SIAM J. OPTIM.
Vol. 20, No. 3, pp. 1364–1377

ON THE COMPLEXITY OF NONNEGATIVE MATRIX FACTORIZATION[*]

STEPHEN A. VAVASIS[†]

Abstract. Nonnegative matrix factorization (NMF) has become a prominent technique for the analysis of image databases, text databases, and other information retrieval and clustering applications. The problem is most naturally posed as continuous optimization. In this report, we define an exact version of NMF. Then we establish several results about exact NMF: (i) that it is equivalent to a problem in polyhedral combinatorics; (ii) that it is NP-hard; and (iii) that a polynomial-time local search heuristic exists.

Key words. nonnegative matrix factorization, nonnegative rank, complexity, NP-hard, data mining, feature detection

AMS subject classifications. 15A23, 15A48, 68T05, 90C60, 90C26

DOI. 10.1137/070709967

1. Nonnegative matrix factorization.

Nonnegative matrix factorization (NMF) has emerged in the past decade as a powerful tool for clustering data and finding features in datasets. Lee and Seung [13] showed that NMF can find features in image databases, and Hofmann [11] showed that probabilistic latent semantic analysis, a variant of NMF, can effectively cluster documents according to their topics. Cohen and Rothblum [6] describe applications for NMF in probability, quantum mechanics, and other fields. The earliest reference to nonnegative factorization known to us is Thomas' solution [14] to a problem posed by Berman and Plemmons (which, according to a remark in the journal, was also solved by Ben-Israel).

NMF is defined as the following problem, which we denote as OPT-NMF. The input is (A, k), where A is an $m \times n$ matrix with nonnegative entries and k is an integer such that $1 \leq k \leq \min(m, n)$. The output is a pair of matrices (W, H) with $W \in \mathbf{R}^{m\times k}$ and $H \in \mathbf{R}^{k\times n}$ such that the distance from A to WH is minimized, subject to the constraints that W and H both have nonnegative entries. The distance can be measured as $N(A, WH)$, where N is a real-valued function such that

$$N(X, Y) \geq 0; \qquad N(X, Y) = 0 \Leftrightarrow X = Y, \tag{1}$$

for example, $N(A, WH) = \|A - WH\|$ for a standard matrix norm.

The precise choice for N may vary from one author to the next. (We could adopt the notation OPT$_N$-NMF for the problem under consideration if we needed to precisely specify the distance function.) Furthermore, some authors seek sparsity in either W or H or both. Sparsity may be imposed as a term in the objective function [12]. We will not pursue sparsity further herein.

The algorithms proposed by [11, 12, 13] and others for OPT-NMF have generally been based on local improvement heuristics. Another class of heuristics is based on greedy rank-one downdating [1, 2, 3, 9]. No algorithm proposed in the literature

[*] Received by the editors December 3, 2007; accepted for publication (in revised form) June 8, 2009; published electronically October 16, 2009.
http://www.siam.org/journals/siopt/20-3/70996.html

[†] Department of Combinatorics and Optimization, University of Waterloo, 200 University Ave. W., Waterloo, Ontario, N2L 3G1, Canada (vavasis@math.uwaterloo.ca). Supported in part by an NSERC award.

comes with a guarantee of optimality. This suggests that solving NMF to optimality may be a difficult problem, although to the best of our knowledge this has never been established formally.

The main purpose of this paper is to provide the proof that NMF is NP-hard. This paper considers a particular version of NMF that we call *exact NMF*, which is defined as follows.

EXACT NMF. The input is a matrix $A \in \mathbf{R}^{m\times n}$ with nonnegative entries whose rank is exactly k, $k \geq 1$. The output is a pair of matrices (W, H), where $W \in \mathbf{R}^{m\times k}$ and $H \in \mathbf{R}^{k\times n}$, W and H both have nonnegative entries and $A = WH$. If no such (W, H) exist, then the output is a statement of nonexistence of a solution. The decision version of EXACT NMF takes the same input and gives as output *yes* if such a W and H exist, else it outputs *no*.

Note that k is no longer explicitly an input to EXACT NMF. This is because the rank of A can be determined efficiently. If A is specified as rational data, then its rank may determined in polynomial time via reduction to row-echelon form [7]. In practice, one would usually prefer singular value decomposition to determine rank(A) [10].

Observe that for the OPT version of NMF, an optimal algorithm when presented with an A whose rank is exactly k must solve the exact NMF problem. This is true for any choice of N satisfying the axiom (1). Thus, the usual NMF problem proposed in the literature is a generalization of EXACT NMF. Therefore, any hardness result that applies to exact NMF (such as our hardness result) would presumably apply to most optimization versions as well.

A different generalization of EXACT NMF is the problem of nonnegative rank determination due to Cohen and Rothblum, which asks, given $A \in \mathbf{R}^{m\times n}$ with nonnegative entries, find the minimum value of k such that $A = WH$, $W \in \mathbf{R}^{m\times k}$, $H \in \mathbf{R}^{k\times n}$, and W, H have nonnegative entries. Cohen and Rothblum give a superexponential time algorithm for finding the rank (but not necessarily W or H). Since nonnegative rank determination is a generalization of EXACT NMF, our result shows that it is also NP-hard.

The proof of NP-hardness of EXACT NMF has two parts: In section 2 we show equivalence between EXACT NMF and a problem in polyhedral combinatorics that we call INTERMEDIATE SIMPLEX, and in section 3 we show the NP-hardness of this problem. A side result emerging from the proof of equivalence of EXACT NMF to INTERMEDIATE SIMPLEX is that a certain local-search heuristic for NMF can be solved with linear programming (section 4).

2. Equivalence to intermediate simplex.

In this section, we show an equivalence between EXACT NMF and a problem in polyhedral combinatorics that we call INTERMEDIATE SIMPLEX. Although the focus in this section is on the decision version of these problems, it is apparent from the proofs that the search-versions could also be reduced to each other. (These reductions, however, do not necessarily preserve the approximation properties of the search version; see the concluding discussion for more remarks.) The reductions use a number of arithmetic operations polynomials in m and n and are therefore polynomial time for both the usual Turing machine model and the real-number model of Blum et al. [4].

A problem related to INTERMEDIATE SIMPLEX was proposed by Cohen and Rothblum [6] and shown to be equivalent to nonnegative rank determination. This was also understood by Thomas [14] and Ben-Israel. Therefore, these earlier results to some extent imply the results of this section. Nonetheless, we present the equivalence here in order to fully support our claim that all reductions are polynomial time.

The equivalence is shown in three steps by first showing an equivalence to a problem denoted P1.

*P*1. Given matrices $W_0 \in \mathbf{R}^{m\times k}$ and $H_0 \in \mathbf{R}^{k\times n}$ such that each has rank k and such that all entries of $W_0 H_0$ are nonnegative, does there exist a nonsingular matrix $Q \in \mathbf{R}^{k\times k}$ such that $W_0 Q^{-1}$ and QH_0 both have all entries nonnegative?

THEOREM 1. *There is a polynomial-time reduction from EXACT NMF to P*1 *and vice versa.*

Proof. First we demonstrate the reduction of EXACT NMF to P1. Suppose that we have an NMF instance, that is, a nonnegative matrix A of rank exactly k. In polynomial time (using, e.g., reduction to row-echelon form) one can factor $A = W_0 H_0$ such that $W_0 \in \mathbf{R}^{m\times k}$ and $H_0 \in \mathbf{R}^{k\times n}$. (This factorization does not solve exact NMF, since the signs of the entries of W_0 and H_0 are unknown.) We claim that the original instance of EXACT NMF is a yes-instance if and only if the instance of P1 is a yes-instance. For one direction, suppose the instance of EXACT NMF is a yes-instance, and suppose W, H are solutions to exact NMF. Then clearly $\text{Range}(A) = \text{Range}(W) = \text{Range}(W_0)$, which is a dimension-$k$ subspace of $\mathbf{R}^n$, and similarly $\text{Range}(A^T) = \text{Range}(H^T) = \text{Range}(H_0^T)$. This means that there exist two nonsingular $k\times k$ nonsingular matrices, say, P, Q, such that $W = W_0 P$ and $H = QH_0$. Thus, the equation $WH = W_0 H_0$ may be rewritten as $W_0 PQH_0 = W_0 H_0$. Notice that W_0 has a left inverse and H_0 has a right-inverse, since W_0 has full column rank and H_0 has full row rank. Thus, the previous equation simplifies to $PQ = I$ (where I denotes the $k \times k$ identity matrix), i.e., $P = Q^{-1}$. Thus, $W_0 Q^{-1}$ and QH_0 both have nonnegative entries, so the instance of P1 is a yes-instance. Conversely, suppose the instance of P1 is a yes-instance. Then there exists Q such that $W = W_0 Q^{-1}$ and $H = QH_0$ both have all nonnegative entries and $WH = W_0 H_0 = A$, so the instance of exact NMF is a yes-instance.

For the opposite reduction, suppose we start with an instance (W_0, H_0) of P1. Let $A = W_0 H_0$; then A is nonnegative and has rank k. We claim that the instance of A is a yes-instance if and only if the instance of P1 is a yes-instance. The proof uses essentially the same arguments as in the previous paragraph. □

In order to simplify the main proof in this section, it is helpful to define a slightly restricted version of P1 as follows.

*RESTRICTED P*1. Given matrices $W_0 \in \mathbf{R}^{m\times k}$ and $H_0 \in \mathbf{R}^{k\times n}$ such that (i) W_0 has rank k; (ii) all entries of $W_0 H_0$ are nonnegative; (iii) the last column of W_0 is all 1's; and (iv) there is no nonzero solution $\mathbf{x} \in \mathbf{R}^{k-1}$ to the inequality $[\mathbf{x}^T, 0]H_0 \geq \mathbf{0}$, does there exist a nonsingular matrix $Q \in \mathbf{R}^{k\times k}$ such that $W_0 Q^{-1}$ and QH_0 both have all nonnegative entries?

Remark 1. Side condition (iv) can be checked in polynomial time by solving a linear programming problem. However, checking this condition is not necessary because the reduction of P1 to RESTRICTED P1 presented below can be modified to produce a certificate that condition (iv) holds (in addition to the instance of RESTRICTED P1).

Remark 2. Note that we dropped the side condition that H_0 has rank k because it is implied by the others. To see this, suppose $\mathbf{g} \in \mathbf{R}^k$ is a solution to $\mathbf{g}^T H_0 = \mathbf{0}$. If the last entry of $\mathbf{g}$ is zero, then $\mathbf{g} = \mathbf{0}$ because of side condition (iv). If the last entry of $\mathbf{g}$ is nonzero, this will lead to a contradiction of the conditions. Assuming the last entry of $\mathbf{g}$ is nonzero, without loss of generality, it may be taken to be 1 (by rescaling). Since W_0 has rank k, there is a row of W_0, say, $\mathbf{w}_i^T$, such that $\mathbf{w}_i \neq \mathbf{g}$. Let $\mathbf{p} = \mathbf{w}_i - \mathbf{g}$. Then $\mathbf{p} \neq \mathbf{0}$, but $p(k) = 0$ (because $w_i(k) = g(k) = 1$). Since $\mathbf{w}_i^T H_0 \geq \mathbf{0}$ by condition (ii) and $\mathbf{g}^T H_0 = \mathbf{0}$, then $\mathbf{p}^T H_0 \geq \mathbf{0}$. The existence of this $\mathbf{p}$, however, contradicts condition (iv). Thus, the conditions imply that the only solution to $\mathbf{g}^T H_0 = \mathbf{0}$ is $\mathbf{g} = \mathbf{0}$, which is the same as saying that $\mathrm{rank}(H_0) = k$.

THEOREM 2. *There is a polynomial-time reduction from* P1 *to RESTRICTED* P1 *and vice versa.*

Proof. Given an instance (W_0, H_0) of P1, we can produce an instance of RESTRICTED P1 as follows. First, delete all rows of W_0 that are identically 0's. This does not affect the rank of W_0, nor does it affect whether the product W_0H_0 is nonnegative. Finally, if Q is a solution problem P1 prior to deletion of identically zero rows, then it is still a solution afterwards and vice versa.

For the next step, let $\hat{Q}$ be a $k\times k$ nonsingular matrix chosen such that $\hat{Q}H_0\mathbf{e} = \mathbf{e}_k$. Here, $\mathbf{e} \in \mathbf{R}^n$ denotes the vector of all 1's, and $\mathbf{e}_k \in \mathbf{R}^k$ denotes the last column of the $k \times k$ identity matrix. Such a $\hat{Q}$ is guaranteed to exist because $H_0\mathbf{e}$ cannot be zero: $W_0H_0\mathbf{e}$ is the sum of columns of W_0H_0, which cannot be zero since the columns of W_0H_0 are all nonnegative and W_0H_0 is not identically zero by the assumption of rank at least 1. Then observe that $(W_0\hat{Q}^{-1}, \hat{Q}H_0)$ is a yes-instance of P1 if and only if (W_0, H_0) is a yes-instance. Such a $\hat{Q}$ may be found in polynomial time; for example, any $k \times k$ nonsingular matrix whose last column is $H_0\mathbf{e}$ may be taken as $\hat{Q}^{-1}$, and matrix inversion is polynomial time in the Turing machine model [7].

Next, we observe that the last column of $W_0\hat{Q}^{-1}$ is $W_0\hat{Q}^{-1}\mathbf{e}_k = W_0\hat{Q}^{-1}\hat{Q}H_0\mathbf{e} = W_0H_0\mathbf{e}$. We already argued above that this vector is nonzero, but now we will argue more strongly that every entry of $W_0H_0\mathbf{e}$ is positive. First, note that $W_0H_0\mathbf{e}$ is the sum of columns of the nonnegative matrix W_0H_0, and hence all its entries are at least nonnegative. Focus on entry i of $W_0H_0\mathbf{e}$; since it is a sum of nonnegative terms, then if it were zero, then the entire ith row of W_0H_0 would have to be zeros. This means that the ith row of W_0 is orthogonal to every column of H_0. But since H_0 has full rank, this is possible only if the ith row of W_0 is identically 0. However, this possibility is ruled out, since we deleted identically zero rows of W_0.

Thus, the last column of $W_0\hat{Q}^{-1}$ contains all positive entries. Therefore, we define an instance of RESTRICTED P1 given by $(DW_0\hat{Q}^{-1}, \hat{Q}H_0)$, where D is an $m \times m$ positive definite diagonal matrix with diagonal entries chosen to make the last column of $DW_0\hat{Q}^{-1}$ equal to all 1's. This instance of P1 is a yes-instance only if the original instance was a yes-instance, because multiplying the first factor by a positive definite diagonal matrix does not affect the signs of W_0H_0 nor of $W_0\hat{Q}^{-1}Q^{-1}$.

We have already verified side conditions (i)–(iii) of the instance produced by the condition, and we can check condition (iv) as follows. Suppose $[\mathbf{x}^T, 0]\hat{Q}H_0 \geq \mathbf{0}$. Multiply on the right by $\mathbf{e}$ to obtain $[\mathbf{x}^T, 0]\hat{Q}H_0\mathbf{e} = [\mathbf{x}^T, 0]\mathbf{e}_k = 0$. Thus, each entry of $[\mathbf{x}^T, 0]\hat{Q}H_0$ is nonnegative, and these entries add to 0, so they must be all zero. Since $\hat{Q}H_0$ has rank k, this possible only if $[\mathbf{x}^T, 0]$ is the zero vector.

The opposite reduction, namely, the one from RESTRICTED P1 to P1, is trivial since any instance of RESTRICTED P1 is also an instance of P1. □

Now finally, we get to the main new problem of this section.

INTERMEDIATE SIMPLEX. We are given a bounded polyhedron $P = \{\mathbf{x} \in \mathbf{R}^{k-1} : A\mathbf{x} \geq \mathbf{b}\}$, where $A \in \mathbf{R}^{n\times(k-1)}$ and $\mathbf{b} \in \mathbf{R}^n$. We are also given a set $S \subset \mathbf{R}^{k-1}$ of m points that are all contained in P and that are not all contained in any hyperplane (i.e., they affinely span $\mathbf{R}^{k-1}$). The question is whether there exists a $(k-1)$-simplex T such that $S \subset T \subset P$.

THEOREM 3. *There is a polynomial-time reduction from RESTRICTED* P1 *to INTERMEDIATE SIMPLEX and vice versa.*

Proof. We will prove that both reductions exist at the same time by exhibiting a bijection between instances of RESTRICTED P1 and instances of INTERMEDIATE SIMPLEX such that both directions of the bijection can be computed in polynomial time.

Given an instance (W_0, H_0) of RESTRICTED P1, we produce an instance of INTERMEDIATE SIMPLEX as follows. The polytope $P \subset \mathbf{R}^{k-1}$ is given by $\{\mathbf{x} \in \mathbf{R}^{k-1} : (H_0(1:k-1,:))^T\mathbf{x} \geq -(H_0(k,:))^T\}$. This constraint may be written more compactly as $H_0^T[\mathbf{x};1] \geq \mathbf{0}$. The set S of m points in P is given by $S = \{(W_0(1,1:k-1))^T, \ldots, (W_0(m,1:k-1))^T\}$. The inverse mapping of this transformation starts with an instance of INTERMEDIATE SIMPLEX given by $P = \{\mathbf{x} : A\mathbf{x} \geq \mathbf{b}\}$, $A \in \mathbf{R}^{n\times(k-1)}$, and $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ and produces an instance of RESTRICTED P1 given by

$$W_0 = \begin{pmatrix} \mathbf{x}_1^T & 1 \\ \vdots & \vdots \\ \mathbf{x}_m^T & 1 \end{pmatrix}$$

and $H_0 = [A^T; -\mathbf{b}^T]$.

We first show that all side constraints present in the statement of RESTRICTED P1 are equivalent under this bijection to the side constraints of INTERMEDIATE SIMPLEX. The side constraint that $\mathbf{x}_1, \ldots, \mathbf{x}_m$ affinely span $\mathbf{R}^{k-1}$ is equivalent to requiring that $[\mathbf{x}_1; 1], \ldots, [\mathbf{x}_m; 1]$ linearly span $\mathbf{R}^k$, i.e., to the side constraint that W_0 has rank k. The side constraint that $S \subset P$ means that $A\mathbf{x}_i \geq \mathbf{b}$ for $i = 1, \ldots, m$, i.e., $[A, -\mathbf{b}][\mathbf{x}_i; 1] \geq 0$, which is equivalent to the side constraint that all entries of W_0H_0 are nonnegative.

Finally, the side constraint that P is bounded is equivalent to requiring that the only solution to $A\mathbf{x} \geq \mathbf{0}$ is $\mathbf{x} = \mathbf{0}$. In turn, this is equivalent to the side constraint of RESTRICTED P1 that there is no nontrivial solution to $(H_0(1:k-1,:))^T\mathbf{x} \geq \mathbf{0}$.

We now show that the above bijection in both directions maps yes-instances to yes-instances. Let (S, P) be a yes-instance of INTERMEDIATE SIMPLEX and (W_0, H_0) the corresponding instance of RESTRICTED P1. Let T be a solution to the instance of INTERMEDIATE SIMPLEX. Let its vertices be $\mathbf{g}_1, \ldots, \mathbf{g}_k$, which are vectors in $\mathbf{R}^{k-1}$. The condition that $T \subset P$ is equivalent to requiring $\mathbf{g}_1, \ldots, \mathbf{g}_k \in P$, i.e., to $H_0^T[\mathbf{g}_i; 1] \geq \mathbf{0}$ for each $i = 1, \ldots, k$. If we let

$$G = \begin{pmatrix} \mathbf{g}_1 & \cdots & \mathbf{g}_k \\ 1 & \cdots & 1 \end{pmatrix}, \tag{2}$$

then we have shown that the condition $T \subset P$ implies that H_0^TG has all nonnegative entries.

The condition that $S \subset T$ means that for all $i = 1, \ldots, m$, $\mathbf{x}_i \in T$. Recall that, by definition, a vector is inside a simplex if it is a convex combination of its vertices. Let $\mathbf{q}_i$ be the putative vector of coefficients of the convex combination that expresses $\mathbf{x}_i$ in the hull of the vertices of T for $i = 1, \ldots, m$. In other words,

$$[\mathbf{g}_1, \ldots, \mathbf{g}_k]\mathbf{q}_i = \mathbf{x}_i, \tag{3}$$

plus the requirements that the entries of $\mathbf{q}_i$ are nonnegative and sum to 1. The latter constraint may be combined with (3) to write $G\mathbf{q}_i = [\mathbf{x}_i; 1]$, where G is as in (2), i.e., $\mathbf{q}_i = G^{-1}(W_0(i,:))^T$. The hypothesis that $S \subset T$ is thus equivalent to the condition that each entry of $G^{-1}W_0^T$ for each $i = 1, \ldots, m$ is nonnegative, i.e., all entries of $G^{-1}W_0^T$ must be nonnegative. Hence, we have shown that if T is a solution to the instance (S, P), then G^T is a solution to the instance (W_0, H_0) of RESTRICTED P1.

For the other direction, let Q be a solution to RESTRICTED P1. Let $\hat{\mathbf{g}}_1, \ldots, \hat{\mathbf{g}}_k$ be the columns of Q^T, none of which can be zero. We claim that the last entry of each $\hat{\mathbf{g}}_i$ is positive. Observe that $H_0^T \hat{\mathbf{g}}_i \geq \mathbf{0}$, since $\hat{\mathbf{g}}_i$ is a solution to RESTRICTED P1. Note that the last entry of $\hat{\mathbf{g}}_i$ cannot be zero because of the side condition that there is no nontrivial solution to $H_0^T \hat{\mathbf{g}}_i \geq \mathbf{0}$ whose last coordinate is zero. We claim the last coordinate of $\hat{\mathbf{g}}_i$ cannot be negative either; if it were, then by rescaling $\hat{\mathbf{g}}_i$, we could take its last entry to be -1, which means that its first $k-1$ entries, say, $\mathbf{g}_i$ (after rescaling) would constitute a solution to $[\mathbf{g}_i^T, -1]^T H_0 \geq \mathbf{0}$, i.e., $A\mathbf{g}_i \geq -\mathbf{b}$. We already have k linearly independent solutions to $A\mathbf{x}' \geq \mathbf{b}$ (namely, the rows of W_0), so if we add $\hat{\mathbf{g}}_i$ to $\mathbf{x}'$, we would have a solution to $A\mathbf{y} \geq \mathbf{0}$. This solution is nontrivial for at least one of the choices of $\mathbf{x}'$, contradicting the hypothesis that there is no nontrivial solution to this equation.

Thus, each $\hat{\mathbf{g}}_i$ has a positive number for its last entry. By rescaling the $\hat{\mathbf{g}}_i$'s if necessary, i.e., replacing the RESTRICTED P1 solution Q by DQ for a positive definite diagonal matrix D, we can assume that each $\hat{\mathbf{g}}_i$ has 1 as its last entry. Then we claim that $\mathbf{g}_1, \ldots, \mathbf{g}_k$ that are defined to be entries 1 to $k-1$ of $\hat{\mathbf{g}}_1, \ldots, \hat{\mathbf{g}}_k$, respectively, constitute a solution to INTERMEDIATE SIMPLEX. It is clear that $A\mathbf{g}_i \geq \mathbf{b}$ for each i, because this is equivalent to $H_0^T \hat{\mathbf{g}}_i \geq \mathbf{0}$. Also, using the same arguments as above, a row of WQ^{-1}, say, $W(i,:)Q^{-1}$, corresponds to the coefficients needed to express $W(i,:)$ as a convex combination of $\hat{\mathbf{g}}_1, \ldots, \hat{\mathbf{g}}_k$. ∎

An easy consequence of the transformation of EXACT NMF to INTERMEDIATE SIMPLEX is the observation that when $\operatorname{rank}(A) = 2$, the NMF instance is always a yes-instance. The reason is that the resulting instance of INTERMEDIATE SIMPLEX is one-dimensional, in which case P is an interval. However, if P is an interval, then it is already a simplex, so one could take $T = P$ to solve the instance. This observation yields a simple linear-time algorithm to find an exact nonnegative factorization of A in the case of $\operatorname{rank}(A) = 2$. This result was first established by Cohen and Rothblum [6], who also propose a simple linear-time algorithm. This observation was also used by Boutsidis and Gallopoulos [5] to develop a heuristic algorithm for NMF.

3. INTERMEDIATE SIMPLEX is NP-hard.

In this section, we will argue that the problem INTERMEDIATE SIMPLEX introduced in the previous section is NP-hard.

Before delving into the statement of the main theorem and its proof, we first state the following simpler lemma. This lemma describes the "gadget" used in the main theorem below to encode a setting of a boolean variable.

LEMMA 1. *Consider the following instance of INTERMEDIATE SIMPLEX: the polyhedron P is given by $P = \{(x, y) \in \mathbf{R}^2 : 0 \leq x, y \leq 1\}$, while the set S is given by $\{(0, 1/2), (1, 1/2), (1/2, 1/4), (1/2, 3/4)\}$. This instance has precisely two solutions T_0 or T_1 defined by $T_0 = \operatorname{hull}\{(0, 0), (0, 1), (1, 1/2)\}$ and $T_1 = \operatorname{hull}\{(1, 0), (1, 1), (0, 1/2)\}$.*

A diagram of the lemma is given in Figure 1. We omit the full proof, since the lemma can be understood from the diagram. The full proof involves taking cases on whether the two points $(0, 1/2)$ and $(1, 1/2)$ are contained in zero- or one-dimensional faces of the simplex T.

We now turn to the main result for this section, namely, the NP-hardness of INTERMEDIATE SIMPLEX. In particular, we reduce 3-SAT [8] to this problem. Our

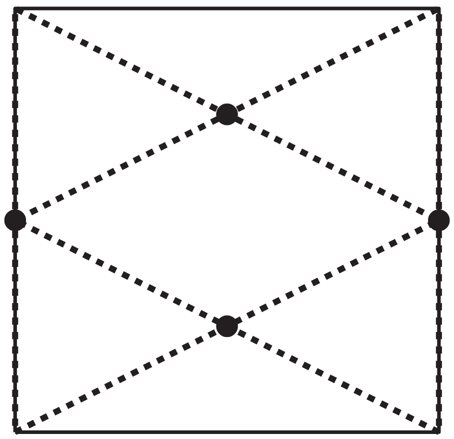

FIG. 1. *Illustration of Lemma* 1. *The four large dots are the points in S; the thin solid line is the boundary of P, and the two triangles indicated with thick dashed lines are the two possible solutions T_0 and T_1.*

reduction uses integers whose magnitude is polynomial in the instance of the 3-SAT instance, and hence our result is "strong" NP-hardness. Recall that an instance of 3-SAT involves p boolean variables denoted $x_1, \ldots, x_p$ and q clauses denoted $c_1, \ldots, c_q$. Each clause is a disjunction of three literals, where a literal is either a variable x_j or its complement $\tilde{x}_j$. An instance of 3-SAT is a yes-instance if and only if there exists a setting of the variables, that is, an assignment of a value of either 0 or 1 to each variable, such that each clause is satisfied, i.e., at least one of its three literals is 1. It is assumed that the same variable does not occur twice (either in complemented or plain form) in any particular clause.

Given such an instance of 3-SAT, we define the following instance of INTERMEDIATE SIMPLEX. It contains $3p+q$ variables (i.e., $k-1 = 3p+q$) denoted s_i, t_i, u_i, $i = 1, \ldots, p$, and v_j, $j = 1, \ldots, q$. These variables are written as $(\mathbf{s}, \mathbf{t}, \mathbf{u}, \mathbf{v})$ for short. The polyhedron P is defined by the following inequalities:

$$P = \left\{ \begin{array}{lll} (\mathbf{s},\mathbf{t},\mathbf{u},\mathbf{v}) : & \mathbf{0} \le \mathbf{s} \le \mathbf{u}, & \\ & \mathbf{0} \le \mathbf{t} \le \mathbf{u}, & \\ & \mathbf{0} \le \mathbf{u} \le \mathbf{e}, & \\ & \mathbf{0} \le \mathbf{v} \le 5q\mathbf{e}, & \\ & s_i - 2t_i \le v_j & \text{whenever } \tilde{x}_i \in c_j, \\ & 2t_i - 2s_i - u_i \le v_j & \text{whenever } x_i \in c_j \end{array} \right\}. \tag{4}$$

Here, $\mathbf{e}$ denotes the vector of all 1's either in $\mathbf{R}^p$ or $\mathbf{R}^q$. Let $\mathbf{e}_i$ denote the ith column of the identity matrix (either the $p \times p$ or $q \times q$ identity). The set of points S is defined as follows. Each of the points in the following equation is also given a name for future reference:

$$S = \left\{ \begin{array}{lll} \mathbf{0}, & & \\ (\mathbf{e}/(4p), \mathbf{e}/(4p), \mathbf{e}/(2p), 2.5\mathbf{e}/(8p)) & (\equiv \mathbf{b}), & \\ (\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{e}_j) & (\equiv \mathbf{h}_j), & j = 1, \ldots, q, \\ (\mathbf{0}, \mathbf{e}_i/4, \mathbf{e}_i/2, \mathbf{e}) & (\equiv \mathbf{r}_i^1), & i = 1, \ldots, p, \\ (\mathbf{e}_i/2, \mathbf{e}_i/4, \mathbf{e}_i/2, \mathbf{e}) & (\equiv \mathbf{r}_i^2), & i = 1, \ldots, p, \\ (\mathbf{e}_i/4, \mathbf{e}_i/8, \mathbf{e}_i/2, \mathbf{e}) & (\equiv \mathbf{r}_i^3), & i = 1, \ldots, p, \\ (\mathbf{e}_i/4, 3\mathbf{e}_i/8, \mathbf{e}_i/2, \mathbf{e}) & (\equiv \mathbf{r}_i^4), & i = 1, \ldots, p. \end{array} \right\}. \tag{5}$$

Let us first confirm that the side constraints of INTERMEDIATE SIMPLEX are satisfied by this instance. Since $\mathbf{0} \in S$, S affinely spans $\mathbf{R}^{3p+q}$ if and only if it linearly spans $\mathbf{R}^{3p+q}$. Points $\mathbf{h}_j$, $j = 1, \ldots, q$, span the subspace defined by the last q coordinate entries. Fix some $i \in \{1, \ldots, p\}$. Subtract $\mathbf{h}_1 + \cdots + \mathbf{h}_q$ from the three points $\mathbf{r}_i^1, \mathbf{r}_i^2, \mathbf{r}_i^3$. This yields three points whose nonzero entries are restricted to the (s_i, t_i, u_i) positions; in these positions the three points have coordinate entries $(0, 1/4, 1/2)$, $(1/2, 1/4, 1/2)$, and $(1/4, 1/8, 1/2)$, which are linearly independent. Thus, the subspace indexed by (s_i, t_i, u_i) is spanned by S. This is true for all i, so therefore the points in S span all of $\mathbf{R}^{3p+q}$.

The next side constraint is that P is bounded. This is clear from the upper and lower bound on the variables. The final side constraint is that $S \subset P$, which is an elementary matter to check.

The main theorem of this section is as follows.

THEOREM 4. *The instance of* 3*-SAT is a yes-instance if and only if the above instance of INTERMEDIATE SIMPLEX is a yes-instance. In other words, the* 3*-SAT instance has a satisfying assignment if and only if there exists a simplex T such that $S \subset T \subset P$.*

Proof. First, let us choose some terminology for the coordinates of $\mathbf{R}^{3p+q}$. The individual coordinates may be denoted by s_i, t_i, u_i, or v_j for $i = 1, \ldots, p$ and $j = 1, \ldots, q$. Collectively, the three coordinates (s_i, t_i, u_i) are called the "x_i coordinates," since they correspond to the ith boolean variable in the 3-SAT instance.

Let T be a solution to the instance of INTERMEDIATE SIMPLEX. From T we will construct a satisfying assignment σ for the 3-SAT instance. Clearly, T has exactly $3p + q + 1$ vertices. Observe first that the point $\mathbf{0}$ is an extreme point of P and also lies in S, and therefore one vertex of T must be $\mathbf{0}$.

Similarly, observe that each $\mathbf{h}_j$, $j = 1, \ldots, q$, lies on extreme edge of P, and therefore T must have q vertices of the form $\lambda_j \mathbf{h}_j$, $j = 1, \ldots, q$ with each $\lambda_j \geq 1$.

This accounts for all but $3p$ of the vertices of T. For an $i \in \{1, \ldots, p\}$, let us say that a vector $(\mathbf{s}, \mathbf{t}, \mathbf{u}, \mathbf{v}) \in \mathbf{R}^{3p+q}$ is *x_i-supported* if it is zero in all the x_j coordinates for all $j \in \{1, \ldots, p\} - \{i\}$. More strongly, say that it is *x_i-positive* if it is x_i-supported and is positive in at least one of the x_i coordinates. Fix a particular $i \in \{1, \ldots, p\}$, and consider the four S-points $\mathbf{r}_i^1, \ldots, \mathbf{r}_i^4$ which are all x_i-positive. Projected into the x_i coordinates, these points are $(0, 1/4, 1/2)$, $(1/2, 1/4, 1/2)$, $(1/4, 1/8, 1/2)$, and $(1/4, 3/8, 1/2)$. Since none of the T-vertices has negative entries, each of $\mathbf{r}_i^1, \ldots, \mathbf{r}_i^4$ must lie in the hull only of T-vertices that are x_i-supported, such as $\mathbf{0}, \lambda_1 \mathbf{h}_1, \ldots, \lambda_q \mathbf{h}_q$. Furthermore, it must lie in the hull of at least one x_i-positive vertex of T. In fact, there must be at least three such x_i-positive T-vertices, since the four points, when projected into the x_i coordinates, are linearly independent. Thus, T must have at least three x_i-positive vertices for each $i = 1, \ldots, p$. Since there are only $3p$ vertices of T not yet enumerated, we conclude that T must have exactly three x_i-positive vertices for each i, which we denote $\mathbf{g}_{i,1}, \mathbf{g}_{i,2}, \mathbf{g}_{i,3}$.

Let $\bar{\mathbf{g}}_{i,1}, \bar{\mathbf{g}}_{i,2}, \bar{\mathbf{g}}_{i,3} \in \mathbf{R}^3$ denote the x_i coordinates of $\mathbf{g}_{i,1}, \mathbf{g}_{i,2}, \mathbf{g}_{i,3}$. By the assumption that T covers the four points $(0, 1/4, 1/2)$, $(1/2, 1/4, 1/2)$, $(1/4, 1/8, 1/2)$, and $(1/4, 3/8, 1/2)$ in the projection into the x_i coordinates, we conclude that there must exist a 3×4 matrix B with nonnegative entries such that

$$(\bar{\mathbf{g}}_{i,1}, \bar{\mathbf{g}}_{i,2}, \bar{\mathbf{g}}_{i,3}) \cdot B = \begin{pmatrix} 0 & 1/2 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/8 & 3/8 \\ 1/2 & 1/2 & 1/2 & 1/2 \end{pmatrix}.$$

As mentioned above, all of $\bar{\mathbf{g}}_{i,1}, \bar{\mathbf{g}}_{i,2}, \bar{\mathbf{g}}_{i,3}$ are nonzero. Because of the inequalities

$\mathbf{0} \le \mathbf{s} \le \mathbf{u}$ and $\mathbf{0} \le \mathbf{t} \le \mathbf{u}$ that define P, it must be the case that the third entries of $\bar{\mathbf{g}}_{i,1}, \bar{\mathbf{g}}_{i,2}, \bar{\mathbf{g}}_{i,3}$ are all positive and no smaller than the first and second entries. Therefore, define new vectors $\hat{\mathbf{g}}_{i,1}, \hat{\mathbf{g}}_{i,2}, \hat{\mathbf{g}}_{i,3}$ that are all exactly $1/2$ in the last coordinate and have other coordinates lying in $[0, 1/2]$ obtained by rescaling each of $\bar{\mathbf{g}}_{i,1}, \bar{\mathbf{g}}_{i,2}, \bar{\mathbf{g}}_{i,3}$ by twice its third coordinate. By rescaling B in a reciprocal manner, we find that there is a nonnegative matrix $\hat{B}$ such that

$$(\hat{\mathbf{g}}_{i,1}, \hat{\mathbf{g}}_{i,2}, \hat{\mathbf{g}}_{i,3}) \cdot \hat{B} = \begin{pmatrix} 0 & 1/2 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/8 & 3/8 \\ 1/2 & 1/2 & 1/2 & 1/2 \end{pmatrix}.$$

By considering the third row of the above system of equations, we conclude that each column of $\hat{B}$ sums to exactly 1. Then dropping the third row on both sides yields the equation

$$(\hat{\mathbf{g}}_{i,1}(1:2), \hat{\mathbf{g}}_{i,2}(1:2), \hat{\mathbf{g}}_{i,3}(1:2)) \cdot \hat{B} = \begin{pmatrix} 0 & 1/2 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/8 & 3/8 \end{pmatrix},$$

where the notation $\mathbf{v}(1:2)$ denotes the first two entries of a vector. Now we observe that this is precisely a half-sized version of the instance of INTERMEDIATE SIMPLEX described in the preliminary lemma of this section, namely, find three points lying in $[0, 1/2]^2$ whose convex hull covers the four points $\{(0, 1/4), (1/2, 1/4), (1/4, 1/8), (1/4, 3/8)\}$. As established by the lemma, there are precisely two solutions to this system, which we will denote $T_0/2$ and $T_1/2$. Let C_0 be the set of i's such that the triangle defined by $(\hat{\mathbf{g}}_{i,1}(1:2), \hat{\mathbf{g}}_{i,2}(1:2), \hat{\mathbf{g}}_{i,3}(1:2))$ is $T_0/2$, while C_1 is the set of i's such that this triangle is $T_1/2$. Thus we conclude that for $i \in C_0$,

$$(\bar{\mathbf{g}}_{i,1}, \bar{\mathbf{g}}_{i,2}, \bar{\mathbf{g}}_{i,3}) = (\mu_{i,1}(0, 0, 1), \mu_{i,2}(0, 1, 1), \mu_{i,3}(1, 1/2, 1)), \tag{6}$$

and, for $i \in C_1$,

$$(\bar{\mathbf{g}}_{i,1}, \bar{\mathbf{g}}_{i,2}, \bar{\mathbf{g}}_{i,3}) = (\mu_{i,1}(1, 0, 1), \mu_{i,2}(1, 1, 1), \mu_{i,3}(0, 1/2, 1)), \tag{7}$$

where $\mu_{i,k} > 0$ for $k = 1, 2, 3$. This determines the x_i entries of $\mathbf{g}_{i,k}$, $k = 1, 2, 3$, and the remaining x_j entries are zeros since $\mathbf{g}_{i,k}$ is x_i-positive. Therefore, it remains only to determine the v_j entries of $\mathbf{g}_{i,k}$, $k = 1, 2, 3$. There are several constraints on these entries as follows. First, we have the inequalities $v_j \ge 0$, and thus all those entries must be nonnegative. Next, we have the constraints $s_i - 2t_i \le v_j$ whenever $\tilde{x}_i \in c_j$ and $2t_i - 2s_i - u_i \le v_j$ whenever $x_i \in c_j$. These inequalities are redundant whenever their left-hand side is nonpositive, since we have already constrained $v_j \ge 0$. Thus, we need only consider the cases when the left-hand sides are positive. We see that the left-hand side of the first inequality $s_i - 2t_i \le v_j$ is positive only in the case of $\bar{\mathbf{g}}_{i,1}$ only for $i \in C_1$, and the left-hand side of the second inequality $2t_i - 2s_i - u_i \le v_j$ is positive only in the case of $\bar{\mathbf{g}}_{i,2}$ only for $i \in C_0$. Thus, for $i \in C_1$, for all j such that $\tilde{x}_i$ occurs as a literal in clause c_j, we must have

$$\mathbf{g}_{i,1}|_{v_j} \ge \mu_{i,1}. \tag{8}$$

(Here, the notation $\mathbf{g}_{i,1}|_{v_j}$ denotes the v_j coordinate entry of $\mathbf{g}_{i,1}$.) Similarly, for $i \in C_0$, for all j such that x_i occurs as a literal in clause c_j, we must have

$$\mathbf{g}_{i,2}|_{v_j} \ge \mu_{i,2}. \tag{9}$$

Next, T must contain the point $\mathbf{b}$ from (5), so there must be coefficients $\alpha_{i,k}$, $i = 1,\ldots,p$, $k = 1,2,3$, and θ_j, $j = 1,\ldots,q$, adding up to at most 1 and all nonnegative such that

$$\mathbf{b} = \sum_{i=1}^{p}\sum_{k=1}^{3}\alpha_{i,k}\mathbf{g}_{i,k} + \sum_{j=1}^{q}\theta_j\lambda_j\mathbf{h}_j. \tag{10}$$

Fix a particular i. The projection of $\mathbf{b}$ into x_i coordinates is $\bar{\mathbf{b}} = (1/(4p), 1/(4p), 1/(2p))$. Referring back to (6) and (7), one can see that regardless of whether $i \in C_0$ or $i \in C_1$, $\bar{\mathbf{b}}$ is expressed uniquely as $\bar{\mathbf{b}} = \bar{\mathbf{g}}_{i,1}/(8p\mu_{i,1}) + \bar{\mathbf{g}}_{i,2}/(8p\mu_{i,2}) + \bar{\mathbf{g}}_{i,3}/(4p\mu_{i,3})$. Therefore,

$$\alpha_{i,1} = 1/(8p\mu_{i,1}); \quad \alpha_{i,2} = 1/(8p\mu_{i,2}); \quad \alpha_{i,3} = 1/(4p\mu_{i,3}). \tag{11}$$

Suppose $i \in C_0$. Then for each j such that x_i occurs as a literal in clause c_j, if we combine (9) and (11), we obtain

$$\left.\sum_{k=1}^{3}\alpha_{i,k}\mathbf{g}_{i,k}\right|_{v_j} \geq 1/(8p).$$

The identical inequality holds when $i \in C_1$ and $\tilde{x}_i \in c_j$.

Now, sum the preceding inequality for $i = 1,\ldots,p$ to obtain

$$\left.\sum_{i=1}^{p}\sum_{k=1}^{3}\alpha_{i,k}\mathbf{g}_{i,k}\right|_{v_j} \geq m_j/(8p), \tag{12}$$

where m_j is the number of literals $x_i \in c_j$ with $i \in C_0$ plus the number of literals $\tilde{x}_i \in c_j$ with $i \in C_1$. Let us now combine these inequalities: From (5), $\mathbf{b}|_{v_j} = 2.5/(8p)$. From (10),

$$\mathbf{b}|_{v_j} \geq \left.\sum_{i=1}^{p}\sum_{k=1}^{3}\alpha_{i,k}\mathbf{g}_{i,k}\right|_{v_j},$$

since the last term of (10) is nonnegative. Finally, from (12), the above summation is at least $m_j/(8p)$. Thus, we conclude that $m_j \leq 2.5$. Since m_j is integral, this means $m_j \leq 2$. Let σ be the setting of the x_i's in the 3-SAT instance defined by taking $x_i = 1$ for $i \in C_1$ and $x_i = 0$ for $i \in C_0$. Then if $x_i \in c_j$ and $i \in C_0$, this literal is falsified in the clause. Similarly, if $\tilde{x}_i \in c_j$ and $i \in C_1$, then this literal is also falsified. In other words, m_j is the number of literals in clause c_j falsified by assignment σ. We have just argued that $m_j \leq 2$ for all $j = 1,\ldots,q$. In other words, for each clause, there are at most two literals falsified by assignment σ. Therefore, σ is a satisfying assignment for the 3-SAT instance.

Summarizing, we have proved that if there is a simplex T solving the instance of INTERMEDIATE SIMPLEX, then there are exactly three vertices of T that are x_i-positive for each $i = 1,\ldots,p$; that, based on these vertices, i can be classified as either C_0 or C_1; and that the assignment σ of the boolean variables in the original 3-SAT instance derived from C_0 and C_1 must be a satisfying assignment.

Conversely, suppose the 3-SAT instance has a satisfying assignment. From this assignment we construct a solution T to the instance of INTERMEDIATE SIMPLEX. The vertices of T will be $\mathbf{0}, 5q\mathbf{h}_1, \ldots, 5q\mathbf{h}_q$ together with $\mathbf{g}_{i,1}, \mathbf{g}_{i,2}, \mathbf{g}_{i,3}$ for each $i = 1, \ldots, p$, defined as follows. Let C_0 index the variables set to 0 by the satisfying assignment and C_1 the variables set to 1. Define $\bar{\mathbf{g}}_{i,1}, \bar{\mathbf{g}}_{i,2}, \bar{\mathbf{g}}_{i,3}$ as in (6) and (7) according to C_0 and C_1. Take $\mu_{i,k} = 5/8$ for all (i,k). When $i \in C_0$ and x_i is a literal in c_j, then take $\mathbf{g}_{i,2}|_{v_j} = 5/8$. When $i \in C_1$ and $\tilde{x}_i$ is a literal in c_j, then take $\mathbf{g}_{i,1}|_{v_j} = 5/8$. In all other cases, take $\mathbf{g}_{i,k}|_{v_j} = 0$. It is easy to see that all the inequalities defining P are satisfied by these choices. Furthermore, all the points in S are covered by convex combinations of the $3p+q+1$ points $\mathbf{0}, 5q\mathbf{h}_1, \ldots, 5q\mathbf{h}_q, \mathbf{g}_{1,1}, \ldots, \mathbf{g}_{p,3}$, which are the vertices of T.

For example, the point $\mathbf{r}_i^1 = (\mathbf{0}, \mathbf{e}_i/4, \mathbf{e}_i/2, \mathbf{e})$ in the case where $i \in C_0$ is expressed as $(2/5)\mathbf{g}_{i,1} + (2/5)\mathbf{g}_{i,2} + \mathbf{h}$, where $\mathbf{h}$ is some linear combination of $5q\mathbf{h}_1, \ldots, 5q\mathbf{h}_q$ chosen to make the v_j entries each equal to 1. (Note that the v_j entries of $(2/5)\mathbf{g}_{i,1} + (2/5)\mathbf{g}_{i,2}$ before $\mathbf{h}$ is added will be either 0 or 1/4). The total sum of the coefficients to express $(\mathbf{0}, \mathbf{e}_i/4, \mathbf{e}_i/2, \mathbf{e})$ is $2/5 + 2/5 + h_1$, where h_1 is the sum of the coefficients needed in the terms of $\mathbf{h}$. These coefficients are bounded as follows. The v_j entry of $\mathbf{h}$ must be either 3/4 or 1. Therefore, the coefficient of $(5q)\mathbf{h}_j$ must be either $1/(5q)$ or $3/(20q)$. The sum of q such coefficients is at most 1/5. Thus, $h_1 \leq 1/5$. If the sum $2/5 + 2/5 + h_1$ is less than 1, then we include a contribution of $\mathbf{0}$, another vertex of T, in the linear combination to make the sum of coefficients exactly 1.

Similarly, as sketched out earlier, to obtain the point $\mathbf{b} = (\mathbf{e}/(4p), \mathbf{e}/(4p), \mathbf{e}/(2p), 2.5\mathbf{e}/(8p))$ in the hull of the vertices of T, we use (10) with coefficients chosen according to (11). This choice of $\alpha_{i,j}$'s yields x_i coordinate entries equal to $(1/(4p), 1/(4p), 1/(2p))$ for each i and has entries less than or equal to $2/(8p)$ in each v_j coordinate entry. Then, as above, one can include additional terms involving $\mathbf{0}$ and $5q\mathbf{h}_1, \ldots, 5q\mathbf{h}_q$ to complete the convex combination. One point to note is that the sum of the $\alpha_{i,k}$ coefficients appearing in (10), assuming $\mu_{i,k} = 5/8$, is equal to 4/5, and hence does not exceed 1. Addition of the θ_j coefficients will make the total higher but still less than 1 because, as in the previous case, the coefficients needed for the points $5q\mathbf{h}_1, \ldots, 5q\mathbf{h}_q$ are all bounded by $1/(5q)$ and hence their sum by 1/5. □

4. Local-search heuristic.

In this section we will describe a heuristic for NMF that arises from consideration of problem P1 (or, equivalently, INTERMEDIATE SIMPLEX).

Consider first EXACT NMF, and rewrite the problem in the form given by P1, i.e., the input is a pair (W_0, H_0). The algorithm initializes Q arbitrarily and then updates Q on each iteration. The update to Q is a rank-one change of the form $\bar{Q} = Q + \mathbf{f}\mathbf{z}^T$, where $\bar{Q}$ is the new value of Q and $\mathbf{f}, \mathbf{z}$ are both k-vectors. We assume that one of $\mathbf{f}, \mathbf{z}$ is chosen according to a fixed rule, while the other is found using optimization. The optimization method is described below.

For example, the fixed rule could be that $\mathbf{f}$ cycles through the columns of the identity matrix denoted $\mathbf{e}_1, \ldots, \mathbf{e}_k$ over successive iterations while using optimization to find $\mathbf{z}$. The interpretation of this rule in the context of INTERMEDIATE SIMPLEX is that the heuristic moves one vertex of the trial simplex per iteration. Another possibility is that $\mathbf{z}$ is taken to be $Q^{-T}\mathbf{e}_i$ as i cycles from 1 to k while using optimization to find $\mathbf{f}$; this corresponds to moving a facet of the trial simplex per iteration.

Let us assume that $\mathbf{f}$ is now determined by the fixed rule and see how to optimize $\mathbf{z}$. The key fact that makes the heuristic feasible is that the optimal selection of $\mathbf{z}$ can be written as linear programming. If we have not found a solution yet, then it may be the case where QH_0 has negative entries, say, the most negative entry is $-\mu$. To diminish the magnitude of the negative entries, we wish to select $\mathbf{z}$ to satisfy a constraint of the form $\bar{Q}H_0 \geq -\mu' E$, where μ' is less than μ and E is the $k \times n$ matrix of all 1's. This can be written $(Q + \mathbf{f}\mathbf{z}^T)H_0 \geq -\mu' E$, which is clearly a linear constraint on $\mathbf{z}$.

Now consider

$$\bar{Q}^{-1} = Q^{-1} - \frac{Q^{-1}\mathbf{f}\mathbf{z}^T Q^{-1}}{1 + \mathbf{z}^T Q^{-1}\mathbf{f}}$$

by the Sherman–Morrison formula. If we wish to impose the constraint $W_0\bar{Q}^{-1} \geq -\mu'' E$, where now E is the $m \times k$ matrix of all 1's, then this can be written

$$(1 + \mathbf{z}^T Q^{-1}\mathbf{f})W_0 Q^{-1} - W_0 Q^{-1}\mathbf{f}\mathbf{z}^T Q^{-1} \geq -\mu'' E(1 + \mathbf{z}^T Q^{-1}\mathbf{f})$$

plus the extra constraint $1 + \mathbf{z}^T Q^{-1}\mathbf{f} > 0$. Both the main and extra constraints are linear in $\mathbf{z}$. In the case where $\mathbf{f} = \mathbf{e}_i$, the extra constraint expresses the geometric condition that when a vertex of the simplex defined by the ith row of Q is moved, it does not cross through the hyperplane defined by the other vertices, i.e., the orientation of the simplex is unchanged.

Thus, for a fixed μ', μ'', the problem of finding $\mathbf{z}$ reduces to linear feasibility. If we wish to optimize μ'', then the problem is nonlinear (because of the cross term $\mu''\mathbf{z}^T Q^{-1}\mathbf{f}$), but the optimal μ'' can be approximated in a straightforward fashion by carrying out a binary search and checking feasibility for each choice of μ''.

Thus, the heuristic consists of iterations in which either $\mathbf{f}$ or $\mathbf{z}$ is determined by a fixed rule and then the other one of $\mathbf{f}$, $\mathbf{z}$ is found by solving linear inequalities to update Q.

Now consider applying this heuristic to OPT-NMF rather than EXACT NMF. Given a nonnegative matrix A and integer k, the singular value decomposition can find $W_0 \in \mathbf{R}^{m\times k}$ and $H_0 \in \mathbf{R}^{k\times n}$ such that W_0H_0 is the optimal approximation to A in both the Frobenius and 2-norms. This does not give rise to an instance of P1, however, because the side constraint $W_0H_0 \geq 0$ will usually not hold for these matrices. On the other hand, the heuristic described above does not require the side constraint, so it is still applicable. Upon termination, the heuristic will yield a factorization $A = (W_0Q^{-1})(QH_0)$, where W_0Q^{-1} and QH_0 are probably closer to being nonnegative than the original W_0 and H_0. This approximation is still optimal in the sense that $A - (W_0Q^{-1})(QH_0)$ has the minimum 2- or Frobenius-norm. If a solution that is truly nonnegative is sought, one can apply this heuristic and then change the negative entries of W_0Q^{-1} and QH_0 to zeros upon termination.

Compared to other local search heuristics mentioned in section 1, the main advantage of this approach is that the search space is $\mathbf{R}^{k\times(k-1)}$ (i.e., a search for Q) rather than $\mathbf{R}^{m\times k} \times \mathbf{R}^{k\times n}$ (i.e., a search for (W, H)). Thus, one might anticipate that the search space of the heuristic described in this section could be explored more thoroughly, since it is of a much lower dimension.

5. Discussion.

We have shown that OPT-NMF, EXACT NMF, INTERMEDIATE SIMPLEX, and nonnegative rank computation are all NP-hard. Some questions left unresolved by these results are as follows.

1. Are any of these problems in NP [8]? The difficulty, of course, is that a certificate of membership in NP is apparently the solution to the relevant problem (e.g., W and H in the case of EXACT NMF), but since the problem

involves high-degree polynomial constraints (e.g., observe the presence of Q^{-1} in the formulation of P1), there is no known polynomial bound on the number of bits needed to write down a solution.

This question is related to an open question posed by Cohen and Rothblum, still unsolved as far as we know. They ask, suppose an $m \times n$ rational matrix A has nonnegative rank k and a corresponding nonnegative factorization $A = WH$, $W \in \mathbf{R}^{m\times k}$, $H \in \mathbf{R}^{k\times n}$. Is it guaranteed that there exist rational W, H with the same properties? Note that the instance of NMF constructed in our NP-hardness proof has the property that if a solution exists, then a rational solution exists, so it does not contribute to progress on this open question.

2. It follows immediately from the NP-hardness of EXACT NMF that the approximation version of the OPT-NMF problem is also NP-hard in the most common sense of approximation. This sense is as follows: We demand that the approximate solution have an objective function value $N(A, WH)$ that is at most a constant multiple larger than the optimal solution. The NP-hardness follows for the trivial reason that a constant multiple of zero is still zero. On the other hand, suppose we demand an approximate solution that differs from the optimum by at most an additive term, perhaps by a term proportional to $\|A\|$. In this case, our hardness result does not apply, but the problem still seems to be difficult.

   Similarly, our result does not say anything about the difficulty of approximate solutions to the nonnegative rank problem.

3. It would be interesting to see if the heuristic proposed in the previous section is competitive with other published and implemented heuristic algorithms.

Acknowledgments. The author acknowledges helpful discussions about this material with Levent Tunçel and Ali Ghodsi (University of Waterloo), Jon Kleinberg and Lillian Lee (Cornell), Chris Ding (Lawrence Berkeley Laboratory), and Vincent Blondel (Louvain). The author also thanks the two referees for their useful suggestions and for catching an error in an earlier draft.

REFERENCES

[1] N. Asgarian and R. Greiner, *Using Rank-1 Biclusters to Classify Microarray Data*, Department of Computing Science and the Alberta Ingenuity Center for Machine Learning, University of Alberta, Edmonton, AB, Canada, 2006.

[2] S. Bergmann, J. Ihmels, and N. Barkai, *Iterative signature algorithm for the analysis of large-scale gene expression data*, Phys. Rev. E (3), 67 (2003), pp. 031902-1–031902-18.

[3] M. Biggs, A. Ghodsi, and S. Vavasis, *Nonnegative matrix factorization via rank-one downdating*, in International Conference on Machine Learning, 2008, available online from http://icml2008.cs.helsinki.fi/papers/667/pdf.

[4] L. Blum, F. Cucker, M. Shub, and S. Smale, *Complexity of Real Computation*, Springer-Verlag, New York, 1998.

[5] C. Boutsidis and E. Gallopoulos, *SVD based initialization: A head start for nonnegative matrix factorization*, Pattern Recognition, 41 (2008), pp. 1350–1362.

[6] J. Cohen and U. Rothblum, *Nonnegative ranks, decompositions and factorizations of nonnegative matrices*, Linear Algebra Appl., 190 (1993), pp. 149–168.

[7] J. Edmonds, *Systems of distinct representatives and linear algebra*, J. Res. Nat. Bureau of Standards, 71B (1967), pp. 241–245.

[8] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco, 1979.

[9] N. Gillis, *Approximation et sous-approximation de matrices par factorisation positive: Algorithmes, complexité et applications*, Master's Thesis, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, 2006 (in French).

[10] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, 1996.

[11] T. Hofmann, *Probabilistic latent semantic analysis*, in Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, 1999, ACM Press, pp. 50–57.

[12] H. Kim and H. Park, *Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis*, Bioinformatics, 23 (2007), pp. 1495–1502.

[13] D. Lee and H. Seung, *Learning the parts of objects by non-negative matrix factorization*, Nature, 401 (1999), pp. 788–791.

[14] L. B. Thomas, *Rank factorization of nonnegative matrices (A. Berman)*, SIAM Rev., 16 (1974), pp. 393–394.