A PROBABILISTIC DISTRIBUTED ALGORITHM FOR SET INTERSECTION AND ITS ANALYSIS

by

Thomas G. Kurtz
and
Udi Manber

Computer Sciences Technical Report #565

November 1984

A PROBABILISTIC DISTRIBUTED ALGORITHM FOR SET INTERSECTION AND ITS ANALYSIS

(Preliminary Version)

Thomas G. Kurtz

Department of Mathematics

University of Wisconsin

Madison, Wisconsin 53706, U.S.A.

and

Udi Manber

Department of Computer Science

University of Wisconsin

1210 W Dayton St.

Madison, Wisconsin 53706, U.S.A.

November 1984

This research was supported in part by the National Science Foundation under Grants DMS-8401360 and MCS-8303134.

ABSTRACT

A Probabilistic algorithm for checking set disjointness and performing set intersection of two sets stored at different machines is presented. The algorithm is intended to minimize the amount of communication between the machines. If n is the number of elements in each set and k is the number of bits required to represent each of the elements, then it is shown that the expected running time of the set disjointness algorithm is $O(\log\log n)$ rounds, each round consisting of exchanging one message with $2n + k$ bits and performing $O(n)$ steps of local computation. The deterministic lower bound on the amount of communication in this case is $\Omega(nk)$. The analysis of the algorithm involves approximating Markov chains by deterministic models.

1. INTRODUCTION

This paper considers the problem of computing set intersection of two sets stored at two different machines. We assume that the sets contain elements whose size is quite large. For example, an element may be a line of text, a picture, or a file. The goal is to avoid sending all of the data to one machine and performing the intersection there. This is essential in cases where communication dominates the computation cost or in cases where there is not enough space in one machine for both sets. We present in this paper a probabilistic distributed algorithm for set intersection that is based on hashing, and in particular, random hash functions [CW79, WC79]. The algorithm efficiently eliminates elements that do not belong to the intersection without sending them over to the other machine. The rate of elimination of elements depends on the relative size of the intersection. We analyze the expected performance of the algorithm and show that if the intersection is small then the improvement in communication cost over any deterministic algorithm is substantial. If the intersection is not small then elements are eliminated at a slower rate. The algorithm can detect this with high probability early and then a deterministic algorithm can be used on the elements that were not eliminated. The additional cost of local computation is not excessive in any case. One of the most important application of set intersection is in the computation of semi-joins in distributed database systems [BG79].

The sequential computational complexity of the set intersection problem under a comparison based model is known. It is straightforward to perform set intersection of two sets of size n, using sorting, with $O(n\log n)$ comparisons. Reingold [Re72] proved that $\Omega(n\log n)$ comparisons are necessary to determine if the two sets are disjoint. Manber and Tompa [MT82] extended Reingold's results to probabilistic and nondeterministic decision trees and proved that the same lower bound holds (see also [MSM84]). Manber

[Ma84] considered the case of sets of different sizes and showed that $\Theta(m \log n)$ comparisons are necessary and sufficient in order to determine set disjointness of two sets of sizes n and m, $m > n$. The same lower bound holds for probabilistic decision trees as well. These results imply that one has to use more than comparisons to improve on the solution using sorting.

In this paper we show that the set disjointness problem can be solved in $O(n \log \log n)$ expected number of operations. The operations include hashing and comparisons. Moreover, the algorithm we present is very suitable to a distributed environment in which the sets are stored at two different machines. It can be divided into $O(\log \log n)$ rounds, each round consists of exchanging one message with $2n + k$ bits (where k is the size of each element) and performing $O(n)$ steps of local computation. (It is possible to modify the algorithm to work in $O(n)$ expected number of bits of communication by reducing the size of the messages as the algorithm progresses. However, more rounds will be required, hence it will be an inferior distributed algorithm. We will not discuss this modification in this paper.)

Analysis of probabilistic algorithms is usually quite complicated. This problem is no exception. We analyze the running time for large n by splitting the evolution of the algorithm into two stages, an essentially deterministic initial stage and a random termination stage. We approximate the behavior of the algorithm in the initial stage by a deterministic model and show that this deterministic approximation is good until most of the elements outside the intersection have been eliminated. We then show that the order of magnitude of the running time of the random termination stage is independent of n. The techniques employed here are applicable to models in a variety of fields (cf. [Ku76]) and further applications to probabilistic algorithms are anticipated.

Several other similar problems have been studied recently under a distributed model. Rodeh [Ro82] showed that exchanging $\Theta(\log n)$ numbers is necessary and sufficient to compute the median of the union of two sets stored at different machines. Mehlhorn and Schmidt[MS82] considered the following very simplified version of set intersection. Given two sequences $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$, such that $X, Y \subseteq \{0, 1, \ldots, 2^n - 1\}$, determine whether there exists i such that $x_i = y_i$. They proved that any deterministic algorithm requires sending n^2 bits, and then showed a probabilistic algorithm whose expected communication cost is only $O(n \log^2 n)$ bits. This serves as another example where probabilistic algorithms are more powerful than deterministic algorithms. In this paper we extend these results to the more general set intersection problem.

Another similar example of the power of probabilistic algorithms is probabilistic counting. Flajolet and Martin [FM83] (see also [St83]) introduced a class of such algorithms to estimate the number of distinct elements in a multiset. They were able to achieve an estimate with typical accuracy of 5-10% by a probabilistic algorithm that runs in linear time and makes only one pass through the data. This is significantly faster than the regular deterministic algorithm which requires sorting.

2. THE ALGORITHM

Let $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_m\}$, such that $X, Y \subseteq \{0, 1, \ldots, 2^k - 1\}$ We assume that X is stored in machine M_1 and Y is stored in machine M_2. We want to compute $X \cap Y$. Both machines can exchange messages and perform local computation on the data they hold. Our main measure of complexity is the number of bits of communication and the number of messages. We are also interested in minimizing the amount of local computation.

The algorithm consists of several identical rounds. In each round the sets are reduced by eliminating elements that are certain not to appear in the intersection. The algorithm terminates when either no more elements are left, in which case the sets are guaranteed to be disjoint, or when we are left with a subset of candidates that belong to the intersection with very high probability. In this case we can either conclude that, with high probability, the sets are not disjoint, or exchange the candidates to ensure that they indeed form the intersection. In section 4 we show that if the sets are disjoint the expected number of rounds to eliminate all elements is $O(\log \log n)$.

Each machine ℓ ($\ell = 1,2$) uses a binary table, called B_ℓ, of size $N > n$. N determines the size of each message. Obviously, the larger N is the less messages we expect to have; it is convenient to consider $N = m + n$. The main part of the algorithm is the use of random hash functions introduced by Carter and Wegman [CW79]. These functions are taken at random from a predetermined class of hash functions. For example, the following class of functions is a good candidate: $H_i \equiv (ax + b \pmod p) \pmod N$, where $a, b < p$ ($a \neq 0$), are chosen at random, $p > 2^k$ is a prime, and N is the size of the table ($p \gg N$) (see [CW79, WC79] for a description of several other good classes and their properties). We denote the elements that are not eliminated after round i by X^i and Y^i respectively ($X^0 = X$ and $Y^0 = Y$); these elements are called *candidates*. In each round $i \geq 0$ one of the machines, say M_1, selects a random hash function H_i from a class of hash functions H. (Since the number of such functions required by the algorithm is small, it is sufficient in practice to select those functions in advance; however, for the analysis we need the fact that the functions are completely random.)

M_1 sends a description of H_i to M_2 (in the case of the function given above the description includes the parameters a and b). M_1 uses H_i to hash the elements of X^i into B_1 in the following way. $B_1[j]$ is set to *true* iff there exists at least one element $x_s \in X^i$ such that $H_i(x_s) = j$ (i.e. x_s is hashed onto the j'th position). M_2 does the same (using the same hash function) for Y^i. The corresponding tables are then exchanged. This requires sending only N bits. M_1 can now eliminate all elements of X^i that were hashed into position j such that $B_2[j] = false$; M_2 does the same for Y^i. (To find those elements one can either leave pointers with the B tables, or rehash.)

There are several ways to terminate the algorithm. If we are only interested in a disjointness test and we are satisfied with a probabilistic algorithm that may make errors then we can run the algorithm for $c(\log_2 \log_2 n)$ steps, where $c > 1$ is a constant. If not all the elements are eliminated then the sets are not disjoint with very high probability (depending on c). If we want to determine the actual intersection with some probability of error we can terminate at round i when either $X^i = \emptyset$ (notice that $X^i = \emptyset$ implies $Y^i = \emptyset$), or $X^{i-q} = X^{i-q+1} = \cdots = X^i$ (and the same for Y) for some predetermined constant q. We have not yet analyzed this case.

Another approach is to run this algorithm for as long as a significant number of elements are expected to be eliminated and then send the remaining elements (most of which belong to the intersection with high probability) to the other machine. The decision when to stop the algorithm depends on the relative costs of the different steps. More precisely, the cost of the computation is

$$C_m = \sum_{i=0}^{m-1} \left[\alpha\left(X^i + Y^i\right) + \beta\right] + \gamma\left(X^m + Y^m\right)$$

if the algorithm is terminated after m rounds. The parameters α, β and γ correspond respectively to the costs of hashing and performing the elimination, sending the bitmaps, and sending the remaining elements at the end (we include both X^m and Y^m for symmetry). Let B be the cardinality of the intersection. We would like to select the first m such that the cost of performing the next round is more than the cost of the extra communication in sending the elements that could have been eliminated in this round, that is

$$
\begin{aligned}
& E\left[\gamma\left(X^{m}+Y^{m}\right)-\gamma\left(X^{m-1}+Y^{m-1}\right) \mid X^{m-1}, Y^{m-1}\right] \\
& =\gamma\left[\left(X^{m-1}-B\right)\left(1-N^{-1}\right)^{Y^{m-1}}+\left(Y^{m-1}-B\right)\left(1-N^{-1}\right)^{X^{m-1}}\right] \\
& \leq \alpha\left(X^{m-1}+Y^{m-1}\right)+\beta .
\end{aligned}
$$

Of course this requires an estimate of B. One such estimate (based on least squares after m rounds) is given by

$$
\frac{\sum_{i=0}^{m-1} W_{im}\left[X^{i+1}+Y^{i+1}-X^{i}\left(1-\left(1-N^{-1}\right)^{Y^{i}}\right)-Y^{i}\left(1-\left(1-N^{-1}\right)^{X^{i}}\right)\right]}{\sum_{i=0}^{m-1} W_{im}\left[\left(1-N^{-1}\right)^{Y^{i}}+\left(1-N^{-1}\right)^{X^{i}}\right]}
$$

where $\{W_{im}\}$ are positive weights. The optimal selection of the weights as well as other estimation methods remained to be studied. Termination rules based on a more careful analysis of the stochastic model also need to be explored.

3. EMPIRICAL RESULTS

We simulated the algorithm for checking set disjointness that was described in section 2. We considered two random disjoint sets and measured the mean and standard deviation (over 200 random inputs) of the number of rounds it took to eliminate all elements. We assume that the hash functions map the elements uniformly onto the bitmap tables. The results are given in table 1.

set size	mean	standard deviation
16	2.845	0.568
32	3.120	0.476
64	3.205	0.452
128	3.270	0.445
256	3.420	0.495
512	3.760	0.428
1024	3.940	0.238
2048	3.990	0.173

4096	4.015	0.122
8192	4.040	0.196
16384	4.065	0.247
32768	4.100	0.301
65536	4.165	0.372

Table 1: Running times for the set disjointness algorithm

4. ANALYSIS

If we assume that the values of H_i for distinct elements are independent and uniformly distributed, then the algorithm in the previous section is probabilistically equivalent to the following "balls in boxes" model. There are N boxes and red, green, and blue balls. At the i'th round let R_i denote the number of red balls, G_i the number of green balls, and B (which does not depend on i) the number of blue balls. The balls are placed at random in the boxes. If a box contains only red or only green balls then those balls are discarded. The remaining R_{i+1} red, G_{i+1} green, and B blue balls are then collected and the process is repeated. Of course, the red balls correspond to elements in X but not in Y, the green balls to elements in Y but not in X, and the blue balls to elements common to X and Y. We are interested in the behavior of the model if N is large and R_0 and G_0 are $O(N)$.

As a first step in the analysis, we show that the probabilistic model is well approximated by the following deterministic model. Given r_0, g_0 and b,

$$\begin{aligned} r_{i+1} &= r_i\left(1 - e^{-(b+g_i)}\right) \\ g_{i+1} &= g_i\left(1 - e^{-(b+r_i)}\right) \end{aligned} \tag{4.1}$$

Theorem 4.1 For $N = 2, 3, \ldots$, let $\left\{\left(R_i^N, G_i^N, B^N\right)\right\}$ be the random process described above with N boxes and starting with R_0^N red balls, G_0^N green balls, and B^N blue balls. Define $r_i^N = N^{-1}R_i^N$, $g_i^N = N^{-1}G_i^N$, and $b^N = N^{-1}B^N$. Let $0 < \alpha < 1/2$, and $\{(r_i, g_i)\}$ satisfy 4.1 . If for each $\epsilon > 0$,

$$\lim_{N\to\infty} \Pr\left\{N^\alpha\left(|r_0^N - r_0| + |g_0^N - g_0| + |b^N - b|\right) > \epsilon\right\} = 0,$$

then for $\epsilon > 0$

$$\lim_{N\to\infty} \Pr\left\{\sup_i N^\alpha\left(\left|r_i^N - r_i\right| + \left|g_i^N - g_i\right|\right) > \epsilon\right\} = 0. \tag{4.2}$$

Remark 4.2 By (4.2), for large N, $|r_i^N - r_i| \leq \epsilon N^{-\alpha}$ and $|g_i^N - g_i| \leq \epsilon N^{-\alpha}$ with high probability.

Proof: The theorem follows from Theorem A.1 in the Appendix. Note, letting $e_N = \left(1 - N^{-1}\right)^{-N}$,

$$\begin{aligned} E\left[r_{i+1}^N \mid r_i^N, g_i^N\right] &= r_i^N \left(1 - e_N{}^{-\left(b^N - g_i^N\right)}\right) \\ &\equiv F_1^N \left(r_i^N, g_i^N\right) \\ E\left[g_{i+1}^N \mid r_i^N, g_i^N\right] &= g_i^N \left(1 - e_N{}^{-\left(b^N + r_i^N\right)}\right) \\ &\equiv F_2^N \left(r_i^N, g_i^N\right) \end{aligned} \tag{4.3}$$

and

$$\begin{aligned} &E\left[\left(r_{i+1}^N - F_1^N\left(r_i^N, g_i^N\right)\right)^2 \mid r_i^N, g_i^N\right] = \\ &N^{-1}\Big[r_i^N\left(1 + r_i^N - N^{-1}\right)\left(e_N^{-\left(b^N + g_i^N\right)} - e_{N/2}^{-2\left(b^N + g_i^N\right)}\right) \\ &+ \left(r_i^N\right)^2\left(e_{N/2}^{-2\left(b^N + g_i^N\right)} - e_N^{-2\left(b^N + g_i^N\right)}\right)\Big] \end{aligned} \tag{4.4}$$

with a similar identity for g_{i+1}^N.

Theorem A.1 gives (4.2) with $\sup_i$ replaced by $\max_{1 \leq i \leq N^\beta}$, with $0 < \beta < 1 - 2\alpha$. However, $r_{N^\beta} \leq r_0\left(1 - e^{-b}\right)^{N^\beta} = o\left(N^{-\alpha}\right)$ and similarly $g_{N^\beta} = o\left(N^{-\alpha}\right)$ so the stronger statement follows from the monotonicity of r_i, r_i^N, g_i, and g_i^N. ∎

Our main interest is in

$$\tau_N = \min\left\{i : R_i^N = G_i^N = 0\right\}. \tag{4.5}$$

We begin by treating the case in which $G_0^N = 0$.

Proposition 4.3 Suppose $G_0^N = 0$, $B^N > 0$, $N = 2, 3, \ldots$, $\sup_N\left(r_0^N + b^N\right) < \infty$, and $R_0^N \to \infty$. Then

$$\lim_{N \to \infty} \sup_{z \in Z+} \left|\Pr\left\{\tau_N \leq z\right\} - \exp\left\{-R_0^N\left(1 - e_N{}^{-b^N}\right)^z\right\}\right| = 0, \tag{4.6}$$

and hence for each $\epsilon > 0$ there exists $\kappa_\epsilon > 0$ (independent of N) such that

$$\limsup_{N \to \infty} \Pr\left\{\left|\tau_N - \frac{\log R_0^N}{\left|\log\left(1 - e_N^{-b^N}\right)\right|}\right| \geq \kappa_\epsilon\right\} \leq \epsilon. \tag{4.7}$$

<u>Remark 4.4</u> Let $\xi_1, \xi_2, \cdots,$ be independent geometrically distributed random variables with $\Pr\{\xi_i \leq z\} = 1-(1-p)^z$. Then for large R

$$\Pr\left\{\max_{1\leq i\leq R} \xi_i \leq z\right\} = \left(1-(1-p)^z\right)^R \approx e^{-R(1-p)^z}.$$

Consequently, τ_N behaves as the maximum of R_0^N independent geometrically distributed random variables with parameter $p = e_N^{-b^N}$.

<u>Proof:</u> Let f_i^N denote the fraction of boxes that contain blue balls at the i'th round. Note that $f_1^N, f_2^N, \ldots$ are i.i.d. with $E\left[f_i^N\right] = 1 - e_N^{-b^N}$ and

$$\begin{aligned}\operatorname{Var}\left(f_i^N\right) &= e_{N/2}^{-2b^N} - e_N^{-2b^N} + N^{-1}\left(e_N^{-b^N} - e_{N/2}^{-2b^N}\right)\\ &\leq N^{-1}\left(e_N^{-b^N} - e_{N/2}^{-2b^N}\right)\end{aligned}$$

Then

$$\begin{aligned}\left(1-\left(b^N\right)^z\right)^{R_0^N} &\geq \Pr\{\tau_N \leq z\}\\ &= E\left[\left(1-\prod_{i=1}^{z} f_i^N\right)^{R_0^N}\right]\\ &\approx E\left[\exp\left\{-R_0^N\prod_{i=1}^{z} f_i^N\right\}\right] \qquad (4.8)\\ &= E\left[\exp\left\{-\left(\prod_{i=1}^{z}\frac{f_i^N}{E\left[f_i^N\right]}\right) R_0^N\left(1-e_N^{-b^N}\right)^z\right\}\right]\\ &\geq \exp\left\{-R_0^N\left(1-e_N^{-b^N}\right)^z\right\}.\end{aligned}$$

The asymptotic validity of the approximation follows from the fact that $R_0^N \to \infty$, and the second inequality by Jensen's inequality. Without loss of generality we may assume $b^N \to b \geq 0$. (Otherwise work with convergent subsequences.) If $b^N \to 0$, the left side and the right side of (4.8) are asymptotically the same and (4.6) follows, so assume $b^N \to b > 0$. By the monotonicity of $\Pr\{\tau_N \leq z\}$ it is enough to show that for arbitrary $0 < k_1 < k_2 < \infty$ the convergence in (4.6) is uniform for z satisfying

$$k_1 \leq R_0^N\left(1-e_N^{-b^N}\right)^z \leq k_2$$

that is

$$\left(\log R_0^N - \log k_2\right) / \left|\log\left(1 - e_N^{-b^N}\right)\right| \leq z \leq \left(\log R_0^N - \log k_1\right) / \left|\log\left(1 - e_N^{-b^N}\right)\right|.$$

By the mean value theorem

$$E\left[\exp\left\{-\left(\prod_{i=1}^{z} \frac{f_i^N}{E\left[f_i^N\right]}\right) R_0^N \left(1 - e_N^{-b^N}\right)^z\right\}\right] - \exp\left\{-R_0^N \left(1 - e_N^{-b^N}\right)^z\right\}$$

$$\leq k_2 E\left[\left|\prod_{i=1}^{z} \frac{f_i^N}{E\left[f_i^N\right]} - 1\right|\right]$$

$$\leq k_2 E\left[\left(\prod_{i=1}^{z} \frac{f_i^N}{E\left[f_i^N\right]} - 1\right)^2\right]^{1/2}$$

$$= k_2 \left(\left(\frac{E\left[\left(f^N\right)^2\right]}{E\left[f^N\right]^2}\right)^z - 1\right)^{1/2} \tag{4.9}$$

$$\leq k_2 \left(\left(1 + \frac{\operatorname{Var}\left(f^N\right)}{E\left[f^N\right]^2}\right)^z - 1\right)^{1/2}$$

$$\leq k_2 \left(\exp\left\{\frac{z}{N} \frac{e_N^{-b^N} - e_{N/2}^{-2b^N}}{\left(1 - e_N^{-b^N}\right)^z}\right\} - 1\right)^{1/2}$$

and (4.6) follows. Finally, to obtain (4.7), solve

$$e^{-R_0^N\left(1 - e_N^{-b^N}\right)^{z_1}} = \epsilon/2 \quad \textit{and} \quad e^{-R_0^N\left(1 - e_N^{-b^N}\right)^{z_2}} = (1 - \epsilon/2),$$

which gives

$$z_1 = \frac{\log R_0^N - \log\left(\left|\log\left(\epsilon/2\right)\right|\right)}{\left|\log\left(1 - e_N^{-b^N}\right)\right|}$$

and

$$z_2 = \frac{\log R_0^N - \log\left(\left|\log\left(1 - \epsilon/2\right)\right|\right)}{\left|\log\left(1 - e_N^{-b^N}\right)\right|}.$$

Define

$$\kappa_\epsilon = 1 + \sup_N \frac{\max\left(\left|\log\left(\left|\log\left(\epsilon/2\right)\right|\right)\right|, \left|\log\left(\left|\log\left(1 - \epsilon/2\right)\right|\right)\right|\right)}{\left|\log\left(1 - e_N^{-b^N}\right)\right|}.$$

■

Corollary 4.5 If $b > 0$ and $\lim_{N\to\infty} \log N\,|b^N - b| = 0$, then

$$\lim_{N\to\infty} \sup_{z\in Z^+} \left|\Pr\{\tau_N \le z\} - \exp\left\{-R_0^N\left(1-e^{-b}\right)^z\right\}\right| = 0, \tag{4.10}$$

and for every $\epsilon > 0$ there is a κ_ϵ such that

$$\limsup_{N\to\infty} \Pr\left\{\left|\tau_N - \frac{\log R_0^N}{\left|\log\left(1-e^{-b}\right)\right|}\right| \ge \kappa_\epsilon\right\} \le \epsilon. \tag{4.11}$$

Proof: As before, by monotonicity, it is sufficient to fix $0 < k_1 < k_2 < \infty$ and consider z satisfying

$$k_1 < R_0^N\left(1-e_N^{-b^N}\right)^z \quad , \quad R_0^N\left(1-e^{-b}\right)^z < k_2,$$

which implies $z \le C \log R_0^N = O(\log N)$. Then by the mean value theorem

$$\begin{aligned} &\left|R_0^N\left(1-e_N^{-b^N}\right)^z - R_0^N\left(1-e^{-b}\right)^z\right| \\ &\le k_2 z \left|e^{-b} - e_N^{-b^N}\right| \left(1-e_N^{-b^N}\right)^{-1}\left(1-e^{-b}\right)^{-1} \\ &\le k_2 z \left(\left|b-b^N\right| + O\left(N^{-1}\right)\right)\left(1-e_N^{-b^N}\right)^{-1}\left(1-e^{-b}\right)^{-1} \end{aligned} \tag{4.12}$$

and the Corollary follows. ■

Next we consider the case $r_0, g_0, b > 0$.

Theorem 4.6 Suppose the conditions of Theorem 4.1 are satisfied with $r_0, g_0, b > 0$. Let $0 < \gamma < \alpha$ and define

$$\sigma_N = \frac{\gamma \log N}{\left|\log\left(1-e^{-b}\right)\right|}.$$

Then

$$\lim_{N\to\infty} \sup_{z\in Z^+} \left|\Pr\{\tau_N \le z\} - \exp\left\{-\left(r_{\sigma_N} + g_{\sigma_N}\right) N \left(1-e^{-b}\right)^{z-\sigma_N}\right\}\right| = 0, \tag{4.13}$$

and for each $\epsilon > 0$ there exists $\kappa_\epsilon > 0$ such that

$$\limsup_{N\to\infty} \Pr\left\{\left|\tau_N - \frac{\log N}{\left|\log\left(1-e^{-b}\right)\right|}\right| \ge \kappa_\epsilon\right\} \le \epsilon. \tag{4.14}$$

Remark 4.7 Note that

$$r_0 + g_0 \le \frac{r_k + g_k}{\left(1 - e^{-b}\right)^k} \le r_0 \exp\left\{ \frac{e^{-b}}{1 - e^{-b}} \sum_{i=1}^{k} g_i \right\} + g_0 \exp\left\{ \frac{e^{-b}}{1 - e^{-b}} \sum_{i=1}^{k} r_i \right\} \le k\left(r_0, g_0, b\right) \tag{4.15}$$

for all k.

Consequently there exist constants c_1 and c_2 depending on r_0, g_0 and b such that

$$c_1 N^{-\gamma} \le r_{\sigma_N} + g_{\sigma_N} \le c_2 N^{-\gamma} \tag{4.16}$$

for all N.

Proof: It follows from Theorem 4.1 that for every $\epsilon > 0$

$$\lim_{N \to \infty} \Pr\left\{ \left| \frac{r^N_{\sigma_N}}{r_{\sigma_N}} - 1 \right| + \left| \frac{g^N_{\sigma_N}}{g_{\sigma_N}} - 1 \right| > \epsilon \right\} = 0. \tag{4.17}$$

Let $\tau^R_N = \min\left\{i : R^N_i = 0\right\}$, $\gamma^R_N = \min\left\{i : \tilde{R}^N_i = 0\right\}$ where for $i > \sigma_N$ $\tilde{R}^N_i$ is the number of red balls that have been in a box with a blue ball at each round j, $\sigma_N < j \le i$. Finally, let $\hat{R}^N_i$, $i > \sigma_N$, be the number of red balls that would remain if at time σ_N all green balls were painted blue, and define $\eta^R_N = \min\left\{i : \hat{R}^N_i = 0\right\}$. Note that $\gamma^R_N \le \tau^R_N$ and that

$$\Pr\left\{\eta^R_N \le z\right\} \le \Pr\left\{\tau^R_N \le z\right\} \le \Pr\left\{\gamma^R_N \le z\right\}. \tag{4.18}$$

After round σ_N, $\left\{\tilde{R}^N_i\right\}$ and $\left\{\hat{R}^N_i\right\}$ behave like the model without green balls analyzed in Proposition 4.3 . Consequently, by Proposition 4.3 , Corollary 4.5 , and the Markov property

$$\begin{aligned} &\lim_{N \to \infty} \sup_{z \in Z+} \left| \Pr\left\{\eta^R_N - \sigma_N \le z\right\} - E\left[\exp\left\{ -R^N_{\sigma_N} \left(1 - e_N^{-\left(b^N + g^N_{\sigma_N}\right)}\right)^z \right\}\right] \right| \\ &= \lim_{N \to \infty} \sup_{z \in Z+} \left| \Pr\left\{\eta^R_N - \sigma_N \le z\right\} - \exp\left\{ -r_{\sigma_N} N \left(1 - e^{-b}\right)^z \right\} \right| = 0 \end{aligned} \tag{4.19}$$

and

$$\begin{aligned} &\lim_{N \to \infty} \sup_{z \in Z+} \left| \Pr\left\{\gamma^R_N - \sigma_N \le z\right\} - E\left[\exp\left\{ -R^N_{\sigma_N} \left(1 - e_N^{-b^N}\right)^z \right\}\right] \right| \\ &= \lim_{N \to \infty} \sup_{z \in Z+} \left| \Pr\left\{\gamma^R_N - \sigma_N \le z\right\} - \exp\left\{ -r_{\sigma_N} N \left(1 - e^{-b}\right)^z \right\} \right| = 0. \end{aligned} \tag{4.20}$$

It follows that $\lim_{N\to\infty} \Pr\left\{\gamma_N^R = \tau_N^R\right\} = 1$.

Let τ_N^G and γ_N^G be defined as τ_N^R and γ_N^R. Note that $\tau_N = \tau_N^R \vee \tau_N^G$ so $\lim_{N\to\infty} \Pr\left\{\tau_N = \gamma_N^R \vee \gamma_N^G\right\} = 1$. Again applying Proposition 4.3

$$\begin{aligned}
&\lim_{N\to\infty} \sup_{z\in Z+} \left|\Pr\left\{\gamma_N^R \vee \gamma_N^G - \sigma_N \leq z\right\} - E\left[\exp\left\{\left(R_{\sigma_N}^N + G_{\sigma_N}^N\right)\left(1 - e_N^{-b^N}\right)^z\right\}\right]\right| \\
(4.21)\qquad &= \lim_{N\to\infty} \sup_{z\in Z+} \left|\Pr\left\{\tau_N - \sigma_N \leq z\right\} - \exp\left\{-\left(r_{\sigma_N} + g_{\sigma_N}\right) N \left(1 - e^{-b}\right)^z\right\}\right| = 0.
\end{aligned}$$

Replacing z by $z - \sigma_N$ gives (4.13).

Finally (4.14) follows from (4.13) and (4.16) by an argument similar to that used for (4.7). ∎

In the final case considered, we assume $B^N = 0$ for all N.

Theorem 4.8 Suppose the conditions of Theorem 4.1 are satisfied with $r_0, g_0 > 0$, and that $B^N = 0$ for all N. Let $\sigma_N = \min\left\{i : r_i g_i \leq N^{-1}\right\}$. Then

$$(4.22)\qquad \lim_{N\to\infty} \Pr\left\{\tau_N \in \left\{\sigma_N, \sigma_N + 1, \sigma_N + 2\right\}\right\} = 1.$$

Specifically, setting $u_N = r_{\sigma_N - 1} g_{\sigma_N - 1}$

$$\begin{aligned}
&\lim_{N\to\infty} \left(\Pr\left\{\tau_N = \sigma_N\right\} - e^{-Nu_N}\right) = 0 \\
(4.23)\qquad &\lim_{N\to\infty} \left(\Pr\left\{\tau_N = \sigma_N + 1\right\} - e^{-Nu_N^2}\left(1 - e^{-Nu_N}\right)\right) = 0 \\
&\lim_{N\to\infty} \left(\Pr\left\{\tau_N = \sigma_N + 2\right\} - \left(1 - e^{-Nu_N^2}\right)\right) = 0.
\end{aligned}$$

Remark 4.9 Note that

$$(4.24)\qquad \lim_{i\to\infty} \frac{r_{i+1}}{r_i g_i} = \lim_{i\to\infty} \frac{1 - e^{-g_i}}{g_i} = 1$$

and

$$(4.25)\qquad \lim_{i\to\infty} \frac{r_i}{g_i} = \frac{r_{i-1}\left(1 - e^{-g_{i-1}}\right)}{g_{i-1}\left(1 - e^{-r_{i-1}}\right)} = 1.$$

In particular $\lim_{N\to\infty} \frac{r_{\sigma_N} g_{\sigma_N}}{u_N^2} = 1$. Note also that $\lim_{N\to\infty} e^{Nu_N}\left(1 - e^{-Nu_N^2}\right) = 0$, so $\lim_{N\to\infty} \Pr\{\tau_N = \sigma_N\} \Pr\{\tau_N = \sigma_N + 2\} = 0$. Note that

$$\begin{aligned} r_{i+1} g_{i+1} &= (r_i g_i)^2 \frac{1-e^{-g_i}}{g_i} \frac{1-e^{-r_i}}{r_i} \\ &= (r_i g_i)^2 C_i \end{aligned}$$

and that $\lim_{i\to\infty} C_i = 1$. It follows that $liminf \sigma_N / \log_2 \log_2 N = 1$.

<u>Proof:</u> From (4.24) and (4.25) it follows that

$$\lim_{N\to\infty} \frac{r_{i+k}}{r_i^{2k}} = \lim_{N\to\infty} \frac{g_{i+k}}{g_i^{2k}} = 1. \tag{4.26}$$

Since $r_{\sigma_N - 1} g_{\sigma_N - 1} > N^{-1}$, (4.25) and (4.26) imply that there exists k such that

$$\lim_{N\to\infty} \frac{N^{-\alpha}}{r_{\sigma_N - k}} = \lim_{N\to\infty} \frac{N^{-\alpha}}{g_{\sigma_N - k}} = 0. \tag{4.27}$$

Then, by Proposition 4.1 , for each $\epsilon > 0$

$$\lim_{N\to\infty} \Pr\left\{ \sup_{i \le \sigma_N - k} \left(\left| \frac{r_i^N}{r_i} - 1 \right| + \left| \frac{g_i^N}{g_i} - 1 \right| \right) > \epsilon \right\} = 0. \tag{4.28}$$

By (4.4)

$$\begin{aligned} & E\left[\left(\frac{r_{i+1}^N}{r_i^N g_i^N} - 1 \right)^2 | r_i^N, g_i^N \right] \\ &= N^{-1} \left(r_i^N g_i^N\right)^{-2} \left[r_i^N \left(1 + r_i^N - N^{-1}\right) \left(e_N^{-g_i^N} - e_{N/2}^{-2g_i^N} \right) \right. \\ & \left. + \left(r_i^N\right)^2 \left(e_{N/2}^{-2g_i^N} - e_N^{-2g_i^N} \right) + \left(r_i^N\right)^2 \left(1 - e_N^{-g_i^N} - g_i^N \right)^2 \right] \\ & \le N^{-1} \left[\left(r_i^N g_i^N\right)^{-1} \left(1 + r_i^N - N^{-1}\right) \frac{2N}{N-2} + 1 \right]. \end{aligned} \tag{4.29}$$

By (4.29) there exists a $c > 0$ such that for each $k > 0$

$$\Pr\left\{ \left| \frac{r_{i+1}^N}{r_i^N g_i^N} - 1 \right| > k \left(N r_i^N g_i^N\right)^{-1/2} \right\} \le c k^{-2} \tag{4.30}$$

and a similar inequality holds for g_{i+1}^N.

By (4.24), (4.28), and the fact that $Nr_{\sigma_N - \ell}\, g_{\sigma_N - \ell} \to \infty$ for $\ell \geq 2$, we can, in (4.28), replace k by 1. It also follows from (4.30) that for every $\epsilon > 0$ there exists a κ_ϵ such that

$$\sup_N \Pr\left\{ N r_{\sigma_N}^N g_{\sigma_N}^N \left(r_{\sigma_N - 1}^N g_{\sigma_N - 1}^N \right)^{-1} > \kappa_\epsilon \right\} \leq \epsilon \tag{4.31}$$

and

$$\sup_N \Pr\left\{ \left| \frac{r_{\sigma_N}^N g_{\sigma_N}^N}{u_N^2} - 1 \right| > \kappa_\epsilon \left(N u_N \right)^{-1/2} \right\} \leq \epsilon. \tag{4.32}$$

Let ξ_i^N denote the fraction of boxes at the i'th round that contain green balls. Then

$$\Pr\left\{ r_{i+1}^N + g_{i+1}^N = 0 | r_i^N, g_i^N \right\} = E\left[\left(1 - \xi_i^N\right)^{R_i^N} |r_i^N, g_i^N \right]. \tag{4.33}$$

Numbering the green balls $1, 2, \ldots G_i^N$, let $\eta_{k\ell} = 1$ if the k'th and ℓ'th balls are in the same box. Then

$$0 \leq g_i^N - \xi_{i+1}^N \leq N^{-1} \sum_{k \leq \ell} \eta_{k\ell} \tag{4.34}$$

and hence

$$E\left[\left| g_i^N - \xi_{i+1}^N \right| |r_i^N, g_i^N \right] \leq g_i^N \left(g_i^N - N^{-1} \right). \tag{4.35}$$

Consequently, using (4.28)

$$\begin{aligned} &\lim_{N \to \infty} \Pr\left\{ r_{\sigma_N - 1}^N + g_{\sigma_N - 1}^N = 0 \right\} \\ &= \lim_{N \to \infty} E\left[\left(1 - \xi_{\sigma_N - 1}^N\right)^{R_{\sigma_N - 2}^N} \right] \\ &= \lim_{N \to \infty} E\left[e^{-N r_{\sigma_N - 2}^N g_{\sigma_N - 2}^N} \right] \\ &= \lim_{N \to \infty} e^{-N r_{\sigma_N - 2} g_{\sigma_N - 2}} = 0, \end{aligned} \tag{4.36}$$

and again by (4.28) (with k replaced by 1),

$$\begin{aligned} &\lim_{N \to \infty} \left| \Pr\left\{ r_{\sigma_N}^N + g_{\sigma_N}^N = 0 \right\} - e^{-N r_{\sigma_N - 1}^N g_{\sigma_N - 1}^N} \right| \\ &= \lim_{N \to \infty} \left| E\left[\left(1 - \xi_{\sigma_N}\right)^{R_{\sigma_N - 1}} \right] - e^{-N r_{\sigma_N - 1} g_{\sigma_N - 1}} \right| \\ &= \lim_{N \to \infty} \left| E\left[e^{-N r_{\sigma_N - 1}^N g_{\sigma_N - 1}^N} \right] - e^{-N r_{\sigma_N - 1} g_{\sigma_N - 1}} \right| = 0, \end{aligned} \tag{4.37}$$

which gives the first limit in (4.23). Verification of the other limits is similar using (4.31) and (4.32). ∎

REFERENCES

[BG79] P. A. Bernstein and N. Goodman, "The Theory of Semi-Joins," Computer Corporation of America Technical Report CCA-79-27, (November 1979).

[CW79] J. L. Carter and M. N. Wegman, "Universal Classes of Hash Functions," *Journal of Computer and System Sciences* **18** (April 1979), 143–154.

[DK] T. Darden and T. G. Kurtz, "Nearly Deterministic Markov Processes Near a Stable Point," (to appear).

[FM83] P. Flajolet and N. G. Martin, "Probabilistic Counting," *24th Annual Symposium on Foundations of Computer Science* (November 1983), 76–82.

[Ku76] T. G. Kurtz, "Limit Theorems and Diffusion Approximations for Density Dependent Markov Chains," *Mathematical Programming Study* **5** (1976) 67–78.

[Ku78] T. G. Kurtz, "Strong Approximations Theorems for Density Dependent Markov Chains." *Stochastic Processes and their Applications* **6** (1978), 223–240.

[Ma84] U. Manber, "A Probabilistic Lower Bound for Checking Disjointness of Sets," *Information Processing Letters* **19** (July 1984), 51–53.

[MT82] U. Manber and M. Tompa, "Probabilistic, Nondeterministic, and Alternating Decision Trees." *Fourteenth Annual ACM Symposium on Theory of Computing* (May 1982), 234–244.

[MS82] K. Mehlhorn and E. M. Schmidt, "Las Vegas is better than Determinism in VLSI and Distributed Computing," *Fourteenth Annual ACM Symposium on Theory of Computing* San Francisco, (May 1982), 330–337.

[MSM84] S. Moran, M. Snir, and U. Manber, "Applications of Ramsey's Theorem to Decision Tree Complexity," *25th Annual Symposium on Foundations of Computer Science* (October 1984), 332–337.

[Re72] E. M. Reingold, "On the Optimality of Some Set Algorithms," *Journal of the ACM* **19** (1972), 649–659.

[Ro82] M. Rodeh, "Finding the Median Distributively," *Journal of Computer and System Sciences* (1982), 162–166.

[St83] L. Stockmeyer, "The Complexity of Approximate Counting," *fifteenth Annual ACM Symposium on Theory of Computing* Boston, (April 1983), 118–126.

[WC79] M. N. Wegman and J. L. Carter, "New Classes and Applications of Hash Functions," *20th Annual Symposium on Foundations of Computer Science* (1979), 175–182.

APPENDIX

Theorem A.1 For $N = 1, 2, \ldots$, let $\{Z_i^N\}$ be a Markov chain with values in R^d. Let $F : R^d \to R^d$ be continuously differentiable and suppose $\{Z_i\}$ satisfies $Z_{i+1} = F(Z_i)$. Define

$$(A.1) \qquad F^N(z) = E\left[Z_{i+1}^N | Z_i^N = z\right]$$

and

$$(A.2) \qquad G^N(z) = NE\left[\left(Z_{i+1}^N - F^N\left(Z_i^N\right)\right)^2 | Z_i^N = z\right].$$

Let $0 < \alpha < 1/2$. Suppose for each compact $K \subset R^d, \sup_N \sup_{z \in K} G^N(z) < \infty$, and

$$(A.3) \qquad \lim_{N \to \infty} \sup_{z \in K} N^\alpha \left|F^N(z) - F(z)\right| = 0,$$

and for each $\epsilon > 0$

$$(A.4) \qquad \lim_{N \to \infty} \Pr\left\{N^\alpha \left|Z_0^N - Z_0\right| > \epsilon\right\} = 0.$$

a) For $M = 1, 2, \ldots$ and $\epsilon > 0$

$$(A.5) \qquad \lim_{N \to \infty} \Pr\left\{\max_{0 \le i \le M} N^\alpha \left|Z_i^N - Z_i\right| > \epsilon\right\} = 0.$$

b) If in addition, $\lim_{i \to \infty} Z_i = Z_\infty$ exists, and $\|\partial F(Z_\infty)\| < 1$ (here ∂F is the matrix of first partial derivatives), then, for $0 < \beta < 1 - 2\alpha$ and $\epsilon > 0$

$$(A.6) \qquad \lim_{N \to \infty} \Pr\left\{\max_{0 \le i \le N^\beta} N^\alpha \left|Z_i^N - Z_i\right| > \epsilon\right\} = 0.$$

Proof: Let $K_M = \sup\{\|\partial F(z)\| : \min_{0 \le i \le M} |z - Z_i| \le \epsilon\}$. Note that for $0 \le i \le M - 1$,

$$(A.7) \quad N^\alpha \left|Z_{i+1}^N - Z_{i+1}\right| = \left|N^\alpha \left(Z_{i+1}^N - F^N(Z_i^N)\right) + N^\alpha \left(F^N(Z_i^N) - F(Z_i^N)\right) + N^\alpha \left(F(Z_i^N) - F(Z_i)\right)\right|,$$

so for $\epsilon > 0$

$$
\begin{aligned}
\Pr\left\{N^\alpha\left|Z_{i+1}^N - Z_{i+1}\right| > \epsilon\right\} \leq\ & \Pr\left\{N^\alpha\left|Z_{i+1}^N - F^N(Z_i^N)\right| > \epsilon/3, N^\alpha\left|Z_i^N - Z_i\right| \leq \epsilon\right\} + \\
& \Pr\left\{N^\alpha\left|F^N(Z_i^N) - F(Z_i^N)\right| > \epsilon/3, N^\alpha\left|Z_i^N - Z_i\right| \leq \epsilon\right\} + \\
& \Pr\left\{N^\alpha\left|Z_i^N - Z_i\right| > \epsilon \wedge \left(3^{-1}K_M^{-1}\epsilon\right)\right\} \\
\leq\ & 9\epsilon^{-2}N^{-(1-2\alpha)}E\left[G^N(Z_i^N)\chi_{\{N^\alpha|Z_i^N - Z_i| \leq \epsilon\}}\right] + \\
& \Pr\left\{N^\alpha\left|F^N(Z_i^N) - F(Z_i^N)\right| > \epsilon/3, N^\alpha\left|Z_i^N - Z_i\right| \leq \epsilon\right\} + \\
& \Pr\left\{N^\alpha\left|Z_i^N - Z_i\right| > \epsilon \wedge \left(3^{-1}K_M^{-1}\epsilon\right)\right\}.
\end{aligned}
\tag{A.8}
$$

Using the uniform (in N) boundedness of G^N on compact sets, (A.3), and (A.4), (A.5) follows by induction.

Under the conditions of (b), let ρ satisfy $\|\partial F(Z_\infty)\| < \rho < 1$. By the continuity of ∂F there exists $\delta_0 > 0$ and $\epsilon_0 > 0$ such that $\|\partial F(z)\| < \rho$ if $|z - Z_\infty| < \delta_0 + \epsilon_0$.

Fix $0 < \epsilon < \epsilon_0$. Let $M = \min\{k : |Z_i - Z_\infty| < \delta_0, i \geq k\}$, and let N_0 be such that for $N \geq N_0$ and $|z - Z_\infty| \leq \delta_0 + \epsilon_0$, $N^\alpha\left|F^N(z) - F(z)\right| \leq \epsilon(1-\rho)/2$. If $i \geq M$, $N \geq N_0$ and $N^\alpha\left|Z_i^N - Z_i\right| \leq \epsilon$, then

$$
N^\alpha\left|Z_{i+1}^N - Z_{i+1}\right| \leq N^\alpha\left|Z_{i+1}^N - F^N(Z_i^N)\right| + \epsilon(1-\rho)/2 + \rho\epsilon. \tag{A.9}
$$

Consequently

$$
\begin{aligned}
\Pr\left\{\max_{0 \leq i \leq N^\beta} N^\alpha\left|Z_i^N - Z_i\right| > \epsilon\right\} \leq\ & \Pr\left\{\max_{0 \leq i \leq M} N^\alpha\left|Z_i^N - Z_i\right| > \epsilon\right\} + \\
& \Pr\left\{\max_{M \leq i \leq N^\beta} N^\alpha\left|Z_i^N - Z_i\right| > \epsilon, N^\alpha\left|Z_M^N - Z_M\right| \leq \epsilon\right\} \\
\leq\ & \Pr\left\{\max_{0 \leq i \leq M} N^\alpha\left|Z_i^N - Z_i\right| > \epsilon\right\} + \\
& \sum_{i=M}^{N^\beta - 1} \Pr\left\{N^\alpha\left|Z_{i+1}^N - Z_{i+1}\right| > \epsilon, N^\alpha\left|Z_i^N - Z_i\right| \leq \epsilon\right\} \\
\leq\ & \Pr\left\{\max_{0 \leq i \leq M} N^\alpha\left|Z_i^N - Z_i\right| > \epsilon\right\} + \\
& \sum_{i=M}^{N^\beta - 1} \Pr\left\{N^\alpha\left|Z_{i+1}^N - F^N(Z_i^N)\right| > \epsilon(1-\rho)/2, N^\alpha\left|Z_i^N - Z_i\right| \leq \epsilon\right\} \\
\leq\ & \Pr\left\{\max_{0 \leq i \leq M} N^\alpha\left|Z_i^N - Z_i\right| > \epsilon\right\} +
\end{aligned}
$$

$$\sum_{i=M}^{N^r-1} 4\epsilon^{-2}(1-\rho)^{-2}N^{-(1-2\alpha)}E\left[G^N(Z_i^N)\chi_{\{N^\alpha|Z_i^N-z_i|\leq\epsilon\}}\right],$$

and the right side goes to zero by part (a) and boundedness of G^N. ∎

For continuous time analogues of the above theorem see Kurtz ([Ku76], [Ku78]), and Darden and Kurtz ([DK]).