SIAM J. DISCRETE MATH.
Vol. 29, No. 4, pp. 1783–1791

INTEGER FACTORIZATION OF A POSITIVE-DEFINITE MATRIX*

JOEL A. TROPP†

Abstract. This paper establishes that every positive-definite matrix can be written as a positive linear combination of outer products of integer-valued vectors whose entries are bounded by the geometric mean of the condition number and the dimension of the matrix.

Key words. conic geometry, convex geometry, discrete geometry, matrix factorization, positive-definite matrix

AMS subject classifications. Primary, 52A99; Secondary, 52C99

DOI. 10.1137/15M1024718

1. Motivation.

This paper addresses a geometric question that arises in the theory of discrete normal approximation [1] and in the analysis of hardware for implementing matrix multiplication [6]. The problem requires us to represent a nonsingular covariance matrix as a positive linear combination of outer products of integer vectors. The theoretical challenge is to obtain an optimal bound on the magnitude of the integers required as a function of the condition number of the matrix. We establish the following result.

THEOREM 1.1. *For positive integers m and d, define a set of bounded integer vectors*

$$\mathbb{Z}_m^d := \left\{ \boldsymbol{z} \in \mathbb{Z}^d : |z_i| \le m \text{ for } i = 1, \dots, d \right\}.$$

Let $\boldsymbol{A}$ be a real $d \times d$ positive-definite matrix with (finite) spectral condition number

$$\kappa(\boldsymbol{A}) := \lambda_{\max}(\boldsymbol{A}) / \lambda_{\min}(\boldsymbol{A}),$$

where $\lambda_{\max}$ and $\lambda_{\min}$ denote the maximum and minimum eigenvalue maps. Every such matrix $\boldsymbol{A}$ can be expressed as

$$\boldsymbol{A} = \sum_{i=1}^{r} \alpha_i \boldsymbol{z}_i \boldsymbol{z}_i^*, \quad \text{where} \quad \boldsymbol{z}_i \in \mathbb{Z}_m^d \quad \text{and} \quad m \le 1 + \frac{1}{2}\sqrt{(d-1)\cdot\kappa(\boldsymbol{A})}.$$

The coefficients α_i are positive, and the number r of terms satisfies $r \le d(d+1)/2$. The symbol * refers to the transpose operation.

This result has an alternative interpretation as a matrix factorization:

$$\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{\Delta}\boldsymbol{Z}^*.$$

In this expression, $\boldsymbol{Z}$ is a $d \times r$ integer matrix with entries bounded by m. The $r \times r$ matrix $\boldsymbol{\Delta}$ is nonnegative and diagonal.

The proof of Theorem 1.1 appears in section 3. Section 4 demonstrates that the dependence on the condition number cannot be improved. We believe that the dependence on the dimension is also optimal, but we did not find an example that confirms this surmise.

*Received by the editors June 8, 2015; accepted for publication (in revised form) August 12, 2015; published electronically October 1, 2015. This work was supported by ONR award N00014-11-1002 and the Gordon and Betty Moore Foundation.
http://www.siam.org/journals/sidma/29-4/M102471.html

†Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125-5000 (jtropp@cms.caltech.edu).

2. Notation and background. This section contains brief preliminaries. The books [5, 3, 2, 4] are good foundational references for the techniques in this paper.

We use lowercase italic letters, such as c, for scalars. Lowercase boldface letters, such as $\boldsymbol{z}$, denote vectors. Uppercase boldface letters, such as $\boldsymbol{A}$, refer to matrices. We write z_i for the ith component of a vector $\boldsymbol{z}$, and a_{ij} for the (i, j) component of a matrix $\boldsymbol{A}$. The jth column of the matrix $\boldsymbol{A}$ will be denoted by $\boldsymbol{a}_j$.

We work primarily in the real linear space $\mathbb{H}^d$ of real $d \times d$ symmetric matrices, equipped with the usual componentwise addition and scalar multiplication:

$$\mathbb{H}^d := \left\{ \boldsymbol{A} \in \mathbb{R}^{d\times d} : \boldsymbol{A} = \boldsymbol{A}^* \right\}.$$

Note that $\mathbb{H}^d$ has dimension $d(d + 1)/2$. The *trace* of a matrix $\boldsymbol{A} \in \mathbb{H}^d$ is the sum of its diagonal entries

$$\operatorname{tr}(\boldsymbol{A}) := \sum_{i=1}^{d} a_{ii}.$$

We equip $\mathbb{H}^d$ with the inner product $(\boldsymbol{B}, \boldsymbol{A}) \mapsto \operatorname{tr}(\boldsymbol{BA})$ to obtain a real inner-product space. All statements about closures refer to the norm topology induced by this inner product.

Define the set of *positive-semidefinite matrices* in $\mathbb{H}^d$:

$$\mathbb{H}^d_+ := \left\{ \boldsymbol{A} \in \mathbb{H}^d : \boldsymbol{u}^* \boldsymbol{A} \boldsymbol{u} \geq 0 \text{ for each } \boldsymbol{u} \in \mathbb{R}^d \right\}.$$

Similarly, the set of *positive-definite matrices* is

$$\mathbb{H}^d_{++} := \left\{ \boldsymbol{A} \in \mathbb{H}^d : \boldsymbol{u}^* \boldsymbol{A} \boldsymbol{u} > 0 \text{ for each } \textit{nonzero } \boldsymbol{u} \in \mathbb{R}^d \right\}.$$

The members of the set $-\mathbb{H}^d_{++}$ are called *negative-definite matrices*.

For a matrix $\boldsymbol{A} \in \mathbb{H}^d$, the decreasingly ordered eigenvalues will be written as

$$\lambda_1^{\downarrow}(\boldsymbol{A}) \geq \lambda_2^{\downarrow}(\boldsymbol{A}) \geq \cdots \geq \lambda_d^{\downarrow}(\boldsymbol{A}).$$

Similarly, the increasingly ordered eigenvalues are denoted as

$$\lambda_1^{\uparrow}(\boldsymbol{A}) \leq \lambda_2^{\uparrow}(\boldsymbol{A}) \leq \cdots \leq \lambda_d^{\uparrow}(\boldsymbol{A}).$$

Note that each eigenvalue map $\lambda(\cdot)$ is positively homogeneous; that is, $\lambda(\alpha\boldsymbol{A}) = \alpha\lambda(\boldsymbol{A})$ for all $\alpha > 0$.

Let us introduce some concepts from conic geometry in the setting of $\mathbb{H}^d$. A *cone* is a subset $K \subset \mathbb{H}^d$ that is positively homogeneous; in other words, $\alpha K = K$ for all $\alpha > 0$. A *convex cone* is a cone that is also a convex set. The *conic hull* of a set $E \subset \mathbb{H}^d$ is the smallest convex cone that contains E:

$$\operatorname{cone}(E) := \left\{ \sum_{i=1}^{r} \alpha_i \boldsymbol{A}_i : \alpha_i \geq 0 \text{ and } \boldsymbol{A}_i \in E \text{ and } r \in \mathbb{N} \right\}. \tag{2.1}$$

The conic hull of a finite set is closed. Since the space $\mathbb{H}^d$ has dimension $d(d + 1)/2$, we can choose the explicit value $r = d(d + 1)/2$ in the expression (2.1). This point follows from a careful application of Carathéodory's theorem [2, Thm. I(2.3)].

The *dual cone* associated with a cone $K \subset \mathbb{H}^d$ is the set

$$K^* := \left\{ \boldsymbol{B} \in \mathbb{H}^d : \operatorname{tr}(\boldsymbol{BA}) \geq 0 \text{ for each } \boldsymbol{A} \in K \right\}. \tag{2.2}$$

This set is always a closed convex cone because it is an intersection of closed halfspaces. It is easy to check that conic duality reverses inclusion; that is, for any two cones $C, K \subset \mathbb{H}^d$,

$$C \subset K \quad \text{implies} \quad K^* \subset C^*.$$

Note that we take the relation $\subset$ to include the possibility that the sets are equal. The bipolar theorem [2, Thm. IV(4.2)] states that the double dual $(K^*)^*$ of a cone K equals the closure of the conic hull of K.

3. Proof of Theorem 1.1. We will establish Theorem 1.1 using methods from the geometry of convex cones. The result is ultimately a statement about the containment of one convex cone in another. We approach this question by verifying the reverse inclusion for the dual cones. To obtain a good bound on the size of the integer vectors, the key idea is to use an averaging argument.

3.1. Step 1: Reduction to conic geometry. Once and for all, fix the ambient dimension d. First, we introduce the convex cone of positive-definite matrices with bounded condition number. For a real number $c \geq 1$, define

$$K(c) := \{\boldsymbol{A} \in \mathbb{H}^d_{++} : \kappa(\boldsymbol{A}) \leq c\}.$$

The set $K(c)$ is a cone because the condition number is scale invariant: $\kappa(\alpha\boldsymbol{A}) = \kappa(\boldsymbol{A})$ for $\alpha > 0$. To see that $K(c)$ is convex, write the membership condition $\kappa(\boldsymbol{A}) \leq c$ in the form

$$\lambda_{\max}(\boldsymbol{A}) - c \cdot \lambda_{\min}(\boldsymbol{A}) \leq 0.$$

On the space of symmetric matrices, the maximum eigenvalue is convex, while the minimum eigenvalue is concave [4, Ex. 3.10]. Since $K(c)$ is a sublevel set of a convex function, it must be convex.

Next, select a positive integer m. We introduce a closed convex cone of positive-semidefinite matrices derived from the outer products of bounded integer vectors:

$$Z(m) := \operatorname{cone}\{\boldsymbol{z}\boldsymbol{z}^* : \boldsymbol{z} \in \mathbb{Z}^d_m\}.$$

It is evident that $Z(m)$ is a closed convex cone because it is the conic hull of a finite set. Note that every element of this cone can be written as

$$\sum_{i=1}^{r} \alpha_i \boldsymbol{z}_i \boldsymbol{z}_i^*, \quad \text{where } \alpha_i \geq 0 \text{ and } \boldsymbol{z}_i \in \mathbb{Z}^d_m.$$

By the Carathéodory theorem, we may take the number r of summands to be $r = d(d+1)/2$.

Therefore, we can prove Theorem 1.1 by verifying that

$$K(c) \subset Z(m) \quad \text{when} \quad m \geq \frac{1}{2}\sqrt{(d-1)\cdot c}. \tag{3.1}$$

Indeed, the formula $1+\frac{1}{2}\sqrt{(d-1)\cdot\kappa(\boldsymbol{A})}$ in the theorem statement produces a positive integer that satisfies the latter inequality when $c = \kappa(\boldsymbol{A})$. Since the operation of conic duality reverses inclusion and $Z(m)$ is closed, the condition (3.1) is equivalent to

$$Z(m)^* \subset K(c)^* \quad \text{when} \quad m \geq \frac{1}{2}\sqrt{(d-1)\cdot c}. \tag{3.2}$$

We will establish the inclusion (3.2).

3.2. Step 2: The dual of $K(c)$. Our next objective is to obtain a formula for the dual cone $K(c)^*$. We claim that

(3.3)
$$K(c)^* = \left\{ \boldsymbol{B} \in \mathbb{H}^d : \sum_{i=s+1}^{d} \lambda_i^{\downarrow}(\boldsymbol{B}) \geq -\frac{1}{c} \sum_{i=1}^{s} \lambda_i^{\downarrow}(\boldsymbol{B}) \text{ where } \lambda_s^{\downarrow}(\boldsymbol{B}) \geq 0 > \lambda_{s+1}^{\downarrow}(\boldsymbol{B}) \right\}.$$

We instate the convention that $s = d$ when $\boldsymbol{B}$ is positive semidefinite. In particular, the set of positive-semidefinite matrices is contained in the dual cone: $\mathbb{H}_+^d \subset K(c)^*$. We also interpret the $s = 0$ case in (3.3) to exclude negative-definite matrices from $K(c)^*$.

Let us establish (3.3). The definition (2.2) of a dual cone leads to the equivalence

$$\boldsymbol{B} \in K(c)^* \quad \text{if and only if} \quad 0 \leq \inf_{\boldsymbol{A} \in K(c)} \operatorname{tr}(\boldsymbol{BA}).$$

To evaluate the infimum, note that the cone $K(c)$ is orthogonally invariant because the condition number of a matrix depends only on the eigenvalues. That is, $\boldsymbol{A} \in K(c)$ implies that $\boldsymbol{QAQ}^* \in K(c)$ for each orthogonal matrix $\boldsymbol{Q}$ with dimension d. Therefore, $\boldsymbol{B} \in K(c)^*$ if and only if

$$0 \leq \inf_{\boldsymbol{A} \in K(c)} \inf_{\boldsymbol{Q}} \operatorname{tr}(\boldsymbol{BQAQ}^*) = \inf_{\boldsymbol{A} \in K(c)} \sum_{i=1}^{d} \lambda_i^{\downarrow}(\boldsymbol{B}) \cdot \lambda_i^{\uparrow}(\boldsymbol{A}). \tag{3.4}$$

The inner infimum takes place over orthogonal matrices $\boldsymbol{Q}$. The identity is a well-known result due to Richter [8, Satz 1]; see the paper [7, Thm. 1] for an alternative proof. This fact is closely related to (a version of) the Hoffman–Wielandt theorem [3, Prob. III.6.15]

Now, the members of the cone $K(c)$ are those matrices $\boldsymbol{A}$ whose eigenvalues satisfy the bounds $0 < \lambda_1^{\uparrow}(\boldsymbol{A})$ and $\lambda_d^{\uparrow}(\boldsymbol{A}) \leq c \cdot \lambda_1^{\uparrow}(\boldsymbol{A})$. Owing to the invariance of the inequality (3.4) and the cone $K(c)$ to scaling, we can normalize $\boldsymbol{A}$ so that $\lambda_1^{\uparrow}(\boldsymbol{A}) = 1$. Thus, the inequality (3.4) holds if and only if

$$0 \leq \inf \left\{ \sum_{i=1}^{d} \lambda_i^{\downarrow}(\boldsymbol{B}) \cdot \mu_i : 1 = \mu_1 \leq \mu_2 \leq \cdots \leq \mu_d \leq c \right\}.$$

If $\boldsymbol{B}$ is positive semidefinite, then this bound is always true. If $\boldsymbol{B}$ is negative definite, then this inequality is always false. Ruling out these cases, let s be the index where $\lambda_s^{\downarrow}(\boldsymbol{B}) \geq 0 > \lambda_{s+1}^{\downarrow}(\boldsymbol{B})$, and observe that $0 < s < d$. The infimum is achieved when we select $\mu_i = 1$ for $i = 1, \ldots s$ and $\mu_i = c$ for $i = s+1, \ldots, d$. In conclusion,

$$\boldsymbol{B} \in K(c)^* \quad \text{if and only if} \quad 0 \leq \sum_{i=1}^{s} \lambda_i^{\downarrow}(\boldsymbol{B}) + c \sum_{i=s+1}^{d} \lambda_i^{\downarrow}(\boldsymbol{B}).$$

With our conventions for $s = 0$ and $s = d$, this inequality coincides with the advertised result (3.3).

3.3. Step 3: The dual of $Z(m)$. Next, we check that

$$Z(m)^* = \left\{ \boldsymbol{B} \in \mathbb{H}^d : \boldsymbol{z}^* \boldsymbol{B} \boldsymbol{z} \geq 0 \text{ for each } \boldsymbol{z} \in \mathbb{Z}_m^d \right\}. \tag{3.5}$$

According to the definition (2.2) of a dual cone,

$$Z(m)^* = \{\boldsymbol{B} \in \mathbb{H}^d : \operatorname{tr}(\boldsymbol{B}\boldsymbol{A}) \geq 0 \text{ for each } \boldsymbol{A} \in Z(m)\}.$$

Since $Z(m)$ is the conic hull of the matrices $\boldsymbol{z}\boldsymbol{z}^*$ where $\boldsymbol{z} \in \mathbb{Z}_m^d$, the matrix $\boldsymbol{B} \in Z(m)^*$ if and only if $\operatorname{tr}(\boldsymbol{B}\boldsymbol{A}) \geq 0$ for each matrix $\boldsymbol{A} = \boldsymbol{z}\boldsymbol{z}^*$. Therefore,

$$Z(m)^* = \{\boldsymbol{B} \in \mathbb{H}^d : \operatorname{tr}(\boldsymbol{B}\boldsymbol{z}\boldsymbol{z}^*) \geq 0 \text{ for each } \boldsymbol{z} \in \mathbb{Z}_m^d\}.$$

Cycling the trace, we arrive at the representation (3.5).

3.4. Step 4: Checking membership. Finally, we need to verify that $Z(m)^* \subset K(c)^*$ under suitable conditions on the parameters m and c.

To that end, select a matrix $\boldsymbol{B} \in Z(m)^*$. If $\boldsymbol{B}$ is positive semidefinite, then $\boldsymbol{B} \in K(c)^*$ because $K(c)^*$ contains the set of positive-semidefinite matrices. It is not possible for $\boldsymbol{B}$ to be negative definite because the expression (3.5) forces $\boldsymbol{z}^*\boldsymbol{B}\boldsymbol{z} \geq 0$ for each nonzero $\boldsymbol{z} \in \mathbb{Z}_m^d$. Therefore, we may exclude these cases.

Let s be the index where $\lambda_s^{\downarrow}(\boldsymbol{B}) \geq 0 > \lambda_{s+1}^{\downarrow}(\boldsymbol{B})$, and note that $0 < s < d$. The formula (3.3) indicates that we should examine the sum of the $d-s$ smallest eigenvalues of $\boldsymbol{B}$ to determine whether $\boldsymbol{B}$ is a member of $K(c)^*$. This sum of eigenvalues can be represented as a trace [5, (4.3.20)]:

$$\sum_{i=s+1}^{d} \lambda_1^{\downarrow}(\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{U}^*\boldsymbol{B}\boldsymbol{U}), \quad \text{where } \boldsymbol{U} \text{ is } d \times (d-s) \text{ with orthonormal columns.}$$

In view of (3.5), we must use the fact that $\boldsymbol{z}^*\boldsymbol{B}\boldsymbol{z} \geq 0$ for $\boldsymbol{z} \in \mathbb{Z}_m^d$ to bound the sum of eigenvalues below.

We will achieve this goal with an averaging argument. For each number $a \in [-1, 1]$, define an integer-valued random variable:

$$R_m(a) := \begin{cases} \lceil ma \rceil & \text{with probability } ma - \lfloor ma \rfloor, \\ \lfloor ma \rfloor & \text{with probability } 1 - (ma - \lfloor ma \rfloor). \end{cases}$$

Each of the random variables $R_m(a)$ is supported on $\{0, \pm 1, \ldots, \pm m\}$. Furthermore, $\mathbb{E}\, R_m(a) = ma$ and $\operatorname{Var}(R_m(a)) \leq \frac{1}{4}$. In other words, we randomly round ma up or down to the nearest integer in such a way that the average value is ma and the variance is uniformly bounded. Note that $R_m(a)$ is a constant random variable whenever ma takes an integer value.

We apply this randomized rounding operation to each entry u_{ij} of the matrix $\boldsymbol{U}$. Let $\boldsymbol{X}$ be a $d \times (d-s)$ random matrix with independent entries X_{ij} that have the distributions

$$X_{ij} \sim \frac{1}{m} R_m(u_{ij}) \quad \text{for } i = 1, \ldots, d \text{ and } j = 1, \ldots, d-s.$$

By construction, $\mathbb{E}\,\boldsymbol{X} = \boldsymbol{U}$ and $\operatorname{Var}(X_{ij}) \leq 1/(4m^2)$ for each pair (i, j) of indices.

Develop the desired quantity by adding and subtracting the random matrix $\boldsymbol{X}$:

$$\begin{aligned} \sum_{i=s+1}^{d} \lambda_1^{\downarrow}(\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{U}^*\boldsymbol{B}\boldsymbol{U}) &= \operatorname{tr}(\boldsymbol{X}^*\boldsymbol{B}\boldsymbol{X}) - \operatorname{tr}((\boldsymbol{X}-\boldsymbol{U})^*\boldsymbol{B}(\boldsymbol{X}-\boldsymbol{U})) \\ &\quad - \operatorname{tr}(\boldsymbol{U}^*\boldsymbol{B}(\boldsymbol{X}-\boldsymbol{U})) - \operatorname{tr}((\boldsymbol{X}-\boldsymbol{U})^*\boldsymbol{B}\boldsymbol{U}). \end{aligned}$$

Take the expectation over $\boldsymbol{X}$ and use the property $\mathbb{E}\,\boldsymbol{X} = \boldsymbol{U}$ to reach

$$\sum_{i=s+1}^{d} \lambda_1^{\downarrow}(\boldsymbol{B}) = \mathbb{E}\,\mathrm{tr}(\boldsymbol{X}^*\boldsymbol{B}\boldsymbol{X}) - \mathbb{E}\,\mathrm{tr}((\boldsymbol{X}-\boldsymbol{U})^*\boldsymbol{B}(\boldsymbol{X}-\boldsymbol{U})). \tag{3.6}$$

It remains to bound the right-hand side of (3.6) below.

Expand the trace in the first term on the right-hand side of (3.6):

$$\mathbb{E}\,\mathrm{tr}(\boldsymbol{X}^*\boldsymbol{B}\boldsymbol{X}) = \mathbb{E}\left[\sum_{j=1}^{d-s} \boldsymbol{x}_j^*\boldsymbol{B}\boldsymbol{x}_j\right] = \frac{1}{m^2}\,\mathbb{E}\left[\sum_{j=1}^{d-s} (m\boldsymbol{x}_j)^*\boldsymbol{B}(m\boldsymbol{x}_j)\right] \geq 0. \tag{3.7}$$

We have written $\boldsymbol{x}_j$ for the jth column of $\boldsymbol{X}$. Each vector $m\boldsymbol{x}_j$ belongs to $\mathbb{Z}_m^d$. Since $\boldsymbol{B} \in Z(m)^*$, it follows from the representation (3.5) of the cone that each of the summands is nonnegative.

Next, we turn to the second term on the right-hand side of (3.6):

$$\begin{aligned}\mathbb{E}\,\mathrm{tr}((\boldsymbol{X}-\boldsymbol{U})^*\boldsymbol{B}(\boldsymbol{X}-\boldsymbol{U})) &= \sum_{j=1}^{d-s} \mathbb{E}\left[(\boldsymbol{x}_j-\boldsymbol{u}_j)^*\boldsymbol{B}(\boldsymbol{x}_j-\boldsymbol{u}_j)\right] \\ &= \sum_{j=1}^{d-s}\sum_{i=1}^{d} \mathbb{E}\left[(X_{ij}-u_{ij})^2\right]\cdot b_{ii} \leq \frac{d-s}{4m^2}\sum_{i=1}^{d}(b_{ii})_+.\end{aligned}$$

In the second identity, we applied the fact that the entries of the vector $\boldsymbol{x}_j - \boldsymbol{u}_j$ are independent, centered random variables to see that there is no contribution from the off-diagonal terms of $\boldsymbol{B}$. The inequality relies on the variance bound $1/(4m^2)$ for each random variable X_{ij}. The function $(\cdot)_+ : a \mapsto \max\{a, 0\}$ returns the positive part of a number.

Schur's theorem [5, Thm. 4.3.26] states that eigenvalues of the symmetric matrix $\boldsymbol{B}$ majorize its diagonal entries. Since $(\cdot)_+$ is convex, the real-valued map $\boldsymbol{a} \mapsto \sum_{i=1}^{d}(a_i)_+$ on $\mathbb{R}^d$ respects the majorization relation [3, Thm. II.3.1]. Thus,

$$\sum_{i=1}^{d}(b_{ii})_+ \leq \sum_{i=1}^{d}(\lambda_i^{\downarrow}(\boldsymbol{B}))_+ = \sum_{i=1}^{s}\lambda_i^{\downarrow}(\boldsymbol{B}).$$

The equality relies on the assumption that the eigenvalues $\lambda_i^{\downarrow}(\boldsymbol{B})$ become negative at index $s+1$.

Merging the last two displays, we obtain the estimate

$$\mathbb{E}\,\mathrm{tr}((\boldsymbol{X}-\boldsymbol{U})^*\boldsymbol{B}(\boldsymbol{X}-\boldsymbol{U})) \leq \frac{d-s}{4m^2}\sum_{i=1}^{s}\lambda_i^{\downarrow}(\boldsymbol{B}). \tag{3.8}$$

This bound has exactly the form that we need.

Combining (3.6), (3.7), and (3.8), we arrive at the inequality

$$\sum_{i=s+1}^{d}\lambda_1^{\downarrow}(\boldsymbol{B}) \geq -\frac{d-s}{4m^2}\sum_{i=1}^{s}\lambda_i^{\downarrow}(\boldsymbol{B}).$$

In view of the representation (3.3) of the dual cone $K(c)^*$, the matrix $\boldsymbol{B} \in K(c)^*$ provided that

$$-\frac{d-s}{4m^2} \geq -\frac{1}{c}.$$

Rearranging this expression, we obtain the sufficient condition

$$m \geq \frac{1}{2}\sqrt{(d-s)\cdot c} \quad \text{implies} \quad \boldsymbol{B} \in K(c)^*.$$

For a general matrix $\boldsymbol{B} \in Z(m)^*$, we do not control the index s where the eigenvalues of $\boldsymbol{B}$ change sign, so we must insulate ourselves against the worst case, $s = 1$. This choice leads to the condition (3.2), and the proof is complete. ∎

4. Optimality. There are specific matrices where the size of the integers in the representation does not depend on the condition number. For instance, let $b \geq 1$, and consider the matrix

$$\boldsymbol{A} = \begin{bmatrix} b & 0 \\ 0 & 1 \end{bmatrix} = b \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}^* + 1 \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix}^*.$$

The condition number $\kappa(\boldsymbol{A}) = b$, which we can make arbitrarily large, but the integers in the representation never exceed one.

Nevertheless, we can show, by example, that the dependence of Theorem 1.1 on the condition number is optimal in dimension $d = 2$. For a number $b \geq 1$, consider the 2×2 matrix

$$\boldsymbol{A} = \begin{bmatrix} b^2+1 & b \\ b & 2 \end{bmatrix} = \begin{bmatrix} b \\ 1 \end{bmatrix} \begin{bmatrix} b \\ 1 \end{bmatrix}^* + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

From this representation, we quickly determine that the eigenvalues of $\boldsymbol{A}$ are 1 and $b^2 + 2$, so the condition number $\kappa(\boldsymbol{A}) = b^2 + 2$.

Suppose that we can represent the matrix $\boldsymbol{A}$ as a positive linear combination of outer products of vectors in $\mathbb{Z}_m^2$. We need at most $d(d+1)/2 = 3$ summands:

$$\boldsymbol{A} = \alpha \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^* + \beta \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}^* + \gamma \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}^*, \tag{4.1}$$

where $\alpha, \beta, \gamma > 0$ and $x_i, y_i, z_i \in \mathbb{Z}_m^1$. The equations in (4.1) associated with the top-left and bottom-right entries of $\boldsymbol{A}$ read as

$$b^2 + 1 = \alpha x_1^2 + \beta y_1^2 + \gamma z_1^2 \quad \text{and} \quad 2 = \alpha x_2^2 + \beta y_2^2 + \gamma z_2^2. \tag{4.2}$$

We consider three cases: (i) all three of x_2, y_2, z_2 are nonzero; (ii) exactly two of x_2, y_2, z_2 are nonzero; and (iii) exactly one of x_2, y_2, z_2 is nonzero.

Let us begin with case (i). Since x_2, y_2, and z_2 take nonzero integer values, the second equation in (4.2) ensures that

$$2 \geq (\alpha + \beta + \gamma) \min\left\{x_2^2, y_2^2, z_2^2\right\} \geq \alpha + \beta + \gamma.$$

Introducing this fact into the first equation in (4.2), we find that

$$b^2 + 1 \leq (\alpha + \beta + \gamma) \max\left\{x_1^2, y_1^2, z_1^2\right\} \leq 2 \max\left\{x_1^2, y_1^2, z_1^2\right\}.$$

We obtain a lower bound on the magnitude m of integers in a representation of $\boldsymbol{A}$ where x_2, y_2, z_2 are all nonzero:

$$m \geq \max\left\{ |x_1|, |y_1|, |z_1| \right\} \geq \frac{1}{\sqrt{2}}\sqrt{b^2+1} = \frac{1}{\sqrt{2}}\sqrt{\kappa(\boldsymbol{A})-1}. \tag{4.3}$$

Since the bound (4.3) is worse than the estimate in Theorem 1.1 for large b, we discover that the optimal integer representation of $\boldsymbol{A}$ has at least one zero among x_2, y_2, z_2.

Next, we turn to case (ii). By symmetry, we may assume that $z_2 = 0$. As before, the second equation in (4.2) shows that $\alpha+\beta \leq 2$. Meanwhile, the representation (4.1) implies that

$$\boldsymbol{A} - \gamma \begin{bmatrix} z_1^2 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} b^2+1-\gamma z_1^2 & b \\ b & 2 \end{bmatrix} \quad \text{is positive semidefinite.}$$

Since the determinant of a positive-semidefinite matrix is nonnegative, we find that $0 \leq 2(b^2+1-\gamma z_1^2) - b^2$. Equivalently, $\gamma z_1^2 \leq \frac{1}{2}(b^2+2)$. The first equation in (4.2) now delivers

$$b^2 + 1 = \alpha x_1^2 + \beta y_1^2 + \gamma z_1^2 \leq 2 \max\left\{x_1^2, y_1^2\right\} + \frac{1}{2}\left(b^2+2\right).$$

It follows that $\max\{x_1^2, y_1^2\} \geq b^2/4$. We obtain a lower bound on the magnitude m of the integers in a representation of $\boldsymbol{A}$ where two of x_2, y_2, z_2 are nonzero:

$$m \geq \max\left\{ |x_1|, |y_1| \right\} \geq \frac{1}{2} b = \frac{1}{2}\sqrt{\kappa(\boldsymbol{A})-2}. \tag{4.4}$$

In case (iii), a similar argument leads to the same lower bound for m.

Examining (4.4), we surmise that the bound from Theorem 1.1,

$$m \leq 1 + \frac{1}{2}\sqrt{(d-1)\cdot\kappa(\boldsymbol{A})},$$

on the magnitude m of integers in a representation of $\boldsymbol{A}$ cannot be improved when $d = 2$ and the condition number $\kappa(\boldsymbol{A})$ becomes large. Considering the $d \times d$ matrix

$$\begin{bmatrix} \boldsymbol{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d-2} \end{bmatrix},$$

an analogous argument proves that the dependence of Theorem 1.1 on the condition number is optimal in every dimension d.

Acknowledgments. The author wishes to thank Andrew Barbour for calling this problem to his attention and to Madeleine Udell for describing the application to analog-to-digital matrix multiplication. This research was undertaken at the Institute for Mathematical Sciences (IMS) at the National University of Singapore (NUS) during the workshop on New Directions in Stein's Method in May 2015.

REFERENCES

[1] A. D. Barbour, M. J. Luczak, and A. Xia, *Multivariate Approximation in Total Variation*, manuscript, 2015.
[2] A. Barvinok, *A Course in Convexity*, Grad. Stud. Math. 54, American Mathematical Society, Providence, RI, 2002.

[3] R. Bhatia, *Matrix Analysis*, Grad. Texts Math. 169, Springer-Verlag, New York, 1997.
[4] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.
[5] R. A. Horn and C. R. Johnson, *Matrix Analysis*, corr. reprint, Cambridge University Press, Cambridge, 1990.
[6] E. H. Lee, M. Udell, and S. S. Wong, *Factorization for analog-to-digital matrix multiplication*, in Proceedings of the 40th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Brisbane, 2015, IEEE, Piscataway, NJ, 2015, pp. 1061–1065.
[7] L. Mirsky, *On the trace of matrix products*, Math. Nachr., 20 (1959), pp. 171–174.
[8] H. Richter, *Zur Abschätzung von Matrizennormen*, Math. Nachr., 18 (1958), pp. 178–187.