Perturbation Analysis of the QR Factor R in the Context of LLL Lattice Basis Reduction

Xiao-Wen Chang, Damien Stehlé, Gilles Villard

PERTURBATION ANALYSIS OF THE QR FACTOR R IN THE CONTEXT OF LLL LATTICE BASIS REDUCTION

XIAO-WEN CHANG, DAMIEN STEHLÉ, AND GILLES VILLARD

Abstract. In 1982, Arjen Lenstra, Hendrik Lenstra Jr. and László Lovász introduced an efficiently computable notion of reduction of basis of a Euclidean lattice that is now commonly referred to as LLL-reduction. The precise definition involves the R-factor of the QR factorisation of the basis matrix. A natural mean of speeding up the LLL reduction algorithm is to use a (floating-point) approximation to the R-factor. In the present article, we investigate the accuracy of the factor R of the QR factorisation of an LLL-reduced basis. Our main contribution is the first fully rigorous perturbation analysis of the R-factor of LLL-reduced matrices under column-wise perturbations. Our results should be very useful to devise LLL-type algorithms relying on floating-point approximations.

1. Introduction

Let $B \in \mathbb{R}^{m\times n}$ be of a full column rank matrix. It has a unique QR factorization $B = QR$, where the Q-factor $Q \in \mathbb{R}^{m\times n}$ has orthonormal columns, i.e., $Q^TQ = I$ (where I is the identity matrix), and the R-factor $R \in \mathbb{R}^{n\times n}$ is upper triangular with positive diagonal entries (see, e.g., [6, §5]). This fundamental tool in matrix computations is central to the LLL reduction algorithm, named after the authors of [12], which aims at efficiently finding reduced bases of Euclidean lattices.

A Euclidean lattice L is a discrete subgroup of $\mathbb{R}^m$ and it can always be represented by a full column rank basis matrix $B \in \mathbb{R}^{m\times n}$: $L = \{B\mathbf{x}, \mathbf{x} \in \mathbb{Z}^n\}$. If $n \geq 2$, L has infinitely many bases. They are related by unimodular transforms, i.e., multiplication on the right of B by an $n \times n$ integer matrix with determinant ± 1. Given a lattice, one is often interested in obtaining a basis whose vectors are short and close to being orthogonal. Refining the quality of a basis is generically called lattice reduction. Among many others, lattice reduction has applications in cryptology [19], algorithmic number theory [4], communications [16], etc. LLL takes as input a basis matrix and returns a basis of the same lattice which is made of fairly short and orthogonal vectors (see Theorem 5.2). The properties satisfied by the output matrix involve its R-factor. The original LLL algorithm [12] used integer arithmetic for the operations on the basis and rational arithmetic for the operations on the R-factor. The bit-size of each rational (the bit-size of a/b with $a, b \in \mathbb{Z}$ is the sum of the bit-sizes of a and b) is bounded by a polynomial in the bit-sizes of the input matrix entries. Nevertheless, the cost of the rational arithmetic grows

2000 *Mathematics Subject Classification.* Primary 11H06, 65F25; Secondary 11Y99, 65F35.
Key words and phrases. lattice reduction, LLL, QR factorization, perturbation analysis.
Xiao-Wen Chang's work was supported by NSERC of Canada Grant RGPIN217191-07.
Damien Stehlé's work was partly funded by the LaRedA ANR project.
Gilles Villard's work was partly funded by the Gecko ANR project.

quickly and dominates the overall cost. Schnorr [21] was the first to use approximations of these rationals in a rigorous way. His algorithm was improved recently by Nguyen and Stehlé [17, 18] who significantly decreased the bit-size required for each approximation, and thus the overall complexity of the LLL-reduction. (Note that contrarily to [17, 18] Schnorr's approximations are not relying on standard floating-point arithmetic.) To further decrease the required precision and therefore the cost, Schnorr [11, 22, 23] suggested using the Householder QR factorization algorithm instead of the Cholesky factorization algorithm as was used in [17, 18], since it is known that the R-factor computed by Householder's algorithm is more accurate than the one computed with the Cholesky factorization of $B^T B$.

The R-factor of the matrix B varies continuously with B. If we consider a perturbed matrix $B + \Delta B$ that is sufficiently close to B (note that in the perturbation matrix ΔB, Δ does not represent anything, i.e., ΔB is not a product of Δ and B), then its R-factor $R + \Delta R$ remains close to R. The goal of the present article is to investigate how ΔB affects ΔR, for LLL-reduced matrices B. This perturbation analysis helps understanding and providing (a priori) guarantees on the quality of numerically computed factors R. The QR-factorization is typically computed by Householder reflections, Givens rotations or the modified Gram-Schmidt orthogonalization. These algorithms are backward stable with respect to the R-factor: if the computations are performed in floating-point arithmetic, then the computed $\widehat{R}$ is the true R-factor of a matrix $\widehat{B}$ which is very close to the input matrix B (see [7, §18]). Along with the backward stability analysis, a perturbation analysis provides accuracy bounds on the computed $\widehat{R}$. In the present paper, we consider a perturbation ΔB that satisfies

$$|\Delta B| \leq \varepsilon C |B|, \tag{1.1}$$

where $c_{i,j} = 1$ for all i, j and $\varepsilon > 0$ is a small scalar (it will be specified in the relevant theorems to be given in the paper how small it needs to be for the results to hold). The motivation for considering such a class of perturbations is that the backward rounding error from a rounding error analysis of the standard QR factorization algorithms fits in this class with $\varepsilon = O(u)$, where we omitted the dependence with respect to the matrix dimensions and u is the unit roundoff (see [7, Th. 18.4] and Theorem 6.4 given later).[1]

Our results. Our main contribution is the first fully rigorous perturbation analysis of the R-factor of LLL-reduced matrices under the perturbation (1.1) (Theorem 5.6). In order to make this result consistent with the LLL-reduction (i.e., the perturbed reduced basis somewhat remains reduced), we introduce *a new notion of LLL-reduction* (Definition 5.3). Matrices reduced in this new sense satisfy essentially the same properties as those satisfied by matrices reduced in the classical sense. But the new notion of reduction is more natural with respect to column-wise perturbations, as the perturbation of a reduced basis remains reduced (this is not the case with the classical notion of reduction). Another important ingredient of the main result, that may be of independent interest, is the improvement of the perturbation analyses of [1] and [27] for general full column rank matrices

[1]Note that the description of the backward error in [7, Th. 18.4] was modified in the newer edition [8, Th. 19.4]. In the latter, the matrix equation (1.1) is replaced by $\|\Delta \mathbf{b}_i\| \leq \varepsilon \|\mathbf{b}_i\|$, for all i. The two formulations are equivalent (up to a small factor that is polynomial in the matrix dimensions), but the matrix equation (1.1) is more suited for our sensitivity analysis.

(section 2). More precisely, all our bounds are fully rigorous, in the sense that no higher order error term is neglicted, and explicit constant factors are provided. Explicit and rigorous bounds are invaluable for *guaranteeing computational accuracy*: one can choose a precision that will be known in advance to provide a certain degree of accuracy in the result. In [1, §6], a rigorous error bound was proved. An improved bound was given in [1, §8], but it is a first-order bound, i.e., high-order terms were neglected. Our rigorous bound is close to this improved bound. Our approach to deriving this rigorous bound is new and has been extended to the perturbation analysis of some other important matrix factorizations [3]. Finally, we give explicit constants in the backward stability analysis of Householder's algorithm from [8, §19], which, along with the perturbation analysis, provides fully rigorous and explicit error bounds for the computed R-factor of a LLL-reduced matrix.

Implications. Our results are descriptive in nature. However, the rigorous and explicit error analysis and the new notion of LLL-reducedness should lead to significant algorithmic improvements. Intuitively, we formalize the idea that only the $O(n)$ most significant bits of the vectors matter for their LLL-reducedness. Such a property has dramatic algorithmic consequences, as it implies that instead of computing with all bits we shall try to make use of only $O(n)$ bits for each matrix entry. For instance, in a context similar to [26], our result implies that in order to check the LLL-reducedness of a matrix, one only needs to consider $O(n)$ most significant bits of each column. This provides a $O(n^5)$-time (resp. $O(n^{4+\varepsilon})$-time) LLL certificate with naive integer arithmetic (with FFT-based arithmetic [25]). Also, our results have been used to devise an efficient algorithm that improves the LLL-reducedness of an already LLL-reduced basis [15]. That algorithm finds a good unimodular transform by looking only at the $O(n)$ most significant bits of each column of the input matrix. Furthermore, the present work is the first step towards achieving Schnorr's goal of an LLL algorithm relying on the floating-point Householder algorithm. This goal has been reached in [14], which relies on the present results. Finally, another possible application could be a quasi-linear time LLL-reduction algorithm in fixed dimension, in the fashion of the Knuth-Schönhage quasi-linear time gcd algorithm [10, 24]. Roughly speaking, the first k bits of the quotients sequence of Euclid's gcd algorithm depends only on the first $2k$ bits of the two input integers. Knuth and Shönhage use that property to compute the quotients sequence by looking only at the first bits of the remainders sequence. Our results could help devising such an algorithm in the context of lattice reduction.

Road map. In section 2, we give our perturbation analysis of the R-factor for general full column matrices. Sections 3, 4 and 5 specialize the analysis to different sets of matrices, including LLL-reduced matrices. Finally, in section 6, we provide explicit backward error bounds for Householder's QR factorization algorithm.

Notation. If $\mathbf{b}$ is a vector, then $\|\mathbf{b}\|_p$ denotes its ℓ_p norm. If $p = 2$, we omit the subscript. The jth column of a matrix $A = (a_{i,j})$ is denoted by $\mathbf{a}_j$ and $|A|$ denotes $(|a_{i,j}|)$. We use the MATLAB notation to denote submatrices: The matrix $A(i_1 : i_2, j_1 : j_2)$ consists of rows i_1 to i_2 and columns j_1 to j_2 of A; If i_1 and i_2 (resp. j_1 and j_2) are omitted, then all the rows (resp. columns) of A are kept; Finally, if $i_1 = i_2$ (resp. $j_1 = j_2$), we will write $A(i_1, j_1 : j_2)$ (resp. $A(i_1 : i_2, j_1)$). The Frobenius norm is $\|A\|_F = (\sum_{i,j} a_{i,j}^2)^{1/2}$. The ℓ_p matrix norm is $\|A\|_p = \sup_{\mathbf{x}\in\mathbb{R}^n} \|A\mathbf{x}\|_p/\|\mathbf{x}\|_p$. We use $\|A\|_{1,\infty}$ to denote either the 1-norm or the ∞-norm. We have $\|A\|_2 \leq \|A\|_F$. If A and B are of compatible sizes, then $\|AB\|_F \leq$

$\|A\|_F\|B\|_2$ (see [8, Pbm. 6.5]) and $\|AB\|_2 \leq \|A\|_2\|B\|_2$. If A is a square matrix, then $\mathrm{up}(A)$ denotes the upper triangular matrix whose ith diagonal entry is $a_{i,i}/2$ and whose upper-diagonal entries match those of A. We let $\mathcal{D}_n \subseteq \mathbb{R}^{n\times n}$ be the set of diagonal matrices with positive diagonal entries. For any nonsingular matrix X we define

$$\mathrm{cond}_2(X) = \left\| |X||X^{-1}| \right\|_2 . \tag{1.2}$$

If a is a real number, then $fl(a)$ denotes the floating-point number closest to a (with even mantissa when a is exactly half-way from two consecutive floating-point numbers). As a side-effect of our bounds being fully explicit, and since we tried not to over-estimate them, some of them involve rather complicated and uninteresting terms. To make the presentation more compact, we encapsulate them in the variables $c_1, c_2, \ldots$. They are of much lower orders of magnitudes than the terms they are used with.

2. Refined Perturbation Analysis of the R Factor

In this section, we first give a general matrix-norm perturbation bound, then derive a column-wise perturbation bound.

2.1. **A matrix-norm perturbation bound.**

We will present a rigorous bound (i.e., without any implicit higher order term) on the perturbation of the R-factor when B is under the perturbation (1.1). In order to do that, we need the following two technical lemmas.

Lemma 2.1. *Let $n > 0$, $X \in \mathbb{R}^{n\times n}$ and $D = \mathrm{diag}(\delta_1, \ldots, \delta_n) \in \mathcal{D}_n$. We define $\zeta_D = 1$ for $n = 1$ and, for $n \geq 2$:*

$$\zeta_D = \sqrt{1 + \max_{1\leq i<j\leq n} (\delta_j/\delta_i)^2}. \tag{2.1}$$

Then we have

$$\|\mathrm{up}(X) + D^{-1}\mathrm{up}(X^T)D\|_F \leq \zeta_D\|X\|_F, \tag{2.2}$$

and in particular, when $X^T = X$ and $D = I$,

$$\|\mathrm{up}(X)\|_F \leq \frac{1}{\sqrt{2}}\|X\|_F. \tag{2.3}$$

Proof. The inequality (2.2) was given in [2, Lemma 5.1]. The inequality (2.3), which was given in [2, Eq. (2.3)], can also be derived from (2.2). □

The following provides a sufficient condition for the rank to be preserved during a continuous change from a full column-rank matrix B to $B + \Delta B$. This ensures that the R-factor is well-defined on the full path. This is of course not true if the matrix B is close to being rank deficient and the perturbation ΔB is not small, but that situation is prevented by assumption (2.4).

Lemma 2.2. *Let $B \in \mathbb{R}^{m\times n}$ be of full column rank with QR factorization $B = QR$. Let the perturbation matrix $\Delta B \in \mathbb{R}^{m\times n}$ satisfy (1.1). If*

$$\mathrm{cond}_2(R)\varepsilon < \frac{c}{m\sqrt{n}}, \tag{2.4}$$

for some constant $0 < c \leq 1$, then the matrix $B + t\Delta B$ has full column rank for any $|t| \leq 1$. Furthermore, $\|\Delta B R^{-1}\|_F < c$.

Proof. The second assertion follows from (2.4). In fact, from (1.1) and (2.4), we obtain

$$\begin{aligned}\|\Delta B R^{-1}\|_F &\le \varepsilon \big\| C|Q||R||R^{-1}| \big\|_F \le \varepsilon \big\| C \big\|_F \big\| Q \big\|_F \big\| |R||R^{-1}| \big\|_2 \\ &= \varepsilon m\sqrt{n}\ \mathrm{cond}_2(R) < c.\end{aligned}$$

We now consider the first assertion. Notice that

$$Q^T(B + t\Delta B) = R + tQ^T\Delta B = \left(I + tQ^T\Delta B R^{-1}\right) R.$$

But $\|tQ^T\Delta B R^{-1}\|_2 \le \|\Delta B R^{-1}\|_2 < 1$, thus $I + tQ^T\Delta B R^{-1}$ is non-singular. So is $Q^T(B + t\Delta B)$, and hence $B + t\Delta B$ must have full column rank. □

Using the above two lemmas, we can prove the following perturbation theorem.

Theorem 2.3. *Let $B \in \mathbb{R}^{m\times n}$ be of full column rank with QR factorization $B = QR$. Let the perturbation matrix $\Delta B \in \mathbb{R}^{m\times n}$ satisfy (1.1). If*

$$\mathrm{cond}_2(R)\varepsilon < \frac{\sqrt{3/2} - 1}{m\sqrt{n}}, \tag{2.5}$$

then $B + \Delta B$ has a unique QR factorization

$$B + \Delta B = (Q + \Delta Q)(R + \Delta R), \tag{2.6}$$

and

$$\frac{\|\Delta R\|_F}{\|R\|_2} \le c_1(m,n)\varkappa(B)\,\varepsilon. \tag{2.7}$$

where, with ζ_D defined in (2.1)

$$c_1(m,n) = (\sqrt{6} + \sqrt{3})mn^{1/2}, \tag{2.8}$$

$$\varkappa(B) = \inf_{D\in\mathcal{D}_n} \varkappa(R,D), \quad \varkappa(R,D) = \frac{\zeta_D \big\| |R||R^{-1}|D \big\|_2 \big\| D^{-1}R \big\|_2}{\|R\|_2}. \tag{2.9}$$

Proof. The condition (2.5) ensures that (2.4) holds with $c = \sqrt{3/2} - 1$. Then, by Lemma 2.2, $B + t\Delta B$ is of full column rank for any $|t| \le 1$. Thus $B + t\Delta B$ has the unique QR factorization

$$B + t\Delta B = (Q + \Delta Q(t))(R + \Delta R(t)), \tag{2.10}$$

which, with $\Delta Q(1) = \Delta Q$ and $\Delta R(1) = \Delta R$, gives (2.6).

From (2.10), we obtain $(B + t\Delta B)^T(B + t\Delta B) = (R + \Delta R(t))^T(R + \Delta R(t))$, leading to

$$R^T\Delta R(t) + \Delta R(t)^T R = tR^TQ^T\Delta B + t\Delta B^TQR + t^2\Delta B^T\Delta B - \Delta R(t)^T\Delta R(t).$$

Multiplying the above by R^{-T} from the left and R^{-1} from the right, we obtain

$$\begin{aligned}&R^{-T}\Delta R(t)^T + \Delta R(t)R^{-1} \\ &= tQ^T\Delta B R^{-1} + tR^{-T}\Delta B^TQ + R^{-T}\left(t^2\Delta B^T\Delta B - \Delta R(t)^T\Delta R(t)\right)R^{-1}.\end{aligned}$$

Since $\Delta R(t)R^{-1}$ is upper triangular, it follows that

$$\begin{aligned}\Delta R(t)R^{-1} = \ &\mathrm{up}(tQ^T\Delta B R^{-1} + tR^{-T}\Delta B^TQ) \\ &+ \mathrm{up}(t^2R^{-T}\Delta B^T\Delta B R^{-1}) - \mathrm{up}\big[R^{-T}\Delta R(t)^T\Delta R(t)R^{-1}\big].\end{aligned} \tag{2.11}$$

Taking the F-norm on both sides of (2.11) and using Lemma 2.1 and the orthogonality of Q, we obtain

$$\|\Delta R(t)R^{-1}\|_F \le \sqrt{2}|t|\cdot\|\Delta BR^{-1}\|_F + \frac{1}{\sqrt{2}}t^2\|\Delta BR^{-1}\|_F^2 + \frac{1}{\sqrt{2}}\|\Delta R(t)R^{-1}\|_F^2. \tag{2.12}$$

Let $\rho(t) = \|\Delta R(t)R^{-1}\|_F$ and $\delta(t) = |t|\cdot\|\Delta BR^{-1}\|_F$. Then from (2.12)

$$\rho(t)(\sqrt{2} - \rho(t)) \le \delta(t)(2 + \delta(t)).$$

Here the left hand side has its maximum of $1/2$ with $\rho(t) = 1/\sqrt{2}$ and is increasing with respect to $\rho(t) \in [0, 1/\sqrt{2}]$. But, by Lemma 2.2, for $|t| \le 1$,

$$0 \le \delta(t) \le \|\Delta BR^{-1}\|_F < c = \sqrt{3/2} - 1. \tag{2.13}$$

This implies that $0 \le \delta(t)(2 + \delta(t)) < 1/2$ and $\rho(t)$, starting from 0, cannot reach its maximum. Hence $\rho(t) < 1/\sqrt{2}$ for any $|t| \le 1$. In particular, when $t = 1$,

$$\|\Delta RR^{-1}\|_F < 1/\sqrt{2}. \tag{2.14}$$

For any matrices $X \in \mathbb{R}^{n\times n}$ and $D \in \mathcal{D}_n$, we have $\mathrm{up}(XD) = \mathrm{up}(X)D$. Thus from (2.11) with $t = 1$ it follows that

$$\begin{aligned}\Delta RR^{-1}D = {} & \mathrm{up}\left[(Q^T\Delta BR^{-1}D) + D^{-1}(DR^{-T}\Delta B^TQ)D\right] \\ & + \mathrm{up}(R^{-T}\Delta B^T\Delta BR^{-1}D) - \mathrm{up}(R^{-T}\Delta R^T\Delta RR^{-1}D).\end{aligned} \tag{2.15}$$

Then, using Lemma 2.1, the inequality $\|\mathrm{up}(X)\|_F \le \|X\|_F$ for any $X \in \mathbb{R}^{n\times n}$ and the orthogonality of Q, we obtain from (2.15) that

$$\begin{aligned}\|\Delta RR^{-1}D\|_F \le {} & \zeta_D\|\Delta BR^{-1}D\|_F + \|\Delta BR^{-1}\|_F\|\Delta BR^{-1}D\|_F \\ & + \|\Delta RR^{-1}\|_F\|\Delta RR^{-1}D\|_F.\end{aligned}$$

Therefore, with (1.1), (2.13) and (2.14), we have

$$\begin{aligned}\|\Delta RR^{-1}D\|_F &\le \frac{\zeta_D + \sqrt{3/2} - 1}{1 - 1/\sqrt{2}}\|C\|_F\|Q\|_F\left\||R||R^{-1}|D\right\|_2 \\ &\le (\sqrt{6} + \sqrt{3})\zeta_D mn^{1/2}\left\||R||R^{-1}|D\right\|_2,\end{aligned}$$

where in deriving the second inequality we used the fact that $\zeta_D \ge 1$. Therefore,

$$\begin{aligned}\|\Delta R\|_F = \|\Delta RR^{-1}DD^{-1}R\|_F &\le \|\Delta RR^{-1}D\|_F\|D^{-1}R\|_2 \\ &\le (\sqrt{6} + \sqrt{3})\zeta_D mn^{1/2}\left\||R||R^{-1}|D\right\|_2\|D^{-1}R\|_2,\end{aligned}$$

leading to the bound (2.7). □

Remark 2.4. Theorem 2.3 is a rigorous version of a first-order perturbation bound given in [1, §8], which also involves $\varkappa(B)$. The new bound given here shows that if (2.4) holds then the high-order terms ignored in [1, §8] are indeed negligible. Numerical tests given in [1, §9] indicated that the first-order bound is a good approximation to the relative perturbation error in the R-factor. This suggests that the rigorous bound (2.7) is a good bound. By taking $D = I$ in (2.9), we obtain $\varkappa(B) \le \sqrt{2}\mathrm{cond}_2(R)$. The quantity $\mathrm{cond}_2(R)$ is involved in the rigorous perturbation bound obtained in [1, §6] and can be arbitrarily larger than $\varkappa(B)$.

Remark 2.5. If the assumptions of Theorem 2.3 hold for B with perturbation ΔB, then they also hold for BS, for any arbitrary column scaling $S \in \mathcal{D}_n$, with perturbation ΔBS. The new R-factor is RS and the corresponding error is ΔRS. However, the quantity $\varkappa(B)$ is not preserved under column scaling.

2.2. **A column-wise perturbation bound.** For $j = 1, \ldots, n$, we define $R_j = R(1\!:\!j, 1\!:\!j)$, $\Delta R_j = \Delta R(1\!:\!j, 1\!:\!j)$, $\mathbf{r}_j = R(1\!:\!j, j)$ and $\Delta\mathbf{r}_j = \Delta R(1\!:\!j, j)$. Using Zha's approach [27, Cor. 2.2], we derive the following result.

Corollary 2.1. *If the assumptions of Theorem 2.3 hold, then for $j = 1, \ldots, n$,*

$$\frac{\|\Delta\mathbf{r}_j\|}{\|\mathbf{r}_j\|} \leq c_1(m, j)\varkappa(B, j)\varepsilon, \tag{2.16}$$

where

$$\varkappa(B, j) = \inf_{D \in \mathcal{D}_j} \varkappa(R, D, j) \geq 1, \quad \varkappa(R, D, j) = \frac{\zeta_D \big\||R_j||R_j^{-1}|D\big\|_2 \|D^{-1}\mathbf{r}_j\|}{\|\mathbf{r}_j\|}. \tag{2.17}$$

Proof. For any $j \leq n$, we define $B_j = B(:, 1\!:\!j)$ and $\Delta B_j = \Delta B(:, 1\!:\!j)$. Note that $|\Delta B_j| \leq \varepsilon C|B_j|$ and $\mathrm{cond}_2(R_j)\varepsilon \leq \mathrm{cond}_2(R)\varepsilon \leq (\sqrt{3/2} - 1)/(m\sqrt{n})$. Thanks to Remark 2.5, we can apply Theorem 2.3 to B_jS for an arbitrary $S \in \mathcal{D}_j$ with the perturbation matrix ΔB_jS. Therefore, for any $D \in \mathcal{D}_j$,

$$\|\Delta R_jS\|_F \leq c_1(m, j)\zeta_D \big\||R_j||R_j^{-1}|D\big\|_2 \big\|D^{-1}R_jS\big\|_2 \varepsilon.$$

Now, let the jth diagonal entry of S be 1 and the others tend to zero. Taking the limit provides (2.16). The lower bound on $\varkappa(B, j)$ in (2.17) follows from $\zeta_D \geq 1$ and

$$\big\||R_j||R_j^{-1}|D\big\|_2 \|D^{-1}\mathbf{r}_j\| \geq \big\||R_j||R_j^{-1}|DD^{-1}|\mathbf{r}_j|\big\| \geq \big\||R_jR_j^{-1}\mathbf{r}_j|\big\| = \|\mathbf{r}_j\|.$$

□

Remark 2.6. The quantity $\varkappa(B, j)$ can be interpreted as an upper bound on the condition number of the jth column of R with respect to the perturbation ΔB of B. It is easy to check that the lower bound 1 on $\varkappa(B, j)$ in (2.17) is reached when $j = 1$, i.e., that $\varkappa(B, 1) = 1$.

In the following sections, we specialize Theorem 2.3 and Corollary 2.1 to several different classes of matrices, that are naturally linked to the LLL reduction.

3. Perturbation Analysis for Size-Reduced Matrices

We now study $\varkappa(B, j)$ for the class of size-reduced matrices, defined as follows.

Definition 3.1. Let $\eta \geq 0$. A full column-rank matrix $B \in \mathbb{R}^{m\times n}$ with R-factor R is η-size reduced if for any $1 \leq i < j \leq n$, we have $|r_{i,j}| \leq \eta \cdot r_{i,i}$.

A matrix is 1-size-reduced if the largest element in magnitude in each row of the R-factor is reached on the diagonal. An example is the QR factorization with standard column pivoting (see, e.g., [6, Sec. 5.4.1]): one permutes the columns of the considered matrix so that for any $j \leq n$, the jth column is the one maximising $r_{j,j}$ among the last $n - j + 1$ columns. If column pivoting is used, then the sorted matrix is 1-size-reduced. The LLL algorithm [12] has a sub-routine usually called size-reduction which aims at computing a 1/2-size-reduced matrix by multiplying the initial matrix on the right by an integer matrix whose determinant is equal to 1

or -1. In the L^2 algorithm from [18], a similar sub-routine, relying on floating-point arithmetic, aims at computing an η-size-reduced matrix, for any specified $\eta > 1/2$.

In subsection 3.1, we establish an upper bound on $\varkappa(B, j)$. That upper bound corresponds to a particular choice of scaling D in $\varkappa(R, D, j)$. In subsection 3.2, we compare our particular scaling with the different scalings discussed in [1, §9]. We then give a geometric interpretation of the result we obtain in subsection 3.3.

3.1. **Perturbation bounds for size-reduced matrices.** We first propose a way of selecting a good diagonal matrix D in (2.9) and in (2.17) to bound $\varkappa(B)$ and $\varkappa(B, j)$, respectively. Combined with Theorem 2.3 and Corollary 2.1, this directly provides matrix-norm and column-wise perturbation bounds.

Theorem 3.2. *Let $B \in \mathbb{R}^{m\times n}$ with full column rank be η-size-reduced and let R be its R-factor. For $j = 1, \ldots, n$, we define $r'_{j,j} = r_{j,j}/\max_{1\le k\le j} r_{k,k}$ and $D'_j = \mathrm{diag}(r'_{1,1}, \ldots, r'_{j,j})$. Then*

$$\varkappa(B) \le 2(1 + (n-1)\eta)(1+\eta)^{n-1}\zeta_{D'_n}, \tag{3.1}$$

$$\varkappa(B, j) \le c_2(j, \eta)(1+\eta)^j \zeta_{D'_j}\big(\max_{1\le k\le j} r_{k,k}\big)/\|\mathbf{r}_j\|, \quad j = 1, \ldots, n, \tag{3.2}$$

where ζ_D is defined in (2.1) *for any arbitrary positive diagonal matrix D, and*

$$c_2(j, \eta) = 2\sqrt{1 + (j-1)\eta^2}/(1+\eta). \tag{3.3}$$

Proof. Let R'_j be obtained from R_j by dividing the kth column by $\max_{1\le i\le k} r_{i,i}$, for $k = 1, \ldots, j$. The diagonal entries of R'_j match $r'_{i,i}$'s from D'_j. Since R_j is η-size-reduced, so is R'_j. Let $T_j = {D'_j}^{-1}R'_j$. We have $t_{i,i} = 1$ and $t_{i,k} \le \eta$ for $k > i$. Therefore, we have $|T_j^{-1}| \le U_j^{-1}$, where $U_j \in \mathbb{R}^{j\times j}$ is upper triangular with $u_{i,i} = 1$ and $u_{i,k} = -\eta$ for $k > i$, see, e.g., [8, Th. 8.12]. Since $V_j = U_j^{-1}$ satisfies $v_{i,i} = 1$ and $v_{i,k} = \eta(1+\eta)^{k-i-1}$ for $k > i$ (see, e.g., [8, Eq. (8.4)]), we obtain

$$\begin{aligned} |R_j||R_j^{-1}|D'_j &= |R'_j||{R'_j}^{-1}|D'_j = D'_j|T_j||T_j^{-1}| \\ &\le D'_j|U_j||V_j| = D'_j \begin{bmatrix} 1 & 2\eta & 2\eta(1+\eta) & \cdot & 2\eta(1+\eta)^{j-2} \\ & 1 & 2\eta & \cdot & 2\eta(1+\eta)^{j-3} \\ & & \cdot & \cdot & \cdot \\ & & & 1 & 2\eta \\ & & & & 1 \end{bmatrix}. \end{aligned}$$

Since $|r'_{i,i}| \le 1$ for any i, we have

$$\big\||R_j||R_j^{-1}|D'_j\big\|_{1,\infty} \le \big(1 + 2\eta\sum_{k=0}^{j-2}(1+\eta)^k\big) \le 2(1+\eta)^{j-1}. \tag{3.4}$$

Notice that $|r_{p,q}|/r'_{p,p} = |r_{p,q}|\max_{1\le k\le p} r_{k,k}/r_{p,p} \le \eta\max_{1\le k\le p} r_{k,k}$. It follows that $|{D'_j}^{-1}R_j| \le (\max_{1\le k\le j} r_{k,k})|U_j|$. Therefore,

$$\|{D'_j}^{-1}R_j\|_{1,\infty} \le (1 + (j-1)\eta)\max_{1\le k\le j} r_{k,k}, \quad \big\|{D'_j}^{-1}\mathbf{r}_j\big\| \le \sqrt{1 + (j-1)\eta^2}\max_{1\le k\le j} r_{k,k}.$$

Then from the above and (3.4), and using the fact that $\|S\|_2 \le (\|S\|_1\|S\|_\infty)^{1/2}$ for any matrix S (see, e.g., [8, Eq. (6.19)]), we obtain

$$\frac{\big\||R||R^{-1}|D'_n\big\|_2\|{D'_n}^{-1}R\|_2}{\|R\|_2} \le 2(1+(n-1)\eta)(1+\eta)^{n-1},$$
$$\frac{\big\||R_j||R_j^{-1}|D'_j\big\|_2\big\|{D'_j}^{-1}\mathbf{r}_j\big\|}{\|\mathbf{r}_j\|} \le \frac{2\sqrt{1+(j-1)\eta^2}(1+\eta)^{j-1}\max_{1\le k\le j} r_{k,k}}{\|\mathbf{r}_j\|}.$$

Thus from (2.9) and (2.17) we conclude that (3.1) and (3.2) hold, respectively. □

Remark 3.3. Suppose we use the standard column pivoting strategy in computing the QR factorization of B. Then $r_{i,i} \ge r_{k,k}$ for $i < k \le j$, implying that $\zeta_{D'_j} \le \sqrt{2}$. Then, if P is the pivoting permutation matrix

$$\varkappa(BP) \le \sqrt{2}n2^n \quad \text{and} \quad \varkappa(BP, j) \le \sqrt{2j}2^j r_{1,1}/\|\mathbf{r}_j\|.$$

A similar bound on $\varkappa(BP)$ was given in [1, Th. 8.2].

3.2. **Choosing the row scaling in $\varkappa(R, D)$.** In [1, §9], Chang and Paige suggest different ways of choosing D in $\varkappa(R, D)$ to approximate $\varkappa(B)$. One way is to choose $D_r := \mathrm{diag}(\|R(i,:)\|)$ and $D = I$ and take $\min\{\varkappa(R, D_r), \varkappa(R, I)\}$ as an approximation to $\varkappa(B)$. The other way is to choose $D = D_e$ (see below for the definition of D_e) and use $\varkappa(R, D_e)$ as an approximation to $\varkappa(B)$.

The following matrix shows that the scaling D' from Theorem 3.2 can provide a much better approximation to $\varkappa(B)$ than $\min(\varkappa(R, D_r), \varkappa(R, I))$. Let

$$B = R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \gamma & \eta\gamma \\ 0 & 0 & 1/\gamma \end{bmatrix}.$$

When γ goes to infinity, both $\varkappa(R, D_r)$ and $\varkappa(R, I)$ tend to infinity, whereas $\varkappa(R, D')$ remains bounded. This also indicates that $\min\{\varkappa(R, D_r), \varkappa(R, I)\}$ can be significantly larger than $\varkappa(B)$.

The scaling D_e is constructed from $D_c R^{-1}$ with $D_c = \mathrm{diag}(\|\mathbf{r}_i\|_1)$. If we assume that B is a generic η-size-reduced matrix (or, more formally, that each $r_{i,j}$ is uniformly and independently distributed in $[-\eta \cdot r_{i,i}, \eta \cdot r_{i,i}]$), then with high probability D_c is the same as $\mathrm{diag}(\max_{1\le k\le i} r_{k,k})$, up to a polynomial factor in n. We have $D_c D^{-1} \le D_c|R^{-1}| \le D_c|V|D^{-1}$, where V is as in the proof of Theorem 3.2 and $D = \mathrm{diag}(r_{i,i})$. This implies that up to a factor exponential in n, $\|(D_c R^{-1})(:, i)\|$ is $1/r'_{i,i}$. The diagonal matrix D_e is defined by $D_e(i,i) = \min_{1\le k\le i} 1/\|(D_c R^{-1})(:,k)\|^2$. Up to factors exponential in n and for generic η-size-reduced matrices, the scaling D_e can be equivalently defined by $D_e(i,i) = \min_{1\le k\le i} r'_{k,k}$. A bound similar to the one of Theorem 3.2 can be derived for the latter scaling. Nevertheless, if R is diagonal, then $D_e = I$ and $\varkappa(R, D_e) = \sqrt{2}$, but $\varkappa(R, D')$ can be significantly larger. Finally, one may note that it is not known how to compute D_e from R in $O(n^2)$ arithmetic operations or less, while computing D' requires only $O(n)$ arithmetic operations.

[2]The description of D_e in [1, §9] has an unintended error.

3.3. **Geometric interpretation of Theorem 3.2.** It is easy to verify that

$$\max_{1\le k\le i\le j}(r'_{i,i}/r'_{k,k}) \le \zeta_{D'_j} \le \sqrt{2}\max_{1\le k\le i\le j}(r'_{i,i}/r'_{k,k}).$$

When $(\max_{1\le i\le j} r_{i,i})/\|\mathbf{r}_j\| = O(1)$, e.g., for a generic η-size-reduced matrix with $|r_{i,j}|$ expected to be somewhat proportional to $r_{i,i}$, we see from (3.2) that the quantity $\max_{1\le k\le i\le j}(r'_{i,i}/r'_{k,k})$ bounds (up to a multiplicative factor that depends only on j) the sensitivity of the jth column of the R-factor. Let $x \mapsto r(x)$ be the piecewise affine interpolating function defined on $[1,n]$ such that $r(j) = r_{j,j}$ for $j = 1,\ldots,n$. For x_1 and x_2 in $[1,n]$ such that $r(x_1) = r(x_2)$, we consider the quantity $\max_{x\in[x_1,x_2]} r(x_1)/r(x) = \max_{x\in[x_1,x_2]} r(x_2)/r(x)$, which, as illustrated by Figure 1, represents the multiplicative depth of the graph of r between x_1 and x_2.

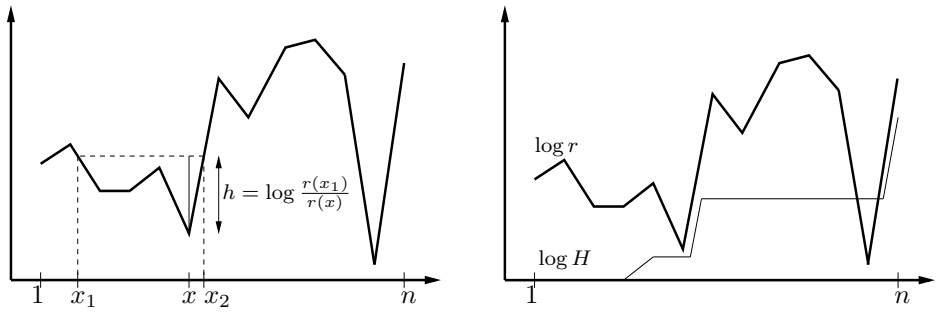

FIGURE 1. A possible graph of $\log r$: on the left hand side, with a depth h between x_1 and x_2 (the multiplicative depth is $\exp(h)$); on the right hand side, with the additive height function $\log H$.

We define the maximum depth before $r_{j,j}$ as:

$$H_j = \max_{1\le x_1\le x_2\le j,\ r(x_1)=r(x_2)} \left(\max_{x\in[x_1,x_2]} \frac{r(x_1)}{r(x)} \right),$$

which is illustrated on the right hand side of Figure 1. We now show the equivalence between $\zeta_{D'_j}$ and H_j. Without loss of generality, we consider only H_n.

Lemma 3.4. *We have* $H_n = \max_{1\le i\le j\le n}(r'_{j,j}/r'_{i,i})$.

Proof. We first prove that for any i and j such that $1 \le i \le j \le n$, $H_n \ge r'_{j,j}/r'_{i,i}$. We distinguish two cases, depending on the smallest index k_0 at which $\max_{1\le k\le j} r_{k,k}$ is reached. If $k_0 \le i$, then $r'_{j,j}/r'_{i,i} = r_{j,j}/r_{i,i}$. If $r_{j,j} \le r_{i,i}$, the result holds since $H_n \ge 1$; otherwise, we have $r_{j,j} > r_{i,i}$, leading to $H_n \ge r_{j,j}/r_{i,i}$ (in the definition of H_n, consider $x = i$, $x_2 = j$ and $x_1 \in [k_0, i]$ such that $r(x_1) = r(x_2)$). Suppose now that $i < k_0$. Since $r'_{j,j} \le 1$, we have $r'_{j,j}/r'_{i,i} \le \max_{1\le k\le i} r_{k,k}/r_{i,i}$. The latter is not greater than H_n (in the definition of H_n, consider $x = i$, $x_1 \le i$ such that $r(x_1) = \max_{1\le k\le i} r_{k,k}$ and $x_2 \in [i, k_0]$ such that $r(x_2) = r(x_1)$).

We now prove that $\max_{1\le i\le j\le n}(r'_{j,j}/r'_{i,i}) \ge H_n$. Let $x_1 \le x \le x_2$ in $[1,n]$ be such that $H_n = r(x_1)/r(x) = r(x_2)/r(x)$. We suppose that $x_1 < x < x_2$ as otherwise $H_n = 1 \le \max_{1\le i\le j\le n}(r'_{j,j}/r'_{i,i})$. By the definition of $r(\cdot)$, the real x must be an integer. Similarly, either x_1 or x_2 is an integer. We consider these two cases separately. Suppose first that $x_1 \in \mathbb{Z}$. Then $r(x_1) \le r(\lceil x_2 \rceil)$. We must have $\max_{1\le k\le \lfloor x_2 \rfloor} r_{k,k} = r_{x_1,x_1}$ and $\max_{1\le k\le \lceil x_2 \rceil} r_{k,k} = r_{\lceil x_2 \rceil, \lceil x_2 \rceil}$. This gives

that $r'_{x,x} = r_{x,x}/r_{x_1,x_1}$ and $r'_{\lceil x_2 \rceil, \lceil x_2 \rceil} = 1$. Thus $H_n = r_{x_1,x_1}/r_{x,x} = r'_{\lceil x_2 \rceil, \lceil x_2 \rceil}/r'_{x,x}$. Suppose now that $x_2 \in \mathbb{Z}$. Then $r(\lfloor x_1 \rfloor) \geq r(x_2)$. Since $\max_{1 \leq k \leq x_2} r_{k,k}$ is reached before x_1, we have $r'_{x_2,x_2}/r'_{x,x} = r_{x_2,x_2}/r_{x,x} = H_n$. □

If $(\max_{i \leq j} r_{i,i})/\|\mathbf{r}_j\| = O(1)$, then from Corollary 2.1 and Theorem 3.2 it follows that $\|\Delta \mathbf{r}_j\|/\|\mathbf{r}_j\| \lesssim (1+\eta)^j H_j \varepsilon$.

4. Perturbation Analysis for Weakly-Size-Reduced Matrices

The perturbation bounds given in Theorem 3.2 does not indicate that size-reducedness, as defined in section 3, is preserved after η-size-reduced B is perturbed to $B + \Delta B$. In fact, from (2.16) and (3.2),

$$|\Delta r_{i,j}| \leq \|\Delta \mathbf{r}_j\| \leq c_1(m,j) c_2(j,\eta) (1+\eta)^j \zeta_{D'_j} \Big(\max_{1 \leq k \leq j} r_{k,k} \Big) \varepsilon, \tag{4.1}$$

and in particular,

$$|\Delta r_{i,i}| \leq c_1(m,i) c_2(i,\eta) (1+\eta)^i \zeta_{D'_i} \Big(\max_{1 \leq k \leq i} r_{k,k} \Big) \varepsilon. \tag{4.2}$$

If the $r_{k,k}$'s are increasing, then the upper bound in (4.1) with $i < j$ can be arbitrarily larger than the upper bound in (4.2). Thus we cannot ensure that $|r_{i,j} + \Delta r_{i,j}| \leq \eta |r_{i,i} + \Delta r_{i,i}|$. Suppose we restrict ourselves to setting ε as a function of n only. Computationally, this corresponds to allowing ourselves to use arbitrary precision arithmetic, but with a precision that shall depend only on the dimension and not on the matrix entries. Then for any $\eta' > \eta$, one may choose $r_{k,k}$'s so that the perturbed basis cannot be guaranteed η'-size-reduced by the perturbation bound. Overall, this means that given a basis that we are told is η-size-reduced, and given $\eta' > \eta$, we cannot always ensure that it is η'-size-reduced, without setting the precision as a function of the matrix entries. This is a very undesirable computational property. For this reason, we modify the notion of size-reducedness. We will not be able to show that this new definition is preserved under the perturbation analysis of the R-factor (altough the counter-example above will not work anymore): to obtain such a property, we will need a LLL-type set of conditions relying on the weakened size-reduction (see section 5).

Definition 4.1. Let $\eta, \theta \geq 0$. A full column-rank matrix $B \in \mathbb{R}^{m \times n}$ with R-factor R is (η,θ)-weakly-size-reduced ((η,θ)-WSR for short) if $|r_{i,j}| \leq \eta r_{i,i} + \theta r_{j,j}$ for any $i < j$.

The following matrix illustrates Definition 4.1: the magnitude of the coefficient $r_{i,j}$ is bounded with respect to both $r_{i,i}$ and $r_{j,j}$.

$$\begin{bmatrix} \ddots & & & & \\ & r_{i,i} & \xleftarrow{\eta} & r_{i,j} & \\ & & \ddots & \downarrow \theta & \\ & & & r_{j,j} & \\ & & & & \ddots \end{bmatrix}$$

As in the previous section, we analyze the quantity $\varkappa(B,j)$ by looking at the diagonal elements of the R-factor, i.e., the sequence of $r_{i,i}$'s.

Theorem 4.2. *Let $B \in \mathbb{R}^{m\times n}$ with full column rank be (η,θ)-WSR for some $\eta \geq 0$ and $\theta \geq 0$. Let R be its R-factor. For $j = 1,\ldots,n$, we let $r'_{j,j} = r_{j,j}/\max_{1\leq k\leq j} r_{k,k}$, $D'_j = \mathrm{diag}(r'_{1,1},\ldots,r'_{j,j})$ and $\xi_{D'_j} = \prod_{1\leq k<j} \max\left(\frac{r'_{k+1,k+1}}{r'_{k,k}}, 1\right)$. Then*

$$\varkappa(B,j) \leq \sqrt{2}c_2(j,\eta+\theta)(1+\eta+\theta)^j \xi_{D'_j}\big(\max_{1\leq k\leq j} r_{k,k}\big)/\|\mathbf{r}_j\|, \quad j = 1,\ldots,n. \tag{4.3}$$

Proof. Without loss of generality, we assume that $r_{1,1} = \max_{1\leq k\leq n} r_{k,k}$. If this is not the case, we divide the jth column of R by $\max_{1\leq k\leq j} r_{k,k}$ for $j = 1,\ldots,n$. Note that $\varkappa(B,j)$ is column-scaling invariant (see (2.17)), and that the quantities $(\max_{1\leq k\leq j} r_{k,k})/\|\mathbf{r}_j\|$ and $\xi_{D'_j}$ are invariant under this particular scaling.

Let $D = \mathrm{diag}(\xi_{D'_1},\ldots,\xi_{D'_n})$ and let $\bar{R} = RD^{-1}$. As $\varkappa(B,j)$ is invariant under column-scaling, we have $\varkappa(B,j) = \varkappa(BD^{-1},j)$. The most important part of the proof is to show that $\bar{R}$ is $\bar{\eta}$-size-reduced with $\bar{\eta} = \eta+\theta$. Once this is established, we will apply Theorem 3.2 to $\bar{R}$ to derive (4.3).

We want to prove that for any $i < j$, we have $|\bar{r}_{i,j}| \leq \bar{\eta}\bar{r}_{i,i}$. Because of the (η,θ)-WSR assumption, this will hold if

$$\eta\frac{r_{i,i}}{\xi_{D'_j}} + \theta\frac{r_{j,j}}{\xi_{D'_j}} \leq (\eta+\theta)\frac{r_{i,i}}{\xi_{D'_i}}.$$

Since $\xi_{D'_j} \geq \xi_{D'_i}$ when $j \geq i$, it suffices to prove that $\frac{r_{j,j}}{\xi_{D'_j}} \leq \frac{r_{i,i}}{\xi_{D'_i}}$, or equivalently that the sequence of the $\bar{r}_{i,i}$'s is non-increasing. This is equivalent to showing that $\frac{r_{j,j}}{\xi_{D'_j}} \leq \frac{r_{j-1,j-1}}{\xi_{D'_{j-1}}}$ holds for any $j \geq 2$, which is a direct consequence of the definition of $\xi_{D'_j}$.

We now apply Theorem 3.2 to BD^{-1}. For any $1 \leq j \leq n$, we have

$$\varkappa(B,j) = \varkappa(BD^{-1},j) \leq c_2(j,\bar{\eta})(1+\bar{\eta})^j \zeta_{\bar{D}'_j}\big(\max_{1\leq k\leq j} \bar{r}_{k,k}\big)/\|\bar{\mathbf{r}}_j\|,$$

with $\bar{D}'_j = \mathrm{diag}\left(\frac{\bar{r}_{i,i}}{\max_{1\leq k\leq i}\bar{r}_{k,k}}\right)_{1\leq i\leq j}$. The fact that the sequence of the $\bar{r}_{i,i}$'s is non-increasing implies that $\bar{D}'_j = \mathrm{diag}\left(\frac{\bar{r}_{i,i}}{\bar{r}_{1,1}}\right)_{1\leq i\leq j}$. For the same reason, we have $\zeta_{\bar{D}'_j} \leq \sqrt{2}$. This also gives that $\max_{1\leq k\leq j}\bar{r}_{k,k} = r_{1,1}$. Finally, we have $\|\bar{\mathbf{r}}_j\| = \|\mathbf{r}_j\|/\xi_{D'_j} = \|\mathbf{r}_j\|/\xi_{D_j}$. Since we assumed that $r_{1,1} = \max_{1\leq k\leq n} r_{k,k}$, this completes the proof. □

Remark 4.3. Naturally, as the assumption on B in Theorem 4.2 is weaker than in Theorem 3.2, the bound obtained for $\varkappa(B,j)$ is weaker as well. Indeed, it is easy to show that we always have $\zeta_{D'_j} \leq \sqrt{2}\xi_{D'_j}$. Furthermore, $\xi_{D'_j}$ can be arbitrarily larger than $\zeta_{D'_j}$. For instance, consider $\{r_{i,i}\}_{1\leq i\leq 5}$ defined by $r_{1,1} = r_{3,3} = r_{5,5} = 1$ and $r_{2,2} = r_{4,4} = \varepsilon$, where $\varepsilon > 0$ tends to 0. In this case, $\zeta_{D'_j} = O(1/\varepsilon)$, whereas $\xi_{D'_j} = O(1/\varepsilon^2)$.

Remark 4.4. Similarly to size-reduced matrices, we cannot argue from the perturbation results given in Corollary 2.1 and Theorem 4.2 that the weak size-reducedness is preserved after the perturbation (cf. the discussion given at the beginning of section 4). However, LLL-reduced matrices, which rely on weak size-reduction and will be introduced in section 5, do not have this drawback.

5. LLL Reduction is a Fix-Point Under Column-Wise Perturbation

In the present section, after some reminders on Euclidean lattices, we will introduce a modification of the LLL reduction [12] which is compatible with the perturbation analysis of the R-factor that we performed in the previous sections.

5.1. **Background on Euclidean lattices.** We give below the background on lattices that is necessary to the upcoming discussion. For more details, we refer to [13]. A *Euclidean lattice* is the set of all integer linear combinations of the columns of a full column rank basis matrix $B \in \mathbb{R}^{m\times n}$: $L = \{B\mathbf{x}, \mathbf{x} \in \mathbb{Z}^n\}$. The matrix B is said to be a basis matrix of L and its columns are a basis of L. If $n \geq 2$, a given lattice has infinitely lattice bases, but they are related to one another by *unimodular transforms*, i.e., by right-multiplication by $n\times n$ integer matrices of determinant ± 1. A lattice invariant is a quantity that does not depend on the particular choice of a basis of a given lattice. The simplest such invariant is the *lattice dimension* n. Let R be the R-factor of the basis matrix B. The *determinant* of the lattice L is defined as the product of the diagonal entries of R: $\det(L) = \prod_{1\leq i\leq n} r_{i,i}$. Since lattice bases are related by unimodular matrices, the determinant is a lattice invariant. Another important invariant is the *minimum* $\lambda(L)$ defined as the norm of a shortest non-zero vector of L.

Lattice reduction is a major paradigm in the theory of Euclidean lattices. The aim is to find a basis of good quality of a lattice given by an arbitrary basis. One usually targets orthogonality and norm properties. A simple reason why one is interested in short vectors is that they require less space to store. One is interested in basis matrices whose columns are fairly orthogonal relatively to their norms (which can be achieved by requiring the off-diagonal $r_{i,j}$'s to be small and the sequence of the $r_{i,i}$'s to not decrease too fast), for several different reasons. For example, it is crucial to bound the complexity of enumeration-type algorithms that find shortest lattice vectors and closest lattice vectors to given targets in the space [9, 5]. As we will see below, basis matrices that have good orthogonality properties also have good numerical properties. In 1982, Lenstra, Lenstra and Lovász [12] described a notion of reduction, called *LLL reduction*, that can be reached in time polynomial in the size of the input basis and that ensures some orthogonality and norm properties. Their algorithm immediately had great impact on various fields of mathematics and computer science (we refer to [20] for an overview).

Definition 5.1. Let $\eta \in [1/2, 1)$ and $\delta \in (\eta^2, 1]$. Let B be a lattice basis matrix and R be its R-factor. The basis matrix B is (δ, η)-LLL-reduced if it is η-size-reduced and if for any i we have $\delta \cdot r_{i,i}^2 \leq r_{i,i+1}^2 + r_{i+1,i+1}^2$.

Originally in [12], the parameter η was set to $1/2$, but this condition was relaxed later by Schnorr [21] to allow inaccuracies in the computation of the entries of the matrix R. Allowing $\eta > 1/2$ does not change significantly the guaranteed quality of LLL-reduced matrices (see below). The parameter δ was chosen to be $3/4$ in [12], because this simplifies the expressions of the constants appearing in the quality bounds of $(\delta, 1/2)$-LLL-reduced matrices (the α in Theorem 5.2 becomes $\sqrt{2}$). The second condition in Definition 5.1 means that after projection onto the orthogonal complement of the first $i-1$ columns, the ith one is approximately shorter (i.e., not much longer) than the $(i+1)$th. Together, the two conditions imply that the $r_{i,i}$'s cannot decrease too quickly and that the norm of the ith column is essentially $r_{i,i}$

(up to a factor that depends only of the dimension). The theorem below gives the main properties of LLL-reduced matrices.

Theorem 5.2. *Let $\eta \in [1/2, 1)$ and $\delta \in (\eta^2, 1]$. Let $\alpha = \frac{1}{\sqrt{\delta - \eta^2}}$. If $B \in \mathbb{R}^{m \times n}$ is a (δ, η)-LLL-reduced basis matrix of a lattice L, then we have:*

$$\begin{aligned}
r_{j,j} &\leq \alpha \cdot r_{j+1,j+1}, \quad j = 1, \ldots, n-1, \\
\|\mathbf{b}_j\| &\leq \alpha^{j-1} \cdot r_{j,j}, \quad j = 1, \ldots, n, \\
\|\mathbf{b}_1\| &\leq \alpha^{n-1} \cdot \lambda(L), \\
\|\mathbf{b}_1\| &\leq \alpha^{\frac{n-1}{2}} \cdot (\det(L))^{\frac{1}{n}}, \\
\prod_{1 \leq j \leq n} \|\mathbf{b}_j\| &\leq \alpha^{\frac{n(n-1)}{2}} \cdot \det(L).
\end{aligned}$$

We do not give a proof, since Theorem 5.2 is a simple corollary of Theorem 5.4.

5.2. **A weakening of the LLL-reduction.** LLL-reduction suffers from the same drawback as size-reduction with respect to column-wise perturbations. If the ε parameter of a column-wise perturbation is set as a function of n, then for any $\eta' > \eta$ and any $\delta' < \delta$, one may choose $r_{k,k}$'s so that the initial basis is (δ, η)-LLL-reduced but the perturbed basis cannot be guaranteed (δ', η')-size-reduced. Indeed, consider the matrix $\begin{bmatrix} 1 & 0 \\ 0 & \gamma \end{bmatrix}$, where γ grows to infinity. We can choose $\Delta r_{1,1} = 0$ and $\Delta r_{1,2} = \varepsilon\gamma$. The latter grows linearly with γ and eventually becomes bigger than any fixed η', thus preventing the perturbed matrix from being size-reduced.

For this reason, we introduce a weakening of LLL-reduction that relies on weak-size-reduction instead of size-reduction. This seems to be more natural with respect to the approximate computation of the R factor of the QR factorization by Householder reflections, Givens rotations or the Modified Gram-Schmidt orthogonalization. The weakening has the nice property that if a basis is reduced according to this definition and the corresponding R-factor is computed by any of these algorithms using floating-point arithmetic, then it suffices to show that the basis is indeed reduced according to this weakening (up to a small additional relaxation of the same type). This relaxation is thus somehow a fix-point with respect to floating-point computation of the R-factor by these algorithms. We will make this statement precise in Corollary 6.1. The need for such a weakening was discovered by Schnorr [22, 23], though he did not define it formally nor proved any quality property.

Definition 5.3. Let $\eta \in [1/2, 1), \theta \geq 0$ and $\delta \in (\eta^2, 1]$. Let B be a lattice basis matrix and R be its R-factor. The basis matrix B is (δ, η, θ)-LLL-reduced if it is (η, θ)-WSR and if for any i we have: $\delta \cdot r_{i,i}^2 \leq r_{i,i+1}^2 + r_{i+1,i+1}^2$.

Figure 2 illustrates the different definitions of LLL-reduction. If the $r_{i,i}$'s are decreasing, then a (δ, η, θ)-LLL-reduced basis matrix is $(\delta, \eta + \theta)$-reduced. The weakening becomes more interesting when the $r_{i,i}$'s do not decrease. In any case, it does not worsen significantly the bounds of Theorem 5.2.

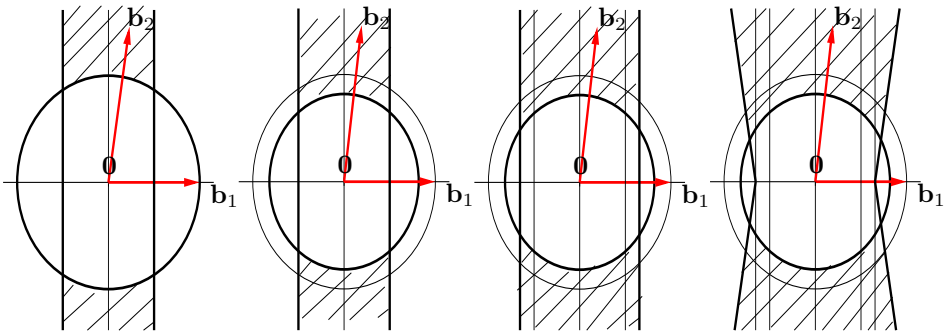

FIGURE 2. The hashed area is the set of vectors $\mathbf{b}_2$ such that $(\mathbf{b}_1, \mathbf{b}_2)$ is (from left to right) $(1,0,0)$-LLL, $(\delta,0,0)$-LLL, $(\delta,\eta,0)$-LLL and (δ,η,θ)-LLL.

Theorem 5.4. *Let $\eta \in [1/2,1), \theta \geq 0$ and $\delta \in (\eta^2, 1]$. Let $\alpha = \frac{\theta\eta + \sqrt{(1+\theta^2)\delta - \eta^2}}{\delta - \eta^2}$. If $B \in \mathbb{R}^{m\times n}$ is a (δ,η,θ)-LLL-reduced basis matrix of a lattice L, then we have*

$$r_{j,j} \leq \alpha \cdot r_{j+1,j+1}, \quad j = 1,\ldots,n-1, \tag{5.1}$$

$$\|\mathbf{b}_j\| \leq \alpha^{j-1} \cdot r_{j,j}, \quad j = 1,\ldots,n, \tag{5.2}$$

$$\|\mathbf{b}_1\| \leq \alpha^{n-1} \cdot \lambda(L), \tag{5.3}$$

$$\|\mathbf{b}_1\| \leq \alpha^{\frac{n-1}{2}} \cdot (\det(L))^{\frac{1}{n}}, \tag{5.4}$$

$$\prod_{1\leq j\leq n} \|\mathbf{b}_j\| \leq \alpha^{\frac{n(n-1)}{2}} \cdot \det(L). \tag{5.5}$$

Here α is always greater than or equal to $\frac{1}{\sqrt{\delta-\eta^2}}$, the value of α defined in Theorem 5.2. However, when θ tends to 0, the former tends to the latter.

Proof. By the given conditions, we have:

$$\delta r_{j,j}^2 \leq (\eta r_{j,j} + \theta r_{j+1,j+1})^2 + r_{j+1,j+1}^2 \leq \eta^2 r_{j,j}^2 + 2\eta\theta r_{j,j} r_{j+1,j+1} + (1+\theta^2) r_{j+1,j+1}^2.$$

This implies that $x := \frac{r_{j,j}}{r_{j+1,j+1}}$ satisfies the following degree-2 inequality:

$$(\delta - \eta^2)x^2 - 2\eta\theta x - (1+\theta^2) \leq 0. \tag{5.6}$$

The discriminant is $4\left((1+\theta^2)\delta - \eta^2\right) > 0$ and the leading coefficient is non-negative. As a consequence, we have:

$$x \leq \frac{\theta\eta + \sqrt{(1+\theta^2)\delta - \eta^2}}{\delta - \eta^2} = \alpha,$$

leading to (5.1).

Now we show (5.2). From (5.6), we have $(\delta-\eta^2)\alpha^2 - 2\eta\theta\alpha - (1+\theta^2) = 0$. But $\delta \leq 1$. Thus $(1-\eta^2)\alpha^2 - 2\eta\theta\alpha - (1+\theta^2) \geq 0$, or equivalently $(\theta+\eta\alpha)^2 \leq \alpha^2 - 1$.

Using this fact and $\alpha \geq 1$ as well, we have

$$\begin{aligned}\|\mathbf{b}_j\|^2 = \sum_{1\leq i\leq j} r_{i,j}^2 &\leq r_{j,j}^2 + \sum_{1\leq i<j} \left(\eta^2 \cdot r_{i,i}^2 + 2\theta\eta \cdot r_{j,j} r_{i,i} + \theta^2 \cdot r_{j,j}^2\right) \\ &\leq \left(1 + \sum_{1\leq i<j} (\eta^2\alpha^{2(j-i)} + 2\theta\eta\alpha^{j-i} + \theta^2)\right) \cdot r_{j,j}^2 \\ &\leq \left(1 + \sum_{1\leq i<j} (\eta^2\alpha^2 + 2\theta\eta\alpha + \theta^2)\alpha^{2(j-i-1)}\right) \cdot r_{j,j}^2 \\ &\leq \left(1 + (\theta+\eta\alpha)^2 \frac{\alpha^{2(j-1)} - 1}{\alpha^2 - 1}\right) \cdot r_{j,j}^2 \leq \alpha^{2(j-1)} \cdot r_{j,j}^2,\end{aligned}$$

leading to (5.2).

From (5.1), we have $r_{j,j} \geq \alpha^{1-j} \cdot r_{1,1}$. Suppose that $\mathbf{z} \in \mathbb{Z}^n$ satisfies $z_i \neq 0$ while $z_j = 0$ for $j = i+1, \ldots, n$. Then

$$\|B\mathbf{z}\| = \|R\mathbf{z}\| \geq |r_{i,i} z_i| \geq r_{i,i} \geq \alpha^{1-i} r_{1,1} = \alpha^{1-i}\|\mathbf{b}_1\|.$$

We thus have $\lambda(L) = \min_{\mathbf{z}\in\mathbb{Z}^n, \mathbf{z}\neq 0} \|B\mathbf{z}\| \geq \alpha^{1-n}\|\mathbf{b}_1\|$, which proves (5.3).

Since $\det(L) = \prod_{1\leq j\leq n} r_{j,j} \geq \prod_{1\leq j\leq n} (\alpha^{1-j} \cdot r_{1,1}) = \alpha^{(n-1)n/2}\|\mathbf{b}_1\|^n$, (5.4) holds. The inequality (5.5) follows from (5.2). □

5.3. **Application to LLL-reduced matrices.** We first show that the assumption of Theorem 2.3 is fulfilled for (δ, η, θ)-reduced basis matrices. To do this, we bound $\mathrm{cond}_2(R)$ for any upper triangular basis matrix R which is reduced.

Lemma 5.5. *Let $\eta, \theta \geq 0$ and $\alpha \geq 1$. Suppose an upper triangular matrix $R \in \mathbb{R}^{n\times n}$ with positive diagonal entries satisfies*

$$|r_{i,j}| \leq \eta r_{i,i} + \theta r_{j,j}, \quad r_{i,i} \leq \alpha r_{i+1,i+1}, \quad j = i+1, \ldots, n, \quad i = 1, \ldots, n-1. \tag{5.7}$$

Then

$$\mathrm{cond}_2(R) \leq \frac{|1-\eta-\theta|\alpha + 1}{(1+\eta+\theta)\alpha - 1}(1+\eta+\theta)^n \alpha^n. \tag{5.8}$$

Proof. In the proof, we will use the following fact a few times: for any strictly upper triangular matrix $U \in \mathbb{R}^{n\times n}$, we have $(I-U)^{-1} = \sum_{0\leq k<n} U^k$.

Write $R = \bar{R} \cdot D$, where $D = \mathrm{diag}(r_{1,1}, \ldots, r_{n,n})$ and $\bar{r}_{i,j} = \frac{r_{i,j}}{r_{j,j}}$ for $i \leq j$. From the assumption (5.7) it follows that $|\bar{r}_{i,j}| \leq (\eta\alpha^{j-i} + \theta)$ for $i < j$. Define T to be the strictly upper triangular matrix with $t_{i,j} = \bar{r}_{i,j}$ for $i < j$. Let J be the matrix whose all entries are 0, except that all $(i, i+1)$ entries are 1's. The matrix T is nilpotent and satisfies

$$|T| \leq (\eta+\theta) \sum_{1\leq k<n} (\alpha J)^k = (\eta+\theta)\alpha J(I-\alpha J)^{-1}.$$

Since $\bar{R} = I + T$, we have

$$|\bar{R}| \leq I + (\eta+\theta)\alpha J(I-\alpha J)^{-1} = (I - (1-\eta-\theta)\alpha J)(I-\alpha J)^{-1}.$$

Since T is strictly upper triangular, $\bar{R}^{-1} = \sum_{0\le k<n}(-T)^k$. As a consequence,

$$\begin{aligned}|\bar{R}^{-1}| \le \sum_{0\le k<n} |T|^k &\le \sum_{0\le k<n} \left[(\eta+\theta)\alpha J(I-\alpha J)^{-1}\right]^k \\ &= \left[I-(\eta+\theta)\alpha J(I-\alpha J)^{-1}\right]^{-1} \\ &= (I-\alpha J)\left(I-(1+\eta+\theta)\alpha J\right)^{-1} \\ &= (I-\alpha J)\sum_{0\le k<n}(1+\eta+\theta)^k\alpha^k J^k.\end{aligned}$$

using the fact that $\|J\|_2 = 1$, we obtain

$$\begin{aligned}\left\||\bar{R}|\cdot|\bar{R}^{-1}|\right\|_2 &\le \|I-(1-\eta-\theta)\alpha J\|_2 \sum_{0\le k<n}\left\|(1+\eta+\theta)^k\alpha^k J^k\right\|_2 \\ &\le (|1-\eta-\theta|\alpha+1)\sum_{0\le k<n}(1+\eta+\theta)^k\alpha^k \\ &\le \frac{|1-\eta-\theta|\alpha+1}{(1+\eta+\theta)\alpha-1}(1+\eta+\theta)^n\alpha^n.\end{aligned}$$

Using the equality $\mathrm{cond}_2(R) = \mathrm{cond}_2(\bar{R})$ allows us to assert that (5.8) holds. □

We now specialize our perturbation analysis of the previous sections to the case of (δ,η,θ)-LLL-reduced basis matrices.

Theorem 5.6. *Let $\eta \in [1/2,1)$, $\theta \ge 0$, $\delta \in (\eta^2,1]$ and $\alpha = \frac{\theta\eta+\sqrt{(1+\theta^2)\delta-\eta^2}}{\delta-\eta^2}$. Let $B \in \mathbb{R}^{m\times n}$ be a (δ,η,θ)-LLL-reduced basis matrix and R be its R-factor. Let $\Delta B \in \mathbb{R}^{m\times n}$ be a perturbation matrix satisfying (1.1), where ε satisfies*

$$c_3(1+\eta+\theta)^n\alpha^n\varepsilon < 1, \tag{5.9}$$

with

$$c_3 = \frac{(|1-\eta-\theta|\alpha+1)m\sqrt{n}}{((1+\eta+\theta)\alpha-1)(\sqrt{3/2}-1)}. \tag{5.10}$$

Then $B+\Delta B$ has a unique R-factor $R+\Delta R$ and

$$\|\Delta\mathbf{r}_j\| \le \sqrt{2}c_1(m,j)c_2(j,\eta+\theta)(1+\eta+\theta)^j\alpha^{k_j}r_{j,j}\varepsilon, \quad j=1,\dots,n, \tag{5.11}$$

where c_1 and c_2 are defined by (2.8) and (3.3), respectively, and k_j is the number of indices i such that $i<j$ and $r_{i,i} > r_{i+1,i+1}$.

Proof. From Lemma 5.5, we see that the condition (5.9) ensures that the assumption (2.5) in Theorem 2.3 is satisfied. From Corollary 2.1 and Theorem 4.2 it follows that

$$\|\Delta\mathbf{r}_j\| \le \sqrt{2}c_1(m,j)c_2(j,\eta+\theta)\xi_{D'_j}(1+\eta+\theta)^j\Big(\max_{1\le i\le j} r_{i,i}\Big)\varepsilon, \quad j=1,\dots,n, \tag{5.12}$$

where $\xi_{D'_j} = \prod_{i=1}^{j-1}\max\left(r'_{i+1,i+1}/r'_{i,i},1\right)$ with $r'_{j,j} = r_{j,j}/\max_{1\le i\le j} r_{i,i}$. If $r_{i,i} > r_{i+1,i+1}$ holds, then with (5.1) we have $r'_{i,i}/r'_{i+1,i+1} = r_{i,i}/r_{i+1,i+1} \le \alpha$, thus $1 \le$

$\alpha \cdot r'_{i+1,i+1}/r'_{i,i}$. Then it follows that

$$\xi_{D'_j} = \left(\prod_{\substack{i=1 \\ r'_{i+1,i+1} \geq r'_{i,i}}}^{j-1} \frac{r'_{i+1,i+1}}{r'_{i,i}} \right) \cdot \left(\prod_{\substack{i=1 \\ r'_{i,i} > r'_{i+1,i+1}}}^{j-1} \alpha \frac{r'_{i+1,i+1}}{r'_{i,i}} \right) \leq \alpha^{k_j} \frac{r'_{j,j}}{r'_{1,1}} = \alpha^{k_j} r'_{j,j},$$

which, combined with (5.12), results in (5.11). □

We can now conclude that the set of LLL-reduced matrices is a fix-point under column-wise perturbations.

Corollary 5.1. *Let $\eta \in [1/2, 1)$, $\theta \geq 0$, $\delta \in (\eta^2, 1]$ and $\alpha = \frac{\theta\eta + \sqrt{(1+\theta^2)\delta - \eta^2}}{\delta - \eta^2}$. Let $B \in \mathbb{R}^{m \times n}$ be a (δ, η, θ)-LLL-reduced basis matrix. Let $\Delta B \in \mathbb{R}^{m \times n}$ be a perturbation matrix satisfying (1.1), where ε is such that*

$$\varepsilon' := c_4(1 + \eta + \theta)^n \alpha^n \varepsilon < 1,$$

with

$$c_4 = \max(c_3, \sqrt{2} c_1(m, n) c_2(n, \eta + \theta)), \tag{5.13}$$

and with c_1, c_2 and c_3 defined by (2.8), (3.3) and (5.10), respectively. Then $B + \Delta B$ is $(\delta', \eta', \theta')$-LLL-reduced with

$$\delta' = \delta \frac{(1 - \varepsilon')^2}{(1 + \varepsilon')^2 (1 + 2\varepsilon'(\eta\alpha + \theta))}, \ \eta' = \frac{\eta}{1 - \varepsilon'} \quad \text{and} \quad \theta' = \frac{\theta + \varepsilon'}{1 - \varepsilon'}.$$

Proof. Let $R' = R + \Delta R$ be the R-factor of $B + \Delta B$. From Theorem 5.6, it follows that for all $1 \leq i \leq j \leq n$, we have $|\Delta r_{i,j}| \leq \varepsilon' r_{j,j}$. Therefore,

$$(1 - \varepsilon') r_{i,i} \leq r'_{i,i} \leq (1 + \varepsilon') r_{i,i} \quad \text{and} \quad |r'_{i,j}| \leq \eta r_{i,i} + (\theta + \varepsilon') r_{j,j}.$$

As a consequence, we have $|r'_{i,j}| \leq \frac{\eta}{1-\varepsilon'} r'_{i,i} + \frac{\theta + \varepsilon'}{1 - \varepsilon'} r'_{j,j}$, which gives us the weak-size-reduction. We also have

$$\begin{aligned} |r'_{i,i+1}| &\geq |r_{i,i+1}| - \varepsilon' r_{i+1,i+1} \\ (r'_{i,i+1})^2 &\geq r_{i,i+1}^2 - 2\varepsilon' |r_{i,i+1}| r_{i+1,i+1} \\ &\geq r_{i,i+1}^2 - 2\varepsilon' \left(\eta r_{i,i} + \theta r_{i+1,i+1}\right) r_{i+1,i+1} \\ &\geq r_{i,i+1}^2 - 2\varepsilon'(\eta\alpha + \theta) r_{i+1,i+1}^2. \end{aligned}$$

Therefore:

$$\begin{aligned} \frac{\delta}{(1 + \varepsilon')^2} \cdot (r'_{i,i})^2 \leq r_{i+1,i+1}^2 + r_{i,i+1}^2 &\leq r_{i+1,i+1}^2 + (r'_{i,i+1})^2 + 2\varepsilon'(\eta\alpha + \theta) r_{i+1,i+1}^2 \\ &\leq \frac{1 + 2\varepsilon'(\eta\alpha + \theta)}{(1 - \varepsilon')^2} \left((r'_{i+1,i+1})^2 + (r'_{i,i+1})^2 \right). \end{aligned}$$

This completes the proof. □

If the initial parameters δ, η and θ are such that $\eta \in (1/2, 1)$, $\theta > 0$, and $\delta \in (\eta^2, 1)$, then ε can be chosen as a function of δ, η, θ, m and n so that the resulting parameters δ', η', θ' also satisfy the domain conditions $\eta' \in (1/2, 1)$, $\theta' > 0$ and $\delta' \in ((\eta')^2, 1)$. Overall, this means that the set of basis matrices that are (δ, η, θ)-LLL-reduced for some parameters $\eta \in (1/2, 1)$, $\theta > 0$, and $\delta \in (\eta^2, 1)$ is stable under column-wise perturbations when ε is limited to a function of the parameters and

the dimensions m and n only. Note that if we fix $\theta = 0$, we cannot guarantee that the perturbed basis is reduced with $\theta' = 0$. This shows that the weakened LLL-reduction is more appropriate with respect to column-wise perturbations.

6. Practical Computation

In many cases, the perturbation matrix considered in a perturbation analysis comes from a backward stability result on some algorithm. In the case of QR factorization, the algorithms for which backward stability is established are the Householder algorithm, the Givens algorithm and the Modified Gram-Schmidt algorithm [8, §19]. In this section, we give a precise backward stability result for Householder's algorithm. We then apply it to LLL reduced bases. Similar results hold for the Givens and Modified Gram-Schmidt algorithms.

6.1. **Backward stability of Householder's algorithm.** Columnwise error analysis of the Householder QR factorization algorithm has been given in [8, §19]. But the constant in the backward error bound is not precisely computed. However, this information is crucial for some applications, such as the LLL reduction, since it will allow one to select floating-point precision to provide correctness guarantees. The purpose of the present section is to give a precisely defined backward error bound. The model of floating-point arithmetic that we use is formally described in [8, Eq. (2.4)].

Suppose we are given an $m \times n$ matrix B that has full column rank and that we aim at computing its R-factor R. Householder's algorithm proceeds column-wise by transforming B to R. Suppose that after j steps we have transformed B into a matrix of the following form:

$$\left(\begin{array}{c|c} B'_{1,1} & B'_{1,2} \\ \hline 0 & B'_{2,2} \end{array}\right),$$

where $B'_{1,1}$ is an $j \times j$ upper triangular matrix with positive entries. In the $(j+1)$th step, we apply a Householder transformation Q_{j+1} (which is orthogonal) to $B'_{2,2}$ from the left such that the first column of $B'_{2,2}$ becomes $[\times, 0, \ldots, 0]^T$. For the computation of the Householder transformation, see Figure 3, which gives two variants and is taken from [8, Lemma 19.1] with some changes. The Householder algorithm computes the full form of the QR factorization: $B = Q \begin{bmatrix} R \\ 0 \end{bmatrix}$, where $Q \in \mathbb{R}^{m \times m}$ is orthogonal and $R \in \mathbb{R}^{n \times n}$ is upper triangular. Some of the diagonal entries of R may be negative, but if we want them to be positive, we can multiply the corresponding rows of R and columns of Q by -1.

The algorithm of Figure 3 is performed with floating-point arithmetic. The computational details are straightforward, except for Step 3 of variant B: the numerator is a term that appears in the computation of Step 2, and thus does not need being re-computed. In our rounding error analysis, all given numbers are assumed to be real numbers (so they may not be floating-point numbers), and all algorithms are assumed to be run with unit roundoff u, i.e., $u = 2^{-p}$, where p is the precision. We use a hat to denote a computed quantity. For convenience, we use δ to denote a quantity satisfying $|\delta| \leq u$. The quantity $\gamma_m := \frac{mu}{1-mu}$ will be used a few times. The computations of some bounds contained in the proofs of the following lemmas

Input: A vector $\mathbf{b} \in \mathbb{R}^m$.
Output: A vector $\mathbf{v} \in \mathbb{R}^m$ such that $Q = I - \mathbf{v}\mathbf{v}^T$ is orthogonal and $Q \cdot \mathbf{b} = (\pm\|\mathbf{b}\|, 0, \ldots, 0)^T$.

1. $\mathbf{v}$:=$\mathbf{b}$.
2. s:=sign$(b_1) \cdot \|\mathbf{b}\|$.
3. v_1:=$b_1 + s$ (variant A) or v_1:=$\frac{-\sum_{i=2}^m b_i^2}{b_1+s}$ (variant B).
4. $\mathbf{v}$:=$\frac{1}{\sqrt{s \cdot v_1}} \cdot \mathbf{v}$ (variant A) or $\mathbf{v}$:=$\frac{1}{\sqrt{-s \cdot v_1}} \cdot \mathbf{v}$ (variant B).

FIGURE 3. Two variants of computing the Householder transformation.

were performed by MAPLE. The corresponding MAPLE work-sheet is available at `http://perso.ens-lyon.fr/damien.stehle/RPERTURB.html`.

The following lemma is a modified version of [8, Lemma 19.1].

Lemma 6.1. *Suppose we run either variant of the algorithm of Figure 3 on a nonzero vector $\mathbf{b} \in \mathbb{R}^m$ with unit roundoff u satisfying $c_5 \cdot u \leq 1$, where:*

$$c_5 = 4(6m + 63) \text{ for variant A, and } c_5 = 8(6m + 39) \text{ for variant B.} \tag{6.1}$$

Let $\hat{\mathbf{v}}$ be the computed vector and $\mathbf{v}$ be the vector that would have been computed with infinite precision. Then $\hat{\mathbf{v}} = \mathbf{v} + \Delta\mathbf{v}$ with $|\Delta\mathbf{v}| \leq (m+11)u \cdot |\mathbf{v}|$ for variant A (resp. $|\Delta\mathbf{v}| \leq \frac{1}{2}(5m + 29)u \cdot |\mathbf{v}|$ for variant B).

Proof. Let $c = \mathbf{b}^T\mathbf{b}$. Then $\hat{c} = fl(\hat{\mathbf{b}}^T\hat{\mathbf{b}})$ where $|\hat{\mathbf{b}} - \mathbf{b}| \leq u|\mathbf{b}|$. By following [8, p. 63], it is easy to verify that

$$\frac{|\hat{c} - c|}{|c|} \leq \gamma_{m+2}. \tag{6.2}$$

Note that the above result is different from [8, Eq. (3.5)], since here b_i's are not assumed to be floating-point numbers. Since $\gamma_{m+2} < 1$, using (6.2) we have

$$\frac{|\sqrt{\hat{c}} - \sqrt{c}|}{\sqrt{c}} = \frac{|\hat{c} - c|}{\sqrt{c}} \frac{1}{\sqrt{\hat{c}} + \sqrt{c}} \leq \frac{|\hat{c} - c|}{2c\sqrt{1 - \gamma_{m+2}}} \leq \frac{\gamma_{m+2}}{2\sqrt{1 - \gamma_{m+2}}} =: \beta_1. \tag{6.3}$$

Then it follows that at Step 2,

$$\frac{|\hat{s} - s|}{|s|} = \frac{|\sqrt{\hat{c}}(1 + \delta) - \sqrt{c}|}{\sqrt{c}} \leq (1 + \beta_1)(1 + u) - 1 =: \beta_2. \tag{6.4}$$

We now consider variants A and B of the algorithm separately. For variant A and at Step 3, the quantities b_1 and s have the same sign, so $|b_1| + |s| = |b_1 + s|$. Thus, using (6.4) we have

$$\begin{aligned} \frac{|\hat{v}_1 - v_1|}{|v_1|} &= \frac{|(\hat{b}_1 + \hat{s})(1 + \delta) - (b_1 + s)|}{|b_1 + s|} \leq \frac{|\hat{b}_1(1 + \delta) - b_1| + |\hat{s}(1 + \delta) - s|}{|b_1 + s|} \\ &\leq \frac{|b_1|[(1 + u)^2 - 1] + |s|[(1 + \beta_2)(1 + u) - 1]}{|b_1 + s|} \leq (1 + \beta_2)(1 + u) - 1 =: \beta_3, \end{aligned} \tag{6.5}$$

Then, using (6.4) and (6.5) we have

$$\frac{|\hat{d} - d|}{|d|} = \frac{|\hat{s}\hat{v}_1(1 + \delta) - sv_1|}{|sv_1|} \leq (1 + \beta_2)(1 + \beta_3)(1 + u) - 1 =: \beta_4. \tag{6.6}$$

The MAPLE work-sheet shows that $\beta_4 < 1$. Let $e = \sqrt{d} = \sqrt{sv_1}$. Then, by the same derivation for (6.4) (see (6.3)), using (6.6) we have

$$\frac{|\hat{e} - e|}{|e|} = \frac{|\sqrt{\hat{d}}\,(1+\delta) - \sqrt{d}|}{\sqrt{d}} \le \left(1 + \frac{\beta_4}{2\sqrt{1-\beta_4}}\right)(1+u) - 1 := \beta_5. \tag{6.7}$$

The MAPLE work-sheet shows that $\beta_5 < 1$. Then from (6.5) and (6.7) we obtain the following componentwise bound:

$$|\hat{\mathbf{v}} - \mathbf{v}| \le \left(\frac{1+\beta_3}{1-\beta_5}(1+u) - 1\right)|\mathbf{v}| = \beta_6|\mathbf{v}|, \tag{6.8}$$

where $\beta_6 = \frac{1+\beta_3}{1-\beta_5}(1+u) - 1 \le (m+11)u$, as indicated by the MAPLE work-sheet.

Now we consider variant B. The quantity $\sum_{i=2}^{m} b_i^2$ from Step 3 has been computed at Step 2. The relative error in the computed value is bounded by γ_{m+1}. Thus, using this fact and (6.5) (for the denominator) we conclude that

$$\frac{|\hat{v}_1 - v_1|}{|v_1|} \le \frac{1+\gamma_{m+1}}{1-\beta_3}(1+u) - 1 =: \beta_3'. \tag{6.9}$$

According to the MAPLE work-sheet, we have $\beta_3' < 1$. The rest analysis is similar to the derivation for (6.6)–(6.8) and we have the following componentwise bound:

$$|\hat{\mathbf{v}} - \mathbf{v}| \le \left(\frac{1+\beta_3'}{1-\beta_5'}(1+u) - 1\right)|\mathbf{v}| = \beta_6'|\mathbf{v}|, \tag{6.10}$$

where $\beta_5' = \left(1 + \frac{\beta_4'}{2\sqrt{1-\beta_4'}}\right)(1+u) - 1$, $\beta_4' = (1+\beta_2)(1+\beta_3')(1+u) - 1$ and $\beta_6' := \frac{1+\beta_3'}{1-\beta_5'}(1+u) - 1$. The MAPLE work-sheet shows that $\beta_6' \le \frac{1}{2}(5m+29)u$. □

At step $j+1$ of the QR factorization, once the Householder vector $\mathbf{v}$ is computed, the Householder matrix is applied to all the remaining column vectors of the matrix $B_{2,2}'$. The following lemma, a modified version of [8, Lemma 19.2], provides a backward analysis for this step.

Lemma 6.2. *Suppose that the assumptions of Lemma 6.1 hold. Let* $\mathbf{c} \in \mathbb{R}^m$, $Q = I - \mathbf{v}\mathbf{v}^T$ *and* $\mathbf{y} = Q\mathbf{c} = \mathbf{c} - \mathbf{v}(\mathbf{v}^T\mathbf{c})$. *In computing* $\mathbf{y}$, *the computed Householder vector* $\hat{\mathbf{v}}$ *is used. Then there exists* $\Delta Q \in \mathbb{R}^{m\times m}$ *such that*

$$\hat{\mathbf{y}} = (Q + \Delta Q)\mathbf{c} \quad \textit{and} \quad \|\Delta Q\|_F \le \frac{1}{4}c_5 u. \tag{6.11}$$

Proof. The proofs for both variants of the algorithm of Figure 3 are the same, so we only consider variant A. Let $t = \mathbf{v}^T\mathbf{c}$. Then $\hat{t} = fl(\hat{\mathbf{v}}^T\hat{\mathbf{c}})$, where $|\hat{\mathbf{c}} - \mathbf{c}| \le u|\mathbf{c}|$ and $|\hat{\mathbf{v}} - \mathbf{v}| \le \beta_6|\mathbf{v}|$, by Lemma 6.1. Then by following the derivation of [8, Eq. (3.4)], we can show that

$$|\hat{t} - t| \le [(1+\beta_6)(1+u)(1+\gamma_m) - 1]\,|\mathbf{v}|^T|\mathbf{c}| = \beta_7|\mathbf{v}|^T|\mathbf{c}|, \tag{6.12}$$

where $\beta_7 := (1+\beta_6)(1+u)(1+\gamma_m) - 1$. Let $\mathbf{w} = \mathbf{v}(\mathbf{v}^T\mathbf{c}) = \mathbf{v}t$. Then $\hat{\mathbf{w}} = \hat{\mathbf{v}}\hat{t}(1+\delta)$. Using (6.8), (6.10) and (6.12) we obtain the following bound:

$$|\hat{\mathbf{w}} - \mathbf{w}| \le ((1+\beta_6)(1+\beta_7)(1+u) - 1)\,|\mathbf{v}||\mathbf{v}|^T|\mathbf{c}| = \beta_8|\mathbf{v}||\mathbf{v}|^T|\mathbf{c}|,$$

where $\beta_8 = (1+\beta_6)(1+\beta_7)(1+u) - 1$. Then it follows that

$$|\hat{\mathbf{y}} - \mathbf{y}| = |fl(\hat{\mathbf{c}} - \hat{\mathbf{w}}) - (\mathbf{c} - \mathbf{w})| \le [(1+u)^2 - 1]|\mathbf{c}| + [(1+\beta_8)(1+u) - 1]|\mathbf{v}||\mathbf{v}|^T|\mathbf{c}|. \tag{6.13}$$

Note that the Householder vector $\mathbf{v}$ satisfies $\|\mathbf{v}\| = \sqrt{2}$. Thus from (6.13) it follows that

$$\|\hat{\mathbf{y}} - \mathbf{y}\| \leq [(1+u)^2 - 1]\|\mathbf{c}\| + 2[(1+\beta_8)(1+u) - 1]\|\mathbf{c}\| = \beta_9\|\mathbf{c}\|,$$

where $\beta_9 = (1+u)^2 + 2(1+\beta_8)(1+u) - 3$. We can write $\hat{\mathbf{y}} = (Q + \Delta Q)\mathbf{c}$ with $\Delta Q = \frac{(\hat{\mathbf{y}}-\mathbf{y})\mathbf{c}^T}{\mathbf{c}^T\mathbf{c}}$. We have $\|\Delta Q\|_F = \frac{\|\hat{\mathbf{y}}-\mathbf{y}\|}{\|\mathbf{c}\|} \leq \beta_9$. In the MAPLE work-sheet, we see that $\beta_9 \leq \frac{1}{4}c_5 u$, and thus (6.11) holds. □

The following lemma is a modified version of [8, Lemma 19.3]. It considers error analysis of a sequence of Householder matrices applied to a given matrix.

Lemma 6.3. *Let $B \in \mathbb{R}^{m\times n}$ and let $Q_i = I - \mathbf{v}_i\mathbf{v}_i^T$ for $i \leq n$ be a sequence of Householder matrices. We consider the sequence of transformations $B_{i+1} = Q_iB_i$, with $B_1 = B$. Suppose that these transformations are performed by using the computed Householder vectors $\hat{\mathbf{v}}_i$ with unit roundoff u. Let*

$$c_6 = \frac{1}{2}nc_5, \tag{6.14}$$

with c_5 defined by (6.1). If $c_6u \leq 1$, then the computed matrix $\widehat{B}_{n+1}$ satisfies

$$\widehat{B}_{n+1} = Q^T(B + \Delta B),$$

where $Q = Q_nQ_{n-1}\dots Q_1$ and

$$\|\Delta\mathbf{b}_j\| \leq c_6u\|\mathbf{b}_j\|, \qquad j = 1,\dots,n. \tag{6.15}$$

Proof. Let $\mathbf{b}_j^{(n+1)}$ be the jth column of B_{n+1}. From Lemma 6.2 it follows that there exist $\Delta Q_1,\dots,\Delta Q_n \in \mathbb{R}^{m\times m}$ such that

$$\hat{\mathbf{b}}_j^{(n+1)} = (Q_n + \Delta Q_n)\dots(Q_1 + \Delta Q_1)\mathbf{b}_j \quad\text{and}\quad \|\Delta Q_i\|_F \leq \frac{1}{4}c_5u.$$

Write $Q^T + \Delta Q^T = (Q_n + \Delta Q_n)\dots(Q_1 + \Delta Q_1)$. Then by [8, Lemma 3.7] we have

$$\|\Delta Q^T\|_F \leq \left(1 + \frac{1}{4}c_5u\right)^n - 1 \leq \frac{\frac{1}{4}c_5nu}{1 - \frac{1}{4}c_5nu} \leq c_6u.$$

Define $\Delta\mathbf{b}_j = Q\Delta Q^T\mathbf{b}_j$. Then

$$\hat{\mathbf{b}}_j^{(n+1)} = Q^T\mathbf{b}_j + \Delta Q^T\mathbf{b}_j = Q^T(\mathbf{b}_j + \Delta\mathbf{b}_j) \quad\text{and}\quad \|\Delta\mathbf{b}_j\| \leq c_6u\|\mathbf{b}_j\|.$$

□

We can now conclude with a more informative version of [7, Th. 18.4] (or [8, Th. 19.4], in the newer edition).

Theorem 6.4. *Let $\widehat{R}$ be the computed R-factor of the QR factorization of a given matrix $B \in \mathbb{R}^{m\times n}$ by the Householder algorithm, with unit roundoff u. If $c_6u \leq 1$ with $c_6 = 2n(6m+63)$ for variant A and $4n(6m+39)$ for variant B, then there exists an orthogonal matrix $Q \in \mathbb{R}^{m\times m}$ such that*

$$B + \Delta B = Q\begin{bmatrix}\widehat{R}\\ 0\end{bmatrix} \quad\text{and}\quad \|\Delta\mathbf{b}_j\| \leq c_6u\|\mathbf{b}_j\|, \quad j = 1,\dots,n.$$

The latter implies that $|\Delta B| \leq c_6uC|B|$, where $c_{i,j} = 1$ for all i,j. The matrix Q is given explicitly by $Q^T = Q_nQ_{n-1}\dots Q_1$, where Q_i is the Householder matrix corresponding to the exact application of the ith step of the Householder algorithm to $\widehat{B}_i$.

Proof. As a direct consequence of Lemma 6.3, we have:

$$B + \Delta B = Q \begin{bmatrix} \widehat{R} \\ 0 \end{bmatrix} \quad \text{and} \quad \|\Delta \mathbf{b}_j\| \leq c_6 u \|\mathbf{b}_j\|, \quad j = 1, \ldots, n,$$

with $Q = Q_1^T Q_2^T \ldots Q_n^T$. Then

$$|\Delta b_{i,j}| \leq c_6 u \|\mathbf{b}_j\| \leq c_6 u \|\mathbf{b}_j\|_1 = c_6 u \mathbf{e}^T |\mathbf{b}_j|,$$

where $\mathbf{e} = [1, \ldots, 1]^T$. We thus have $|\Delta \mathbf{b}_j| \leq c_6 u \mathbf{e}\mathbf{e}^T |\mathbf{b}_j|$ for all j, which gives $|\Delta B| \leq c_6 u C |B|$ since $C = \mathbf{e}\mathbf{e}^T$. □

6.2. **Application to LLL-reduced matrices.** By using Theorem 6.4 and Corollary 5.1, we have the following result on LLL-reduced bases.

Corollary 6.1. *Let $\eta \in [1/2, 1], \theta \geq 0, \delta \in (\eta^2, 1]$ and $\alpha = \frac{\theta\eta + \sqrt{(1+\theta^2)\delta - \eta^2}}{\delta - \eta^2}$. Let $B \in \mathbb{R}^{m \times n}$ be a (δ, η, θ)-LLL-reduced basis matrix. Let u be such that*

$$u' := c_7 (1 + \eta + \theta)^n \alpha^n u < 1,$$

where $c_7 = c_4 c_6$ and with c_4 defined by (5.13) *and c_6 defined by* (6.14)*. Suppose we compute the R-factor of B with the algorithm described in Subsection 6.1. Then the computed matrix $\widehat{R}$ is $(\delta', \eta', \theta')$-LLL-reduced with*

$$\delta' = \delta \frac{(1 - u')^2}{(1 + u')^2 (1 + 2u'(\eta\alpha + \theta))}, \quad \eta' = \frac{\eta}{1 - u'}, \quad \theta' = \frac{\theta + u'}{1 - u'}.$$

Proof. From Theorem 6.4, we know that (1.1) holds with $\varepsilon = c_6 u$. The result directly follows from Corollary 5.1. □

The weakening of the LLL-reduction is stable under Householder's algorithm: if the input basis is reduced, then so is the output basis (with slightly relaxed factors).

7. Concluding Remarks

We investigated the sensitivity of the R-factor of the QR-factorisation under column-wise perturbations, which correspond to the backward stability results of the standard QR factorization algorithms. We focused on the case of LLL-reduced matrices, and showed that if the classical definition of LLL-reducedness is sligthly modified, then LLL-reducedness is preserved under column-wise perturbations. This implies that by computing the R-factor of a reduced matrix with a standard floating-point QR factorization algorithm (e.g., Householder's algorithm), then one can numerically check that the LLL conditions (5.3) are indeed satisfied, for slightly degraded parameters. These certified reduction parameters can be made arbitrarily close to the actual reduction parameters by setting the precision sufficiently high. Importantly, the required precision for the above to be valid is linear with respect to the dimension, and does not depend on the magnitudes of the matrix entries. This study was motivated by its algorithmic implications: the results may be used to efficiently check the LLL-reducedness of a basis and to speed up the LLL-reduction process.

References

1. X.-W. Chang and C. C. Paige, *Componentwise perturbation analyses for the QR factorization*, Numerische Mathematik **88** (2001), 319–345.
2. X.-W. Chang, C. C. Paige, and G. W. Stewart, *Perturbation analyses for the QR factorization*, SIAM J. Matrix Anal. Appl. **18** (1997), 775–791.
3. X.-W. Chang and D. Stehlé, *Rigorous perturbation bounds for some matrix factorization.*
4. H. Cohen, *A course in computational algebraic number theory, 2nd edition*, Springer, Berlin, 1995.
5. U. Fincke and M. Pohst, *A procedure for determining algebraic integers of given norm*, Proceedings of EUROCAL, Lecture Notes in Computer Science, vol. 162, 1983, pp. 194–202.
6. G. H. Golub and C. F. Van Loan, *Matrix computations, 3rd edition*, The John Hopkins University Press, Baltimore, MD, 1996.
7. N. J. Higham, *Accuracy and stability of numerical algorithms, 1st edition*, Society for Industrial and Applied Mathematics, 1996.
8. ———, *Accuracy and stability of numerical algorithms, 2nd edition*, Society for Industrial and Applied Mathematics, 2002.
9. R. Kannan, *Improved algorithms for integer programming and related lattice problems*, Proceedings of the 15th Symposium on the Theory of Computing (STOC 1983), ACM Press, 1983, pp. 99–108.
10. D. Knuth, *The analysis of algorithms*, Actes du Congrès International des Mathématiciens de 1970, vol. 3, Gauthiers-Villars, 1971, pp. 269–274.
11. H. Koy and C. P. Schnorr, *Segment LLL-reduction of lattice bases with floating-point orthogonalization*, Proceedings of the 2001 Cryptography and Lattices Conference (CALC'01), Lecture Notes in Computer Science, vol. 2146, Springer, Berlin, 2001, pp. 81–96.
12. A. K. Lenstra, H. W. Lenstra, Jr., and L. Lovász, *Factoring polynomials with rational coefficients*, Math. Ann. **261** (1982), 515–534.
13. L. Lovász, *An Algorithmic Theory of Numbers, Graphs and Convexity*, SIAM Publications, 1986, CBMS-NSF Regional Conference Series in Applied Mathematics.
14. I. Morel, D. Stehlé, and G. Villard, *H-LLL: Using Householder inside LLL*, Proceedings of the 2009 International Symposium on Symbolic and Algebraic Computation (ISSAC'09), ACM Press, 2009, pp. 271–278.
15. ———, *From an LLL-reduced basis to another*, Work in progress, 2010.
16. W. H. Mow, *Maximum likelihood sequence estimation from the lattice viewpoint*, IEEE Transactions on Information Theory **40** (1994), 1591–1600.
17. P. Nguyen and D. Stehlé, *Floating-point LLL revisited*, Proceedings of Eurocrypt 2005, Lecture Notes in Computer Science, vol. 3494, Springer, Berlin, 2005, pp. 215–233.
18. ———, *An LLL algorithm with quadratic complexity*, SIAM Journal on Computing **39** (2009), no. 3, 874–903.
19. P. Nguyen and J. Stern, *The two faces of lattices in cryptology*, Proceedings of the 2001 Cryptography and Lattices Conference (CALC'01), Lecture Notes in Computer Science, vol. 2146, Springer, Berlin, 2001, pp. 146–180.
20. P. Q. Nguyen and B. Vallée (editors), *The lll algorithm: survey and applications*, Information Security and Cryptography, Springer, Berlin, 2009, Published after the LLL25 conference held in Caen in June 2007, in honour of the 25-th anniversary of the LLL algorithm.
21. C. P. Schnorr, *A more efficient algorithm for lattice basis reduction*, Journal of Algorithms **9** (1988), no. 1, 47–62.
22. ———, *Fast LLL-type lattice reduction*, Inform. Comput. **204** (2006), 1–25.
23. ———, *Progress on LLL and lattice reduction*, In [20], 2009.
24. A. Schönhage, *Schnelle Berechnung von Kettenbruchentwicklungen*, Acta Informatica **1** (1971), 139–144.
25. A. Schönhage and V. Strassen, *Schnelle Multiplikation grosser Zahlen*, Computing **7** (1971), 281–292.
26. G. Villard, *Certification of the QR factor R and of lattice basis reducedness*, Proceedings of the 2007 International Symposium on Symbolic and Algebraic Computation (ISSAC'07), ACM Press, 2007, pp. 143–150.
27. H. Zha, *A componentwise perturbation analysis of the QR decomposition*, SIAM J. Matrix Anal. Appl. **14** (1993), no. 4, 1124–1131.

School of Computer Science, McGill University Montreal, Quebec, Canada H3A 2A7, http://www.cs.mcgill.ca/~chang
E-mail address: chang@cs.mcgill.ca

CNRS, Université de Lyon, Laboratoire LIP, CNRS-ENSL-INRIA-UCBL, 46 Allée d'Italie, 69364 Lyon Cedex 07, France, http://perso.ens-lyon.fr/damien.stehle
E-mail address: damien.stehle@ens-lyon.fr

CNRS, Université de Lyon, Laboratoire LIP, CNRS-ENSL-INRIA-UCBL, 46 Allée d'Italie, 69364 Lyon Cedex 07, France, http://perso.ens-lyon.fr/gilles.villard
E-mail address: gilles.villard@ens-lyon.fr