On Complexity of Matrix Scaling

Arkadi Nemirovski and Uriel Rothblum
nemirovs@ie.technion.ac.il rothblum@ie.technion.ac.il
Faculty of Industrial Engineering and Management at Technion – Israel Institute of Technology

Abstract

The Line Sum Scaling problem for a nonnegative matrix A is to find positive definite diagonal matrices Y, Z which result in prescribed row and column sums of the scaled matrix YAZ. The Matrix Balancing problem for a nonnegative square matrix A is to find a positive definite diagonal matrix X such that the row sums in the scaled matrix XAX are equal to the corresponding column sums. We demonstrate that ϵ-versions of both these problems, same as those of other scaling problems for nonnegative multiindex arrays, can be reduced to a specific Geometric Programming problem. For the latter problem, we develop a polynomial-time algorithm, thus deriving polynomial time solvability of a number of generic scaling problems for nonnegative multiindex arrays. Our results extend those previously known for the problems of matrix balancing [3] and of double-stochastic scaling of a square nonnegative matrix [2].

Key words: matrix scaling, matrix balancing, polynomial-time complexity.

1 Introduction

The Line Sum Scaling problem is as follows:

(LSS): Given two positive vectors $r \in \mathbf{R}^m$, $c \in \mathbf{R}^n$ and an $m \times n$ matrix $A = [A_{ij}]$ with nonnegative entries and without zero rows and columns, find positive $m \times m$ diagonal matrix Y and positive $n \times n$ diagonal matrix Z such that the row sums in the matrix YAZ form the vector r, and the column sums form the vector c:

$$YAZ\mathbf{1}_n = r, \ (YAZ)^T\mathbf{1}_m = c,$$

where $\mathbf{1}_k = (\underbrace{1,...,1}_{k})^T$.

It is convenient (and, of course, does not restrict generality) to assume once for ever that the data r, c of the problem are normalized by

$$\sum_{i=1}^{m} r_i + \sum_{j=1}^{n} c_j = 2. \tag{1}$$

The "ϵ-relaxation" of (LSS) is the problem

> (LSS$_\epsilon$): Given the same data as in (LSS) and a positive ϵ, find positive $m \times m$ diagonal matrix Y and positive $n \times n$ diagonal matrix Z such that the row sums in the matrix YAZ are ϵ-close to r, and the column sums are ϵ-close to c:
> $$\|YAZ\mathbf{1}_n - r\|_1 + \|(YAZ)^T\mathbf{1}_m - c\|_1 \leq \epsilon;$$

from now on, for a vector $x \in \mathbf{R}^N$ $\|x\|_1 = \sum_{i=1}^N |x_i|$.

The data (A, r, c) of (LSS) are called *proper*, if (LSS) is solvable, and are called *semi-proper*, if all problems (LSS$_\epsilon$), $\epsilon > 0$, are solvable. The goal of this paper is to prove polynomial time complexity bound for the following problem:

> (LSS$^*_+$): Given the same data as in (LSS) and a positive ϵ, find positive $m \times m$ diagonal matrix Y and positive $n \times n$ diagonal matrix Z such that the row sums in the matrix YAZ are ϵ-close to r, and the column sums are ϵ-close to c:
> $$\|YAZ\mathbf{1}_n - r\|_1 + \|(YAZ)^T\mathbf{1}_m - c\|_1 \leq \epsilon,$$
> or detect correctly that the data (A, r, c) are not semi-proper.

The main result of our paper is that if $\epsilon \in (0, 1)$ and the data r, c of (LSS) are normalized to have $\|r\|_1 + \|c\|_1 = 2$, then problem (LSS$^*_+$) can be solved in no more than

$$O(1)(m+n)^4 \ln\left(2 + \frac{mn\sqrt{m^3+n^3}\ln(mn\beta)}{\epsilon^3}\right) \tag{2}$$

real arithmetic operations, where β is the ratio of the largest and the smallest positive entries of A.

There is a significant literature devoted to LSS; see [7, 2] and references therein. Most of this literature concerns existence, characterization and reductions; in particular, semi-properness is characterized in [7]. Here our goal is to obtain complexity bounds on solving (LSS$^*_+$), that is, on computing approximate scalings to prescribed accuracy. Our approach follows [2] which considered the important special case when A is square and $r = c = \mathbf{1}_n$ (the "double-stochastic scaling"). It is shown in this reference that "ϵ-double-stochastic" scaling of a nonnegative matrix A for which a double-stochastic scaling exists can be found in polynomial time, specifically, in $O\left(n^4 \ln\left(\frac{n \ln \beta}{\epsilon}\right)\right)$ operations (essentially same as implied by (2) for the case of $m = n$).[1)]

[1)] After the work on this paper was finished, we became aware of the "in process" paper [5] where the authors announce strongly polynomial algorithm for LSS with complexity bound $O(n^7 \ln n \ln(1/\epsilon))$ (in [5], $n = m$). The advantage of the latter bound is that it is free of the "number-dependent" quantity β; note, however, that (2) is proportional to $\ln\ln\beta$ and that the dependence on the sizes n, m in our bound is much better than the one in [5].

The paper is organized as follows: in Section 2 we start with formulating a specific geometric programming problem (problem (GS) – (GS$^*_+$) below) which covers LSS as a special case and study (GS) – (GS$^*_+$) in the case of "standard data" – those satisfying a "standard" side condition (which is automatically satisfied for the LSS problem). We identify necessary and sufficient conditions for solvability of (GS) with standard data and obtain an explicit upper bound on the norm of a solution to a solvable (GS). Equipped with this result, we present in Section 3 a simple polynomial time algorithm for (GS) with standard data. Section 4 contains applications of the results to the LSS problem; in particular, the application of our algorithm to the (GS)-reformulation of the LSS problem implies the complexity bound of (2). In concluding Section 5 we illustrate our results on polynomial time solvability of (GS) by a pair of other applications. The first is the *balancing* problem for a nonnegative matrix; here we demonstrate that the best known so far polynomial time complexity bound for the matrix balancing problem from [3] can be straightforwardly derived from our results on (GS). The second application is a "multi-index sum scaling problem". In this problem, one is given a p-dimensional nonnegative array A (say, 3D array $\{A_{ijk}\}$) and is allowed to multiply the entries by q positive "scaling arrays", the entries of each array depending on a given part of the indices (in our 3D example this could be the transformations $A_{ijk} \mapsto B_{ijk} = X_i Y_j Z_k A_{ijk}$ with positive X_i, Y_j, Z_k). The goal is to find a scaling of this type which results in prescribed partial sums of the entries of the scaled array (in the example we have specified, these are the planar sums $\sum\limits_{j,k} B_{ijk}$, $\sum\limits_{i,k} B_{ijk}$, $\sum\limits_{ij} B_{ijk}$). It turns out that a problem of this type can be easily converted to the form of (GS), so that our results on the latter problem imply straightforwardly polynomial time solvability of the multi-index sum scaling problem.

2 Convex Programming reformulation of (LSS)

2.1 Reformulation and solvability issues

Let A be a nonnegative matrix, K be the total number of nonzero entries in A, and let $(i(k), j(k))$, $k = 1, ..., K$, be an enumeration of the corresponding cells. Let us set

- $a = (a_1, ..., a_K)^T$, $a_k = A_{i(k)j(k)}$,
- $\sigma_k = e_{i(k)} + f_{j(k)} \in \mathbf{R}^N \equiv \mathbf{R}^m \times \mathbf{R}^n$, $k = 1, ..., K$, where e_i and f_j are the natural extensions (by adding zeros) to $\mathbf{R}^N$ of the basic unit vectors in $\mathbf{R}^m$, $\mathbf{R}^n$, respectively;
- $\sigma = \begin{pmatrix} r \\ c \end{pmatrix} \in \mathbf{R}^N$
- $e = \frac{1}{2}\mathbf{1}_N$.

Passing in (LSS) from the unknowns Y, Z to unknowns $x \in \mathbf{R}^{m+n}$ according to

$$Y_{ii} = \exp\{x_i\},\ i = 1, ..., m;\ Z_{jj} = \exp\{x_{m+j}\},\ j = 1, ..., n,$$

we reformulate (LSS) equivalently as the problem

$$\text{Find } x\text{:}\quad \sum_{k=1}^{K} a_k \exp\{\sigma_k^T x\}\sigma_k = \sigma, \tag{GS}$$

while (LSS_+^*) becomes the problem

$$\begin{array}{l}\text{Given } \epsilon > 0, \text{ find } x \text{ such that } \left\|\sum_{k=1}^K a_k \exp\{\sigma_k^T x\}\sigma_k - \sigma\right\|_1 \leq \epsilon \\ \text{or detect correctly that } \inf_x \left\|\sum_{k=1}^K a_k \exp\{\sigma_k^T x\}\sigma_k - \sigma\right\|_1 > 0.\end{array} \tag{GS_+^*}$$

Note also that

$$e^T\sigma_k = 1,\ k = 1, ..., K;\ e^T\sigma = 1 \tag{3}$$

(the latter equation is given by the normalization (1)).

2.1.1 Generalization

Problems (GS) and (GS_+^*) with data $a > 0, \{\sigma_k \in \mathbf{R}^N\}_{k=1}^K$ and $\sigma \in \mathbf{R}^N$ not necessarily derived from (LSS) were introduced and studied in [8]. In particular, for applications of these problems beyond (LSS) see [8] and Section 5. In the remainder of the Section and in the following Section we deal with problems (GS) and (GS_+^*) independently of the origin of the data, but with the assumption that the data admits a vector e such that the relation (3) is satisfied; in such cases we call the data of (GS) *standard*. Of course, testing whether or not a particular data of (GS) is standard and computing a vector e satisfying (3) if the answer is affirmative is a simple Linear Algebra problem.

Note that for *solvable* (GS) the assumption that the data of (GS) is standard is "basically equivalent" to the assumption that

(a) The affine hull of $\sigma_1, ..., \sigma_K$ does not contain the origin.

Indeed, if the data of (GS) is standard, then (a) of course is satisfied. Now let (a) be satisfied. An elementary result in Linear Algebra then assures the existence of a vector e with $e^T\sigma_k = 1,\ k = 1, ..., K$; further, in this case, solving a simple Linear Algebra problem, we may find such a vector e. After e is identified, we may check whether $e^T\sigma > 0$. If it is not the case, (GS) clearly is unsolvable, otherwise we may multiply σ by a positive constant λ to get $e^T(\lambda\sigma) = 1$. It remains to note that in the case of (a) the problems (GS) with proportional to each other, with positive coefficients, vectors σ are equivalent to each other: if $\sum_k a_k \exp\{\sigma_k^T x\}\sigma_k$ is equal/close to σ, then $\sum_k a_k \exp\{\sigma^k(x + (\ln\lambda)e)\}\sigma_k$ is equal, respectively, close to $\lambda\sigma$. Thus, in the case of a we either can detect that (GS) is unsolvable, or pass to equivalent "normalized data" satisfying (3) (this is what we are doing when imposing normalization condition (1) on the LSS data).

From now on, speaking about (GS) – (GS_+^*), we exclude the trivial case when all σ_k, $k = 1, ..., K$, are equal to each other; for the LSS problem, it means that we assume that $\min[m, n] > 1$.

2.1.2 Solvability conditions

The data $(a > 0, \{\sigma_k\}_{k=1}^K, \sigma)$ of (GS) are called *proper*, if (GS) is solvable, and are called *semi-proper*, if

$$\inf_x \left\| \sum_{k=1}^K a_k \exp\{\sigma_k^T x\}\sigma_k - \sigma \right\|_1 = 0. \tag{4}$$

Note that when the data $(a > 0, \{\sigma_k\}_{k=1}^K, \sigma)$ of (GS) are proper, with solution x, then (multiplying both sides of

$$\sum_{k=1}^K a_k \exp\{\sigma_k^T x\}\sigma_k = \sigma$$

by e^T and using (3)) $\sum_{k=1}^K a_k \exp\{\sigma_k^T x\} = 1$, thus, σ is a convex combination, with *positive* weights, of $\sigma_1, ..., \sigma_K$. Similarly, assuming that the data of (GS) are semi-proper we conclude, by using a limiting argument, that σ belongs to the convex hull of $\sigma_1, ..., \sigma_K$. Thus, we see that the *necessary* condition for properness of the data of (GS) is

C. σ is a convex combination, with positive weights, of $\sigma_1, ..., \sigma_K$

while the necessary condition for the data of (GS) to be semi-proper is

C′. σ is a convex combination of $\sigma_1, ..., \sigma_K$.

Variants of **C** and **C′** which consider positive and nonnegative linear combinations (without asserting that the corresponding coefficients sum to 1) were considered in [8]; specifically these variants were shown to be equivalent, respectively, to properness and semi-properness of the data of (GS) without the assumption that there exists e satisfying (3). In the same spirit, we will show below that **C** and **C′** themselves are sufficient, and not just necessary, for properness and semi-properness of the standard data of (GS). Our proof is a byproduct of results we develop in the next subsection.

2.1.3 Convex Programming reformulation and bounds on the norm of a solution

Let E be the linear span of the vectors $\sigma_k - \sigma_\ell$, $k, \ell = 1, ..., K$, and F be the orthogonal complement to E in $\mathbf{R}^N$. Consider the convex function

$$f(x) = \phi(x) - \sigma^T x,\ \phi(x) = \ln(\sum_{k=1}^K a_k \exp\{\sigma_k^T x\})$$

(to check that f is convex, see, e.g., [1], Lemma 7.12, p. 197). We start with the following simple observation:

Lemma 2.1 *Let σ be an affine combination of $\sigma_1, ..., \sigma_K$ (as it is the case under assumption $\mathbf{C}'$). Then f is constant along F:*

$$f(x+v) = f(x) \quad \forall x \in \mathbf{R}^N \ \forall v \in F.$$

Proof. If $x \in \mathbf{R}^N$ and $v \in F$, then

$$\sigma_1^T v = \sigma_2^T v = ... = \sigma_K^T v = \sigma^T v$$

(the first $K-1$ equalities are readily given by the fact that v is orthogonal to all differences $\sigma_k - \sigma_\ell$, $k, \ell = 1, ..., K$, and the last equality follows from the first $K-1$ of them since σ is an affine combination of $\sigma_1, ..., \sigma_K$). Consequently,

$$\begin{array}{rcl} f(x+v) & = & \ln\left(\sum_{k=1}^K a_k \exp\{\sigma_k^T(x+v)\}\right) - \sigma^T(x+v) \\ & = & \left[\ln\left(\sum_{k=1}^K a_k \exp\{\sigma_k^T x\}\right)\right] - \sigma^T x + \left[\sigma_1^T v - \sigma^T v\right] \\ & = & f(x). \end{array} \qquad \blacksquare$$

Our next observation is as follows:

Lemma 2.2 *Assume that there exists e satisfying* (3)*. Then the set of solutions to* (GS) *is exactly the set of minimizers x of f satisfying the condition $\phi(x) = 0$. Further, if σ is an affine combination of $\sigma_1, ..., \sigma_K$, as is the case under assumption $\mathbf{C}'$, f attains a minimum over $\mathbf{R}^N$ if and only if* (GS) *is solvable.*

Proof. We first observe that

$$\nabla f(x) = \frac{\sum_k a_k \exp\{\sigma_k^T x\}\sigma_k}{\sum_k a_k \exp\{\sigma_k^T x\}} - \sigma. \tag{5}$$

Now, if x is a solution to (GS), then

$$\sum_k a_k \exp\{\sigma_k^T x\}\sigma_k - \sigma = 0 \tag{6}$$

and, as we have seen in the previous subsection,

$$\sum_k a_k \exp\{\sigma_k^T x\} = 1, \text{ or, equivalently, } \phi(x) = 0, \tag{7}$$

These two conditions combine with (5) to show that $\nabla f(x) = 0$, thus, every solution x to (GS) is a global minimizer of f satisfying $\phi(x) = 0$. Alternatively, if x is a global minimizer of f, then $\nabla f(x) = 0$; hence, if in addition $\phi(x) = 0$, the equivalence in (7) combines with (5) to show that x is a solution to (GS). To complete the proof, we should demonstrate that if σ is an affine combination of $\sigma_1, ..., \sigma_K$ and f attains its minimum, then among minimizers of f there are points with $\phi = 0$, which is immediate:

indeed, if f attains its minimum at a point x, then, by Lemma 2.1, all points from the affine plane $x+F$ also are minimizers of f. By (3), $e \in F$, so that the point

$$\bar{x} = x - \phi(x)e$$

is a global minimizer of f. It remains to note that

$$\begin{array}{rcll} \phi(\bar{x}) & = & \ln\left(\sum_k a_k \exp\{\sigma_k^T(x-\phi(x)e)\}\right) & \\ & = & \phi(x) - \phi(x) & \text{[by (3)]} \\ & = & 0. & \end{array} \qquad \blacksquare$$

We are about to demonstrate that under assumption **C** f attains its minimum on $\mathbf{R}^N$ (so that, by Lemma 2.2, (GS) is solvable) and to get an upper bound on the distance from the origin to the set of minima of f. This bound will be expressed in terms of four data-dependent quantities we are about to introduce.

Observe that $x \in \mathbf{R}^N$ satisfies

$$\max_k \sigma_k^T x = \min_k \sigma_k^T x$$

if and only if $\sigma_k^T x$ is a constant over k, that is if and only if $x \in F$; hence such x is in E if and only if $x = 0$. Consequently, the following quantity is well-defined:

$$\alpha \equiv \alpha(\sigma_1, ..., \sigma_K) = \max_{x \in E, \|x\|_2 = 1} \frac{1}{\max_k \sigma_k^T x - \min_k \sigma_k^T x} \tag{8}$$

with $\|\cdot\|_2$ being the standard Euclidean norm of a vector. Note that by homogeneity reasons one has

$$x \in E \Rightarrow \max_k \sigma_k^T x - \min_k \sigma_k^T x \geq \alpha^{-1}\|x\|_2. \tag{9}$$

Assuming that the data of (GS) satisfy **C**, let us set

$$\begin{array}{rcl} \gamma & \equiv & \gamma(\sigma_1, ..., \sigma_K, \sigma) = \min\limits_{\lambda \in \Lambda} \max\limits_{k \leq K} \lambda_k^{-1}, \\ \Lambda & \equiv & \Lambda(\sigma_1, ..., \sigma_K, \sigma) = \left\{\lambda \in \mathbf{R}^K \mid \lambda > 0, \sum_k \lambda_k = 1, \sum_k \lambda_k \sigma_k = \sigma\right\}. \end{array} \tag{10}$$

(It is straightforward to check that when **C** is satisfied, the minimum in (10) is attained.) If **C** is not satisfied, we set $\gamma(\sigma_1, ..., \sigma_K, \sigma) = +\infty$.
Finally, let

$$\beta \equiv \beta(a) = \frac{\max_k a_k}{\min_k a_k}, \tag{11}$$

and

$$\delta \equiv \delta(\sigma_1, ..., \sigma_K) = \max_{k \leq K} \|\sigma_k\|_1. \tag{12}$$

We are ready to formulate one of our main results:

Proposition 2.1 *Let the data* $a > 0, \{\sigma_k \in \mathbf{R}^N\}_{k=1}^K$, $\sigma \in \mathbf{R}^N$ *of* (GS) *satisfy* $\mathbf{C}$, *and let there exist* $e \in \mathbf{R}^N$ *satisfying* (3). *Then problem* (GS) *is solvable, and there exists a solution* x_* *to this problem such that*

$$\|x\|_2 \leq R \equiv R(a, \sigma_1, ..., \sigma_K, \sigma) = \alpha\gamma \ln(K\beta), \tag{13}$$

with α, β, γ *given by* (8), (10), (11), *respectively.*

In particular, $\mathbf{C}$ *is a necessary and sufficient condition for the solvability of problem* (GS).

Proof. We have already seen that $\mathbf{C}$ is necessary for solvability of (GS). Now assume that $\mathbf{C}$ is satisfied. As we have seen in Lemma 2.2, solvability of (GS) is equivalent to the fact that f attains its minimum on $\mathbf{R}^N$; thus, all we need to prove is that f attains its minimum on $\mathbf{R}^N$, and that at least one of the minimizers of f satisfies (13). Since $\mathbf{C}$ clearly implies the premise of Lemma 2.1, f is constant along F, so that it suffices to verify that f attains its minimum on E at a point satisfying (13). To this end, in turn, it suffices to demonstrate that

$$x \in E, \|x\|_2 > R \Rightarrow f(x) > f(0). \tag{14}$$

To establish (14), observe first that

$$f(0) = \ln\left(\sum_k a_k\right) \leq \ln(K \max_k a_k). \tag{15}$$

On the other hand, by definition of γ in (10) there exists representation

$$\sigma = \sum_{k=1}^K \lambda_k \sigma_k$$

with $\sum_k \lambda_k = 1$ and $\min_k \lambda_k = \gamma^{-1}$. Let $x \in E$. We clearly have, with $c \equiv \min_k \ln a_k$ and k^*, k_* as the maximizer and minimizer of $\sigma_k^T x$ over k, respectively,

$$\phi(x) \geq \ln(a_{k^*} \exp\{\sigma_{k^*}^T x\}) \geq c + \max_k \sigma_k^T x,$$

whence

$$\begin{array}{rcll}
f(x) & \geq & c + \max_k \sigma_k^T x - \sigma^T x & \\
& = & c + \max_k \sigma_k^T x - \sum_k \lambda_k \sigma_k^T x & \\
& = & c + \sum_\ell \lambda_\ell [\max_k \sigma_k^T x - \sigma_\ell^T x] & \\
& \geq & c + \lambda_{k_*} [\max_k \sigma_k^T x - \sigma_{k_*}^T x] & \\
& \geq & c + (\min_k \lambda_k)[\max_k \sigma_k^T x - \min_k \sigma_k^T x] & \\
& \geq & c + (\min_k \lambda_k)\alpha^{-1}\|x\|_2 & \text{[by (9)]} \\
& = & c + \gamma^{-1}\alpha^{-1}\|x\|_2. &
\end{array}$$

Combining the resulting inequality with (15), we get

$$\begin{array}{rcl}
x \in E \Rightarrow f(x) - f(0) & \geq & c - \ln(K \max_k a_k) + \alpha^{-1}\gamma^{-1}\|x\|_2 \\
& = & \min_k \ln a_k - \ln(K \max_k a_k) + \alpha^{-1}\gamma^{-1}\|x\|_2 \\
& = & \ln(\min_k a_k) - \ln(K\beta \min_k a_k) + \alpha^{-1}\gamma^{-1}\|x\|_2 \\
& = & \alpha^{-1}\gamma^{-1}\left[\|x\|_2 - \alpha\gamma \ln(K\beta)\right],
\end{array}$$

and the concluding quantity is positive when $\|x\|_2 > R$. ■

Corollary 2.1 *Let the data $a > 0, \{\sigma_k \in \mathbf{R}^N\}_{k=1}^K$, $\sigma \in \mathbf{R}^N$ of* (GS) *satisfy* $\mathbf{C}'$, *and let there exist $e \in \mathbf{R}^N$ satisfying* (3)*. Then the data are semi-proper, i.e.,* (4) *is satisfied. Thus,* $\mathbf{C}'$ *is a necessary and sufficient condition for semi-properness of the data of* (GS)*.*

Proof. $\mathbf{C}'$ implies that $\|\sigma\|_1 \leq \max_k \|\sigma_k\|_1 = \delta(\sigma_1, ..., \sigma_k)$. Now, given $\epsilon > 0$, let us set

$$\begin{array}{rcl} \theta_\epsilon & = & \frac{\epsilon}{\epsilon+2\delta}, \\ \sigma_\epsilon & = & (1-\theta_\epsilon)\sigma + \frac{\theta_\epsilon}{K}\sum_{k=1}^K \sigma_k. \end{array} \tag{16}$$

Under assumption $\mathbf{C}'$ the data $a, \{\sigma_k\}_{k=1}^K$, σ_ϵ clearly satisfy $\mathbf{C}$, so that by Proposition 2.1 there exists $x_\epsilon \in \mathbf{R}^N$ such that

$$\sum_{k=1}^K a_k \exp\{\sigma_k^T x_\epsilon\}\sigma_k = \sigma_\epsilon$$

and consequently

$$\begin{array}{rcll} \left\|\sigma - \sum_{k=1}^K a_k \exp\{\sigma_k^T x\}\sigma_k\right\|_1 & = & \|\sigma - \sigma_\epsilon\|_1 & \\ & = & \theta_\epsilon \left\|\sigma - \frac{1}{K}\sum_k \sigma_k\right\|_1 & \\ & \leq & \theta_\epsilon \left(\|\sigma\|_1 + \max_{k\leq K}\|\sigma_k\|_1\right) & \\ & \leq & 2\theta_\epsilon\delta & [\text{by } \mathbf{C}'] \\ & \leq & \epsilon. & \end{array}$$

$\blacksquare$

3 Polynomial complexity of (GS$^*_+$)

We are about to demonstrate that problem (GS$^*_+$) can be solved in polynomial time. Given $\epsilon > 0$, let us choose somehow an a priori upper bound $\widehat{\alpha}$ on $\alpha(\sigma_1, ..., \sigma_K)$ and set (cf. the proof of Corollary 2.1)

$$\begin{array}{rcll} \beta & = & \beta(a) & [\text{see (11)}] \\ \delta & = & \delta(\sigma_1, ..., \sigma_K) & [\text{see (12)}] \\ \theta_\epsilon & = & \frac{\epsilon}{\epsilon+4\delta} & \\ \psi & = & \theta_\epsilon^2 & \\ \widehat{\gamma} & = & \frac{K}{\theta_\epsilon} & \\ \sigma_\epsilon & = & (1-\theta_\epsilon)\sigma + \frac{\theta_\epsilon}{K}\sum_{k=1}^K \sigma_k & \\ \widehat{R} & = & \widehat{\alpha}\widehat{\gamma}\ln(K\beta(a)) & \\ f_\epsilon(x) & = & \phi(x) - \sigma_\epsilon^T x = \ln\left(\sum_{k=1}^K a_k \exp\{\sigma_k^T x\}\right) - \sigma_\epsilon^T x, & \end{array} \tag{17}$$

Our key observation is given by

Proposition 3.1 *Assume that the data* $a > 0, \{\sigma_k\}_{k=1}^K, \sigma$ *of problem* (GS) *satisfy* $\mathbf{C}'$ *and that there exists* e *satisfying* (3). *Given* $\epsilon > 0$, *define the quantities* (17), *and let* x_ψ *be a* ψ*-minimizer of* f_ϵ *in the ball* $V = \{x \in \mathbf{R}^N \mid \|x\|_2 \le \widehat{R}\}$*:*

$$x_\psi \in V, \;\; f_\epsilon(x_\psi) - \min_V f_\epsilon \le \psi. \tag{18}$$

Then the point

$$\bar{x}_\epsilon = x_\psi - \phi(x_\psi)e \tag{19}$$

satisfies:

$$\left\| \sum_{k=1}^K a_k \exp\{\sigma_k^T x_\psi\}\sigma_k - \sigma \right\|_1 \le \epsilon. \tag{20}$$

Proof. Observe, first, that independently of any assumptions on the data (except $a > 0$) for *all* $x \in \mathbf{R}^N$ one has

$$\begin{array}{rcl} \nabla\phi(x) & = & (\sum_k a_k \exp\{\sigma_k^T x\})^{-1} \sum_k a_k \exp\{\sigma_k^T x\}\sigma_k, \\ \nabla^2 f_\epsilon(x) & = & \nabla^2\phi(x) \\ & = & (\sum_k a_k \exp\{\sigma_k^T x\})^{-1} \sum_k a_k \exp\{\sigma_k^T x\}\sigma_k\sigma_k^T - [\nabla\phi(x)][\nabla\phi(x)]^T. \end{array}$$

Thus, with δ given by (12) we immediately conclude that

$$\|\nabla\phi(x)\|_1 \le \delta \tag{21}$$

and

$$\|\nabla^2 f_\epsilon(x)\|_1 \equiv \sum_{k,\ell=1}^K \left| \frac{\partial^2 f_\epsilon(x)}{\partial x_k \partial x_\ell} \right| \le 2\delta^2. \tag{22}$$

Now let the data $a, \{\sigma_k\}_{k=1}^K, \sigma$ of (GS) satisfy $\mathbf{C}'$. The function f_ϵ is exactly the function f from the previous Section associated with the perturbed data $a, \{\sigma_k\}_{k=1}^K, \sigma_\epsilon$. Same as in the proof of Corollary 2.1, these data satisfy $\mathbf{C}$. Moreover, if $\sigma = \sum_{k=1}^K \lambda_k\sigma_k$ is a representation of σ as a convex combination of $\sigma_1, ..., \sigma_K$ (such a representation exists in view of $\mathbf{C}'$), then σ_ϵ can be represented as the convex combination

$$\sigma_\epsilon = \sum_{k=1}^K \lambda_k^\epsilon \sigma_k$$

of σ_k with the weights

$$\lambda_k^\epsilon = (1-\theta_\epsilon)\lambda_k + \frac{\theta_\epsilon}{K} \ge \frac{\theta_\epsilon}{K},$$

whence (see (10))

$$\gamma(\sigma_1, ..., \sigma_K, \sigma_\epsilon) \le \widehat{\gamma}.$$

Applying Proposition 2.1 to the data $a, \{\sigma_k\}_{k=1}^K, \sigma_\epsilon$ and taking into account that the corresponding parameters α, β are exactly the same as those for the original data, we conclude that f_ϵ attains its global minimum at a point $x_* \in V$.

Now let x_ψ be a ψ-minimizer of f_ϵ in V. Since V contains a global minimizer of f_ϵ, we have

$$f_\epsilon(x_\psi) - \min f_\epsilon \leq \psi,$$

and since $\bar{x}_\epsilon = x_\psi - \phi(x_\psi)e$ differs from x_ψ by a vector proportional to the vector e which is in F (see (3)) and f_ϵ is constant along F by Lemma 2.1, we have

$$f_\epsilon(\bar{x}_\epsilon) - \min f_\epsilon \leq \psi \tag{23}$$

as well. Also, as in the proof of Lemma 2.2, we have $\phi(\bar{x}_\epsilon) = 0$, whence

$$\sum_k a_k \exp\{\sigma_k^T \bar{x}_\epsilon\} = 1$$

and

$$g \equiv \nabla f_\epsilon(\bar{x}_\epsilon) = \nabla f(\bar{x}_\epsilon) - \sigma_\epsilon = \sum_k a_k \exp\{\sigma_k^T \bar{x}_\epsilon\}\sigma_k - \sigma_\epsilon, \tag{24}$$

By (22) and the standard approximation bound, for each $h \in \mathbf{R}^N$ we have

$$f_\epsilon(\bar{x}_\epsilon + h) \leq f_\epsilon(\bar{x}_\epsilon) + g^T h + \delta^2 \|h\|_\infty^2 \qquad [\|h\|_\infty = \max_i |h_i|].$$

Let $d \in \mathbf{R}^N$ be given by $d_i = -\text{sign}\left(\frac{\partial f(\bar{x}_\epsilon)}{x_i}\right)$, so that $g^T d = -\|g\|_1$ and $\|d\|_\infty \leq 1$, and let $h = \frac{\|g\|_1}{2\delta^2} d$. From the above bound,

$$f_\epsilon(\bar{x}_\epsilon + h) \leq f_\epsilon(\bar{x}_\epsilon) - \frac{\|g\|_1^2}{4\delta^2},$$

whence

$$\min f_\epsilon \leq f_\epsilon(\bar{x}_\epsilon + h) \leq f_\epsilon(\bar{x}_\epsilon) - \frac{\|g\|_1^2}{4\delta^2}.$$

Combining the latter inequality with (23), we get

$$\|g\|_1 \leq 2\delta\sqrt{\psi} = 2\delta\theta_\epsilon \leq \frac{\epsilon}{2} \tag{25}$$

(see (17)). Combining this result with (24), we get

$$\left\| \sum_k a_k \exp\{\sigma_k^T \bar{x}_\epsilon\} - \sigma_\epsilon \right\|_1 \leq \frac{\epsilon}{2}. \tag{26}$$

On the other hand, as in the last string of inequalities in the proof of Corollary 2.1 it holds

$$\|\sigma - \sigma_\epsilon\|_1 \leq 2\theta_\epsilon\delta \leq \frac{\epsilon}{2},$$

the last inequality following from the definition of θ_ϵ in (17). Combining the latter inequality with (26), we come to (20). ■

Now we are ready to present a polynomial time algorithm for solving (GS_+^*) and to evaluate its complexity. For the sake of simplicity, we restrict ourselves with an algorithm based on the Ellipsoid method[2]. For our purposes it suffices to outline the following properties of the Ellipsoid method (for a detailed description and proofs, see, e.g., [6]): as applied to an optimization program

$$g(x) \to \min \mid x \in V = \{x \in \mathbf{R}^N \mid \|x\|_2 \leq \widehat{R}\} \tag{P}$$

with a convex continuous objective g on V, the method generates a ψ-solution with $\psi > 0$ being a prescribed accuracy, that is, a point $x_\psi \in V$, $g(x_\psi) \leq \min_V g + \psi$, in no more than

$$\mathcal{I}_{\rm Ell}(P, \epsilon) = O(1)N^2 \ln\left(\frac{2\psi + {\rm Var}_V(g)}{\psi}\right), \qquad {\rm Var}_V(g) = \max_V g - \min_V g \tag{27}$$

iterations. An iteration requires a single computation of the value and a subgradient of g at a given point plus $O(N^2)$ operations of exact real arithmetic to run the method itself.

In order to find a solution to $({\rm GS}^*_+)$, we first check whether

$$\|\sigma\|_1 \leq \delta(\sigma_1, ..., \sigma_K) \equiv \max_k \|\sigma_k\|_1. \tag{28}$$

If it is not the case, then $\mathbf{C}'$ definitely is not satisfied, and we terminate reporting that the data are not semi-proper. Otherwise we define ψ, $\widehat{R}$ and f_ϵ according to (17) and apply the Ellipsoid method to problem (P), the objective being f_ϵ. After a ψ-solution x_ψ to (P) is found, we convert it into $\bar{x}_\epsilon$ according to (19) and check whether $\bar{x}_\epsilon$ indeed solves (20). If it is not the case, we announce that the data $a, \{\sigma_k\}_{k=1}^K, \sigma$ are not semi-proper for (GS).

The correctness and the complexity of the outlined algorithm are given by the following

Theorem 3.1 *Let $\epsilon > 0$, $a > 0, \{\sigma_k\}_{k=1}^K, \sigma$, e and an a priori upper bound $\widehat{\alpha}$ on the quantity $\alpha(\sigma_1, ..., \sigma_K)$ defined in (8) be given and let (3) be satisfied. Then the outlined algorithm is correct, i.e., it either produces a solution to $({\rm GS}^*_+)$, or recognizes correctly that $\mathbf{C}'$ is not satisfied. The result is obtained in no more than*

$$\begin{array}{c}\mathcal{I} = O(1)N^2 \ln\left(2 + \frac{4K(\epsilon+4\delta)^3\delta\widehat{\alpha}\ln(K\beta)}{\epsilon^3}\right) \\ \left[\begin{array}{rcl} \beta & = & \beta(a) = \frac{\max_k a_k}{\min_k a_k}, \\ \delta & = & \delta(\sigma_1, ..., \sigma_K) = \max_k \|\sigma_k\|_1 \end{array}\right]\end{array} \tag{29}$$

iterations with no more than

$$O(1)(N^2 + L)$$

operations of real arithmetic (including taking exp *and* log*) per iteration, where L is the total number of nonzero entries in $\sigma_1, ..., \sigma_K$.*

[2] An alternative would be exploiting interior-point techniques; however, for the LSS problem with $m = O(n)$ these techniques have no advantages as compared to the Ellipsoid method, see [2].

Proof. By construction of the algorithm, its output either is a solution to (20), or is a claim that the data is not semi-proper. Proposition 3.1 shows that if $\mathbf{C}'$ is satisfied, then the second of these alternatives cannot take place, so that the algorithm indeed solves (GS_+^*). To evaluate the complexity of the algorithm, note that by (21) (this bound is valid independently of any assumptions on the data except $a > 0$) and (28) (recall that the Ellipsoid method is run only when this inequality is satisfied) we have $\|\nabla f_\epsilon(x)\|_1 \leq \|\nabla\phi(x)\|_1 + \|\sigma_\epsilon\|_1 \leq 2\delta$, whence

$$\mathrm{Var}_V(f_\epsilon) \leq 4\delta\widehat{R}.$$

Combining the latter bound, (17) and (27), we come to (29). The upper bound on the arithmetic cost of an iteration is readily given by the above remark on the complexity of an iteration in the Ellipsoid method. ■

Remark 3.1 The complexity bounds in Theorem 3.1 deal with idealized *precise real arithmetic* implementation of the algorithm. They, however, remain valid for finite-precision computations as well. Namely, assume that all a_k are nonnegative integers, and let, with $\ln_+(s) = \max\{\ln s, 0\}$,

$$L(\epsilon) = 1 + \ln(\max_k a_k) + \ln_+\left(\frac{1}{\epsilon}\right) + \max_k \ln_+(\|\sigma_k\|_1) + \ln K + \ln_+(\widehat{\alpha}),$$

It can be seen that the algorithm underlying Theorem 3.1 admits an implementation in which the number of bit-wise operations sufficient to produce a solution to (GS_+^*), or to detect correctly that $\mathbf{C}'$ is not satisfied is polynomial in $NL(\epsilon)$.

An upper bound on $\alpha(\sigma_1, ..., \sigma_k)$. The only quantity appearing in our construction and complexity bound (see (17), (29)) which is not readily given by the data is an a priori upper bound $\widehat{\alpha}$ on the quantity $\alpha(\sigma_1, ..., \sigma_K)$. Our current goal is to build a "universal" bound of this type.

We start with the simple and, essentially, well-known fact as follows:

Lemma 3.1 *Let $\mu_1, ..., \mu_q$ be integer linearly independent vectors in $\mathbf{R}^N$, $N > q$, with $\|\mu_i\|_\infty \leq L$, $i = 1, ..., q$, and let E be the linear span of these vectors. Then*

$$\min_{\substack{\|x\|_\infty \geq 1 \\ x \in E}} \max_i |\mu_i^T x| \geq \frac{1}{(L^2 N^{3/2})^N}. \tag{30}$$

Proof. The proof to follow originates from [4]. Consider $2N$ Linear Programming programs

$$\begin{array}{c} t \to \min \\ -t \leq \mu_\ell^T \sum\limits_{i=1}^q \xi_i \mu_i \leq t, \ \ell = 1, ..., q, \\ \eta \left(\sum\limits_{i=1}^q \xi_i \mu_i \right)_j \geq 1 \end{array} \tag{$P[\eta, j]$}$$

in design variables $t, \xi_1, ..., \xi_q$, the parameters of a problem being $\eta = \pm 1$ and j, $1 \leq j \leq N$. Let us fix a problem (P$[\eta, j]$), and assume that it is feasible. The optimal value t_* of the problem clearly is nonnegative, and in fact it is positive, since from $\mu_\ell^T \sum_{i=1}^{q} \xi_i \mu_i = 0$, $\ell = 1, ..., q$, it would follow that $\sum_{i=1}^{q} \xi_i \mu_i = 0$, which is forbidden by the last constraint of the problem. From the below boundedness of the problem and the fact that $\{\mu_i\}$ are linearly independent it follows immediately that the feasible set does not contain lines; thus, there exists an optimal solution (ξ_*, t_*) to the problem which is an extreme point of the feasible set. By the standard characterization of the extreme points of a polyhedral set, it means that $q+1$ linearly independent inequalities from those defining (P$[\eta, j]$) at the point (ξ_*, t_*) become equalities, so that (ξ_*, t_*) is a solution of a nonsingular system of linear equations with integral coefficients of the matrix and of the right hand side, modulae of the coefficients not exceeding NL^2. By Cramer's rule combined with the Hadamard upper bound on a determinant, it follows that every coordinate of (ξ_*, t_*), in particular, t_*, is the ratio of two integers not exceeding in absolute value the quantity $(NL^2\sqrt{q+1})^{q+1} \leq (L^2 N^{3/2})^N$. Since t_* is positive, we have $t_* \geq (L^2 N^{3/2})^{-N}$. Thus, whenever $(P[\eta, j])$ is feasible, the optimal value in the problem is $\geq (L^2 N^{3/2})^{-N}$.

Now consider a point $x = \sum_{i=1}^{q} \xi_i \mu_i \in E$ such that $\|x\|_\infty \geq 1$, and let $\mu(x) = \max_{i \leq q} |\mu_i^T x|$. There exists $j \leq N$ such that $|x_j| \geq 1$; specifying η as the sign of x_j, we see that the collection $\xi_1, ..., \xi_q, \mu(x)$ is a feasible solution of the problem $(P[\eta, j])$; since the optimal value in this problem, as we just have seen, is $\geq (L^2 N^{3/2})^{-N}$, we get $\mu(x) \geq (L^2 N^{3/2})^{-N}$. ■

Proposition 3.2 *Let the vectors $\sigma_1, ..., \sigma_K$ be integral, and let the absolute values of the coordinates of these vectors be $\leq L$. Then*

$$\alpha(\sigma_1, ..., \sigma_K) \leq \widehat{\alpha} = (2L)^{2N} N^{\frac{3N+1}{2}}. \tag{31}$$

Proof. Let $\mu_1, ..., \mu_q$ be a maximal linearly independent subset of the set of differences $\sigma_k - \sigma_\ell$, $1 \leq k, \ell \leq K$; then $\mu_1, ..., \mu_q$ are integral vectors with $\|\mu_i\|_\infty \leq 2L$, $i = 1, ..., q$, which form a basis in E. By Lemma 3.1, it follows that for every $x \in E$ such that $\|x\|_\infty \geq 1$ there exists $i \leq q$ such that $|\mu_i^T x| \geq \theta \equiv (2L^2 N^{3/2})^{-N}$. It follows that whenever $x \in E$ satisfies $\|x\|_\infty \geq 1$, one has

$$\max_{k \leq K} \sigma_k^T x - \min_{k \leq K} \sigma_k x \geq \max_{i \leq q} |\mu_i^T x| \geq \theta,$$

whence, by homogeneity reasons, for every $x \in E$ it holds

$$\max_{k \leq K} \sigma_k^T x - \min_{k \leq K} \sigma_k x \geq \theta \|x\|_\infty \geq \theta N^{-1/2} \|x\|_2, \quad \theta = (2L^2 N^{3/2})^{-N}.$$

The resulting inequality, in view of the definition of $\alpha(\sigma_1, ..., \sigma_K)$ (see (8)), implies (31). ■

4 The LSS case

We are about to specify our results for the case when (GS) comes form (LSS). Basically all we need is to bound from above the quantity $\alpha(\sigma_1, ..., \sigma_K)$. A bound of this type is readily given by Proposition 3.2 (note that in the case in question the vectors σ_k are integral with 0-1 entries), and already this bound implies polynomial time solvability of (LSS$^*_+$). However, in the LSS case α admits an incomparably better bound:

Proposition 4.1 *In the LSS problem*

$$\alpha \leq 4mn\sqrt{m+n}. \tag{32}$$

Proof. It is well-known that a nonnegative $m \times n$ matrix A without zero rows and columns by permutations of rows and columns can be converted to the form

$$\begin{pmatrix} {\scriptstyle I_1}\{\overbrace{A_1}^{\bar{J}_1} & & & \\ & {\scriptstyle I_2}\{\overbrace{A_2}^{\bar{J}_2} & & \\ & & \ddots & \\ & & & {\scriptstyle I_q}\{\overbrace{A_q}^{\bar{J}_q} \end{pmatrix}, \tag{33}$$

where every block A_ℓ is *chainable*. The latter property is defined as follows. Let B be a nonnegative $p \times q$ matrix without zero rows and columns; we say that a row i of B *intersects* a column j of the matrix, if $B_{ij} > 0$. We can associate with B two graphs $G_{\text{row}} = (\{1, ..., p\}, E_{\text{row}})$ and $G_{\text{col}} = (\{1, ..., q\}, E_{\text{col}})$ as follows: a pair (i, i'), $i \neq i'$ of nodes of G_{row} is adjacent if and only if the ith and the i'th rows in B are intersected by a common column (i.e., $B_{ij} > 0, B_{i'j} > 0$ for some j). Similarly, a pair (j, j'), $j \neq j'$, of nodes of G_{col} is adjacent if and only if the columns j, j' in B are intersected by a common row, i.e., $B_{ij} > 0$, $B_{ij'} > 0$ for some i. B is called *chainable*, if both the graphs G_{row} and G_{col} are connected. It is immediately seen that an equivalent definition of chainability of B is as follows:

> For every $i, i' \in \{1, ..., p\}$ there exists a "chain"
>
> $$(i_1 = i, j_1), (i_2, j_1), (i_2, j_2), (i_3, j_2), ..., (i_{r-1}, j_{r-1}), (i_r = i', j_{r-1})$$
>
> of pairs of indices (μ, ν) with $r \leq p$ such that $B_{\mu\nu} > 0$ for every pair from the chain. Similarly, for every $j, j' \in \{1, ..., q\}$ there exists a chain
>
> $$(i_1, j_1 = j), (i_1, j_2), (i_2, j_2), (i_2, j_3), ..., (i_{s-1}, j_{s-1}), (i_{s-1}, j_s = j')$$
>
> of pairs of indices (μ, ν) with $s \leq q$ such that $B_{\mu\nu} > 0$ for every pair from the chain.

It is immediately seen that in the LSS case the quantity $\alpha(\sigma_1, ..., \sigma_K)$ we are interested in remains unchanged under permutations of rows and columns of the underlying matrix; thus, we may assume w.l.o.g. that the matrix A in question is in the form of (33) with chainable blocks $A_1, ..., A_q$.

Let $a, \sigma_1, ..., \sigma_K, \sigma$ be the data of the (GS)-reformulation of the LSS problem with matrix (33) (see the beginning of Section 2.1). Recall that in the case in question $N = m + n$. The index sets I_ℓ, $\bar{J}_\ell$ appearing in (33) induce a partition of the set $\{1, ..., N\}$ of entry indices of a vector from $\mathbf{R}^N$ into $2q$ sets I_ν, J_ν, $\nu = 1, ..., q$, where

$$J_\nu = m + \bar{J}_\nu = \{m + j \mid j \in \bar{J}_\nu\}.$$

We denote the cardinalities of I_ν, J_ν by m_ν, n_ν, respectively.

Let $x \in E$, and let

$$r = \max_{k \le K} \sigma_k^T x - \min_{k \le K} \sigma_k^T x.$$

1^0. Let us fix ν, $1 \le \nu \le q$. We claim that

$$\begin{array}{llrcl} (a) & & S_\nu \equiv \sum\limits_{i \in I_\nu} x_i & = & \sum\limits_{j \in J_\nu} x_j; \\ (b) & \forall i \in I_\ell : & |x_i - m_\nu^{-1} S_\nu| & \le & m_\nu r; \\ (c) & \forall j \in J_\ell : & |x_j - n_\nu^{-1} S_\nu| & \le & n_\nu r. \end{array} \tag{34}$$

Indeed, (34.a) is evident when x is of the form $\sigma_k - \sigma_\ell$, $1 \le k, \ell \le K$; since the relation is linear in x, it holds true on the linear span E of the vectors $\sigma_k - \sigma_\ell$.

To prove (34.b), let $i_+, i_- \in I_\nu$ be the indices of (one of) the largest, respectively, the smallest of the entries of x with indices from I_ν. Since A_ν is chainable, there exists a chain

$$(i_1 = i_+, j_1), (i_2, j_1), (i_2, j_2), ..., (i_{p-1}, j_{p-1}), (i_p = i_-, j_{p-1})$$

of pairs of indices with $p \le m_\nu$ such that for every pair (α, β) from the chain one has $A_{\alpha\beta} > 0$. For each such pair, $x_\alpha + x_{m+\beta} = \sigma_k^T x$ for certain k, consequently,

$$\begin{array}{rcl} x_{i_+} - x_{i_-} & = & [(x_{i_1} + x_{m+j_1}) - (x_{i_2} + x_{m+j_1})] \\ & & +[(x_{i_2} + x_{m+j_2}) - (x_{i_3} + x_{m+j_2})] + ... \\ & & +[(x_{i_{p-1}} + x_{m+j_{p-1}}) - (x_{i_p} + x_{m+j_{p-1}})] \\ & \le & m_\nu r. \end{array}$$

Thus, $\max\limits_{i \in I_\nu} x_i - \min\limits_{i \in I_\nu} x_i \le m_\nu r$, and therefore the distance of every one of x_i's, $i \in I_\nu$, from their mean $m_\nu^{-1} S_\nu$ does not exceed $m_\nu r$, as required in (34.b). Relation (34.c) is proved by the "symmetric" reasoning (taking into account (34.a) as well).

2^0. Let $\nu, \nu' \le q$. There exists k such that $\sigma_k = e_i + f_j$ with $i \in I_\nu$, $j \in \bar{J}_\nu$, same as there exists k' such that $\sigma_{k'} = e_{i'} + f_{j'}$ with $i' \in I_{\nu'}$, $j' \in \bar{J}_{\nu'}$. By definition of r we have

$$|(x_i + x_{m+j}) - (x_{i'} + x_{m+j'})| = |\sigma_k^T x - \sigma_{k'}^T x| \le r,$$

whence, in view of (34),

$$|S_\nu(m_\nu^{-1}+n_\nu^{-1}) - S_{\nu'}(m_{\nu'}^{-1}+n_{\nu'}^{-1})| \le r(1+m_\nu+r_\nu+m_{\nu'}+n_{\nu'}). \tag{35}$$

At the same time,

$$\sum_{\nu=1}^{q} S_\nu = \sum_{i=1}^{m} x_i = 0, \tag{36}$$

the concluding relation being readily given by the fact that it is valid when x is of the form $\sigma_k - \sigma_\ell$ and thus, by linearity – for all $x \in E$. It follows that

$$\forall \nu: \quad |S_\nu| \le 3mnr$$

(choose as ν, ν' in (35) the indices of the largest, respectively, the smallest of S_ℓ's and take into account that in view of (36) the resulting S_ν, $S_{\nu'}$ are of opposite signs). Combining the latter inequality and (34), we come to

$$\|x\|_\infty \le 4mnr,$$

whence

$$\|x\|_2 \le 4mn\sqrt{m+n}r.$$

Thus, whenever $x \in E$ is such that $\|x\|_2 = 1$, we have

$$r \equiv \max_k \sigma_k^T x - \min_k m\sigma_k^T x \ge \frac{1}{4mn\sqrt{m+n}},$$

and (32) follows (cf. (8)). ■

Remark 4.1 *It is easily seen that in the case when A has at least one positive row and at least one positive column, the bound* (32) *can be replaced with*

$$\alpha(\sigma_1,...,\sigma_K) \le \sqrt{m+n}.$$

Combining Theorem 3.1 and Proposition 4.1, we reach the following conclusion:

Corollary 4.1 *In the case of the LSS problem with chainable matrix A and r, c normalized according to* (1) *for every $\epsilon > 0$ the algorithm from Section 3 with the setup*

$$\widehat{\alpha} = 4mn\sqrt{m+n}$$

solves problem (LSS_+^*) *in no more than*

$$\mathcal{I} = O(1)(m+n)^2 \ln\left(2 + \tfrac{32K(\epsilon+8)^3 mn\sqrt{m+n}\ln(K\beta)}{\epsilon^3}\right)$$
$$\text{with} \tag{37}$$
$$\beta = \tfrac{\max_{i,j} A_{ij}}{\min\{A_{ij}\,|\,i,j: A_{ij}>0\}}$$

iterations with no more than

$$O(1)(m+n)^2$$

operations of real arithmetic per iteration, where K is the total number of nonzero entries in A.

Proof. To get the result from the one of Theorem 3.1, note that for the LSS problem one has $\delta(\sigma_1,...,\sigma_K) = 2$.

5 Extensions

We have demonstrated that the Line Sum Scaling problem for a nonnegative chainable matrix A that can be scaled to arbitrary prescribed accuracy can be solved, within prescribed accuracy $\epsilon \in (0,1)$, in no more than

$$O(1)(m+n)^4 \ln \left(2 + \frac{32m^2n^2\sqrt{m+n}\ln\left(\frac{mn\max_{i,j} A_{ij}}{\min\{A_{ij}\,|\,i,j: A_{ij}>0\}}\right)}{\epsilon^3} \right)$$

real arithmetic operations.

Note that the LSS problem is not the only interesting case of general setting (GS), (GS$^*_+$); see the examples in [8]. In particular, our analysis can be applied to some other incidents of (GS) as long as they admit a vector e satisfying (3). Let us consider two examples – *matrix balancing* and *multi-index sum scaling*.

5.1 Matrix balancing

The *matrix balancing* problem as follows:

> (MB) Given an $n \times n$ matrix A with nonnegative entries, find a diagonal matrix X with positive diagonal entries such that the row sums in the scaled matrix XAX^{-1} are equal to the respective column sums:
>
> $$XAX^{-1}\mathbf{1}_n = X^{-1}A^TX\mathbf{1}_n$$

along with the following approximate version of this problem:

> (MB$^*_+$) Given an $n \times n$ matrix A with nonnegative entries and $\epsilon > 0$, find a diagonal matrix X with positive diagonal entries such that
>
> $$\frac{\|XAX^{-1}\mathbf{1}_n - X^{-1}A^TX\mathbf{1}_n\|_1}{\mathbf{1}_n^TXAX^{-1}\mathbf{1}_n} \leq \epsilon$$
>
> or detect correctly that
>
> $$\inf\left\{\|XAX^{-1}\mathbf{1}_n - X^{-1}A^TX\mathbf{1}_n\|_1 \mid X = \mathrm{Diag}(x), x > 0\right\} > 0.$$

It is well-known (for details, see [3] and references therein) that (MB$^*_+$) can be easily reduced to the case when the matrix $A + A^T$ is chainable, which is assumed from now on. To represent (MB), (MB$^*_+$) in the form of (GS), (GS$^*_+$), it suffices to enumerate the pairs of indices (i,j) of the nonzero entries of A as $(i(1),j(1)),...,(i(K),j(K))$ and to set

$$\begin{array}{rcl} a & = & (A_{i(1)j(1)},...,A_{i(K)j(K)})^T, \\ N & = & n+1, \\ \sigma_k & = & \begin{pmatrix} 1 \\ e_{i(k)}-e_{j(k)} \end{pmatrix} \in \mathbf{R}^N,\ k=1,...,K, \\ \sigma & = & \begin{pmatrix} 1 \\ 0_n \end{pmatrix} \in \mathbf{R}^N, \end{array}$$

where the vectors $e_1,...,e_n$ form the standard basis of $\mathbf{R}^n$. With this setup, problem (GS) becomes

$$\text{Find } x \text{ such that } \sum_{k=1}^K A_{i(k)j(k)} \exp\{x_0+x_{i(k)}-x_{j(k)}\} \begin{pmatrix} 1 \\ e_{i(k)}-e_{j(k)} \end{pmatrix} = \sigma \equiv \begin{pmatrix} 1 \\ 0_n \end{pmatrix}, \quad (38)$$

which is nothing but problem (MB), the correspondence between X and x being given by $X_{ii}=\exp\{x_i\}$, $i=1,...,n$.

The associated problem (GS$^*_+$) is

$$\begin{array}{l} \text{Given } \epsilon>0, \text{ find } x \text{ such that} \\ \left\| \sum_{k=1}^K A_{i(k)j(k)} \exp\{x_0+x_{i(k)}-x_{j(k)}\} \begin{pmatrix} 1 \\ e_{i(k)}-e_{j(k)} \end{pmatrix} - \begin{pmatrix} 1 \\ 0_n \end{pmatrix} \right\|_1 \leq \epsilon \\ \text{or detect correctly that} \\ \inf_x \left\| A_{i(k)j(k)} \exp\{x_0+x_{i(k)}-x_{j(k)}\} \begin{pmatrix} 1 \\ e_{i(k)}-e_{j(k)} \end{pmatrix} - \begin{pmatrix} 1 \\ 0_n \end{pmatrix} \right\|_1 > 0 \end{array} \quad (39)$$

Note that if x is such that

$$\left\| \sum_{k=1}^K A_{i(k)j(k)} \exp\{x_0+x_{i(k)}-x_{j(k)}\} \begin{pmatrix} 1 \\ e_{i(k)}-e_{j(k)} \end{pmatrix} - \begin{pmatrix} 1 \\ 0_n \end{pmatrix} \right\|_1 \leq \epsilon < 1,$$

and $X=\mathrm{Diag}\{\exp\{x_1\},...,\exp\{x_n\}\}$, then

$$\begin{array}{rl} & \left\| \sum_{k=1}^K A_{i(k)j(k)} \exp\{x_0+x_{i(k)}-x_{j(k)}\} \begin{pmatrix} 1 \\ e_{i(k)}-e_{j(k)} \end{pmatrix} - \begin{pmatrix} 1 \\ 0_n \end{pmatrix} \right\|_1 \\ = & \left\| \begin{pmatrix} \exp\{x_0\}\mathbf{1}_n^T XAX^{-1}\mathbf{1}_n - 1 \\ \exp\{x_0\}\left[XAX^{-1}\mathbf{1}_n - X^{-1}A^TX\mathbf{1}_n\right] \end{pmatrix} \right\|_1 \\ \leq & \epsilon, \end{array}$$

whence

$$\frac{\left\| XAX^{-1}\mathbf{1}_n - X^{-1}A^TX\mathbf{1}_n \right\|_1}{\mathbf{1}_n^T XAX^{-1}\mathbf{1}_n} \leq \frac{\epsilon}{1-\epsilon},$$

so that (39) is, basically, (MB$^*_+$).

Note that the data of (38) clearly satisfy (3) (one should take $e=\sigma$). Thus, we can apply the results of the previous sections to get a solution to (39), or, which is the same, to (MB$^*_+$). The only element of the construction which is missing for the moment is an upper bound $\widehat{\alpha}$ on the quantity $\alpha(\sigma_1,...,\sigma_K)$ for our new situation. To get polynomial time results, it would be sufficient for us to use the universal bound from Proposition 3.2 (our vectors σ_k are integral with entries 0,1,-1). However, we, same as in the LSS case, can bet a much better upper bound on α:

Proposition 5.1 *Let $A+A^T$ be chainable, and let the data of problem* (38) *satisfy* $\mathbf{C}'$. *Then for $\sigma_1,...,\sigma_K$ associated with* (38) *one has*

$$\alpha(\sigma_1,...,\sigma_K) \leq 2n^{3/2}. \tag{40}$$

Proof. The linear span E of the vectors $\sigma_k - \sigma_\ell$, $k,\ell = 1,...,K$, clearly is contained in the space $E^+ = \{x \in \mathbf{R}^{n+1} \mid x_0 = 0, \sum_{i=1}^n x_i = 0\}$. Let $x \in E$. Since $\mathbf{C}'$ is satisfied, the vector $\sigma = \begin{pmatrix} 1 \\ 0_n \end{pmatrix}$ belongs to the convex hull of $\sigma_1,...,\sigma_K$, so that the segment $\Delta = [\min_k \sigma_k^T x, \max_k \sigma_k^T x]$ contains $0 = \sigma^T x$. It follows that if

$$\theta = \max_k \sigma_k^T x - \min_k \sigma_k^T x$$

is the length of Δ, then

$$|\sigma_k^T x| \leq \theta \quad \forall k. \tag{41}$$

Now let i_+ be the index of (one of) the largest, and i_- be the index of (one of) the smallest of the reals $x_1,...,x_n$. Since $A+A^T$ is chainable, there exists a chain

$$(i_1 = i_+, j_1), (i_2, j_1), (i_2, j_2), ..., (i_{p-1}, j_{p-1}), (i_p = i_-, j_{p-1})$$

with $p \leq n$ such that for every pair (μ,ν) from the chain either $A_{\mu\nu} > 0$, or $A_{\nu\mu} > 0$, or both. Denoting $y = (x_1,...,x_n)^T \in \mathbf{R}^n$, we have

$$\begin{array}{rcl} x_{i_+} - x_{i_-} &=& \left[\underbrace{e_{i_1} - e_{j_1}}_{d_1}\right]^T y - \left[\underbrace{e_{i_2} - e_{j_1}}_{d_2}\right]^T y \\ && + \left[\underbrace{e_{i_2} - e_{j_2}}_{d_3}\right]^T y - \left[\underbrace{e_{i_3} - e_{j_2}}_{d_4}\right]^T y + ... \\ && + \left[\underbrace{e_{i_{p-1}} - e_{j_{p-1}}}_{d_{2p-3}}\right]^T y - \left[\underbrace{e_{i_p} - e_{j_{p-1}}}_{d_{2p-2}}\right]^T y \end{array} \tag{42}$$

Now, for every ℓ $d_\ell^T y$ is either $\sigma_k^T x$ or $-\sigma_k^T x$ for some $k = k(\ell)$, so that (42), (41) imply that

$$x_{i_+} - x_{i_-} \leq (2p-2)\theta \leq 2(n-1)\theta. \tag{43}$$

Since $\sum_{i=1}^n x_i = 0$ and $x_0 = 0$, we have

$$\|x\|_2 \leq \sqrt{n}(x_{i_+} - x_{i_-}) \leq 2n^{3/2}\theta,$$

whence

$$x \in E, x \neq 0 \Rightarrow \frac{\|x\|_2}{\max_k \sigma_k^T x - \min_k \sigma_k^T x} \leq 2n^{3/2} \qquad \blacksquare$$

According to Proposition 5.1, when solving (39) via the scheme of Section 3, we can use, as an upper bound $\widehat{\alpha}$ on $\alpha(\sigma_1,...,\sigma_K)$, the quantity $2n^{3/2}$. Indeed, if the data of the problem satisfy $\mathbf{C}'$, then this is a valid bound on the true value of α, otherwise we should not bother at all whether this bound is valid or not, since the result generated by the algorithm, independently of its setup, is either an ϵ-balancing of A, or the conclusion that the data of A are not semi-proper, and in the case when $\mathbf{C}'$ is not satisfied (i.e., when the data are not semi-proper) both possibilities are acceptable. With $\widehat{\alpha} = 2n^{3/2}$, Theorem 3.1 states that for every $\epsilon \in (0,1)$ a solution to (BM$^*_\epsilon$) can be obtained by the algorithm from Section 3 in no more than

$$O(1)n^2 \ln\left(2 + \frac{n^{7/2}\ln(n^2\beta)}{\epsilon^3}\right)$$

iterations with no more than $O(1)n^2$ real arithmetic operations per iteration, where β is the ratio of the largest and the smallest positive entries of A. This is exactly the result established for the matrix balancing problem in [3]. As we see, one can obtain this result quite straightforwardly from Theorem 3.1.

5.2 Multi-index sum scaling

The problem we intend to address is as follows. Assume we are given a $n[1] \times n[2] \times \ldots \times n[p]$ nonnegative array

$$A = \{A_\iota\}_{\iota \in \mathcal{I}}, \mathcal{I} = \{\iota = (\iota[1], ..., \iota[p]), 1 \leq \iota[i] \leq n_i, i = 1, ..., p\},$$

along with q distinct nonempty subsets I_ℓ, $\ell = 1, ..., q$, of the set $I = \{1, ..., p\}$:

$$I_\ell = \{i[1,\ell]; i[2,\ell]; ...; i[p_\ell,\ell]\},\ 1 \leq i[1,\ell] < i[2,\ell] < ... < i[p_\ell,\ell] \leq p.$$

For a p-dimensional multiindex $\iota = (\iota[1], ..., \iota[p])$, let its projection $\iota^{(\ell)}$ on I_ℓ be defined as the p_ℓ-dimensional multiindex $(\iota[i[1,\ell]], \iota[i[2,\ell]], ..., \iota[i[p_\ell,\ell]])$. Finally, let for each $\ell \leq q$ an array

$$R^\ell = \{R_\omega\}_{\omega \in \mathcal{I}_\ell}, \quad \mathcal{I}_\ell = ell = \{\omega = (\omega[1], ..., \omega[p_\ell]), 1 \leq \omega[j] \leq n[i[j,\ell]], j = 1, ..., p_\ell\}$$

be given. The data $(A, I_1, R^1, I_2, R^2, ..., I_q, R^q)$ define a scaling problem as follows:

(MIS) Find positive arrays $X^1, ..., X^q$ of the same structure as $R^1, ..., R^q$ in such a way that for the "X-scaling of A" – the p-dimensional array

$$B = B(X^1, ..., X^q, A) \equiv \{B_\iota = X^1_{\iota^{(1)}} X^2_{\iota^{(2)}} ... X^q_{\iota^{(q)}} A_\iota\}_{\iota \in \mathcal{I}}$$

for every $\ell \leq q$ and every $\omega \in \mathcal{I}_\ell$ it holds

$$R^\ell_\omega = \sum_{\iota \in \mathcal{I}: \iota^{(\ell)} = \omega} B_\iota.$$

Note that (MIS) covers a lot of different scaling problems with nonnegative arrays. E.g.,

• when $p = q$ and $I_\ell = \{\ell\}$, $\ell = 1, 2, ..., p$, (MIS) becomes the problem of a diagonal scaling of a nonnegative p-dimensional array to prescribed "hyperplane" sums:

> *Given a nonnegative* $n[1] \times ... \times n[p]$ *array* $A = \{A_{i_1,...,i_p}\}$ *and vectors* $R^\ell \in \mathbf{R}^{[n[\ell]}$, $\ell = 1, ..., p$, *find positive vectors of scales* $X^\ell \in \mathbf{R}^{n[\ell]}$, $\ell = 1, ..., p$, *such that*
>
> $$R^\ell_{i_\ell} = \sum_{i_1,...,i_{\ell-1},i_{\ell+1},...,i_p} X^1_{i_1} X^2_{i_2} ... X^p_{i_p} A_{i_1,...,i_p}$$
>
> *for all* ℓ *and all* i_ℓ, $1 \le i_\ell \le n[\ell]$.

Note that when $p = 2$, we get the usual LSS problem.

• when $p = q$ and $I_\ell = \{1, ..., p\} \backslash \{\ell\}$, $\ell = 1, ..., p$, (MIS) becomes the problem of "codiagonal" scaling of a nonnegative p-dimensional array to prescribed line sums:

> *Given a nonnegative* $n[1] \times ... \times n[p]$ *array* $A = \{A_{i_1,...,i_p}\}$ *and* $n[1] \times ... \times n[\ell-1] \times n[\ell+1] \times ... \times n[p]$ *arrays* R^ℓ, $\ell = 1, ..., p$, *find positive* $n[1] \times ... \times n[p]$ *arrays* $X^\ell = \{X^\ell_{i_1,...,i_p}\}$ *with* $X^\ell_{i_1,...,i_p}$ *independent of* i_ℓ, $\ell = 1, ..., p$, *such that*
>
> $$R^\ell_{i_1,...,i_{\ell-1},i_{\ell+1},...,i_p} = \sum_{i_\ell} X^1_{i_1,...,i_p} X^2_{i_1,...,i_p} ... X^p_{i_1,...,i_p} A_{i_1,...,i_p}$$
>
> *for all* ℓ *and all* $i_1, ..., i_{\ell-1}, i_{\ell+1}, ..., i_p$.

Observe that an evident necessary condition for (MIS) to be solvable is

$$R^\ell \ge 0, \ell = 1, ..., q; \quad \sum_{\omega \in \mathcal{I}_\ell} R^\ell_\omega = \sum_{\omega' \in \mathcal{I}_{\ell'}} R^{\ell'}_{\omega'}, \ \ell, \ell' = 1, ..., q.$$

Besides this, a normalization $R^\ell \mapsto tR^\ell$, $t > 0$, converts an instance of (MIS) into an equivalent instance. Thus, when speaking about (MIS), we without loss of generality may normalize the data to satisfy the condition

$$R^\ell \ge 0 \ \& \ \sum_{\omega \in \mathcal{I}_\ell} R^\ell_\omega = 1, \ \ell = 1, ..., q. \tag{44}$$

In the discussion to follow, we assume that this condition holds true.

Note that (MIS) can be easily reformulated in the form of (GS). Indeed, let

$$\mathbf{E}_\ell = \mathbf{R}^{n[i[1,\ell]]} \otimes \mathbf{R}^{n[i[2,\ell]]} \otimes ... \otimes \mathbf{R}^{n[i[p_\ell,\ell]]},$$

($\otimes$ stands for the tensor product), so that the vectors from the natural basis of $\mathbf{E}_\ell$ are indexed by multiindices $\omega \in \mathcal{I}_\ell$. Let us set

$$\mathbf{R}^N = \mathbf{E}_1 \times ... \times \mathbf{E}_q,$$

and let e^ℓ_ω, $1 \le \ell \le q$, $\omega \in \mathcal{I}_\ell$, be the elements of the natural basis in the direct product (so that the only nonzero component of e^ℓ_ω is the basis vector, indexed by ω, of the direct factor $\mathbf{E}_\ell$). Now, let $\mathcal{J} \subset \mathcal{I}$ be the set of indices of nonzero elements of the array A, and let $\iota[k] = (\iota[1,k], ..., \iota[p,k])$, $k = 1, 2, ..., K = \mathrm{Card}\mathcal{J}$, be a enumeration of $\mathcal{J}$. For $1 \le k \le K$, let us set

$$\sigma_k = \sum_{\ell=1}^q e^\ell_{(\iota[k])^{(\ell)}} \in \mathbf{R}^N,$$

and let

$$\sigma = \sum_{\ell=1}^q \sum_{\omega \in \mathcal{I}_\ell} R^\ell_\omega e^\ell_\omega \in \mathbf{R}^N.$$

It is immediately seen that (MIS) is equivalent to the problem

$$\text{Find } x \in \mathbf{R}^N\text{:} \quad \sum_{k=1}^K A_{\iota[k]} \exp\{\sigma_k^T x\}\sigma_k = \sigma, \tag{45}$$

which is an instance of (GS). Moreover, (45) is a standard instance of (GS), since with $e = \frac{1}{q}\mathbf{1}_N$ we clearly have

$$e^T\sigma_1 = e^T\sigma_2 = ... = e^T\sigma_K = e^T\sigma = 1.$$

Thus, we can apply the machinery from Section 3 to solve ϵ-version of (45) and thus – ϵ-version of (MIS). Note that the vectors σ_k arising in (45) are integral with modulae of entries not exceeding 1, so that by Proposition 3.2 we have $\alpha(\sigma_1, ..., \sigma_K) \le 2^{2N} N^{(3N+1)/2}$. Applying Theorem 3.1, we get the following result:

Proposition 5.2 *Let the data* $(p, q, n[1], ..., n[p], A, I_1, R^1, I_2, R^2, ..., I_q, R^q)$, $A \ge 0$, *of an instance of* (MIS) *satisfy* (44), *and let* $\epsilon > 0$ *be given. Then in no more than*

$$\begin{array}{c} N^2 \ln\left(2 + \frac{2^{2N+2}N^{(3N+1)/2}K(\epsilon+4q)q\ln(K\beta)}{\epsilon^3}\right) \\ \left[\begin{array}{c} N = \sum\limits_{\ell=1}^q \prod\limits_{j=1}^{p_\ell} n[i[j,\ell]], \\ K = \mathrm{Card}\{\iota : A_\iota \neq 0\}, \quad \beta = \frac{\max_\iota A_\iota}{\min_{\iota:A_\iota>0} A_\iota} \end{array}\right] \end{array} \tag{46}$$

iterations of certain algorithm, with no more than $O(1)(N^2 + qK)$ *operations of real arithmetic per iteration, one can either find positive "scalings"* $\{X^\ell_\omega\}_{\omega\in\mathcal{I}_\ell}$, $\ell = 1, ..., q$, *forming an* ϵ*-solution to* (MIS)*:*

$$\sum_{\omega\in\mathcal{I}_\ell} \left| R^\ell_\omega - \sum_{\iota:\iota^{(\ell)}=\omega} X^1_{\iota^{(1)}} ... X^q_{\iota^{(q)}} A_\iota \right| \le \epsilon, \ \ell = 1, ..., q,$$

or detect correctly that (MIS) *has no solutions.*

Note that for a once for ever fixed "dimensionality" p of (MIS), the operations count given by Proposition 5.2 is polynomial in $\ln\frac{1}{\epsilon}$, $\ln\ln\beta$ and the sizes $n[1], ..., n[p]$ of the problem.

References

[1] Avriel, M. *Nonlinear programming: Analysis and methods.* Prentice Hall, Englewood Cliffs, New Jersey, 1976.

[2] Kalantari, B., L. Khachiyan. On the complexity of nonnegative-matrix scaling. *Linear Algebra Appl.* **240** (1996), 87-103.

[3] Kalantari, B., L. Khachiyan, A. Shokoufandeh. On the complexity of matrix balancing. *SIAM J. Matrix Anal. Appl.* **16**:2 (1997), 450-463.

[4] Khachiyan, L. Polynomial time algorithm in Linear Programming. *Soviet Math. Doklady* **244** (1979), 1093-1096.

[5] Linial, N., A. Samorodnitsky, A. Wigderson. A deterministic strongly polynomial algorithm for matrix scaling and approximate permanents. Proceedings of STOC 98.

[6] Nemirovski, A. Polynomial time methods in Convex Programming. in: J. Renegar, M. Shub and S. Smale, Eds., *The Mathematics of Numerical Analysis*, 1995 AMS-SIAM Summer Seminar on Mathematics in Applied Mathematics, July 17 – August 11, 1995, Park City, Utah. – Lectures in Applied Mathematics, v. 32 (1996): AMS, Providence, 543-589.

[7] Rothblum, U.G., H. Schneider. Scaling of matrices which have prescribed row sums and column sums via optimization. *Linear Algebra Appl.* **114/115** (1989), 737-764.

[8] Rothblum, U.G. Generalized scalings satisfying linear equations. *Linear Algebra Appl.* **114/115** (1989), 765-783.