma s_{0,j} + \left(2\gamma\right)^2 \frac{s_{0,j}^2}{2!} + ...\right) \\ &\leq [\mathbf{s}_0]_1 \exp\left(\frac{2\gamma[\mathbf{s}_0^2]_1}{[\mathbf{s}_0]_1} + O\left(\frac{\gamma^2[\mathbf{s}_0^3]_1}{[\mathbf{s}_0]_1}\right)\right). \end{aligned}$$

We then conclude that

$$\begin{aligned} & \binom{[\mathbf{s}_0]_1}{\rho_{0,1}}^{-1} \widetilde{w}^{\rho_{0,1}} \\ \leq & \binom{[\mathbf{s}_0]_1}{\rho_{0,1}}^{-1} [\mathbf{s}_0]_1^{\rho_{0,1}} \times \exp\left(\frac{2\gamma\rho_{0,1}[\mathbf{s}_0^2]_1}{[\mathbf{s}_0]_1} + O\left(\frac{\gamma^2[\mathbf{s}_0^3]_1}{[\mathbf{s}_0]_1}\right)\right). \\ = & \prod_{k=1}^{\rho_{0,1}-1} \left(1 - \frac{k}{[\mathbf{s}_0]_1}\right)^{-1} \times \exp\left(\frac{2\gamma\rho_{0,1}[\mathbf{s}_0^2]_1}{[\mathbf{s}_0]_1} + O\left(\frac{\gamma^2[\mathbf{s}_0^3]_1}{[\mathbf{s}_0]_1}\right)\right). \end{aligned} \tag{14}$$

Assuming $\rho_{0,1}/[\mathbf{s}_0]_1 \leq 1/2$ we have

$$-\sum_{k=1}^{\rho_{0,1}-1} \log\left(1 - \frac{k}{[\mathbf{s}_0]_1}\right) \leq \frac{[\rho_{0,1}]_2}{2[\mathbf{s}_0]_1} + \frac{\rho_{0,1}\left(\rho_{0,1}+1\right)\left(2\rho_{0,1}+1\right)}{6[\mathbf{s}_0]_1^2}$$

and therefore we have

$$\begin{aligned} & \log\left(\binom{[\mathbf{s}_0]_1}{\rho_{0,1}}^{-1} \widetilde{w}^{\rho_{0,1}}\right) \\ \leq & \frac{2\gamma\rho_{0,1}[\mathbf{s}_0^2]_1}{[\mathbf{s}_0]_1} + \frac{[\rho_{0,1}]_2}{[\mathbf{s}_0]_1} + \frac{\rho_{0,1}\left(\rho_{0,1}+1\right)\left(2\rho_{0,1}+1\right)}{6[\mathbf{s}_0]_1^2} + O\left(\frac{\gamma^2[\mathbf{s}_0^3]_1\rho_{0,1}}{[\mathbf{s}_0]_1}\right). \end{aligned}$$

We now estimate $\widetilde{P}(A^c)$ using the inclusion-exclusion principle and Bonferrioni's inequalities. We have that

$$\widetilde{P}(A^c) \leq \binom{\rho_{0,1}}{2} \frac{1}{\widetilde{w}^2} \sum_{j=1}^{m} s_{0,j}^2 \exp(4\gamma s_{0,j}),$$

this corresponds to the union bound taking all possible ways in which $\{J_{i_1} = J_{i_2}\}$ for $i_1 \neq i_2$. Next we obtain the following lower bound corresponding to the cases in which $\{J_{i_1} = J_{i_2},\, J_{i_3} = J_{i_4}\}$,

$$\begin{aligned}
\widetilde{P}(A^c) \;\geq\; & \binom{\rho_{0,1}}{2} \frac{1}{\widetilde{w}^2} \sum_{j=1}^{m} s_{0,j}^2 \exp(4\gamma s_{0,j}) \\
& - \binom{3}{1} \binom{\rho_{0,1}}{3} \frac{1}{\widetilde{w}^3} \sum_{j=1}^{m} s_{0,j}^3 \exp(12\gamma s_{0,j}) \\
& - \binom{4}{2} \binom{\rho_{0,1}}{4} \frac{1}{\widetilde{w}^4} \left(\sum_{j=1}^{m} s_{0,j}^2 \exp(4\gamma s_{0,j}) \right)^2 .
\end{aligned}$$

We then conclude

$$\begin{aligned}
\widetilde{P}(A) \leq 1 & - \binom{\rho_{0,1}}{2} \frac{1}{\widetilde{w}^2} \sum_{j=1}^{m} s_{0,j}^2 \exp(4\gamma s_{0,j}) \\
& + \binom{3}{1} \binom{\rho_{0,1}}{3} \frac{1}{\widetilde{w}^3} \sum_{j=1}^{m} s_{0,j}^3 \exp(12\gamma s_{0,j}) \\
& + \binom{4}{2} \binom{\rho_{0,1}}{4} \frac{1}{\widetilde{w}^4} \left(\sum_{j=1}^{m} s_{0,j}^2 \exp(4\gamma s_{0,j}) \right)^2 .
\end{aligned}$$

Now, we have that

$$\begin{aligned}
& \binom{\rho_{0,1}}{2} \frac{1}{\widetilde{w}^2} \sum_{j=1}^{m} s_{0,j}^2 \exp(4\gamma s_{0,j}) \\
= \; & \binom{\rho_{0,1}}{2} \frac{[\mathbf{s}_0^2]_1}{[\mathbf{s}_0]_1} \times \frac{\left(1 + 4\gamma \, [\mathbf{s}_0^3]_1 / [\mathbf{s}_0^2]_1 + (4\gamma)^2 \, [\mathbf{s}_0^4]_1 / (2! \, [\mathbf{s}_0^2]_1) + ...\right)}{\left(1 + 2\gamma \, [\mathbf{s}_0^2]_1 / [\mathbf{s}_0]_1 + (2\gamma)^2 \, [\mathbf{s}_0^3]_1 / (2! \, [\mathbf{s}_0]_1) + ...\right)} \\
= \; & \binom{\rho_{0,1}}{2} \frac{[\mathbf{s}_0^2]_1}{[\mathbf{s}_0]_1} \left(1 + 4\gamma \frac{[\mathbf{s}_0^3]_1}{[\mathbf{s}_0^2]_1} + O\left(\gamma^2 \frac{[\mathbf{s}_0^4]_1}{[\mathbf{s}_0^2]_1}\right)\right) \\
& \times \left(1 - 2\gamma \frac{[\mathbf{s}_0^2]_1}{[\mathbf{s}_0]_1} + O\left(\gamma^2 \frac{[\mathbf{s}_0^3]_1}{[\mathbf{s}_0]_1}\right)\right) \\
= \; & \binom{\rho_{0,1}}{2} \frac{[\mathbf{s}_0^2]_1}{[\mathbf{s}_0]_1} \left(1 + 4\gamma \frac{[\mathbf{s}_0^3]_1}{[\mathbf{s}_0^2]_1} - 2\gamma \frac{[\mathbf{s}_0^2]_1}{[\mathbf{s}_0]_1} + O\left(\gamma^2 \frac{[\mathbf{s}_0^4]_1}{[\mathbf{s}_0]_1}\right)\right).
\end{aligned}$$

Note that

$$\begin{aligned}
\frac{2\left[\mathbf{s}_0^3\right]_1}{\left[\mathbf{s}_0^2\right]_1}-\frac{\left[\mathbf{s}_0^2\right]_1}{\left[\mathbf{s}_0\right]_1} &= \frac{2\left[\mathbf{s}_0^3\right]_1\left[\mathbf{s}_0\right]_1-\left[\mathbf{s}_0^2\right]_1^2}{\left[\mathbf{s}_0\right]_1\left[\mathbf{s}_0^2\right]_1} \\
&= \frac{\left[\mathbf{s}_0^4\right]_1}{\left[\mathbf{s}_0\right]_1\left[\mathbf{s}_0^2\right]_1}+2\frac{\sum_{i<j}\left(s_{0,i}^3 s_{0,j}+s_{0,i}s_{0,j}^3-s_{0,i}^2 s_{0,j}^2\right)}{\left[\mathbf{s}_0\right]_1\left[\mathbf{s}_0^2\right]_1} \\
&\geq \frac{\left[\mathbf{s}_0^4\right]_1}{\left[\mathbf{s}_0\right]_1\left[\mathbf{s}_0^2\right]_1}+2\frac{\sum_{i<j}s_{0,i}s_{0,j}\min(s_{0,j}^2,s_{0,i}^2)}{\left[\mathbf{s}_0\right]_1\left[\mathbf{s}_0^2\right]_1} \\
&\geq \frac{\left[\mathbf{s}_0^4\right]_1}{\left[\mathbf{s}_0\right]_1\left[\mathbf{s}_0^2\right]_1}+\frac{\left[\rho_{0,1}\right]_2}{\left[\mathbf{s}_0\right]_1\left[\mathbf{s}_0^2\right]_1}.
\end{aligned}$$

As a consequence,

$$\begin{aligned}
&\binom{\rho_{0,1}}{2}\frac{1}{\widetilde{w}^2}\sum_{j=1}^m s_{0,j}^2\exp\left(4\gamma s_{0,j}\right) \\
\geq\ &\binom{\rho_{0,1}}{2}\frac{\left[\mathbf{s}_0^2\right]_1}{\left[\mathbf{s}_0\right]_1^2}\times\left(1+\frac{2\gamma\left[\mathbf{s}_0^4\right]_1}{\left[\mathbf{s}_0\right]_1\left[\mathbf{s}_0^2\right]_1}+\frac{2\gamma\left[\rho_{0,1}\right]_2}{\left[\mathbf{s}_0\right]_1\left[\mathbf{s}_0^2\right]_1}+O\left(\gamma^2\frac{\left[\mathbf{s}_0^4\right]_1}{\left[\mathbf{s}_0\right]_1}\right)\right)^2 \\
\geq\ &\binom{\rho_{0,1}}{2}\frac{\left[\mathbf{s}_0^2\right]_1}{\left[\mathbf{s}_0\right]_1^2}+O\left(\frac{\left[\rho_{0,1}\right]_2^2}{\left[\mathbf{s}_0\right]_1^4}+\frac{\left[\rho_{0,1}\right]_2}{\left[\mathbf{s}_0\right]_1^3}\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\widetilde{P}\left(A\right) &\leq 1-\binom{\rho_{0,1}}{2}\frac{\left[\mathbf{s}_0^2\right]_1}{\left[\mathbf{s}_0\right]_1^2}+3\binom{\rho_{0,1}}{3}\frac{\left[\mathbf{s}_0^3\right]_1}{\left[\mathbf{s}_0\right]_1^3} \\
&\quad+\binom{4}{2}\binom{\rho_{0,1}}{4}\frac{\left[\mathbf{s}_0^2\right]_1^2}{\left[\mathbf{s}_0\right]_1^4}\left(1+4\gamma\frac{\left[\mathbf{s}_0^3\right]_1}{\left[\mathbf{s}_0^2\right]_1}\right)^2+\binom{\rho_{0,1}}{2}\frac{\left[\mathbf{s}_0^2\right]_1}{\left[\mathbf{s}_0\right]_1^2} \\
&\quad+O\left(\frac{\left[\rho_{0,1}\right]_2^2}{\left[\mathbf{s}_0\right]_1^4}+\frac{\left[\rho_{0,1}\right]_2}{\left[\mathbf{s}_0\right]_1^3}\right) \\
&= 1-\binom{\rho_{0,1}}{2}\frac{\left[\mathbf{s}_0^2\right]_1}{\left[\mathbf{s}_0\right]_1^2}+O\left(\frac{\left[\rho_{0,1}\right]_4}{\left[\mathbf{s}_0\right]_1^2}+\frac{\left[\rho_{0,1}\right]_2^2}{\left[\mathbf{s}_0\right]_1^4}+\frac{\left[\rho_{0,1}\right]_2}{\left[\mathbf{s}_0\right]_1^3}\right) \\
&\leq \exp\left(-\binom{\rho_{0,1}}{2}\frac{\left[\mathbf{s}_0^2\right]_1}{\left[\mathbf{s}_0\right]_1^2}+O\left(\frac{\left[\rho_{0,1}\right]_4}{\left[\mathbf{s}_0\right]_1^2}\right)\right).
\end{aligned}$$

We now group all of our terms together in a convenient way in order to estimate the ratio $w(\mathbf{s}_0, \boldsymbol{\rho}_0)/v(\mathbf{s}_0, \boldsymbol{\rho}_0)$. In order to do this we define the terms χ_1, χ_2 and χ_3 as

$$\begin{aligned}
\chi_1 &= -\frac{2\gamma\rho_{0,1}[\mathbf{s}_0]_2}{[\boldsymbol{\rho}_0]_1} + \frac{2\gamma\rho_{0,1}[\mathbf{s}_0^2]_1}{[\mathbf{s}_0]_1} - 2\gamma\rho_{0,1}, \\
\chi_2 &= -\binom{\rho_{0,1}}{2}\frac{[\mathbf{s}_0^2]_1}{[\mathbf{s}_0]_1^2} + \frac{\left[\rho_{0,1}\right]_2[\mathbf{s}_0]_2}{2[\boldsymbol{\rho}_0]_1^2} + \frac{[\rho_{0,1}]_2}{2[\mathbf{s}_0]_1}, \\
\chi_3 &= \frac{\rho_{0,1}\left(\rho_{0,1}+1\right)\left(2\rho_{0,1}+1\right)}{6[\mathbf{s}_0]_1^2} + \frac{\gamma\rho_{0,1}^2[\mathbf{s}_0]_2}{[\boldsymbol{\rho}_0]_1^2} + O\left(\frac{\left[\rho_{0,1}\right]_4}{[\mathbf{s}_0]_1^2}\right) \\
&= O\left(\frac{\left[\rho_{0,1}\right]_4}{[\mathbf{s}_0]_1^2}\right)
\end{aligned}$$

We have that if $\rho_{0,1}/[\boldsymbol{\rho}_0]_1 \leq 1/2$ then

$$\log\left(\frac{w(\mathbf{s}_0, \boldsymbol{\rho}_0)}{v(\mathbf{s}_0, \boldsymbol{\rho}_0)}\right) \leq \chi_1 + \chi_2 + O\left(\frac{\left[\rho_{0,1}\right]_4}{[\mathbf{s}_0]_1^2}\right).$$

Now we compute χ_1 and χ_2, first we have that

$$\begin{aligned}
-\frac{2\gamma\rho_{0,1}[\mathbf{s}_0]_2}{[\boldsymbol{\rho}_0]_1} + \frac{2\gamma\rho_{0,1}[\mathbf{s}_0^2]_1}{[\mathbf{s}_0]_1} &= \frac{2\gamma\rho_{0,1}}{[\mathbf{s}_0]_1}\left(-[\mathbf{s}_0]_2 + [\mathbf{s}_0^2]_1\right) \\
&= \frac{2\gamma\rho_{0,1}}{[\mathbf{s}_0]_1}[\mathbf{s}_0]_1 = 2\gamma\rho_{0,1}.
\end{aligned}$$

Therefore, we have that $\chi_1 = 0$. A similar computation yields $\chi_2 = 0$, indeed

$$\begin{aligned}
&-\binom{\rho_{0,1}}{2}\frac{[\mathbf{s}_0^2]_1}{[\mathbf{s}_0]_1^2} + \frac{\left[\rho_{0,1}\right]_2[\mathbf{s}_0]_2}{2[\boldsymbol{\rho}_0]_1^2} + \frac{[\rho_{0,1}]_2}{[\mathbf{s}_0]_1} \\
= \ & \frac{\left[\rho_{0,1}\right]_2}{2[\mathbf{s}_0]_1^2}\left(-\left[\mathbf{s}_0^2\right]_1 + [\mathbf{s}_0]_2\right) + \frac{[\rho_{0,1}]_2}{2[\mathbf{s}_0]_1} = -\frac{\left[\rho_{0,1}\right]_2}{2[\mathbf{s}_0]_1^2}[\mathbf{s}_0]_1 + \frac{[\rho_{0,1}]_2}{2[\mathbf{s}_0]_1} = 0.
\end{aligned}$$

The result of the lemma then follows. ■

A consequence of the previous result is the following corollary.

Corollary 3 *Assuming that* $\max_{i\leq m} r_i = O(d)$, $[\mathbf{c}^2]_1 = O(d)$ *and* $\max_{j\leq n} c_j = o\left(d^{1/4-\delta_0}\right)$ *as* $d \nearrow \infty$, *then there exists a constant* $\lambda^* \in (0,\infty)$ *independent of* d *such that*

$$R_d\left((\mathbf{S}_0, \boldsymbol{\rho}_0), ..., \left(\mathbf{S}_{n-1}, \boldsymbol{\rho}_{n-1}\right)\right) \leq \lambda^*.$$

Proof. Iterating the estimate obtained in Lemma 3 we obtain that

$$R_d\left(\left(\mathbf{S}_0,\boldsymbol{\rho}_0\right),...,\left(\mathbf{S}_{n-1},\boldsymbol{\rho}_{n-1}\right)\right)\leq\kappa_0\exp\left(\lambda\sum_{k=0}^{n-1}\frac{\left[\rho_{k,1}\right]_4}{\left[\boldsymbol{\rho}_k\right]_1^2}\right).$$

The result then follows as a consequence of (12) in Lemma 2. ■

With Corollary 3 at hand we have all what is needed to establish exponential efficiency. However, before we put all the pieces together let us continue with the basic elements behind the strong efficiency properties indicated in part i) of Theorem 2. We then will conclude with a summary of all our results and the complete proof of Theorem 2.

In order to establish strong efficiency we must study the function

$$g\left(\mathbf{s}_0,\boldsymbol{\rho}_0\right)\triangleq E^Q_{\mathbf{s}_0,\boldsymbol{\rho}_0}\left(R_d^2\right)=E^Q_{\mathbf{s}_0,\boldsymbol{\rho}_0}\left(\prod_{k=0}^{n-1}\frac{w^2\left(\mathbf{S}_k,\boldsymbol{\rho}_k\right)}{v^2\left(\mathbf{S}_k,\boldsymbol{\rho}_k\right)}I\left(\mathbf{S}_n=0\right)\right).$$

In particular, we must show that $g\left(\mathbf{s}_0,\boldsymbol{\rho}_0\right)$ remains bounded as $\left[\mathbf{s}_0\right]_1\nearrow\infty$. Our strategy is to derive a linear inequality for $g\left(\cdot\right)$ and show that one can satisfy this inequality with a convenient Lyapunov function $f\left(\cdot\right)$ that remains bounded as $d\nearrow\infty$. The next result provides sufficient conditions for the construction of an appropriate Lyapunov function. The corresponding proof is given at the end of the section.

Proposition 1 *Assume $f\geq 1$ is a function that satisfies*

$$f\left(\mathbf{s}_0,\boldsymbol{\rho}_0\right)\geq\frac{w^2\left(\mathbf{s}_0,\boldsymbol{\rho}_0\right)}{v^2\left(\mathbf{s}_0,\boldsymbol{\rho}_0\right)}E^Q_{\mathbf{s}_0,\boldsymbol{\rho}_0}f\left(\mathbf{S}_1,\boldsymbol{\rho}_1\right)\tag{15}$$

as long as $\left[\mathbf{s}_0\right]_1\geq d_0$ (for some $d_0\in(0,\infty)$ fixed). Then,

$$g\left(\mathbf{s}_0,\boldsymbol{\rho}_0\right)\leq\kappa_{d_0}f\left(\mathbf{s}_0,\boldsymbol{\rho}_0\right)$$

where $\kappa_{d_0}=\sup_{\left[\mathbf{s}_0\right]_1\leq d_0}g\left(\mathbf{s}_0,\boldsymbol{\rho}_0\right)<\infty$.

A function $f\left(\cdot\right)$ satisfying the hypothesis of Proposition 1 is typically called a Lyapunov function in the context of stability of Markov chains (see Meyn and Tweedie (1993)). Our goal is to build a bounded Lyapunov function $f\left(\cdot\right)$. Similar Lyapunov-type bounds have been studied in the rare-event simulation literature, see for instance Blanchet and Glynn (2007) for applications in the context of rare-event estimation problems related to first passage time probabilities.

Constructing an appropriate Lyapunov function $f\left(\cdot\right)$ is typically not a simple task. Nevertheless, such construction is often guided by a solid understanding of the

estimates involved in the ratio of $w(\mathbf{s}_0, \boldsymbol{\rho}_0)/v(\mathbf{s}_0, \boldsymbol{\rho}_0)$. This is precisely the strategy that we use to construction our Lyapunov function. In particular, we first put for $\theta_0 > 0$

$$f_0(\boldsymbol{\rho}) = \exp\left(\theta_0 \sum_{j=1}^{size(\boldsymbol{\rho})} \frac{\rho_j}{\left(\sum_{k=j}^{n} \rho_k\right)^2}\right),$$

then set

$$f_1(\mathbf{s}, \boldsymbol{\rho}) = \exp\left(\theta_1 \alpha(\mathbf{s}, \boldsymbol{\rho})\right),$$

for some $\theta_1 > 0$ and finally define

$$f(\mathbf{s}, \boldsymbol{\rho}) = f_0(\boldsymbol{\rho}) f_1(\mathbf{s}, \boldsymbol{\rho}).$$

The form of this function was obtained by inspecting carefully the analysis behind Lemma 3. We first tried a Lyapunov function such as f_1 and then, after doing some computations, recognized the need for a term such as f_0 as the proof of the next lemma indicates.

Lemma 4 *There exists $\theta_1, \theta_2 > 0$ such that $f(\cdot)$ satisfies the conditions of Proposition 1.*

Proof. The proof proceeds along the same lines as that of Lemma 3. Given $(\mathbf{s}_0, \boldsymbol{\rho}_0)$ let us denote $(\mathbf{s}_1, \boldsymbol{\rho}_1)$ an admissible transition step (so that $(\mathbf{s}_0, \boldsymbol{\rho}_0) \rightarrow (\mathbf{s}_1, \boldsymbol{\rho}_1)$). In particular, we have that there exists a set of subindexes $\Gamma = \{i_1, ..., i_{\rho_{0,1}}\}$ such that $s_{1,j} = s_{0,j} - 1 \, (j \in \Gamma)$. We write $\gamma = [\boldsymbol{\rho}_1]_2/(2[\boldsymbol{\rho}_1]_1)$ and introduce $\rho_{0,1}$ iid random variables $J_1, ..., J_{\rho_{0,1}}$ with distribution

$$\widetilde{P}(J_1 = k) = \frac{\exp(2\gamma s_k) s_k}{\widetilde{w}}$$

where

$$\widetilde{w} = \sum_{i=1}^{m} \exp(2\gamma s_i) s_i.$$

We have that

$$\begin{aligned}
\frac{f_1(\mathbf{s}_1, \boldsymbol{\rho}_1)}{f_1(\mathbf{s}_0, \boldsymbol{\rho}_0)} &= \exp\left(\theta_1 \left(\alpha(\mathbf{s}_1, \boldsymbol{\rho}_1) - \alpha(\mathbf{s}_0, \boldsymbol{\rho}_0)\right)\right) \\
&= \exp\left(\theta_1 \frac{2\gamma\rho_{0,1}[\mathbf{s}_0]_2}{[\boldsymbol{\rho}_0]_1} + 2\theta_1\gamma\rho_{0,1} - \theta_1 \frac{\gamma\rho_{0,1}^2[\mathbf{s}_0]_2}{[\boldsymbol{\rho}_0]_1^2}\right) \\
&\quad \times \exp\left(-\theta_1 \frac{\left[\rho_{0,1}\right]_2 [\mathbf{s}_0]_2}{2[\boldsymbol{\rho}_0]_1^2} - 2\theta_1\gamma [\mathbf{s}_{0,\Gamma}]_1\right),
\end{aligned}$$

where $\mathbf{s}_{0,\Gamma} = (s_{0,j_1}, ..., s_{0,j_{\rho_{0,1}}})$. We need to show that there exists d_0, θ_1 and θ_2 such that for $[\mathbf{s}]_1 \geq d_0$ we have

$$\begin{aligned}
& \exp\left(\theta_2 \frac{\rho_{0,1}}{[\boldsymbol{\rho}_0]_1^2}\right) \\
\geq & \frac{w(\mathbf{s}_0, \boldsymbol{\rho}_0)^2}{v(\mathbf{s}_0, \boldsymbol{\rho}_0)^2} E^Q_{\mathbf{s},\boldsymbol{\rho}}\left(\frac{f_1(\mathbf{S}_1, \boldsymbol{\rho}_1)}{f_1(\mathbf{s}_0, \boldsymbol{\rho}_0)}\right) \\
= & \frac{w(\mathbf{s}_0, \boldsymbol{\rho}_0)^2}{v(\mathbf{s}_0, \boldsymbol{\rho}_0)^2} \sum_{(\mathbf{s}_0, \boldsymbol{\rho}_0) \to (\mathbf{s}_1, \boldsymbol{\rho}_1)} \binom{m}{\rho_{0,1}}^{-1} \frac{f_1(\mathbf{s}_1, \boldsymbol{\rho}_1)}{f_1(\mathbf{s}_0, \boldsymbol{\rho}_0)} \frac{v(\mathbf{s}_1, \boldsymbol{\rho}_1)}{w(\mathbf{s}_0, \boldsymbol{\rho}_0)} \\
= & \frac{w(\mathbf{s}_0, \boldsymbol{\rho}_0)}{v(\mathbf{s}_0, \boldsymbol{\rho}_0)} \sum_{(\mathbf{s}_0, \boldsymbol{\rho}_0) \to (\mathbf{s}_1, \boldsymbol{\rho}_1)} \binom{m}{\rho_{0,1}}^{-1} \frac{f_1(\mathbf{s}_1, \boldsymbol{\rho}_1)}{f_1(\mathbf{s}_0, \boldsymbol{\rho}_0)} \frac{v(\mathbf{s}_1, \boldsymbol{\rho}_1)}{v(\mathbf{s}_0, \boldsymbol{\rho}_0)}.
\end{aligned}$$

As in Lemma 3 we have that

$$\begin{aligned}
& \sum_{(\mathbf{s}_0, \boldsymbol{\rho}_0) \to (\mathbf{s}_1, \boldsymbol{\rho}_1)} \binom{m}{\rho_{0,1}}^{-1} \frac{f_1(\mathbf{s}_1, \boldsymbol{\rho}_1)}{f_1(\mathbf{s}_0, \boldsymbol{\rho}_0)} \frac{v(\mathbf{s}_1, \boldsymbol{\rho}_1)}{v(\mathbf{s}_0, \boldsymbol{\rho}_0)} \\
= & \binom{[\mathbf{s}_0]_1}{\rho_{0,1}}^{-1} \widetilde{w}^{\rho_{0,1}} \exp\left((\theta_1 - 1)\frac{2\gamma\rho_{0,1}[\mathbf{s}_0]_2}{[\boldsymbol{\rho}_0]_1} + 2(\theta_1 - 1)\gamma\rho_{0,1}\right) \\
& \times \exp\left(-(\theta_1 - 1)\frac{\gamma\rho_{0,1}^2[\mathbf{s}_0]_2}{[\boldsymbol{\rho}_0]_1^2} - (\theta_1 - 1)\frac{\left[\rho_{0,1}\right]_2[\mathbf{s}_0]_2}{2[\boldsymbol{\rho}_0]_1^2}\right) \\
& \times \widetilde{E}\left(\exp\left(\sum_{i=1}^{\rho_{0,1}} -2\theta\gamma s_{0,J_i}\right); A\right)
\end{aligned}$$

where A is the event that consists that all the J_i's are distinct. During the proof of Lemma 3 we obtained that if $\rho_{0,1}/[\mathbf{s}_0]_1 \leq 1/2$ then

$$\begin{aligned}
& \log\left(\binom{[\mathbf{s}_0]_1}{\rho_{0,1}}^{-1} \widetilde{w}^{\rho_{0,1}}\right) \\
\leq & \frac{2\gamma\rho_{0,1}[\mathbf{s}_0^2]_1}{[\mathbf{s}_0]_1} + \frac{[\rho_{0,1}]_2}{[\mathbf{s}_0]_1} + O\left(\frac{\left[\rho_{0,1}\right]_4}{[\boldsymbol{\rho}_0]_1^2}\right).
\end{aligned}$$

Now, evidently we have that

$$\begin{aligned}
& \widetilde{E}\left(\exp\left(\sum_{i=1}^{\rho_{0,1}} -2\theta_1\gamma s_{0,J_i}\right); A\right) \\
\leq & \left(\widetilde{E}\left(\exp\left(-2\theta_1\gamma s_{0,J_i}\right)\right)\right)^{\rho_{0,1}} \leq \exp\left(-2\theta_1\gamma\rho_{0,1}\widetilde{E}s_{0,J_i} + O\left(\theta_1\rho_{0,1}\gamma^2\right)\right).
\end{aligned}$$

Therefore, combining all these estimates together with Lemma 3 we have that there exists a constant $\lambda > 0$ such that

$$\begin{aligned}
&\frac{w\left(\mathbf{s}_0, \boldsymbol{\rho}_0\right)^2}{v\left(\mathbf{s}_0, \boldsymbol{\rho}_0\right)^2} E^Q_{\mathbf{s},\boldsymbol{\rho}}\left(\frac{f_1\left(\mathbf{S}_1, \boldsymbol{\rho}_1\right)}{f_1\left(\mathbf{s}_0, \boldsymbol{\rho}_0\right)}\right) \\
\leq\ & \exp\left(\lambda \frac{\rho_{0,1}^4}{[\boldsymbol{\rho}_0]_1^2} + \frac{2\gamma\rho_{0,1}[\mathbf{s}_0^2]_1}{[\mathbf{s}_0]_1} + \frac{[\rho_{0,1}]_2}{[\mathbf{s}_0]_1}\right) \\
&\times \exp\left(\left(\theta_1 - 1\right)\frac{2\gamma\rho_{0,1}[\mathbf{s}_0]_2}{[\boldsymbol{\rho}_0]_1} + 2\left(\theta_1 - 1\right)\gamma\rho_{0,1}\right) \\
&\times \exp\left(-\left(\theta_1 - 1\right)\frac{\gamma\rho_{0,1}^2[\mathbf{s}_0]_2}{[\boldsymbol{\rho}_0]_1^2} - \left(\theta_1 - 1\right)\frac{\left[\rho_{0,1}\right]_2[\mathbf{s}_0]_2}{2\left[\boldsymbol{\rho}_0\right]_1^2}\right) \\
&\times \exp\left(-\frac{2\theta_1\gamma\rho_{0,1}[\mathbf{s}_0^2]}{[\mathbf{s}_0]_1} + O\left(\theta_1\rho_{0,1}\gamma^2\right)\right). \qquad (16a)
\end{aligned}$$

Note that in the last line of the previous display we have used the fact that

$$\widetilde{E}s_{J_1}^2 = \frac{[\mathbf{s}_0^2]}{[\mathbf{s}_0]_1} + O\left(\gamma\right).$$

Now we note (just as we did in Lemma 3) that

$$-\frac{2\gamma\rho_{0,1}[\mathbf{s}_0]_2}{[\boldsymbol{\rho}_0]_1} + \frac{2\gamma\rho_{0,1}[\mathbf{s}_0^2]_1}{[\mathbf{s}_0]_1} - 2\gamma\rho_{0,1} = 0,$$

which implies that logarithm of the right hand side of (16a) equals

$$\begin{aligned}
&\lambda\frac{\left[\rho_{0,1}\right]_4}{[\boldsymbol{\rho}_0]_1^2} + \frac{[\rho_{0,1}]_2}{[\mathbf{s}_0]_1} - \left(\theta_1 - 1\right)\frac{\gamma\rho_{0,1}^2[\mathbf{s}_0]_2}{[\boldsymbol{\rho}_0]_1^2} \\
&- \left(\theta_1 - 1\right)\frac{\left[\rho_{0,1}\right]_2[\mathbf{s}_0]_2}{2\left[\boldsymbol{\rho}_0\right]_1^2} + O\left(\theta_1\rho_{0,1}\gamma^2\right).
\end{aligned}$$

It is immediate from the previous expression that one can select first $\theta_1 > 0$ and then θ_2 depending on θ_1 so that the previous quantity is less or equal to $\theta_2\rho_{0,1}/[\boldsymbol{\rho}_0]_1^2$ as long as $[\mathbf{s}_0]_1 \geq d_0$ so that $\rho_{0,1}/[\mathbf{s}_0]_1 \leq 1/2$. The conclusion of the lemma then follows. ■

It is time to summarize all the previous estimates and to complete the proof of Theorem 2.

Proof of Theorem 2. We first establish part ii). By virtue of Theorem 1 and Corollary 3 we have that

$$\frac{L_d}{u\left(\mathbf{r}, \mathbf{c}\right)} = \frac{v\left(\mathbf{r}, \mathbf{c}\right)}{u\left(\mathbf{r}, \mathbf{c}\right)} R_d \leq \frac{v\left(\mathbf{r}, \mathbf{c}\right)}{u\left(\mathbf{r}, \mathbf{c}\right)}\lambda^* = O\left(1\right)$$

as $d \nearrow \infty$. Therefore, because of our observations in Section 2, L_d is exponentially efficient and part ii) follows. Part i) is established similarly, thanks to Lemma 4. Note that

$$\frac{E^Q_{\mathbf{r},\mathbf{c}} L_d^2}{u(\mathbf{r},\mathbf{c})^2} = \frac{v(\mathbf{r},\mathbf{c})^2}{u(\mathbf{r},\mathbf{c})^2} g(\mathbf{r},\mathbf{c}) \leq \frac{v(\mathbf{r},\mathbf{c})^2}{u(\mathbf{r},\mathbf{c})^2} f(\mathbf{r},\mathbf{c}).$$

Theorem 1 guarantees that $v(\mathbf{r},\mathbf{c})^2 / u(\mathbf{r},\mathbf{c})^2 \longrightarrow 1$ as $d \nearrow \infty$. On the other hand Lemma 2 implies that $f(\mathbf{r},\mathbf{c}) = O(1)$ as $d \nearrow \infty$. This concludes the proof of Theorem 2. ■

Finally, before providing the proof of the pending results it is worth discussing the practical implications of the previous bounds. The previous results imply a bound on the coefficient of variation that involves an exponential function to a power that depends on the maximum degree of the row sums. In practical situations this bound can quickly become large, so the bounds given here, although computable, may be far too pessimistic in practical applications. Improving these bounds is particularly interesting given that, empirically according to Chen et al (2005), the estimated coefficient of variation of the estimator given by Algorithm 1 is consistently small (they report values that are even less than 1). The key issue involves controlling the behavior of the row sums during the course of the algorithm under $Q(\cdot)$. The techniques here can be adapted to deal with situations when the row sums may grow and this will be illustrated elsewhere in the future.

Proof of Lemma 2. We have that

$$\frac{y^{(2)}_{k+1,n}}{y^{(1)}_{k+1,n}} - \frac{y^{(2)}_{k,n}}{y^{(1)}_{k,n}} = \frac{y^{(2)}_{k+1,n} y^{(1)}_{k,n} - y^{(2)}_{k,n} y^{(1)}_{k+1,n}}{y^{(1)}_{k+1,n} y^{(1)}_{k,n}}.$$

Now

$$\begin{aligned} y^{(2)}_{k+1,n} y^{(1)}_{k,n} - y^{(2)}_{k,n} y^{(1)}_{k+1,n} &= \left(y^{(2)}_{k,n} - x^2_{k+1,n}\right) y^{(1)}_{k,n} - y^{(2)}_{k,n} \left(y^{(1)}_{k,n} - x_{k+1,n}\right) \\ &= x_{k+1,n} \left(y^{(2)}_{k,n} - x_{k+1,n} y^{(1)}_{k,n}\right). \end{aligned}$$

The result then follows from the fact that

$$y^{(2)}_{k,n} = \sum_{j=k+1}^{n} x^2_{j,n} \leq x_{k+1,n} \sum_{j=k+1}^{n} x_{j,n} = x_{k+1,n} y^{(1)}_{k,n}.$$

For part ii) we note that, by assumption there exists $a > 0$ such that $y^{(2)}_{0,n} \leq a^{1/2} y^{(1)}_{0,n}$. Using Cauchy-Schwartz inequality and part i) it follows that

$$y^{(2)}_{k,n} \leq a^{1/2} y^{(1)}_{k,n} \leq a^{1/2} \left((n-k)\, y^{(2)}_{k,n}\right)^{1/2},$$

which implies $y_{k,n}^{(2)} \leq a\,(n-k)$. Finally, combining part i) and the assumption that $y_{0,n}^{(2)}/y_{0,n}^{(1)} = O\,(1)$, we can write

$$\frac{x_{j,n}}{y_{j-1,n}^{(1)}} \leq a^{1/2}\frac{x_{j,n}}{y_{j-1,n}^{(2)}} = \frac{a^{1/2}}{x_{j,n} + x_{j+1,n}/x_{j,n} + ... + x_{n,n}/x_{j,n}}.$$

Now, it follows that

$$x_{j,n} + x_{j+1,n}/x_{j,n} + ... + x_{n,n}/x_{j,n} \geq 1 + (n-j)\,/x_{1,n-j}.$$

We conclude that

$$\frac{x_{j,n}}{y_{j-1,n}^{(1)}} \leq \frac{a^{1/2}}{1 + (n-j)^{1-\beta_0+\delta_0}}$$

which yields (11).

For part iii) we use (10) and (11). In particular, we have that

$$x_{j,n} \leq \frac{a^{1/2} y_{j-1,n}^{(1)}}{1 + (n-j)^{1-\beta_0+\delta_0}} \leq a\,(n-j)^{\beta_0-\delta_0}$$

and therefore

$$\sum_{j=1}^{n} \frac{x_{j,n}^{1/\beta_0}}{\left(y_{j-1,n}^{(1)}\right)^2} = O\left(\sum_{j=1}^{n-1} \frac{1}{(n-j)^{1+\delta_0/\beta_0}}\right),$$

which yields (12). ■

Proof of Proposition 1. Define $\tau_{d_0} = \inf\{k \geq 0 : \mathbf{S}_k < d_0\}$ and let $\mathcal{F}_k = \sigma\,(\mathbf{S}_j : 0 \leq j \leq k)$ be the σ-field generated by the process $\mathbf{S}_\cdot$ up to time k. As in Section 4 we let τ be the first time $k \leq n$ for which the number of strictly positive components of the vector $\mathbf{S}_k$ is less than c_{k+1} (and set $c_{n+1} = 1$). Note that (using the notation $a \wedge b$ for the minimum between a and b)

$$\begin{aligned} g\,(\mathbf{s}_0, \boldsymbol{\rho}_0) &= E_{\mathbf{s}_0,\boldsymbol{\rho}_0}^{Q}\left(\prod_{k=0}^{n-1} \frac{w^2\,(\mathbf{S}_k, \boldsymbol{\rho}_k)}{v^2\,(\mathbf{S}_k, \boldsymbol{\rho}_k)} I\,(\mathbf{S}_n = 0)\right) \\ &= E_{\mathbf{s}_0,\boldsymbol{\rho}_0}^{Q}\left(\prod_{k=0}^{\tau_{d_0}\wedge\tau-1} \frac{w^2\,(\mathbf{S}_k, \boldsymbol{\rho}_k)}{v^2\,(\mathbf{S}_k, \boldsymbol{\rho}_k)} g\left(\mathbf{S}_{\tau_d\wedge\tau}, \boldsymbol{\rho}_{\tau_d\wedge\tau}\right)\right). \end{aligned}$$

Clearly, the dynamics of $Q\,(\cdot)$ imply

$$g\left(\mathbf{S}_{\tau_{d_0}\wedge\tau}, \boldsymbol{\rho}_{\tau_{d_0}\wedge\tau}\right) 1\,(\tau_{d_0} > \tau) = 0,$$

therefore

$$\begin{aligned} g(\mathbf{s}_0, \boldsymbol{\rho}_0) &= E^Q_{\mathbf{s}_0,\boldsymbol{\rho}_0}\left(\prod_{k=0}^{\tau_{d_0}-1} \frac{w^2(\mathbf{S}_k, \boldsymbol{\rho}_k)}{v^2(\mathbf{S}_k, \boldsymbol{\rho}_k)} g\left(\mathbf{S}_{\tau_d}, \boldsymbol{\rho}_{\tau_d}\right); \tau_d < \tau\right) \\ &\leq \left(\sup_{[\mathbf{s}_0]_1 \leq d_0} g(\mathbf{s}_0, \boldsymbol{\rho}_0)\right) E^Q_{\mathbf{s}_0,\boldsymbol{\rho}_0}\left(\prod_{k=0}^{\tau_{d_0}-1} \frac{w^2(\mathbf{S}_k, \boldsymbol{\rho}_k)}{v^2(\mathbf{S}_k, \boldsymbol{\rho}_k)}\right). \end{aligned} \tag{17}$$

Now, define the stochastic process $(Z_k : k \geq 0)$ via

$$Z_k = f(\mathbf{S}_k, \boldsymbol{\rho}_k) \prod_{j=0}^{k-1} \frac{w^2(\mathbf{S}_j, \boldsymbol{\rho}_j)}{v^2(\mathbf{S}_j, \boldsymbol{\rho}_j)}$$

and consider the stopped process $M_k = Z_{k \wedge \tau_{d_0}}$. Note that $(M_k : k \geq 0)$ is a non-negative supermartingale, that is

$$\begin{aligned} & E^Q(M_{k+1} | \mathcal{F}_k) \\ = \; & E(M_{k+1}; \tau_{d_0} > k | \mathcal{F}_k) + E(M_{k+1}; \tau_{d_0} \leq k | \mathcal{F}_k) \\ = \; & 1(\tau_{d_0} > k) \prod_{j=0}^{k} \frac{w^2(\mathbf{S}_j, \boldsymbol{\rho}_j)}{v^2(\mathbf{S}_j, \boldsymbol{\rho}_j)} E(f\left(\mathbf{S}_{k+1}, \boldsymbol{\rho}_{k+1}\right) \big| \mathbf{S}_k) \\ & + 1(\tau_{d_0} \leq k) \prod_{j=0}^{\tau_{d_0}-1} \frac{w^2(\mathbf{S}_j, \boldsymbol{\rho}_j)}{v^2(\mathbf{S}_j, \boldsymbol{\rho}_j)} f(\mathbf{S}_{\tau_{d_0}}, \boldsymbol{\rho}_{k+1}) \\ \leq \; & 1(\tau_{d_0} > k) \prod_{j=0}^{k-1} \frac{w^2(\mathbf{S}_j, \boldsymbol{\rho}_j)}{v^2(\mathbf{S}_j, \boldsymbol{\rho}_j)} f(\mathbf{S}_k, \boldsymbol{\rho}_k) \\ & + 1(\tau_{d_0} \leq k) \prod_{j=0}^{\tau_{d_0}-1} \frac{w^2(\mathbf{S}_j, \boldsymbol{\rho}_j)}{v^2(\mathbf{S}_j, \boldsymbol{\rho}_j)} f(\mathbf{S}_{\tau_{d_0}}, \boldsymbol{\rho}_{k+1}) \\ = \; & M_k. \end{aligned}$$

Therefore,

$$\begin{aligned} h(\mathbf{s}_0, \boldsymbol{\rho}_0) &\geq E^Q_{\mathbf{s}_0,\boldsymbol{\rho}_0}\left(h(\mathbf{S}_{\tau_{d_0}}, \boldsymbol{\rho}_k) \prod_{j=0}^{\tau_{d_0}-1} \frac{w^2(\mathbf{S}_j, \boldsymbol{\rho}_j)}{v^2(\mathbf{S}_j, \boldsymbol{\rho}_j)}\right) \\ &\geq E^Q_{\mathbf{s}_0,\boldsymbol{\rho}_0}\left(\prod_{j=0}^{\tau_{d_0}-1} \frac{w^2(\mathbf{S}_j, \boldsymbol{\rho}_j)}{v^2(\mathbf{S}_j, \boldsymbol{\rho}_j)}\right). \end{aligned}$$

This estimate, together with (17), implies the conclusion of the proposition. ■

Acknowledgement The author is grateful to Persi Diaconis and Jun S. Liu for helpful discussions on contingency tables, to Dan Rudoy for valuable conversations on approximate counting, to Alistair Sinclair and Alexandre Stauffer for their very insightful comments on an earlier version of this paper, and to the referee for his careful reading and suggestions that helped improve the presentation. This research was partially supported by NSF grant DMS 0595595.

References

[1] Asmussen, S. and Glynn, P. (2007) *Stochastic Simulation: Algorithms and Analysis*. Springer, New York.

[2] Bayati, M., Kim, J., and Saberi, A. (2007) A Sequential Algorithm for Generating Random Graphs. Preprint.

[3] Bekessy, A., Bekessy, P., and Komlos, J. (1972) Asymptotic Enumeration of Regular Matrices. *Studia Scientiarum Mathematicarum Hungarica*, 7, 343-353.

[4] Bezakova, I., Bhatnagar, N., and Vigoda, E. (2006) Sampling Binary Contingency Tables with a Greedy Start. *Symp. of Discrete. Algorithms.* 414-423.

[5] Bezakova, I., Sinclair, A., Stefankovic, D., Vigoda, E. (2007) Negative Examples for Sequential Importance Sampling of Binary Contingency Tables. Submitted.

[6] Blanchet, J. and Glynn, P. (2007) Efficient Rare-event Simulation for the Maximum of a Random Walk with Heavy-tailed Increments. To appear in *Ann. of Appl. Prob.*

[7] Blanchet, J. and Liu, J. C. (2007) Efficient Rare-event Simulation for Large Deviation Probabilities of Regularly Varying Random Walks. Submitted.

[8] Blitzstein, J. and Diaconis, P. (2006) A Sequential Importance Sampling Algorithm for Generating Random Graphs with Prescribed Degrees. Preprint.

[9] Bucklew, J. (2004) *Introduction to Rare-event Simulation.* Springer, New York.

[10] Chen, S. X., and Liu, J. S. (1997) Statistical Applications of the Poisson-Binomial and Conditional Bernoulli Distributions. *Statistica Sinica*, 7, 875-892.

[11] Chen, X. H., Dempster, A., and Liu, J. S. (1994) Weighted Finite Population Sampling to Maximize Entropy. *Biometrika*, 81, 457-469.

[12] Chen, Y. Diaconis, P. Holmes, S. and Liu, J. (2005) Sequential Monte Carlo Method for Statistical Analysis of Tables. *Journal of the American Statistical Association*, 100, 109-120.

[13] Doob, J. (1957) Conditional Brownian motion and the Boundary of Harmonic Functions. *Bull. Soc. Math. France*, 85, 431-458.

[14] Glynn, P. and Iglehart, D. (1989) Importance Sampling for Stochastic Simulations. *Management Science*, 35, 1367-1392.

[15] Greenhill, C., Mckay, B. D., and Wang, X. (2006) Asymptotic Enumeration of Sparse 0-1 Matrices with Irregular Row and Column Sums. *J. Combinatorial Theory, Ser. A.*, 113, 291-324.

[16] Jerrum, M. (2003) *Counting, Sampling and Integrating: Algorithms and Complexity*. Birkhauser Verlag, Basel.

[17] Juneja, S. and Shahabuddin, P. (2006) Rare Event Simulation Techniques: An Introduction and Recent Advances. *Handbook on Simulation.* Elsevier. Editors: Shane Henderson and Barry Nelson p. 291-350.

[18] Kannan, R., Tetali, P., and Vempala, S. (1997) Simple Markov-chain Algorithms for Generating Bipartite Graphs and Tournaments. In Proceed. of the 8th Annual ACM-SIAM Symp. of Discrete Algorithms, 193-200.

[19] Kim, J., and Vu, V. (2003) Generating Random Regular Graphs. In Proc. of the 35th Annual ACM Symposium on Theory of Computing, 213-222.

[20] Liu, J. (2001) *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, New York..

[21] McKay, B. D. (1984) Asymptotics for 0-1 Matrices with Prescribed Line Sums, Enumeration and Design, *Academic Press*, Canada, 225-238.

[22] Meyn, S. and Tweedie, R. (1993) *Markov Chains and Stochastic Stability*. Springer-Verlag, London.

[23] Mitzenmacher, M. (2003) *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press. Cambridge.

[24] Rubinstein, R. Y. (2007) How Many Needles Are in a Hay Stack, or How to Solve Fast #P-Complete Counting Problems. Methodology and Computing in Applied Probability, 11, 5-49.

[25] Sinclair, A. (1993). *Algorithms for Random Generation and Counting*. Birkhauser, Boston.

[26] Valiant, L. (1979). The Complexity of Computing the Permanent. Theoretical Computer Science, 8, 189-201.