A Genuinely Polynomial-Time Algorithm for Sampling Two-Rowed Contingency Tables

Martin Dyer and Catherine Greenhill

School of Computer Studies, University of Leeds
Leeds, LS2 9JT, United Kingdom

Abstract. In this paper a Markov chain for contingency tables with two rows is defined. The chain is shown to be rapidly mixing using the path coupling method. The mixing time of the chain is quadratic in the number of columns and linear in the logarithm of the table sum. Two extensions of the new chain are discussed: one for three-rowed contingency tables and one for m-rowed contingency tables. We show that, unfortunately, it is not possible to prove rapid mixing for these chains by simply extending the path coupling approach used in the two-rowed case.

1 Introduction

A contingency table is a matrix of nonnegative integers with prescribed positive row and column sums. Contingency tables are used in statistics to store data from sample surveys (see for example [3, Chapter 8]). For a survey of contingency tables and related problems, see [8]. The data is often analysed under the assumption of independence. If the set of contingency tables under consideration is small, this assumption can be tested by applying a chi-squared statistic to each such table (see for example [1, 7]). However, this approach becomes computationally infeasible as the number of contingency tables grows. Suppose that we had a method for sampling almost uniformly from the set of contingency tables with given row and column sums. Then we may proceed by applying the statistic to a sample of contingency tables selected almost uniformly.

The problem of almost uniform sampling can be efficiently solved using the Markov chain Monte Carlo method (see [13]), provided that there exists a Markov chain for the set of contingency tables which converges to the uniform distribution in polynomial time. Here 'polynomial time' means 'in time polynomial in the number of rows, the number of columns and the logarithm of the table sum'. If the Markov chain converges in time polynomial in the table sum itself, then we shall say it converges in pseudopolynomial time. Approximately counting two-rowed contingency tables is polynomial-time reducible to almost uniform sampling, as can be proved using standard methods. Moreover, the problem of exactly counting the number of contingency tables with fixed row and column sums is known to be \#P-complete, even when there are only two rows (see [11]).

The first Markov chain for contingency tables was described in [9] by Diaconis and Saloff-Coste. We shall refer to this chain as the Diaconis chain. For fixed
dimensions, they proved that their chain converged in pseudopolynomial time. However, the constants involved grow exponentially with the number of rows and columns. Some Markov chains for restricted classes of contingency tables have been defined. In [14], Kannan, Tetali and Vempala gave a Markov chain with polynomial-time convergence for the 0-1 case (where every entry in the table is zero or one) with nearly equal margin totals, while Chung, Graham and Yau [6] described a Markov chain for contingency tables which converges in pseudopolynomial time for contingency tables with large enough margin totals. An improvement on this result is the chain described by Dyer, Kannan and Mount [11]. Their chain converges in polynomial time whenever all the row and column sums are sufficiently large, this bound being smaller than that in [6].

In [12], Hernek analysed the Diaconis chain for two-rowed contingency tables using coupling. She showed that this chain converges in time which is quadratic in the number of columns and in the table sum (i.e. pseudopolynomial time). In this paper, a new Markov chain for two-rowed contingency tables is described, and the convergence of the chain is analysed using the path coupling method [4]. We show that the new chain converges to the uniform distribution in time which is quadratic in the number of columns and linear in the logarithm of the table sum. Therefore our chain runs in (genuinely) polynomial time, whereas the Diaconis chain does not (and indeed cannot). In the final section we discuss two extensions of the new chain. The first applies to three-rowed contingency tables and the second applies to general contingency tables with $m$ rows. It is not known whether these chains converge rapidly to the uniform distribution, and it is quite possible that they do. However, we show that it is seemingly not possible to prove this simply by extending the path coupling approach used in the two-rowed case.

The structure of the remainder of the paper is as follows. In the next section the path coupling method is reviewed. In Section 3 we introduce notation for contingency tables and describe the Diaconis chain, which converges in pseudopolynomial time. We then outline a procedure which can perform exact counting for two-rowed contingency tables in pseudopolynomial time. A new Markov chain for two-rowed contingency tables is described in Section 4 and the mixing time is analysed using path coupling. The new chain is the first which converges in genuinely polynomial time for all two-rowed contingency tables. In Section 5 two extensions of this chain are introduced and discussed.

## 2 A review of path coupling

In this section we present some necessary notation and review the path coupling method. Let $\Omega$ be a finite set and let $M$ be a Markov chain with state space $\Omega$, transition matrix $P$ and unique stationary distribution $\pi$. If the initial state of the Markov chain is $x$ then the distribution of the chain at time $t$ is given by $P^t(x, y) = P^t(x, y)$. The total variation distance of the Markov chain from $\pi$ at