Fast Approximation of the Permanent for Very Dense Problems*

Mark Huber† Jenny Law‡

Abstract

Approximation of the permanent of a matrix with nonnegative entries is a well studied problem. The most successful approach to date for general matrices uses Markov chains to approximately sample from a distribution on weighted permutations, and Jerrum, Sinclair, and Vigoda developed such a method they proved runs in polynomial time in the input. The current bound on the running time of their method is $O(n^7(\log n)^4)$. Here we present a very different approach using sequential acceptance/rejection, and show that for a class of dense problems this method has an $O(n^4 \log n)$ expected running time.

1 The Permanent.

DEFINITION 1.1. The permanent of an n by n matrix $A = (A(i,j))$ is

$$\text{per}(A) := \sum_{\sigma \in S_n} \prod_{i=1}^{n} A(i, \sigma(i)), \tag{1.1}$$

where S_n is the set of permutations on $\{1, \ldots, n\}$.

While similar in form to the determinant, finding the permanent of a matrix is a $\#P$ complete problem [27]. Even approximation is only possible when the entries of the matrix are nonnegative [17]. Aside from its interest as a $\#P$ problem, approximating the permanent of a matrix with nonnegative entries has applications in nonparametric correlation tests [7] and problems in computer vision [24].

Various approaches have been tried to approximate the permanent, such as problem decomposition [16] and using determinants of random matrices [2]. Linial et al. [20] used scaling together with the van der Waerden lower bound on the permanent of a doubly stochastic matrix. All of these methods either had exponential running time or only approximated the permanent to an exponential factor.

An alternate approach is to sample from the set of permutations in S_n according to distribution π with probability mass function

$$\pi(\sigma) = \text{per}(A)^{-1} \prod_{i=1}^{n} A(i, \sigma(i)). \tag{1.2}$$

Since permutations are self-reducible (fixing an element of the permutation leaves a smaller permutation problem of the same form), any technique for generating samples approximately from π can be used to approximate the permanent efficiently (see [15] for details.)

This gives rise to the Markov chain approach: construct a Markov chain whose state space Ω contains S_n and whose stationary distribution on S_n is π. First consider the case where $A(i,j) \in \{0,1\}$ for all i and j (call A a 0-1 matrix). The problem remains $\#P$ complete with this restriction. Such a 0-1 matrix can be viewed as an adjacency matrix for a bipartite graph, where permutations with weight 1 are just perfect matchings in the graph.

Broder [5] designed such a Markov chain. Jerrum and Sinclair showed this chain could be used to generate approximate samples from π in time $O(n^8 \log n)$ [13, 6, 14] when the problem was (1/2)-dense, meaning that the row and column sums of the 0-1 matrix are at least $(1/2)n$. This gives an $O(n^9 \log n)$ method for approximating the permanent. Broder showed that even when restricted to γ-dense problems for fixed $\gamma \in (0,1)$, finding the permanent remains $\#P$ hard [5].

Jerrum, Sinclair and Vigoda [17] developed a different Markov chain where parameters of the chain were fine tuned to approach the desired π. This method allowed approximation of the permanent in $O(n^{10}(\log n)^3)$ time, and was the first fully polynomial randomized approximation scheme (FPRAS) for finding the permanent of a matrix with nonnegative entries. and was later improved by Bezáková et al. [3] to $O(n^7(\log n)^4)$.

Given that a FPRAS exists for this problem, two questions remain.

1. What is the smallest polynomial running time for restricted classes of matrices?
2. When is it possible to sample from π exactly?

For example, Kasteleyn showed that for 0-1 matrices whose associated graph is planar, the permanent can be found exactly in $O(n^3)$ time [19].

*Both authors supported by NSF CAREER grant DMS-05-48153.

†Duke University, mhuber@math.duke.edu

‡Duke University, waijlaw@math.duke.edu

The purpose of this work is to present a *perfect sampling* algorithm that generates random variates exactly from π where the running time of the algorithm is itself a random variable. In addition, the time used to generate these variates can be used to approximate the permanent of the matrix at no additional cost. Finally, the expected running time of the method is polynomial when the original problem is dense.

THEOREM 1.1. *For any $\epsilon \geq 0$ and $\delta \in (0,1]$, there exists a randomized approximation algorithm whose output comes within a factor of $1+\delta$ of the permanent of a nonnegative matrix A with probability at least $1-\epsilon$ with random running time T satisfying: for a function $R(n)$,*

$$\mathbb{E}[T] = O(n^4 \log n + R(n)\delta^{-2} \log \epsilon^{-1}), \tag{1.3}$$

$$\mathbb{P}(T > s\mathbb{E}[T]) \leq 2^{1-s/2}, \text{ for all } s > 0. \tag{1.4}$$

When A is a 0-1 matrix such that all the row and column sums are at least γn and $\gamma \in (0.5, 1]$, then $R(n) = O(n^{1.5+0.5/(2\gamma-1)})$. In particular, if $\gamma \geq .6$, then the running time is $O(n^4[\log n + \delta^{-2} \log \epsilon^{-1}])$.

The primary tool in this algorithm is a version of Bregman's Theorem similar in form to an inequality of Soules [26]. Let us first define:

$$h(r) = \begin{cases} r + (1/2)\ln(r) + e - 1, & r \geq 1 \\ 1 + (e-1)r, & r \in [0,1] \end{cases} \tag{1.5}$$

Our bound is as follows:

THEOREM 1.2. *Let A be a matrix with entries in $[0,1]$. Let $r(i)$ be the sum of the ith row of the matrix.*

$$per(A) \leq \prod_{i=1}^{n} \frac{h(r(i))}{e}. \tag{1.6}$$

The next section describes the algorithm in detail. Section 3 describes the history of the original Bregman's Theorem and proves Theorem 1.2. In Section 4, the bound on the running time in Theorem 1.1 is proven. In Section 5 further applications are considered.

2 The Algorithm.

The algorithm is an extension of ideas in [12]. There the algorithm only ran on 0-1 matrices, and had a provably polynomial running time only when the matrix was γ dense (defined below) for some $\gamma \in (0,1]$ and regular in the sense that all the row and column sums were identical. The algorithm presented here dispenses with the regularity requirement by first scaling the matrix so that all the row and column sums are close to 1, then scaling the matrix to make the row sums as large as possible, then applying the idea in [12] using the new generalization of Bregman's theorem.

For simplicity, in this section the algorithm is presented only for 0-1 matrices, and where the δ of Theorem 1.1 lies in $(0,1]$. For arbitrary nonnegative matrices, see Section 5. Let A be a 0-1 matrix. Then to determine if the permanent is zero or nonzero, the Hopcroft and Karp algorithm [11] can be used to find a permutation in $O(n^{2.5})$ time (for weighted bipartite graphs, the Hungarian Algorithm [1] can be employed, but takes $O(n^3)$ time.) When the permanent is nonzero, this method finds a permutation σ with $A(i,\sigma(i)) = 1$ for all i, so $\text{per}(A) \geq 1$. Then changing any zeros in A to $\alpha_1 = (\delta/3)(n!)^{-1}$ increases the permanent by at most a factor of $1 + \delta/3$.

DEFINITION 2.1. *A 0-1 matrix A is γ-dense if every row and column sum is at least γn.*

When a matrix is (1/2)-dense the choice of α_1 need not be so extreme. Work in [5, 14] implies that for (1/2)-dense matrices, when $\alpha_1 = (\delta/3)n^{-3}$, the permanent increases by at most a factor of $1 + \delta/3$.

Phase I: Nearly Doubly Stochastic Scaling. For diagonal matrices X and Y, XAY is a scaling of the matrix A where each row i is multiplied by $X(i,i)$ and each column j is multiplied by $Y(j,j)$. In our work A must be scaled to be nearly doubly stochastic so that the rows and columns each sum to almost 1. To be precise, following the same definition in [18], we say diagonal matrices X and Y scale A to *accuracy* α_2 if

$$||XAY\vec{1} - \vec{1}||_\infty < \alpha_2, \; ||YA^TX\vec{1} - \vec{1}||_\infty < \alpha_2, \tag{2.7}$$

where $\vec{1}$ is the vector of all 1's. In [18], the ellipsoid method was used to show that accuracy α_2 could be achieved after $O(n^4 \log(n\alpha_2^{-1} \log \alpha_1^{-1}))$ arithmetic operations on numbers with $O(\log(n/(\alpha_1\alpha_2)))$ digits, where α_1 is the minimum nonzero element of the matrix. In our algorithm, α_2 will be $\Theta(n^{-2})$.

The drawback of the ellipsoid method is its complexity. A slightly slower but much easier method to implement is the Sinkhorn method [25]. First all the rows are normalized by dividing by their row sum, then the columns, and this two step procedure is repeated until the accuracy falls below α_2. This requires $O((\alpha_2^{-1} + \log n)\sqrt{n} \log \alpha_1^{-1})$ steps, each of which take $\Theta(n^2)$ time. This makes the Sinkhorn method $O(n^{4.5}[n \log n + \log \delta^{-1}])$ for general matrices, and $O(n^{4.5}[\log n + \log \delta^{-1}])$ for (1/2)-dense matrices. Therefore in the dense case this method is slower than the ellipsoid method by a factor of $n^{1/2}$.

Phase II: row scaling. Let $m(i)$ denote the maximum entry of row i of the nearly doubly stochastic

matrix. Then each row can be scaled once more by dividing by $m(i)$. Note that after this phase, each element of the matrix is still in $[0,1]$, but the row sums are now $r(i) \in [(1-\alpha_2)m(i)^{-1}, (1+\alpha_2)m(i)^{-1}]$. Note Phases I and II need only be done once.

Phase III: acceptance/rejection. In [15], self-reducibility of permutations is used to turn a method for sampling into a method for counting. The self-reducible property can also be employed to construct an acceptance/rejection algorithm for a particular problem.

The idea is as follows. The goal of the algorithm is to randomly generate variates W over Ω that satisfy $\mathbb{P}(W = x) = w(x)/Z$, where $w(x)$ is a nonnegative weight for all x, and $Z = \sum_{x\in\Omega} w(x)$ is a difficult to compute normalizing function. Furthermore, suppose that Ω can be partitioned into $\Omega_1, \ldots, \Omega_n$, each with its own normalizing constant $Z_i = \sum_{x\in\Omega_i} w(x)$.

Also, assume that there is an upper bound function U satisfying $Z_i \le U(\Omega_i)$, and the upper bound respects the partition in the following way:

$$\sum_{i=1}^{n} U(\Omega_i) \le U(\Omega). \tag{2.8}$$

Finally, if $|\Omega_i| = 1$, then $U(\Omega_i) = Z_i$.

Then at the first step of the algorithm, consider the random variable I where $\mathbb{P}(I = i) = U(\Omega_i)/U(\Omega)$. By (2.8) these sum to at most 1, so let $\mathbb{P}(I = 0) = 1 - \sum_{i=1}^{n} \mathbb{P}(I = i)$. Draw a random variate I from this distribution. If $I = 0$, reject and start over, otherwise recursively sample from Ω_I. Either: 1) at some point the procedure rejects and starts over, or 2) eventually the sets are reduced to a single state that is the output. This procedure is summarized in pseudocode as follows:

Self-reducible Acceptance/Rejection
Input: $w(\cdot)$, Ω, $U(\cdot)$ *Output:* W

```
1)  Repeat
2)    Let Ω_temp ← Ω
3)    Repeat
4)      Partition Ω_temp into Ω_1, ..., Ω_n
5)      For i from 1 to n,
6)        Let p(i) ← U(Ω_i)/U(Ω_temp)
7)      Let p(0) ← 1 − Σ_{i=1}^n p(i)
8)      Choose I using P(I = i) = p(i)
9)      If I > 0 let Ω_temp ← Ω_I
10)   Until (I = 0) or (|Ω_temp| = 1)
11) Until |Ω_temp| = 1
12) Let W be the single element of Ω_temp
```

The above is a version of acceptance/rejection, and so has the correct output. To be complete we reprove it here for this particular variation.

THEOREM 2.1. *The above procedure for self-reducible acceptance/rejection outputs* $W \sim w(x)/Z$.

Proof. Let $x \in \Omega$, and consider $\mathbb{P}(W = x)$. Suppose k uses of the repeat loop in steps 3 through 10 are necessary to get from Ω down to $\{x\}$, and let $I(1)$ through $I(k)$ be the indices of the partition chosen at line 8. So

$$\{x\} = \Omega_{I(k)} \subset \Omega_{I(k-1)} \subset \Omega_{I(k-2)} \subset \cdots \subset \Omega_{I(0)} = \Omega.$$

Since the choice at each level was independent of the other, the probability of making these particular choices in lines 3 through 10 are the telescoping product:

$$\frac{U(\Omega_{I(1)})}{U(\Omega_{I(0)})} \cdot \frac{U(\Omega_{I(2)})}{U(\Omega_{I(1)})} \cdots \frac{U(\Omega_{I(k)})}{U(\Omega_{I(k-1)})} = \frac{w(x)}{U(\Omega)}. \tag{2.9}$$

So the probability of hitting x in one run of lines 3 through 10 is $w(x)/U(\Omega)$. This makes the probability of accepting $\sum_y w(y)/U(\Omega) = Z/U(\Omega)$. The outer repeat loop (from line 1 to line 11) ensures that the final output is conditioned on acceptance, and so

$$\mathbb{P}(W = x) = \frac{\mathbb{P}(\{x\} \text{ in lines 3 to 10})}{\mathbb{P}(\text{acceptance})} = \frac{w(x)/U(\Omega)}{Z/U(\Omega)}$$

which means $\mathbb{P}(W = x) = w(x)/Z$ as desired. □

Note this method (like standard acceptance/rejection) can be easily modified to deal with continuous problems and a target distribution defined by an unnormalized density, however, for estimating the permanent we only need the discrete version.

The set of permutations can be partitioned into permutations with $\sigma^{-1}(1) = 1$, $\sigma^{-1}(1) = 2, \ldots, \sigma^{-1}(1) = n$. In other words, the choice of which row goes with column 1 provides a natural way of partitioning the set of permutations. At the first step of the algorithm column 1 will be assigned a row $\sigma^{-1}(1)$. At the next step, column 2 will be assigned $\sigma^{-1}(2)$, and so on until the entire permutation σ has been generated.

Suppose that column j is assigned $\sigma^{-1}(j)$ as its row. Then that row and column cannot be used again, and so it helps to have notation where the matrix is altered to reflect this. Let $f(A, i, j)$ be a matrix equal to A except that row i and column j are all zero, except entry (i, j) which is 1 (illustrated in Figure 1.)

For a matrix with row sums $r(i)$, define

$$M(A) := \prod_{i=1}^{n} \frac{h(r_i)}{e}. \tag{2.10}$$

Then Theorem 1.2 states that per(A) $\le M(A)$. Set the weight of a permutation on A to be $w(\sigma) = \prod_{i=1}^{n} A(i, \sigma(i))$. Let $\Omega = \{\sigma : w(\sigma) > 0\}$. So if columns 1 through j have already been assigned to rows, and the

$$A = \begin{bmatrix} .2 & 0 & .7 \\ .5 & .9 & 0 \\ .3 & .6 & .1 \end{bmatrix}$$

$$f(A,1,1) = \begin{bmatrix} \mathbf{1} & 0 & 0 \\ 0 & .9 & 0 \\ 0 & .6 & .1 \end{bmatrix} \quad f(A,2,1) = \begin{bmatrix} 0 & 0 & .7 \\ \mathbf{1} & 0 & 0 \\ 0 & .6 & .1 \end{bmatrix} \quad f(A,3,1) = \begin{bmatrix} 0 & 0 & .7 \\ 0 & .9 & 0 \\ \mathbf{1} & 0 & 0 \end{bmatrix}$$

Figure 1: Example of reduced matrices.

matrix has been reduced to A_j, then an upper bound on the normalizing constant is

$$\begin{aligned} U(A_j) &:= \left[\prod_{k=1}^{j} A(\sigma^{-1}(k),k)\right] M(A_j) \\ &\geq \left[\prod_{k=1}^{j} A(\sigma^{-1}(k),k)\right] \mathrm{per}(A_j). \end{aligned}$$

If we are considering the step where column j is assigned, then if $\sigma^{-1}(j) = i$, $A_j = f(A_{j-1}, i, j)$. Note that many factors in $U(A_j)/U(A_{j-1})$ cancel out, so:

$$\frac{U(A_j)}{U(A_{j-1})} = \frac{A(i,j)M(f(A_{j-1},i,j))}{M(A_{j-1})},$$

and inequality (2.8) is satisfied if for all matrices with entries in $[0,1]$:

$$\sum_{i=1}^{n} A(i,j)M(f(A,i,j)) \leq M(A) \text{ for all } j. \tag{2.11}$$

This will be proved as Lemma 3.2 in the next section.

So in the first two phases, the algorithm scales the matrix so that it is almost regular and the maximum entry is 1. Putting this in pseudocode:

Scale Matrix
Input: A, *Output:* X, Y, Z, C

If A is not $(1/2)$-dense,
  let $\alpha_1 \leftarrow (\delta/3)(n!)^{-1}$
Else
  let $\alpha_1 \leftarrow (\delta/3)n^{-3}$
For all i, j
  If $A(i,j) = 0$, **let** $A(i,j) \leftarrow \alpha_1$
Let $B = XAY$ so row and col. sums of B in $(1-(.1)n^{-2}, 1+(.1)n^{-2})$ via ellipsoid
For all i,j,
  If $i = j$,
    let $Z(i,j) \leftarrow \min_k B(i,k)^{-1}$
  Else
    let $Z(i,j) \leftarrow 0$
Let $C \leftarrow ZB$

Once scaled, use self-reducible acceptance/rejection to generate variates. The algorithm will run through the columns one at a time from 1 to n. At column j, row $\sigma^{-1}(j) = i$ is chosen with probability $A(i,j)M(f(A_{j-1},i,j))/M(A_{j-1})$. Putting this in pseudocode gives us:

Generate Samples/Estimate Permanent
Input: C, X, Y, Z, ϵ, δ *Output:* $\sigma_1, \ldots, \sigma_k$, $\widehat{\mathrm{per}}(A)$

Let $d \leftarrow 0$, **let** $k \leftarrow 14\delta^{-2}\ln(2/\epsilon)$
For c from 1 to k
  Repeat
    Let $D \leftarrow C$
    Let σ_c be undefined everywhere
    Let $d \leftarrow d+1$
    For j from 1 to n do
      For i from 1 to n do
        Let $p(i) \leftarrow D(i,j)M(f(D,i,j))/M(D)$
      Let $p(0) \leftarrow 1 - \sum_{i>0} p(i)$
      Choose $i \in \{0,1,\ldots,n\}$ randomly, using $p(i)$ as probability mass function
      If $i > 0$, **let** $\sigma_c(i) \leftarrow j$, **let** $D \leftarrow f(D,i,j)$
  Until $\sigma_c(i)$ defined for all i
Let $s \leftarrow \prod_{i=1}^{n}[X(i,i)Z(i,i)Y(i,i)]$
Let $\widehat{\mathrm{per}}(A) \leftarrow M(C)kd^{-1}s^{-1}$

It follows from Chernoff bounds (see [23]) that setting $k = 14\delta^{-2}\ln(2/\epsilon)$ will return a result within a factor of $1+\delta/2$ of the answer with probability at least $1-\epsilon$. Combined with the fact that the modified matrix has permanent within a factor of $1+\delta/3$ of the orginal ensures that the final estimate will be within a factor of $1+\delta$ of the correct answer for $\delta \leq 1$.

By Markov's inequality $\mathbb{P}(d > 2\mathbb{E}[d]) \leq 1/2$. Now consider $\mathbb{P}(d > s\mathbb{E}[d])$. Break the $s\mathbb{E}[d]$ steps into blocks of size $2\mathbb{E}[d]$. Then the chance that each block fails is at most $1/2$, and the chance that they all fail is at most $(1/2)^{\lfloor s\mathbb{E}[d]/(2\mathbb{E}[d])\rfloor}$ which is at most $2(1/2)^{s/2}$, and so $\mathbb{P}(d > s\mathbb{E}[d]) \leq 2^{1-s/2}$.

To prove Theorem 1.1, it remains to show inequality (2.11), and verify $R = R(n)$ for γ-dense matrices.

3 Generalizing Bregman's Theorem.

Minc made the following conjecture [22], later proved by Bregman [4], which states given a 0-1 matrix A:

$$\mathrm{per}(A) \leq \prod_{i=1}^{n} (r(i)!)^{1/r(i)}, \tag{3.12}$$

where $r(i)$ is the sum of the ith row of the matrix. Even when starting with a matrix that is 0-1, after scaling the entries will lie between 0 and 1, and so a more general form of Bregman's Inequality is necessary.

Soules [26] proved one such inequality, unfortunately, Soules' bound cannot be proved by direct in-

duction on the size of the matrix, and so it cannot be used within the framework of our algorithm.

Therefore we create a new version of Bregman's Theorem that is slightly weaker than Soules' Inequality, but which can be proven by direct induction on the size of the matrix. Note $(1!)^{1/1} = h(1)/e = 1$ and approximating $r(i)!$ by Stirling's formula yields $\lim_{r\to\infty}[h(r)/e]/(r!)^{1/r} = 1$. Therefore, the factor of $h(r)/e$ as given in (1.6) in Theorem 1.2 is a generalization of the Bregman factors to noninteger matrices.

To prove Theorem 1.2, let A be a matrix with entries in $[0,1]$. Our goal is to show (1.6):

$$\mathrm{per}(A) \le \prod_{i=1}^{n} \frac{h(r(i))}{e}.$$

It suffices to prove (2.11) for $M(A) = \prod_i (h(r(i))/e)$, since then an induction on n yields the theorem.

Fix i and j, and consider how $M(A)$ and $A(i,j)M(f(A,i,j))$ are related. First, in $M(f(A,i,j))$ all rows $i' \neq i$ have their row sum reduced by $A(i',j)$. Second, the factor $h(r(i))$ gets removed entirely. Last, a factor of e is removed. Hence

(3.13)
$$\frac{M(f(A,i,j))}{M(A)} = \frac{e}{h(r(i)-A(i,j))} \prod_{i'=1}^{n} \frac{h(r(i')-A(i',j))}{h(r(i'))}.$$

It is easier to deal with the product using exponentials, which is the motivation behind the next lemma.

LEMMA 3.1. *For matrices with entries in* $[0,1]$*, and* $A(i',j) \le r(i')$*,*

(3.14)
$$\exp\left(-\frac{A(i',j))}{h(r(i'))-A(i',j)}\right) \ge \frac{h(r(i')-A(i',j))}{h(r(i'))}.$$

The proof, while technical, is straightforward, and left to the appendix. The three graphs in Figure 2 illustrates Lemma 3.1 for $A(i',j)$ equal to either 1, .5 or .1 by plotting the left hand side of inequality (3.14) minus the right hand side. Note that 0 is achieved only when $A(i',j) = 1$ and $r(i') = 1$.

LEMMA 3.2. *For matrices* A *with* $A(i,j) \in [0,1]$ *for all* i *and* j*:*

$$M(A) \ge \sum_{i=1}^{n} A(i,j)M(f(A,i,j)) \text{ for all } j.$$

Proof. Multiply (3.13) by $A(i,j)$ and then sum both sides to get:

$$(3.15) \qquad \frac{\sum_{i''} A(i'',j)M(f(A,i'',j))}{M(A)}$$

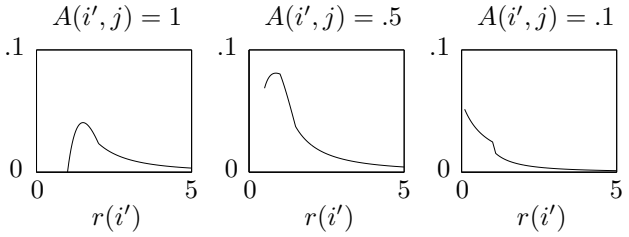

Figure 2: Graphs of LHS - RHS for inequality (3.14).

$$\le e\left[\sum_{i''=1}^{n}\left(\prod_{i'=1}^{n}\frac{h(r(i')-A(i',j))}{h(r(i'))}\right)\frac{A(i'',j)}{h(r(i'')-A(i'',j))}\right].$$

Notice the product in the sum does not depend on i'' and can be brought out of the sum. Now use (3.14) to replace the factors in the product with exponentials, and the product turns into a sum:

$$(3.16) \qquad \frac{\sum_{i''} A(i'',j)M(f(A,i'',j))}{M(A)}$$

$$\le e\exp\left(\sum_{i'=1}^{n}\frac{-A(i',j)}{h(r(i')-A(i',j))}\right)\sum_{i''=1}^{n}\frac{A(i'',j)}{h(r(i'')-A(i'',j))}.$$

The last line is of the form ye^{1-y}, which is easily shown to have a global maximum of 1 at $y = 1$. But $y = 1$ is equivalent to inequality (2.11), and that is indentical to the statement of Lemma 3.2. □

As noted above, (2.11) is exactly what is needed to show by induction that $\mathrm{per}(A) \le \prod_{i=1}^{n}[h(r(i))/e]$, thereby proving Theorem 1.2.

4 Dense Matrices.

The $R(n)$ in Theorem 1.1 equals $n^2M(C)/\mathrm{per}(C)$, where C is the scaled version of the matrix A, and $M(C) = \prod(h(r_C(i))/e)$ for r_C the vector of row sums of C. The n^2 factor comes from the time needed to generate a single sample, and the $M(C)/\mathrm{per}(C)$ is the expected number of samples generated before one is accepted. In this section we show that when A is a dense 0-1 matrix, that $R(n)$ is polynomial in n.

LEMMA 4.1. *Let* $\gamma \in (1/2, 1]$*, and* A *be an* n *by* n *matrix with entries in* $[0,1]$ *and at least* γn *entries equal to 1 in every row and column. Let* $B = XAY$ *with* X *and* Y *diagonal matrices so that* B *is doubly stochastic with accuracy* n^{-2}*, and let* $C = ZB$ *where* Z *is a diagonal matrix whose* i*th element is the inverse of*

the maximum entry of the ith row of B. Then

$$R(n) = n^2 \frac{M(C)}{per(C)} = O(n^{1.5+0.5/(2\gamma-1)}). \tag{4.17}$$

In order to lower bound the permanent of C, first study the permanent of B. Van der Waerden conjectured the following result: matrices whose row and column sums are exactly 1 (called a *doubly stochastic* matrix) have permanent bounded below by $n!/n^n$. This was later independently proved by Egorychev [8] and Falikman [9].

Linial, Samordinisky, and Wigderson [20] presented an argument that the permanent of matrices whose row and column sums are close to 1 is similarly bounded below. Here we recast their argument using the notion of accuracy in (2.7). The first step is to put limits on accuracy when the permanent of a matrix is 0.

LEMMA 4.2. *Any nonnegative matrix B such that $|r_B(i) - 1| < 1/(2n-1)$ and $|c_B(j) - 1| < 1/(2n-1)$ for all i and j has per$(B) > 0$.*

Proof. Consider the contrapositive. Let B be a matrix with $\mathrm{per}(B) = 0$ and accuracy α_2. Then the König-Hall Theorem (see [21]) states B is of the form

$$B = \begin{pmatrix} O & B_1 \\ B_2 & B_3 \end{pmatrix}, \tag{4.18}$$

where O is an s by $(n-s+1)$ zero submatrix. Given B has accuracy α_2, the sum of entries in B_1 and B_3 is at most $(s-1)(1+\alpha_2)$. Similarly the sum of entries in B_3 and B_2 is at most $(n-s)(1+\alpha_2)$. Hence the sum of entries in the entire matrix is at most $(n-1)(1+\alpha_2)$.

However, the same accuracy requires the sum of entries in the matrix to be at least $n(1-\alpha_2)$. Hence $n(1-\alpha_2) \le (n-1)(1+\alpha_2)$, and $\alpha_2 \ge 1/(2n-1)$.

LEMMA 4.3. *Let $B = XAY$ be A scaled to accuracy $\alpha_2 \le .79/(2n)$. Then*

$$per(B) \ge \frac{n!}{n^n} \exp(-4n^2\alpha_2). \tag{4.19}$$

Proof. Linial et al. [20] noted that any nonnegative matrix B can be written as $\lambda D + W$, where $\lambda \ge 0$, D is exactly doubly stochastic, W is nonnegative, and $\mathrm{per}(W) = 0$. Since $W \ge 0$, $\mathrm{per}(B) \ge \lambda^n \mathrm{per}(D) \ge \lambda^n n!/n^n$. If $\lambda \ge 1$, then $\lambda^n \ge 1$, so this case is proved.

Say $\lambda \in [0,1)$, and consider $W' = W/(1-\lambda)$. Then $r_B(i) = \lambda + (1-\lambda) r_{W'}(i)$ and $r_{W'}(i) = (r_B(i) - \lambda)/(1-\lambda)$. Note $r_B(i)$ is doubly stochastic to accuracy α_2, so:

$$\frac{1-\lambda-\alpha_2}{1-\lambda} < r_{W'}(i) <= \frac{1-\lambda+\alpha_2}{1-\lambda}, \tag{4.20}$$

and a similar pair of inequalities holds for $c_{W'}$. Therefore W' has accuracy $\alpha_2/(1-\lambda)$. Since W' is just a scaled version of W, $\mathrm{per}(W') = 0$, which means from Lemma 4.2 that $\alpha_2/(1-\lambda) \ge 1/(2n-1)$, so $\lambda \ge 1 - \alpha_2(2n-1) \ge 1 - 2n\alpha_2$.

Since $1 - x \ge e^{-2x}$ for all $x \in [0, .79]$, this yields $\mathrm{per}(B) \ge \exp[-n(2)(2n\alpha_2)]n!/n^n$ which is just (4.19).

Now suppose that B is A scaled to accuracy $(.1)n^{-2}$ and consider $\mathrm{per}(C)$. The matrix C is just B scaled by the diagonal elements of Z. Let $z(i) = Z(i,i)$, so

$$\mathrm{per}(C) = \mathrm{per}(B) \prod_{i=1}^{n} z(i) \ge \frac{n!}{n^n} \exp(-.4) \prod_{i=1}^{n} z(i). \tag{4.21}$$

On the other hand, the row sums of C will be at most $z(i)(1+\alpha_2)$, so we have $M(C) \le \prod_i h(z(i)(1+\alpha_2))/e$. By Stirling's formula $n! > \sqrt{2\pi n}(n/e)^n$, so

$$\begin{aligned} & \frac{M(C)}{\mathrm{per}(C)} \left(e^{.4}(2\pi n)^{-1/2}\right)^{-1} \\ \le & \prod_i \frac{(z(i)(1+\alpha_2) + .5\ln(z(i)(1+\alpha_2)) + e - 1)}{z(i)} \\ \le & \prod_i \exp\left[\alpha_2 + \frac{.5\ln z(i) + .5\ln(1+\alpha_2) + e - 1}{z(i)}\right]. \end{aligned}$$

Let $m = \max\{1, n\max_{i,j} B(i,j)\}$ so $B(i,j) \le m/n$ for all i and j. Then for all i, $z(i) \ge n/m$ and it is straightfoward to show $(.5\ln z(i) + e - 1)/z(i) \le (.5\ln(n/m) + e - 1)(m/n)$. Since $\alpha_2 = .1n^{-2}$, $[\exp(\alpha_2 + \ln(1+\alpha_2))]^n \le 1.22$ and

$$\begin{aligned} & M(C)/\mathrm{per}(C) \qquad (4.22) \\ \le & e^{.4}(2\pi n)^{-1/2} 1.22 \exp((.5\ln n - .5\ln m + e - 1)m) \\ \le & 4.1 \left(\frac{31.1}{m} n\right)^{.5(m-1)}. \end{aligned}$$

It remains to show that if A is a 0-1 γ-dense matrix with $\gamma > 1/2$, then m is small in the scaled matrix B.

LEMMA 4.4. *Let A be a matrix with entries in $(0,1]$, $\gamma \in (1/2, 1]$, and suppose every row and column of A has at least γn entries equal to 1. Let $B = XAY$ (with X and Y diagonal) be doubly stochastic to accuracy α_2. Then for $2\gamma - 1 - 3\alpha_2 > 0$:*

$$\max_{i,j} B(i,j) \le \frac{(1+\alpha_2)^2}{(2\gamma - 1 - 3\alpha_2)n}. \tag{4.23}$$

Proof. For convenience let $x(i) := X(i,i)$ and $y(i) := Y(i,i)$. Let i and j be indices from 1 to n. Then

$B(i,j) = x(i)A(i,j)y(j)$. Since $B(i,j)$ is α_2 accurate as in (2.7), $\sum_{i'=1}^{n} x(i')A(i',j)y(j) \leq 1+\alpha_2$, and $\sum_{j'=1}^{n} x(i)A(i,j')y(j') \leq 1+\alpha_2$. Hence

$$\sum_{i' \neq i} x(i')A(i',j) \leq y(j)^{-1}(1+\alpha_2 - B(i,j)), \tag{4.24}$$

$$\sum_{j' \neq j} A(i,j')y(j') \leq x(i)^{-1}(1+\alpha_2 - B(i,j)). \tag{4.25}$$

Let $S_r = \{j' \neq j : A(i,j') = 1\}$, and $S_c = \{i' \neq i : A(i',j) = 1\}$. Under our assumptions both $|S_r|$ and $|S_c|$ are at least $\gamma n - 1$. Use S_r and S_c to break A into four submatrices: $A_1 := \{i' \in S_c, j' \in S_r\}$, $A_2 := \{i' \in S_c, j' \notin S_r\}$, $A_3 := \{i' \notin S_c, j' \in S_r\}$, and $A_4 := \{i' \notin S_c, j' \notin S_r\}$, with B_1, B_2, B_3, and B_4 corresponding to the same submatrices in B.

For a matrix D, let $s(D)$ denote the sum of the entries in D. Then since B is doubly stochastic to accuracy α_2,

$$\begin{aligned} s(B_1) + s(B_2) &\geq |S_c|(1-\alpha_2) \\ s(B_1) + s(B_3) &\geq |S_r|(1-\alpha_2). \end{aligned}$$

Let $n_{\alpha_2} := n(1+\alpha_2)$. Then

$$\begin{aligned} n_{\alpha_2} &\geq s(B_1) + s(B_2) + s(B_3) + s(B_4) \\ &\geq B(i,j) + |S_r|(1-\alpha_2) + |S_c|(1-\alpha_2) - s(B_1). \end{aligned}$$

To lower bound $-s(B_1)$, note

$$\begin{aligned} s(B_1) &\leq \sum_{i' \in S_c} \sum_{j' \in S_r} x(i')A(i',j')y(j') \\ &\leq \left[\sum_{i' \in S_c} x(i')\right] \left[\sum_{j' \in S_r} y(j')\right] \end{aligned}$$

each of which factors is bounded in (4.24) and (4.25), respectively. Using this bound together with $|S_r| + |S_c| \geq 2(\gamma n - 1)$ yields:

$$n_{\alpha_2} \geq B(i,j) + 2(\gamma n - 1)(1-\alpha_2) - \frac{(1+\alpha_2 - B(i,j))^2}{x(i)y(j)}. \tag{4.26}$$

Since $B(i,j) \leq x(i)y(j)$, inverting and negating gives $-B(i,j)^{-1} \leq -[x(i)y(j)]^{-1}$, which means

$$\begin{aligned} n_{\alpha_2} B(i,j) &\geq B(i,j)^2 + 2(\gamma n - 1)(1-\alpha_2)B(i,j) \\ &\quad -(1+\alpha_2 - B(i,j))^2. \end{aligned}$$

In the RHS the $B(i,j)^2$ terms cancel, and solving for $B(i,j)$ and using $2\gamma + 1 \leq 3$ yields (4.23).

Proof of Lemma 4.1 Let A be a $[0,1]$ matrix with at least γn 1's in each row and column, and B a scaling of A that is doubly stochastic to accuracy $.1n^{-2}$. From the previous lemma $m \leq (1+\alpha_2)^2(2\gamma - 1 - 3\alpha_2)^{-1}$, so by (4.22), we have the following

$$\begin{aligned} M(C)/\mathrm{per}(C) &\leq 4.1(31.1n)^{.5[(1+\alpha_2)^2(2\gamma-1-3\alpha_2)^{-1}-1]} \\ &\leq O(n^{-0.5+0.5(2\gamma-1)^{-1}}). \end{aligned}$$

Multiplying both sides by n^2 finishes the proof. □

5 Permanents for Matrices That Are Not 0-1.

To deal with matrices with arbitrary nonnegative entries, we employ a method similar to that used in [17]. First solve the maximum assignment problem in $O(n^3)$ time to find the highest weight permutation. This gives a lower bound on the permanent α_3. Then scale the matrix by dividing each entry by the largest entry in the matrix. Finally, change any entries whose value is less than

$$\alpha_1 := (\delta/3)\alpha_3(n!)^{-1}, \tag{5.27}$$

to α_1. This changes line 6) in the **Scale Matrix** pseudocode of Section 2. While this increases the permanent by a factor of at most $1 + \delta/3$, it ensures that the ratio between the the maximum and minimum entry of the matrix has at most $n \ln n \delta^{-1}$ digits. The rest of the algorithm continues as before.

6 Conclusions.

This work shows that for very dense matrices, approximating the permanent or generating weighted perfect matchings exactly from their correct distribution can be accomplished in $O(n^4 \log n + n^{1.5+.5(2\gamma-1)^{-1}})$ time. This is an improvement in running time by a factor of n^3 for matrices that are .6-dense [3]. This work extends earlier work in [12] that required the matrix to be both dense and regular in the sense of having equal row and column sums. This is an important step on the way to an efficient perfect sampling method applicable to all nonnegative matrices.

For estimating the permanent of a 0-1 matrix to within a factor of $1+\delta$ with probability at least $1-\epsilon$, the expected running time is bounded above by $O(n^4 \log n + n^{1.5+.5/(2\gamma-1)}\delta^{-2} \log \epsilon^{-1})$.

It is important to note that our method can be applied to generating perfect matchings from any matrix with nonnegative entries, but there is currently no a priori bound on the running time.

References

[1] R. Ahuja, T. Magnanti, and J. Orlin, *Network Flows*, Prentice Hall, 1993.
[2] A. Barvinok, *Polynomial time algorithms to approximate permanents and mixed discriminants within a simply exponential factor*, Random Structures Algorithms, 14 (1999), pp. 29–61.
[3] I. Bezáková, D. Stefankovic, V. V. Vazirani, and E. Vigoda, *Accelerating simulated annealing for the permanent and combinatorial counting problems*, SODA '06: Proc. 17th ACM-SIAM Sympos. on Discrete Algorithms, 2006, pp. 900–907.
[4] L. M. Bregman, *Some properties of nonnegative matrices and their permanents*, Soviet. Math. Dokl., 14 (1973), pp. 945–949.
[5] A. Z. Broder, *How hard is it to marry at random? (On the approximation of the permanent)*, STOC '86: Proc. 18th Annual ACM Sympos. on Theory of Computing, 1986, pp. 50–58.
[6] P. Diaconis and D. Stroock, *Geometric bounds for eigenvalues of Markov chains*, Ann. Appl. Probab., 1 (1991), pp. 36–61.
[7] B. Efron and V. Petrosian, *Nonparametric Methods for Doubly Truncated Data*, J. Amer. Statist. Assoc., 94 (1999), pp. 824–834.
[8] G. P. Egorychev, *The solution of van der Waerden's problem for pemanents*, Advances in Math., 42 (1981), pp. 299–305.
[9] D. I. Falikman, *Proof of the van der Waerden's conjecture on the permanent of a doubly stochastic matrix*, Mat. Zametki, 29 (1981), pp. 931–938.
[10] G. S. Fishman, *Choosing sample path length and number of sample paths when starting in the steady state*, Oper. Res. Letters, 16 (1994), pp. 209–219.
[11] J. E. Hopcroft and R. M. Karp, *An $n^{5/2}$ Algorithm for Maximum Matchings in Bipartite Graphs*, SIAM J. Comput., 2 (1973), pp. 225–231.
[12] M. Huber, *Exact Sampling from Perfect Matchings of Dense Regular Bipartite Graphs*, Algorithmica, 44 (2006), pp. 183–193.
[13] M. Jerrum and A. Sinclair, *Approximating the permanent*, J. Comput., 18 (1989), pp. 1149–1178.
[14] M. Jerrum and A. Sinclair, *The Markov Chain Monte Carlo Method: An Approach to Approximate Counting and Integration*, PWS Publishing, 1996.
[15] M. Jerrum, L. Valiant, and V. Vazirani, *Random generation of combinatorial structures from a uniform distribution*, Theoret. Comput. Sci., 43 (1986), pp. 169–188.
[16] M. Jerrum and U. V. Vazirani, *A Mildly Exponential Approximation Algorithm for the Permanent*, Algorithmica, 16 (1996), pp. 392–401.
[17] M. R. Jerrum, A. Sinclair, and E. Vigoda, *A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries*, J. of the ACM, 51 (2004), pp. 671–697.
[18] B. Kalantari and L. Khachiyan, *On the Complexity of Nonnegative-Matrix Scalings*, Linear Algebra Appl., 240 (1996), pp. 87–103.
[19] P. W. Kasteleyn, *The statistics of dimers on a lattice, I., The number of dimer arrangements on a quadratic lattice*, Physica, 27 (1961), pp. 1664–1672.
[20] N. Linial, A. Samorodnitsky, and A. Wigderson, *A deterministic strongly polynomial algorithm for matrix scaling and approximate permanents*, Combinatorica, 20 (2000), pp. 545–568.
[21] J. H. Van Lint and R. M. Wilson, *A Course in Combinatorics*, Cambridge University Press, 1992.
[22] H. Minc, *Upper bounds for permanents of (0,1)-matrices*, Bull. Amer. Math. Soc., 69 (1963), pp. 789–791.
[23] R. Motwani and P. Raghavan, *Randomized Algorithms*, Cambridge University Press, 1995.
[24] H. Pasula, S. J. Russell, M. Ostland, and Y. Ritov, *Tracking many objects with many sensors*, 16th Int. Joint Conf. on Artificial Intelligence (IJCAI),1999, pp. 1160–1171.
[25] R. Sinkhorn, *A relatiohship between arbitrary positive matrices and double stochastic matrices*, A.. Math. Statist., 35 (1964), pp. 876–879.
[26] G. W. Soules, *Extending the Minc-Bregman upper bound for the permanent*, Linear and Multilinear Algebra, 47 (2000), pp. 77–91.
[27] L. G. Valiant, *The complexity of computing the permanent*, Theoret. Comput. Sci., 8 (1979), pp. 189–201.

7 Appendix

In this appendix we prove Lemma 3.1. First we restate the lemma by inverting both sides and removing references to row sums and entries of a matrix.

LEMMA 7.1. *Let $a \in (0,1]$ and $b \geq 0$ satisfy $a \leq b$. Then*

$$\frac{h(b)}{h(b-a)} \geq \exp\left(\frac{a}{h(b-a)}\right). \tag{7.28}$$

Proof. Set $g(x) := 1 + x + .5x^2 + (e - 2.5)x^3$. Then

$$(\forall x \in [0,1])(e^x \leq g(x)). \tag{7.29}$$

To see this, let $f_1(x) := g(x) - e^x$ so $f_1(0) = f_1(1) = 0$. Now $f_1''(x) = 1 + 6(e - 2.5)x - e^x$. The zeros of this function occur where the line $1 + 6(e - 2.5)x$ and the curve e^x intersect. Since e^x is convex this happens either 0, 1, or 2 times. Now $f_1''(0) = 0$, $f_1''(.5) > .006$, $f_1''(1) < -.4$ so there are two zeros, one at 0, and one at α_4, where $\alpha_4 \in [.5, 1]$. Since $f_1'(0) = 0$, and f_1'' is nonnegative on $[0, \alpha_4]$, then $f_1'(0) \geq 0$ on $[0, \alpha_4]$. Also, f_1'' is negative on $(\alpha_4, 1]$, and so f_1' has at most one zero in $(0, 1]$. By Rolle's Theorem, this means $f_1([0,1])$ is either in $[0, \infty)$ or $(-\infty, 0]$. Since $f_1(.5) > 0.003$, $f_1(x) \geq 0$ for all $x \in [0, 1]$.

To show (7.28), consider the LHS - RHS:

$$f_2(b,a) = \frac{h(b)}{h(b-a)} - \exp\left(\frac{a}{h(b-a)}\right). \tag{7.30}$$

So (7.28) is true if and only if $f_2(b,a) \geq 0$ for all $b \geq a$ and $a \in (0,1]$.

Case I: $b - a \geq 1$**.** We can use (7.29) to bound the exponential term since $0 \leq a/h(b-a) \leq 1$. To bound the first term, first note that

$$h'(x) = \begin{cases} 1 + 1/(2x), & x > 1 \\ e - 1, & x \in [0,1] \end{cases}$$

is nonincreasing. This means that

(7.31)
$$h(b) = h(b-a) + \int_{b-a}^{b} h'(x)\, \mathrm{d}x \geq h(b-a) + a \cdot h'(b).$$

Let $\tilde{a} = a/h(b-a)$, then

$$f_2(b,a) \geq 1 + \tilde{a}h'(b) - g(\tilde{a}). \tag{7.32}$$

Since $b > 1$, $h'(b) = 1 + 1/(2b)$. So

$$f_2(b,a) \geq \tilde{a}[1/(2b) - (1/2)\tilde{a} - (e-2.5)\tilde{a}^2], \tag{7.33}$$

which is nonnegative if and only if the bracketed term is nonnegative.

Now $\tilde{a} = a(b - a + (1/2)\ln(b-a) + e - 1)^{-1} \leq (b + e - 2)^{-1}$, and

(7.34)
$$\frac{1}{2b} - \frac{1}{2(b+e-2)} - \frac{e-2.5}{(b+e-2)^2} = \frac{(3-e)b + (e-2)^2}{2b(b+e-2)^2}.$$

Since $(3 - e) > 0$, the RHS is greater than 0 and $f_2(b,a) > 0$.

Case II: $b - a < 1$**.** Fix $a \in [0,1]$, and again let $\tilde{a} := a/h(b-a) = a/(1 + (e-1)(b-a))$. Rearranging gives $a = \tilde{a}[1 + b(e-1)]/[1 + \tilde{a}(e-1)]$, so

$$\frac{1}{h(b-a)} = \frac{\tilde{a}}{a} = \frac{1 + \tilde{a}(e-1)}{1 + b(e-1)}. \tag{7.35}$$

So (7.28) is equivalent to

$$\frac{h(b)(1 + \tilde{a}(e-1))}{1 + b(e-1)} \geq \exp(\tilde{a}). \tag{7.36}$$

The LHS is just a line in $\tilde{a}$, and the RHS is a convex function, and so it is sufficient to show that the equation holds for the endpoints of an interval containing all the values of $\tilde{a}$. Consider $\tilde{a}$ as a function of a. Then this has the form:

$$\tilde{a} = \frac{a}{c_1 - c_2 a}, \tag{7.37}$$

where $c_1 - c_2 a \geq 1$. The derivative becomes:

$$\frac{\mathrm{d}\tilde{a}}{\mathrm{d}a} = \frac{(c_1 - c_2 a) + c_2 a}{(c_1 - c_2 a)^2} > 0. \tag{7.38}$$

So $\tilde{a}$ is strictly increasing in a, and 1-1 over any interval. That means to evaluate (7.36) at the extreme points for $\tilde{a}$, we need only evaluate (7.28) at the extreme points of a. These extreme points vary depending on the value of b, and so two subcases are necessary.

Subcase IIa: $b \leq 1$**.** Then a lies in $[0,b]$, so evaluating (7.28) at 0 and b gives:

(7.39)
$$1 = \frac{h(b)}{h(b)} = \exp(0) = 1, \quad \frac{h(b)}{h(0)} = 1 + b(e-1) \geq \exp(b).$$

(The second inequality holds for all $b \in [0,1]$ since $\exp(b)$ is convex and the inequality is satisfied at $b = 0$ and $b = 1$.)

Subcase IIb: $b > 1$. Then a lies in $[b-1, 1]$. Evaluating at $a = b - 1$ gives:

$$\frac{h(b)}{h(1)} = \frac{b + .5\ln b + e - 1}{e} = 1 + \frac{b-1}{e} + \frac{.5\ln b}{e}$$

Using $g((b - 1/e) \leq \exp((b-1)/e)$ on the first two terms of the RHS means it suffices to show $.5(\ln b)/e \geq .5(b-1)^2/e^2 + (e-2.5)(b-1)^3/e^3$ for $b \in [1,2]$. The inequality holds with equality at $b = 1$ and strict inequality at $b = 1.5$ and $b = 2$. The derivative of $.5(\ln b)/e$ is strictly decreasing on $(1,2)$, while that of $.5(b-1)^2/e^2 + (e-2.5)(b-1)^3/e^3$ is strictly increasing, so the derivative of the difference has at most one zero in $(1,2)$, and so the inequality holds on the entire interval.

Now suppose $a = 1$. Then set $f_3(b) = f_2(b,1)$, and define $f_4(b)$ as: $\frac{\mathrm{d}}{\mathrm{d}b} f_3(b) = f_4(b)/(h(b-1)^2)$, where

$$\begin{aligned} f_4(b) &:= h(b-1)h'(b) - h(b)h'(b-1) \\ &\quad + \exp\left(\frac{1}{h(b-1)}\right) h'(b-1). \end{aligned}$$

For $b \in (1,2)$ and $b - 1 \in (0,1)$, $h(b) = b + .5\ln b + e - 1$, $h'(b) = 1 + 1/(2b)$, $h(b-1) = 1 + (e-1)(b-1)$, and $h'(b-1) = e - 1$. After simplifying

$$\begin{aligned} & h(b-1)h'(b) - h(b)h'(b-1) \\ = \;& .5 + 1.5e - e^2 - .5(e-2)b^{-1} - .5(e-1)\ln b, \end{aligned}$$

we have

$$f_4'(b) = \frac{e-2}{2b} - \frac{e-1}{2b} - \left(\frac{e-1}{h(b-1)}\right)^2 \exp\left(\frac{1}{h(b-1)}\right).$$

This makes $f_4'(b) < 0$ for $b \in [1,2]$, and $f_4(b)$ a strictly decreasing function on $[1,2]$. Since $\mathrm{sgn}(f_3') = \mathrm{sgn}(f_4)$, f_3' has at most one zero in $[1,2]$.

Now $f_3(1) = 0$, $f_3(2) > .05$ and $f_3'(1) = 1.5$, so the fact that $f_3'(b) = 0$ at most once in $[1,2]$ means $f_3(b) \geq 0$ for all $b \in [1,2]$.

Hence for any value of a and b, $f_2(b,a) \geq 0$, and the lemma is proved.