arXiv:2303.03547v3 [math.NA] 20 Feb 2024

Small singular values can increase in lower precision

Christos Boutsikas
Purdue University
cboutsik@purdue.edu

Petros Drineas
Purdue University
pdrineas@purdue.edu

Ilse C.F. Ipsen
North Carolina State University
ipsen@ncsu.edu

Abstract

We perturb a real matrix A of full column rank, and derive lower bounds for the smallest singular values of the perturbed matrix, in terms of normwise absolute perturbations. Our bounds, which extend existing lower-order expressions, demonstrate the potential increase in the smallest singular values, and represent a qualitative model for the increase in the small singular values after a matrix has been downcast to a lower arithmetic precision. Numerical experiments confirm the qualitative validity of this model and its ability to predict singular values changes in the presence of decreased arithmetic precision.

1 Introduction

Given a real, full column-rank matrix $\boldsymbol{A}$, we present lower bounds for the smallest singular values of a perturbed matrix $\boldsymbol{A} + \boldsymbol{E}$.

1.1 Motivation

We investigate the change in the computed singular values of a tall and skinny matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ with $m \geq n$ and $\text{rank}(\boldsymbol{A}) = n$ when $\boldsymbol{A}$ is demoted, that is, downcast to a lower arithmetic precision.

We have observed that demotion to lower precision can improve the conditioning of $\boldsymbol{A}$ by significantly increasing the computed small singular values, while leaving large singular values mostly unharmed. For instance, if the smallest singular value of $\boldsymbol{A}$ is on the order of double precision roundoff,

$$\sigma_{\min}\left(\texttt{double}(\boldsymbol{A})\right) \approx 10^{-16},$$

then downcasting $\boldsymbol{A}$ to single precision can increase the smallest singular value to single precision roundoff,

$$\sigma_{\min}\left(\texttt{double}(\texttt{single}(\boldsymbol{A}))\right) \approx 10^{-8}.$$

This phenomenon has been observed before, as the following quotes illustrate:

> *... small singular values tend to increase [SS90, page 266]*
> *... even an approximate inverse of an arbitrarily ill-conditioned matrix does, in general, contain useful information [Rum09, page 260]*
> *This is due to a kind of regularization by rounding to working precision [Rum09, page 261]*

1.2 Modelling demotion to lower precision in terms of perturbations

We model the downcasting of a matrix to lower precision in terms of normwise absolute perturbations.

The accumulated error from typical singular value algorithms in Matlab and Julia is a normwise absolute error [GV13, section 8.6]. Thus, if we downcast a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ with $m \geq n$ to single precision and compute the singular values of the demoted matrix, the resulting error can be represented as an absolute

perturbation $\boldsymbol{E}$. According to Weyl's inequality [GV13, Corollary 8.6.2], corresponding singular values[1] of $\boldsymbol{A}$ change by at most $\|\boldsymbol{E}\|_2$,

$$|\sigma_j(\boldsymbol{A}+\boldsymbol{E}) - \sigma_j(\boldsymbol{A})| \leq \|\boldsymbol{E}\|_2 \approx 10^{-8}, \qquad 1 \leq j \leq n.$$

The bound implies that singular values larger than single precision roundoff, i.e. $\sigma_j(\boldsymbol{A}) \gg \|\boldsymbol{E}\|_2$, remain essentially the same,

$$\sigma_j(\boldsymbol{A}) \approx \sigma_j(\boldsymbol{A}) - \underbrace{\|\boldsymbol{E}\|_2}_{10^{-8}} \leq \sigma_j(\boldsymbol{A}+\boldsymbol{E}) \leq \sigma_j(\boldsymbol{A}) + \underbrace{\|\boldsymbol{E}\|_2}_{10^{-8}} \approx \sigma_j(\boldsymbol{A}),$$

while it is inconclusive about small singular values on the order of double precision roundoff,

$$\underbrace{\sigma_\ell(\boldsymbol{A})}_{10^{-16}} - \underbrace{\|\boldsymbol{E}\|_2}_{10^{-8}} \leq \sigma_\ell(\boldsymbol{A}+\boldsymbol{E}) \leq \underbrace{\sigma_\ell(\boldsymbol{A})}_{10^{-16}} + \underbrace{\|\boldsymbol{E}\|_2}_{10^{-8}}.$$

1.3 Our contributions

Our main results are normwise absolute lower bounds on the smallest singular value cluster of a perturbed matrix. The bounds, compactly summarized in Theorem 1 below, testify to a definitive increase in the perturbed small singular values. The qualitative validity of the bounds is confirmed by the numerical experiments in section 4. Our assumptions are not restrictive and merely require the smallest singular value cluster to be separated by a small gap from the remaining singular values.

Theorem 1 improves the second-order perturbation expansions in [Ste84; SS90; Ste06], because it is a true lower bound and, unlike [Ste84], [SS90, Lemma V.4.5], it needs no assumptions on the size of the offdiagonal blocks. Its proof (Sections 2 and 3) follows from Theorem 7 and a restatement of Assumptions 3.1. The special case where $r = 1$ reduces to Assumptions 2.1 and Theorem 4.

Theorem 1. *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with $m \geq n$ have $\operatorname{rank}(\boldsymbol{A}) \geq n - r$ for some $r \geq 1$. Let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be a full singular value decomposition, where $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times n}$ is diagonal, and $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ are orthogonal matrices. Partition commensurately,*

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_2 \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}, \qquad \boldsymbol{U}^T\boldsymbol{E}\boldsymbol{V} = \begin{bmatrix} \boldsymbol{E}_{11} & \boldsymbol{E}_{12} \\ \boldsymbol{E}_{21} & \boldsymbol{E}_{22} \\ \boldsymbol{E}_{31} & \boldsymbol{E}_{32} \end{bmatrix},$$

where $\boldsymbol{\Sigma}_1, \boldsymbol{E}_{11} \in \mathbb{R}^{(n-r)\times(n-r)}$ with $\boldsymbol{\Sigma}_1$ nonsingular diagonal; and $\boldsymbol{\Sigma}_2, \boldsymbol{E}_{22} \in \mathbb{R}^{r\times r}$ with $\boldsymbol{\Sigma}_2$ diagonal.

If $1/\|\boldsymbol{\Sigma}_1^{-1}\|_2 > 4\|\boldsymbol{E}\|_2$ and $\|\boldsymbol{\Sigma}_2\|_2 < \|\boldsymbol{E}\|_2$, then[2]

$$\sigma_{n-r+j}(\boldsymbol{A}+\boldsymbol{E})^2 \geq \lambda_j(\boldsymbol{E}_{32}^T\boldsymbol{E}_{32} + (\boldsymbol{\Sigma}_2 + \boldsymbol{E}_{22})^T(\boldsymbol{\Sigma}_2 + \boldsymbol{E}_{22}) - \boldsymbol{R}_3) - r_4, \quad 1 \leq j \leq r,$$

where $\boldsymbol{R}_3$ contains terms of order 3,

$$\boldsymbol{R}_3 \equiv \boldsymbol{E}_{12}^T\boldsymbol{W} + \boldsymbol{W}^T\boldsymbol{E}_{12}, \qquad \boldsymbol{W} \equiv (\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})^{-T}\begin{bmatrix}\boldsymbol{E}_{21}^T & \boldsymbol{E}_{31}^T\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_2 + \boldsymbol{E}_{22} \\ \boldsymbol{E}_{32}\end{bmatrix}$$

and r_4 contains terms of order 4 and higher,

$$r_4 \equiv \|\boldsymbol{W}\|_2^2 + 4\frac{\|\boldsymbol{E}\|_2^2\,\|(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})^{-1}(\boldsymbol{E}_{12} + \boldsymbol{W})\|_2^2}{1 - 4\|\boldsymbol{E}\|_2^2\|(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})^{-1}\|_2^2}.$$

Future work, sketched in section 5, will refine the above results towards a quantitative analysis that predicts the order of magnitude of the increase, and the influential matrix properties, in particular, the role of the singular value gap. We also note that our theorem does not directly account for computational precision. However, our experiments in Section 4 demonstrate that when performing calculations in double precision, the "exact" singular values exhibit complete overlap with those computed in double precision. This suggests that perturbations arising from computational precision should not affect our results.

[1] The singular values of a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ are labelled in non-increasing order, by $\sigma_1(\boldsymbol{A}) \geq \sigma_2(\boldsymbol{A}) \geq \ldots \geq \sigma_{\min\{m,n\}}(\boldsymbol{A})$.

[2] The eigenvalues of a symmetric matrix $\boldsymbol{H} \in \mathbb{R}^{k\times k}$ are labelled in non-increasing order, by $\lambda_1(\boldsymbol{H}) \geq \cdots \geq \lambda_k(\boldsymbol{H})$.

1.4 Existing work

There are many bounds for the smallest singular value of general, unstructured matrices. The bounds for nonsingular matrices $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ in [LX21; Shu22; Zou12] involve the factor $|\det(\boldsymbol{A})|^2 \left(\frac{n-1}{\|\boldsymbol{A}\|_F^2}\right)^{n-1}$, while the ones in [HP92; YG97] contain factors like $|\det(\boldsymbol{A})|^2 \left(\frac{n-1}{n}\right)^{(n-1)/2}$ and row and column norms. The Schur complement-based bounds for strictly diagonally dominant matrices in [Hua08; Li20; Ois23; San21; Var75] depend on the degree of diagonal dominance, as do the Gerschgorin type bounds for rectangular matrices in [Joh89; JS98].

In contrast, we are bounding the smallest singular values of *perturbed* matrices. The expressions for small singular values in [Ste84, Theorem], [Ste06, Theorem 8], [SS90, Section V.4.2] are second-order perturbation expansions rather than bounds, and require assumptions on the singular vectors.

1.5 Overview

Our deterministic lower bounds for small singular values of $\boldsymbol{A} + \boldsymbol{E}$ are based on eigenvalue bounds for $(\boldsymbol{A} + \boldsymbol{E})^T(\boldsymbol{A} + \boldsymbol{E})$. We present normwise absolute bounds for a single smallest singular value (section 2) and for a cluster of small singular values (section 3). The numerical experiments (section 4) confirm the qualitative increase in small singular values resulting from the demotion of the matrix to lower precision. A brief discussion of future work (section 5) concludes the paper.

2 A single smallest singular value

We perturb a matrix that has a single smallest singular value, and derive a lower bound for the smallest singular value of the perturbed matrix in terms of normwise absolute perturbations (Section 2.2), based on eigenvalue bounds (Section 2.1).

2.1 Auxiliary eigenvalue results

We square the singular values of $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and consider instead the eigenvalues of the symmetric positive semi-definite matrix $\boldsymbol{B} \equiv \boldsymbol{A}^T\boldsymbol{A} \in \mathbb{R}^{n\times n}$.

For a symmetric positive semi-definite matrix $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ with a single smallest eigenvalue $\lambda_{\min}(\boldsymbol{B})$, we present two expressions for $\lambda_{\min}(\boldsymbol{B})$ with different assumptions (Lemmas 1 and 2), and two lower bounds in terms of normwise absolute perturbations (Theorems 2 and 3).

We assume that $\lambda_{\min}(\boldsymbol{B})$ is separated from the remaining eigenvalues, in the sense that it is strictly smaller than the smallest eigenvalue of the leading principal submatrix $\boldsymbol{B}_{11}$ of order $n-1$. The equality below expresses $\lambda_{\min}(\boldsymbol{B})$ in terms of itself.

Lemma 1 (Exact expression)**.** *Let $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ be symmetric positive semi-definite with* $\operatorname{rank}(\boldsymbol{B}) \geq n-1$*, and partition*

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{b} \\ \boldsymbol{b}^T & \beta \end{bmatrix} \qquad \text{where} \quad \boldsymbol{B}_{11} \in \mathbb{R}^{(n-1)\times(n-1)}.$$

Then

$$0 \leq \lambda_{min}(\boldsymbol{B}) \leq \beta. \tag{1}$$

If also $\lambda_{min}(\boldsymbol{B}) < \lambda_{min}(\boldsymbol{B}_{11})$ then

$$\lambda_{min}(\boldsymbol{B}) = \beta - \boldsymbol{b}^T(\boldsymbol{B}_{11} - \lambda_{min}(\boldsymbol{B})\,\boldsymbol{I})^{-1}\boldsymbol{b},$$

Proof. Abbreviate $\tilde{\lambda}_{\min} \equiv \lambda_{\min}(\boldsymbol{B})$. The positive semi-definiteness of $\boldsymbol{B}$ implies the lower bound in (1), while the variational inequalities imply the upper bound,

$$0 \leq \tilde{\lambda}_{\min} = \min_{\|\mathbf{x}\|_2=1} \mathbf{x}^T\boldsymbol{B}\mathbf{x} \leq \boldsymbol{e}_n^T\boldsymbol{B}\boldsymbol{e}_n = \beta.$$

To show the expression for $\tilde{\lambda}_{\min}$, observe that the shifted matrix

$$\boldsymbol{B} - \tilde{\lambda}_{\min}\,\boldsymbol{I} = \begin{bmatrix} \boldsymbol{B}_{11} - \tilde{\lambda}_{\min}\,\boldsymbol{I} & \boldsymbol{b} \\ \boldsymbol{b}^T & \beta - \tilde{\lambda}_{\min} \end{bmatrix}$$

is singular. From the assumption $\tilde{\lambda}_{\min} < \lambda_{\min}(\boldsymbol{B}_{11})$ follows that $\boldsymbol{B}_{11} - \tilde{\lambda}_{\min}\,\boldsymbol{I}$ is nonsingular. So we can determine the block LU decomposition $\boldsymbol{B} - \tilde{\lambda}_{\min}\,\boldsymbol{I} = \boldsymbol{L}\widehat{\boldsymbol{U}}$ with

$$\begin{aligned} \boldsymbol{L} &\equiv \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{b}^T(\boldsymbol{B}_{11} - \tilde{\lambda}_{\min}\,\boldsymbol{I})^{-1} & 1 \end{bmatrix}, \\ \widehat{\boldsymbol{U}} &\equiv \begin{bmatrix} \boldsymbol{B}_{11} - \tilde{\lambda}_{\min}\,\boldsymbol{I} & \boldsymbol{b} \\ \boldsymbol{0} & \beta - \tilde{\lambda}_{\min} - \boldsymbol{b}^T(\boldsymbol{B}_{11} - \tilde{\lambda}_{\min}\,\boldsymbol{I})^{-1}\boldsymbol{b} \end{bmatrix}. \end{aligned}$$

Since $\boldsymbol{B} - \tilde{\lambda}_{\min}\,\boldsymbol{I}$ is singular and the unit triangular matrix $\boldsymbol{L}$ is nonsingular, the block upper triangular matrix $\widehat{\boldsymbol{U}}$ has no choice but to be singular. Its leading principal submatrix $\boldsymbol{B}_{11} - \tilde{\lambda}_{\min}\,\boldsymbol{I}$ is nonsingular by assumption, which leaves the (2,2) element to be singular, but it being a scalar implies

$$\beta - \tilde{\lambda}_{\min} - \boldsymbol{b}^T(\boldsymbol{B}_{11} - \tilde{\lambda}_{\min}\,\boldsymbol{I})^{-1}\boldsymbol{b} = 0.$$

This gives the expression for $\tilde{\lambda}_{\min}$. □

If $\boldsymbol{b} = \boldsymbol{0}$ then Lemma 1 correctly asserts that $\lambda_{\min}(\boldsymbol{B}) = \beta$.

Lemma 2 below presents the same expression for $\lambda_{\min}(\boldsymbol{B})$ as in Lemma 1, but under a stronger albeit more useful assumption.

Lemma 2 (Exact expression with stronger assumption)**.** *Let $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ be symmetric positive semi-definite with* $\operatorname{rank}(\boldsymbol{B}) \geq n-1$*, and partition*

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{b} \\ \boldsymbol{b}^T & \beta \end{bmatrix} \qquad \textit{where} \quad \boldsymbol{B}_{11} \in \mathbb{R}^{(n-1)\times(n-1)}.$$

If $\beta < \lambda_{min}(\boldsymbol{B}_{11})$ then

$$\lambda_{min}(\boldsymbol{B}) = \beta - \boldsymbol{b}^T(\boldsymbol{B}_{11} - \lambda_{min}(\boldsymbol{B})\,\boldsymbol{I})^{-1}\boldsymbol{b} \geq 0.$$

Proof. The upper bound (1) combined with the assumption $\beta < \lambda_{\min}(\boldsymbol{B}_{11})$ implies the assumption in Lemma 1,

$$0 \leq \lambda_{\min}(\boldsymbol{B}) \leq \beta < \lambda_{\min}(\boldsymbol{B}_{11}). \tag{2}$$

□

The subsequent lower bounds for $\lambda_{\min}(\boldsymbol{B})$ are informative if the offdiagonal part has small norm.

Theorem 2 (First lower bound)**.** *Let $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ be symmetric positive semi-definite with* $\operatorname{rank}(\boldsymbol{B}) \geq n-1$*, and partition*

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{b} \\ \boldsymbol{b}^T & \beta \end{bmatrix} \qquad \textit{where} \quad \boldsymbol{B}_{11} \in \mathbb{R}^{(n-1)\times(n-1)}.$$

If $\beta < \lambda_{min}(\boldsymbol{B}_{11})$ then

$$\lambda_{min}(\boldsymbol{B}) \geq \beta - \boldsymbol{b}^T\boldsymbol{B}_{11}^{-1}\boldsymbol{b} - \frac{\beta\,\|\boldsymbol{B}_{11}^{-1}\boldsymbol{b}\|_2^2}{1 - \beta\,\|\boldsymbol{B}_{11}^{-1}\|_2}.$$

Proof. Abbreviate $\tilde{\lambda}_{\min} \equiv \lambda_{\min}(\boldsymbol{B})$. From (2) follows that $\boldsymbol{B}_{11}$ and $\boldsymbol{B}_{11} - \tilde{\lambda}_{\min}\boldsymbol{I}$ are nonsingular. Combined with the symmetric positive semi-definiteness of $\boldsymbol{B}_{11}$ this gives

$$\tilde{\lambda}_{\min} < \lambda_{\min}(\boldsymbol{B}_{11}) = 1/\|\boldsymbol{B}_{11}^{-1}\|_2,$$

hence

$$\|\,\tilde{\lambda}_{\min}\,\boldsymbol{B}_{11}^{-1}\|_2 < 1. \tag{3}$$

Thus we can apply the Sherman-Morrison formula [GV13, Section 2.1.4],

$$(\boldsymbol{B}_{11} - \tilde{\lambda}_{\min}\boldsymbol{I})^{-1} = \boldsymbol{B}_{11}^{-1} + \tilde{\lambda}_{\min}\,\boldsymbol{B}_{11}^{-1}(\boldsymbol{I} - \tilde{\lambda}_{\min}\,\boldsymbol{B}_{11}^{-1})^{-1}\boldsymbol{B}_{11}^{-1},$$

and substitute the above into the expression for $\tilde{\lambda}_{\min}$ from Lemma 1,

$$\tilde{\lambda}_{\min} = \beta - \boldsymbol{b}^T\boldsymbol{B}_{11}^{-1}\boldsymbol{b} - \tilde{\lambda}_{\min}\,\boldsymbol{b}^T\boldsymbol{B}_{11}^{-1}(\boldsymbol{I} - \tilde{\lambda}_{\min}\,\boldsymbol{B}_{11}^{-1})^{-1}\boldsymbol{B}_{11}^{-1}\boldsymbol{b}. \tag{4}$$

The symmetric positive semi-definiteness of $\boldsymbol{B}$ implies that $\beta \geq 0$ and $\boldsymbol{b}^T\boldsymbol{B}_{11}^{-1}\boldsymbol{b} \geq 0$, hence it remains to bound the norm of the remaining summand. From the symmetry of $\boldsymbol{B}_{11}$ and the invariance of the two-norm under transposition follows

$$\begin{aligned}\|\boldsymbol{b}^T\boldsymbol{B}_{11}^{-1}(\boldsymbol{I} - \tilde{\lambda}_{\min}\,\boldsymbol{B}_{11}^{-1})^{-1}\boldsymbol{B}_{11}^{-1}\boldsymbol{b}\|_2 &\leq \|\boldsymbol{b}^T\boldsymbol{B}_{11}^{-1}\|_2\|(\boldsymbol{I} - \tilde{\lambda}_{\min}\,\boldsymbol{B}_{11}^{-1})^{-1}\|_2\|\boldsymbol{B}_{11}^{-1}\boldsymbol{b}\|_2 \\ &\leq \|\boldsymbol{B}_{11}^{-1}\boldsymbol{b}\|_2^2\|(\boldsymbol{I} - \tilde{\lambda}_{\min}\,\boldsymbol{B}_{11}^{-1})^{-1}\|_2. \end{aligned} \tag{5}$$

The inequality (3) allows us to apply the Banach lemma [GV13, Lemma 2.3.3] to bound the norm of the inverse by

$$\|(\boldsymbol{I} - \tilde{\lambda}_{\min}\,\boldsymbol{B}_{11}^{-1})^{-1}\|_2 \leq \frac{1}{1 - \|\,\tilde{\lambda}_{\min}\,\boldsymbol{B}_{11}^{-1}\|_2} = \frac{1}{1 - \tilde{\lambda}_{\min}\,\|\boldsymbol{B}_{11}^{-1}\|_2}.$$

Substitute this into (5) and the resulting bound into the expression for $\tilde{\lambda}_{\min}$ in (4),

$$\tilde{\lambda}_{\min} \geq \beta - \boldsymbol{b}^T\boldsymbol{B}_{11}^{-1}\boldsymbol{b} - \frac{\tilde{\lambda}_{\min}\,\|\boldsymbol{B}_{11}^{-1}\boldsymbol{b}\|_2^2}{1 - \tilde{\lambda}_{\min}\,\|\boldsymbol{B}_{11}^{-1}\|_2},$$

and at last apply the upper bound (1). □

The lower bound in Theorem 2 is positive if $\|\boldsymbol{b}\|_2$ is sufficiently small, in which case $\lambda_{\min}(\boldsymbol{B}) \geq \beta - \mathcal{O}(\|\boldsymbol{b}\|_2^2)$. If $\boldsymbol{b} = \boldsymbol{0}$ then (1) and Theorem 2 imply $\lambda_{\min}(\boldsymbol{B}) = \beta$.

The slightly weaker bound below focusses on a 'dominant part' of $\boldsymbol{B}_{11}$.

Theorem 3 (Second lower bound)**.** *Let $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ be symmetric positive semi-definite with* $\operatorname{rank}(\boldsymbol{B}) \geq n-1$*, and partition*

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{b} \\ \boldsymbol{b}^T & \beta \end{bmatrix} \qquad \textit{where} \quad \boldsymbol{B}_{11} \in \mathbb{R}^{(n-1)\times(n-1)}.$$

If $\boldsymbol{B}_{11} = \boldsymbol{C}_{11} + \boldsymbol{C}_{12}$ where $\boldsymbol{C}_{11}$ is symmetric positive definite with $\lambda_{min}(\boldsymbol{C}_{11}) > \beta$, and $\boldsymbol{C}_{12}$ is symmetric positive semi-definite then

$$\lambda_{min}(\boldsymbol{B}) \geq \beta - \boldsymbol{b}^T\boldsymbol{C}_{11}^{-1}\boldsymbol{b} - \frac{\beta\,\|\boldsymbol{C}_{11}^{-1}\boldsymbol{b}\|_2^2}{1 - \beta\,\|\boldsymbol{C}_{11}^{-1}\|_2}.$$

Proof. Abbreviate $\tilde{\lambda}_{\min} \equiv \lambda_{\min}(\boldsymbol{B})$. From (1) and the assumption follows $\tilde{\lambda}_{\min} \leq \beta < \lambda_{\min}(\boldsymbol{C}_{11})$, hence $\boldsymbol{C}_{11}$ and $\boldsymbol{C}_{11} - \tilde{\lambda}_{\min}\boldsymbol{I}$ are nonsingular. Write

$$\boldsymbol{B}_{11} - \tilde{\lambda}_{\min}\boldsymbol{I} = \underbrace{\boldsymbol{C}_{11} - \tilde{\lambda}_{\min}\boldsymbol{I}}_{\boldsymbol{G}} + \boldsymbol{C}_{12} = \boldsymbol{G}^{1/2}\,(\boldsymbol{I} + \underbrace{\boldsymbol{G}^{-1/2}\,\boldsymbol{C}_{12}\,\boldsymbol{G}^{-1/2}}_{\boldsymbol{H}})\,\boldsymbol{G}^{1/2},$$

where $\boldsymbol{G}$ is symmetric positive definite, $\boldsymbol{G}^{1/2}$ is its symmetric positive definite square root, and $\boldsymbol{H}$ is symmetric positive semi-definite. The Loewner partial ordering[3] implies $\boldsymbol{I} \preceq \boldsymbol{I} + \boldsymbol{H}$. From [HJ13, Corollary 7.7.4] follows $(\boldsymbol{I} + \boldsymbol{H})^{-1} \preceq \boldsymbol{I}^{-1} = \boldsymbol{I}$. Thus

$$\begin{aligned}(\boldsymbol{B}_{11} - \tilde{\lambda}_{\min}\,\boldsymbol{I})^{-1} &= \boldsymbol{G}^{-1/2}(\boldsymbol{I} + \boldsymbol{H})^{-1}\boldsymbol{G}^{-1/2} \\ &\preceq \boldsymbol{G}^{-1/2}\boldsymbol{I}\boldsymbol{G}^{-1/2} = \boldsymbol{G}^{-1} = (\boldsymbol{C}_{11} - \tilde{\lambda}_{\min}\,\boldsymbol{I})^{-1}.\end{aligned}$$

Substituting $(\boldsymbol{B}_{11} - \tilde{\lambda}_{\min}\,\boldsymbol{I})^{-1} \preceq (\boldsymbol{C}_{11} - \tilde{\lambda}_{\min}\,\boldsymbol{I})^{-1}$ into the expression for $\tilde{\lambda}_{\min}$ in Lemma 1 gives

$$\tilde{\lambda}_{\min} \geq \beta - \boldsymbol{b}^T(\boldsymbol{C}_{11} - \tilde{\lambda}_{\min}\,\boldsymbol{I})^{-1}\boldsymbol{b}.$$

We continue as in the proof of Theorem 2 with the Sherman-Morrison formula [GV13, Section 2.1.4],

$$\begin{aligned}\tilde{\lambda}_{\min} &\geq \beta - \boldsymbol{b}^T\boldsymbol{C}_{11}^{-1}\boldsymbol{b} - \tilde{\lambda}_{\min}\;\boldsymbol{b}^T\boldsymbol{C}_{11}^{-1}(\boldsymbol{I} - \tilde{\lambda}_{\min}\;\boldsymbol{C}_{11}^{-1})^{-1}\boldsymbol{C}_{11}^{-1}\boldsymbol{b} \\ &\geq \beta - \boldsymbol{b}^T\boldsymbol{C}_{11}^{-1}\boldsymbol{b} - \frac{\tilde{\lambda}_{\min}\;\|\boldsymbol{C}_{11}^{-1}\boldsymbol{b}\|_2^2}{1 - \tilde{\lambda}_{\min}\;\|\boldsymbol{C}_{11}^{-1}\|_2},\end{aligned}$$

and at last apply (1). □

If $\boldsymbol{C}_{12} = \boldsymbol{0}$, then Theorem 3 reduces to Theorem 2.

2.2 A lower bound for the smallest singular value

We consider a matrix with a distinct smallest singular value. Based on the eigenvalue bounds in section 2.1, we derive a lower bound for the smallest singular value of a perturbed matrix (Theorem 4) in terms of normwise absolute perturbations. We start with a summary of all assumptions (Assumptions 2.1), and end with a discussion of their generality (Remark 2.1).

Assumptions 2.1. *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with $m \geq n$ have $\operatorname{rank}(\boldsymbol{A}) \geq n-1$. Let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be a full singular value decomposition, where $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times n}$ is diagonal, and $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ are orthogonal matrices. Partition commensurately,*

$$\boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_{min} \\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}, \qquad \boldsymbol{E} = \boldsymbol{U}\begin{bmatrix}\boldsymbol{E}_{11} & \boldsymbol{e}_{12} \\ \boldsymbol{e}_{21}^T & e_{22} \\ \boldsymbol{E}_{31} & \boldsymbol{e}_{32}\end{bmatrix}\boldsymbol{V}^T,$$

where $\boldsymbol{\Sigma}_1 \in \mathbb{R}^{(n-1)\times(n-1)}$ is nonsingular diagonal, and $\sigma_{min} \geq 0$.

For a matrix with a single smallest singular value, we corroborate the observation that 'small singular values tend to increase' [SS90, page 266]. Motivated by the second-order expressions in terms of absolute perturbations [SS90, Section V.4.2] and [Ste06, Theorem 8], we derive a true lower bound.

Theorem 4. *Let $\boldsymbol{A}, \boldsymbol{E} \in \mathbb{R}^{m\times n}$ satisfy Assumptions 2.1. If $1/\|\boldsymbol{\Sigma}_1^{-1}\|_2 > 4\|\boldsymbol{E}\|_2$ and $\sigma_{min} < \|\boldsymbol{E}\|_2$, then*

$$\sigma_{min}(\boldsymbol{A} + \boldsymbol{E})^2 \geq \|\boldsymbol{e}_{32}\|_2^2 + (\sigma_{min} + e_{22})^2 - r_3 - r_4,$$

where r_3 contains terms of order 3,

$$r_3 \equiv 2\boldsymbol{e}_{12}^T\mathbf{w} \qquad \mathbf{w} \equiv (\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})^{-T}\begin{bmatrix}\boldsymbol{e}_{21} & \boldsymbol{E}_{31}^T\end{bmatrix}\begin{bmatrix}e_{22} + \sigma_{min} \\ \boldsymbol{e}_{32}\end{bmatrix},$$

and r_4 contains terms of order 4 and higher,

$$r_4 \equiv \|\mathbf{w}\|_2^2 + 4\frac{\|\boldsymbol{E}\|_2^2\,\|(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})^{-1}(\boldsymbol{e}_{12} + \mathbf{w})\|_2^2}{1 - 4\|\boldsymbol{E}\|_2^2\|(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})^{-1}\|_2^2}.$$

[3] For Hermitian matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A} \preceq \boldsymbol{B}$ means that $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite.

Proof. We square the singular values of $\boldsymbol{A}+\boldsymbol{E}$, and consider the eigenvalues of

$$\boldsymbol{B} \equiv (\boldsymbol{A}+\boldsymbol{E})^T(\boldsymbol{A}+\boldsymbol{E}) = \boldsymbol{V}\begin{bmatrix}\boldsymbol{B}_{11} & \boldsymbol{b}\\ \boldsymbol{b}^T & \beta\end{bmatrix}\boldsymbol{V}^T$$

where

$$\boldsymbol{B}_{11} = \underbrace{(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^T(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})}_{\boldsymbol{C}_{11}} + \underbrace{\boldsymbol{e}_{21}\boldsymbol{e}_{21}^T+\boldsymbol{E}_{31}^T\boldsymbol{E}_{31}}_{\boldsymbol{C}_{12}} \tag{6}$$

$$\begin{aligned}\beta &= \|\boldsymbol{e}_{12}\|_2^2 + (\sigma_{\min}+e_{22})^2 + \|\boldsymbol{e}_{32}\|_2^2\\ \boldsymbol{b} &= (\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^T\boldsymbol{e}_{12} + \boldsymbol{e}_{21}(\sigma_{\min}+e_{22}) + \boldsymbol{E}_{31}^T\boldsymbol{e}_{32}.\end{aligned} \tag{7}$$

From $\sigma_{\min}(\boldsymbol{\Sigma}_1) > 4\|\boldsymbol{E}\|_2$ follows that $\boldsymbol{C}_{11}$ is symmetric positive definite, while $\boldsymbol{C}_{12}$ is symmetric positive semi-definite and contains only second order terms. Abbreviate $\tilde{\lambda}_{\min} \equiv \lambda_{\min}(\boldsymbol{B}) = \sigma_{\min}(\boldsymbol{A}+\boldsymbol{E})^2$.

The proof proceeds in two steps:

1. Confirming that $\boldsymbol{C}_{11}$ satisfies the assumptions of Theorem 3.
2. Deriving the lower bound for $\tilde{\lambda}_{\min}$ from Theorem 3.

1. Confirm that $\boldsymbol{C}_{11}$ satisfies the assumptions of Theorem 3 We show that $\lambda_{\min}(\boldsymbol{C}_{11}) > \beta$, by bounding β from above and $\lambda_{\min}(\boldsymbol{C}_{11})$ from below.

Regarding the upper bound for β, the expression (7) and the assumption $\sigma_{\min} < \|\boldsymbol{E}\|_2$ imply

$$\beta = \left\| \begin{bmatrix}\boldsymbol{e}_{12}^T & e_{22}+\sigma_{\min} & \boldsymbol{e}_{32}^T\end{bmatrix}^T \right\|_2^2 \leq (\sigma_{\min}+\|\boldsymbol{E}\boldsymbol{e}_n\|_2)^2 \leq 4\|\boldsymbol{E}\|_2^2. \tag{8}$$

Regarding the lower bound for $\lambda_{\min}(\boldsymbol{C}_{11})$, view $\boldsymbol{C}_{11} = (\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^T(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})$ as a singular value problem, so that $\lambda_{\min}(\boldsymbol{C}_{11}) = \sigma_{\min}(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^2$. The well-conditioning of singular values [GV13, Corollary 8.6.2] implies

$$|\,\sigma_{\min}(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11}) - \sigma_{\min}(\boldsymbol{\Sigma}_1)| \leq \|\boldsymbol{E}_{11}\|_2 \leq \|\boldsymbol{E}\|_2.$$

Adding the assumption $\sigma_{\min}(\boldsymbol{\Sigma}_1) = 1/\|\boldsymbol{\Sigma}_1^{-1}\|_2 > 4\|\boldsymbol{E}\|_2$ gives

$$\sigma_{\min}(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11}) \geq \sigma_{\min}(\boldsymbol{\Sigma}_1) - \|\boldsymbol{E}\|_2 > 4\,\|\boldsymbol{E}\|_2 - \|\boldsymbol{E}\|_2 = 3\|\boldsymbol{E}\|_2.$$

Now combine this lower bound for $\lambda_{\min}(\boldsymbol{C}_{11})$ with (8),

$$\lambda_{\min}(\boldsymbol{C}_{11}) = \sigma_{\min}(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^2 > 9\|\boldsymbol{E}\|_2^2 > 4\|\boldsymbol{E}\|_2^2 \geq \beta.$$

Hence $\lambda_{\min}(\boldsymbol{C}_{11}) > \beta$, and $\boldsymbol{C}_{11}$ satisfies the assumptions of Theorem 3.

2. Derive the lower bound for $\tilde{\lambda}_{\min}$ from Theorem 3 In this bound,

$$\tilde{\lambda}_{\min} \geq \beta - \boldsymbol{b}^T\boldsymbol{C}_{11}^{-1}\boldsymbol{b} - \frac{\beta\,\|\boldsymbol{C}_{11}^{-1}\boldsymbol{b}\|_2^2}{1-\beta\,\|\boldsymbol{C}_{11}^{-1}\|_2}, \tag{9}$$

where the key term is $\boldsymbol{C}_{11}^{-1}\boldsymbol{b}$. Insert the expression for $\boldsymbol{b}$ from (7),

$$\begin{aligned}(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^{-T}\boldsymbol{b} &= \boldsymbol{e}_{12} + (\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^{-T}\left(\boldsymbol{e}_{21}(\sigma_{\min}+e_{22}) + \boldsymbol{E}_{31}^T\boldsymbol{e}_{32}\right)\\ &= \boldsymbol{e}_{12}+\mathbf{w}.\end{aligned} \tag{10}$$

Combine the expression for $\boldsymbol{C}_{11}$ from (6) with (10),

$$\boldsymbol{C}_{11}^{-1}\boldsymbol{b} = (\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^{-1}\underbrace{(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^{-T}\boldsymbol{b}}_{\boldsymbol{e}_{12}+\mathbf{w}} = (\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^{-1}(\boldsymbol{e}_{12}+\mathbf{w}) \tag{11}$$

Multiply the above by $\boldsymbol{b}^T$ on the left, and use (10)

$$\begin{aligned}\boldsymbol{b}^T\boldsymbol{C}_{11}^{-1}\boldsymbol{b} &= \boldsymbol{b}^T(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^{-1}(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^{-T}\boldsymbol{b} = (\boldsymbol{e}_{12}+\mathbf{w})^T(\boldsymbol{e}_{12}+\mathbf{w})\\ &= \|\boldsymbol{e}_{12}+\mathbf{w}\|_2^2 = \|\boldsymbol{e}_{12}\|_2^2 + 2\boldsymbol{e}_{12}^T\mathbf{w} + \|\mathbf{w}\|_2^2.\end{aligned}$$

Substitute the above, and β from (7) into the first two summands of (9),

$$\begin{aligned}\beta - \boldsymbol{b}^T\boldsymbol{C}_{11}^{-1}\boldsymbol{b} &= (\|\boldsymbol{e}_{12}\|_2^2 + (\sigma_{\min}+e_{22})^2 + \|\boldsymbol{e}_{32}\|_2^2) - (\|\boldsymbol{e}_{12}\|_2^2 + 2\boldsymbol{e}_{12}^T\mathbf{w} + \|\mathbf{w}\|_2^2)\\ &= \|\boldsymbol{e}_{32}\|_2^2 + (\sigma_{\min}+e_{22})^2 - \underbrace{2\boldsymbol{e}_{12}^T\mathbf{w}}_{r_3} - \|\mathbf{w}\|_2^2. \qquad (12)\end{aligned}$$

Substitute the bound for β in (8), and (11) into the third summand of (9),

$$\frac{\beta\,\|\boldsymbol{C}_{11}^{-1}\boldsymbol{b}\|_2^2}{1-\beta\,\|\boldsymbol{C}_{11}^{-1}\|_2} \le 4\frac{\|\boldsymbol{E}\|_2^2\,\|(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^{-1}(\boldsymbol{e}_{12}+\mathbf{w})\|_2^2}{1-4\|\boldsymbol{E}\|_2^2\|(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^{-1}\|_2^2}. \qquad (13)$$

Inserting (12) and (13) into (9) gives

$$\begin{aligned}\tilde{\lambda}_{\min} \ge\ & \|\boldsymbol{e}_{32}\|_2^2 + (\sigma_{\min}+e_{22})^2\\ & - r_3 - \underbrace{\left(\|\mathbf{w}\|_2^2 + 4\frac{\|\boldsymbol{E}\|_2^2\,\|(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^{-1}(\boldsymbol{e}_{12}+\mathbf{w})\|_2^2}{1-4\|\boldsymbol{E}\|_2^2\|(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^{-1}\|_2^2}\right)}_{r_4}.\end{aligned}$$

□

Remark 2.1. *The assumptions in Theorem 4 are not restrictive. Only a small gap of* $3\|\boldsymbol{E}\|_2$ *is required to separate the smallest singular value of* $\boldsymbol{A}$ *from the remaining singular values,*

$$\sigma_{min}(\boldsymbol{A}) < \|\boldsymbol{E}\|_2 < 4\|\boldsymbol{E}\|_2 \le 1/\|\boldsymbol{\Sigma}_1^{-1}\|_2.$$

3 A cluster of small singular values

We extend the results in Section 2 from a single smallest singular value to a cluster of small singular values. To this end, we derive lower bounds for the small singular values of the perturbed matrix in terms of normwise absolute perturbations (Section 3.2), based on eigenvalue bounds (Section 3.1).

3.1 Auxiliary eigenvalue results

We square the singular values of $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and consider instead the eigenvalues of the symmetric positive semi-definite matrix $\boldsymbol{B} \equiv \boldsymbol{A}^T\boldsymbol{A} \in \mathbb{R}^{n\times n}$.

For a symmetric positive semi-definite matrix $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ with a cluster of r small eigenvalues, we present an expression for these eigenvalues (Lemma 3), and two lower bounds in terms of normwise absolute perturbations (Theorems 5 and 6).

We assume that the r small eigenvalues are separated from the remaining ones, in the sense that they are strictly smaller than the smallest eigenvalue of the leading principal submatrix $\boldsymbol{B}_{11}$ of order $n-r$. The eigenvalues are labelled so that

$$\lambda_n(\boldsymbol{B}) \le \cdots \le \lambda_{n-r+1}(\boldsymbol{B}) < \lambda_{n-r}(\boldsymbol{B}) \le \cdots \le \lambda_1(\boldsymbol{B}).$$

The equality below expresses the smallest eigenvalues in terms of themselves, and represents an extension of Lemma 2 to clusters.

Lemma 3 (Exact expression)**.** *Let* $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ *be symmetric positive semi-definite with* $\operatorname{rank}(\boldsymbol{B}) \ge n-r$ *for some* $r \ge 1$*, and partition*

$$\boldsymbol{B} = \begin{bmatrix}\boldsymbol{B}_{11} & \boldsymbol{B}_{12}\\ \boldsymbol{B}_{12}^T & \boldsymbol{B}_{22}\end{bmatrix} \qquad \text{where} \quad \boldsymbol{B}_{11} \in \mathbb{R}^{(n-r)\times(n-r)}, \quad \boldsymbol{B}_{22} \in \mathbb{R}^{r\times r}.$$

If $\|\boldsymbol{B}_{22}\|_2 < \lambda_{min}(\boldsymbol{B}_{11})$ *then*

$$\lambda_{n-r+j}(\boldsymbol{B}) = \lambda_j\left(\boldsymbol{B}_{22} - \boldsymbol{B}_{12}^T(\boldsymbol{B}_{11} - \lambda_{n-r+j}(\boldsymbol{B})\,\boldsymbol{I})^{-1}\boldsymbol{B}_{12}\right), \qquad 1 \le j \le r,$$

where

$$0 \le \lambda_{n-r+j}(\boldsymbol{B}) \le \|\boldsymbol{B}_{22}\|_2, \qquad 1 \le j \le r. \tag{14}$$

Proof. Abbreviate $\tilde{\lambda}_{n-r+j} \equiv \lambda_{n-r+j}(\boldsymbol{B})$, $1 \le j \le r$. The lower bound in (14) follows from the positive semi-definiteness of $\boldsymbol{B}$, and the upper bound from the Cauchy interlace theorem [Par80, Section 10.1]

$$\tilde{\lambda}_{n-r+j} \le \lambda_j(\boldsymbol{B}_{22}) \le \lambda_{\max}(\boldsymbol{B}_{22}) = \|\boldsymbol{B}_{22}\|_2, \qquad 1 \le j \le r.$$

Combining this with the assumption $\|\boldsymbol{B}_{22}\|_2 < \lambda_{\min}(\boldsymbol{B}_{11})$ shows

$$\tilde{\lambda}_{n-r+j} \le \|\boldsymbol{B}_{22}\|_2 < \lambda_{\min}(\boldsymbol{B}_{11}), \qquad 1 \le j \le r. \tag{15}$$

Hence is $\boldsymbol{B}_{11} - \tilde{\lambda}_{n-r+j}\,\boldsymbol{I}$ is nonsingular.

To derive the expression for $\tilde{\lambda}_{n-r+j}$, we start as in the proof of [Par80, Theorem (10-1-2)]. The shifted matrix $\boldsymbol{B} - \tilde{\lambda}_{n-r+j}\,\boldsymbol{I}$ has at most $n-r+j-1$ positive eigenvalues, at least one zero eigenvalue, and at most $r-j$ negative eigenvalues, $1 \le j \le r$. Perform the congruence transformation

$$\boldsymbol{B} - \tilde{\lambda}_{n-r+j}\,\boldsymbol{I} = \boldsymbol{L}\begin{bmatrix}\boldsymbol{B}_{11} - \tilde{\lambda}_{n-r+j}\,\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}\end{bmatrix}\boldsymbol{L}^T, \qquad \boldsymbol{L} \equiv \begin{bmatrix}\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{B}_{12}^T(\boldsymbol{B}_{11} - \tilde{\lambda}_{n-r+j}\,\boldsymbol{I})^{-1} & \boldsymbol{I}\end{bmatrix}$$

where

$$\boldsymbol{S} \equiv \boldsymbol{B}_{22} - \tilde{\lambda}_{n-r+j}\,\boldsymbol{I} - \boldsymbol{B}_{12}^T(\boldsymbol{B}_{11} - \tilde{\lambda}_{n-r+j}\,\boldsymbol{I})^{-1}\boldsymbol{B}_{12}, \qquad 1 \le j \le r. \tag{16}$$

From (15) follows that $\boldsymbol{B}_{11} - \tilde{\lambda}_{n-r+j}\,\boldsymbol{I}$ has $n-r$ positive eigenvalues. Combining this with the inertia preservation of congruence transformations implies that $\boldsymbol{S}$ has at most $r-j$ positive eigenvalues, at least one zero eigenvalue $\lambda_j(\boldsymbol{S}) = 0$, and at least $j-1$ negative eigenvalues, $1 \le j \le r$. Insert (16) into $\lambda_j(\boldsymbol{S}) = 0$, and exploit the fact that the shift $\tilde{\lambda}_{n-r+j}\,\boldsymbol{I}$ does not change the algebraic eigenvalue ordering, to obtain the expression for $\tilde{\lambda}_{n-r+j}$, $1 \le j \le r$. □

By restricting ourselves to a 'dominant part' of $\boldsymbol{B}_{11}$, we weaken the expression in Lemma 3 to a lower bound, which allows the eigenvalues to be negative.

Lemma 4 (Lower bound)**.** *Let* $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ *be symmetric positive semi-definite with* $\operatorname{rank}(\boldsymbol{B}) \ge n-r$ *for some* $r \ge 1$*, and partition*

$$\boldsymbol{B} = \begin{bmatrix}\boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{12}^T & \boldsymbol{B}_{22}\end{bmatrix} \qquad \text{where} \quad \boldsymbol{B}_{11} \in \mathbb{R}^{(n-r)\times(n-r)}, \quad \boldsymbol{B}_{22} \in \mathbb{R}^{r\times r}.$$

Let $\boldsymbol{B}_{11} = \boldsymbol{C}_{11} + \boldsymbol{C}_{12}$ *where* $\boldsymbol{C}_{11} \in \mathbb{R}^{(n-r)\times(n-r)}$ *is symmetric positive definite and* $\boldsymbol{C}_{12} \in \mathbb{R}^{(n-r)\times(n-r)}$ *is symmetric positive semi-definite. If*

$$\widehat{\boldsymbol{B}} \equiv \begin{bmatrix}\boldsymbol{C}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{12}^T & \boldsymbol{B}_{22}\end{bmatrix}$$

with $\lambda_{min}(\boldsymbol{C}_{11}) > \|\boldsymbol{B}_{22}\|_2$*, then*

$$\lambda_{n-r+j}(\boldsymbol{B}) \ge \lambda_{n-r+j}(\widehat{\boldsymbol{B}}) \tag{17}$$

$$= \lambda_j\left(\boldsymbol{B}_{22} - \boldsymbol{B}_{12}^T(\boldsymbol{C}_{11} - \lambda_{n-r+j}(\widehat{\boldsymbol{B}})\,\boldsymbol{I})^{-1}\boldsymbol{B}_{12}\right), \quad 1 \le j \le r, \tag{18}$$

where

$$\lambda_{n-r+j}(\widehat{\boldsymbol{B}}) \le \|\boldsymbol{B}_{22}\|_2, \tag{19}$$

$$\left\|\left(\boldsymbol{C}_{11} - \lambda_{n-r+j}(\widehat{\boldsymbol{B}})\,\boldsymbol{I}\right)^{-1}\right\|_2 \le \frac{\|\boldsymbol{C}_{11}^{-1}\|_2}{1 - \|\boldsymbol{B}_{22}\|_2\,\|\boldsymbol{C}_{11}^{-1}\|_2}, \qquad 1 \le j \le r. \tag{20}$$

Proof. The proof proceeds in four steps.

Proof of (17) The symmetric positive semi-definiteness of $\boldsymbol{C}_{12}$ and Weyl's monotonicity theorem [HJ13, Corollary 4.3.3] imply

$$\lambda_j(\boldsymbol{B}) \geq \lambda_j(\widehat{\boldsymbol{B}}), \qquad 1 \leq j \leq n.$$

Now we concentrate on the eigenvalues of $\widehat{\boldsymbol{B}}$, and abbreviate $\widehat{\lambda}_{n-r+j} \equiv \lambda_{n-r+j}(\widehat{\boldsymbol{B}})$, $1 \leq j \leq r$.

Proof of (19) Apply the Cauchy interlace theorem [Par80, Section 10.1] to $\widehat{\boldsymbol{B}}$,

$$\widehat{\lambda}_{n-r+j} \leq \lambda_j(\boldsymbol{B}_{22}) \leq \lambda_{\max}(\boldsymbol{B}_{22}) = \|\boldsymbol{B}_{22}\|_2, \qquad 1 \leq j \leq r.$$

Combining this with the assumption $\|\boldsymbol{B}_{22}\|_2 < \lambda_{\min}(\boldsymbol{C}_{11})$ shows

$$\widehat{\lambda}_{n-r+j} \leq \|\boldsymbol{B}_{22}\|_2 < \lambda_{\min}(\boldsymbol{C}_{11}), \qquad 1 \leq j \leq r.$$

Hence $\boldsymbol{C}_{11} - \widehat{\lambda}_{n-r+j}\, \boldsymbol{I}$ is nonsingular, which holds in particular if $\widehat{\lambda}_{n-r+j} < 0$.

Proof of (18) To derive the expression for $\widehat{\lambda}_{n-r+j}$, apply the proof of Lemma 3 to the eigenvalues of $\widehat{\boldsymbol{B}}$. This proof relies only on the signs of eigenvalues of shifted matrices, and does not require positive semi-definiteness of the host matrix $\widehat{\boldsymbol{B}}$.

Proof of (20) Fix some $1 \leq j \leq r$ for the inverse in (18). Then factor out $\boldsymbol{C}_{11}^{-1}$,

$$(\boldsymbol{C}_{11} - \widehat{\lambda}_{n-r+j}\, \boldsymbol{I})^{-1} = \boldsymbol{C}_{11}^{-1}\boldsymbol{D} \qquad \text{where} \quad \boldsymbol{D} \equiv (\boldsymbol{I} - \widehat{\lambda}_{n-r+j}\, \boldsymbol{C}_{11}^{-1})^{-1},$$

and take norms,

$$\left\| (\boldsymbol{C}_{11} - \widehat{\lambda}_{n-r+j}\, \boldsymbol{I})^{-1} \right\|_2 \leq \|\boldsymbol{C}_{11}^{-1}\|_2\, \|\boldsymbol{D}\|_2.$$

To bound $\|\boldsymbol{D}\|_2$, consider the eigenvalue decomposition $\boldsymbol{C}_{11} = \boldsymbol{W}\boldsymbol{\Lambda}\boldsymbol{W}^T$, where $\boldsymbol{W}$ is an orthogonal matrix, and the diagonal matrix

$$\boldsymbol{\Lambda} = \operatorname{diag}\begin{pmatrix}\gamma_1 & \cdots & \gamma_{n-r}\end{pmatrix} \in \mathbb{R}^{(n-r)\times(n-r)}$$

has positive diagonal elements $\gamma_\ell > 0$. Thus $\boldsymbol{D}$ has an eigenvalue decomposition $\boldsymbol{D} = \boldsymbol{W}(\boldsymbol{I} - \widehat{\lambda}_{n-r+j}\, \boldsymbol{\Lambda}^{-1})^{-1}\boldsymbol{W}^T$ with eigenvalues

$$\lambda_\ell(\boldsymbol{D}) = 1/\left(1 - \frac{\widehat{\lambda}_{n-r+j}}{\gamma_\ell}\right), \qquad 1 \leq \ell \leq n-r.$$

Case 1: If $\widehat{\lambda}_{n-r+j} \geq 0$, then (19) implies

$$0 \leq \frac{\widehat{\lambda}_{n-r+j}}{\gamma_\ell} \leq \frac{\|\boldsymbol{B}_{22}\|_2}{\lambda_{\min}(\boldsymbol{C}_{11})} = \|\boldsymbol{B}_{22}\|_2\|\boldsymbol{C}_{11}^{-1}\|_2 < 1, \qquad 1 \leq \ell \leq n-r.$$

Hence

$$\|\boldsymbol{D}\|_{22} = \max_{1\leq\ell\leq n-r} |\lambda_j(\boldsymbol{D})| \leq \frac{1}{1 - \|\boldsymbol{B}_{22}\|_2\|\boldsymbol{C}_{11}^{-1}\|_2}. \tag{21}$$

Case 2: If $\widehat{\lambda}_{n-r+j} < 0$, then $\gamma_\ell > 0$ and (19) imply

$$1 - \frac{\widehat{\lambda}_{n-r+j}}{\gamma_\ell} = 1 + \frac{|\,\widehat{\lambda}_{n-r+j}\,|}{\gamma_\ell} > 1 > 1 - \|\boldsymbol{B}_{22}\|_2\|\boldsymbol{C}_{11}^{-1}\|_2, \qquad 1 \leq \ell \leq n-r.$$

Again, as in (21) we conclude

$$\|\boldsymbol{D}\|_{22} = \max_{1\leq\ell\leq n-r} |\lambda_j(\boldsymbol{D})| \leq \frac{1}{1 - \|\boldsymbol{B}_{22}\|_2\|\boldsymbol{C}_{11}^{-1}\|_2}.$$

Since we fixed an arbitrary j to show (20), it holds for all $1 \leq j \leq r$. □

The subsequent lower bounds are informative if the offdiagonal part has small norm. We start with an extension of Theorem 2.

Theorem 5 (First lower bound)**.** *Let $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ be symmetric positive semi-definite with* $\operatorname{rank}(\boldsymbol{B}) \geq n-r$ *for some $r \geq 1$, and partition*

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{12}^T & \boldsymbol{B}_{22} \end{bmatrix} \qquad \text{where} \quad \boldsymbol{B}_{11} \in \mathbb{R}^{(n-r)\times(n-r)}, \quad \boldsymbol{B}_{22} \in \mathbb{R}^{r\times r}.$$

If $\|\boldsymbol{B}_{22}\|_2 < \lambda_{min}(\boldsymbol{B}_{11})$ then $\lambda_{n-r+j}(\boldsymbol{B}) \geq \lambda_j(\boldsymbol{Z}_j)$, $1 \leq j \leq r$, where

$$\boldsymbol{Z}_j \equiv \boldsymbol{B}_{22} - \boldsymbol{B}_{12}^T\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12} - \|\boldsymbol{B}_{22}\|_2\, \boldsymbol{B}_{12}^T\boldsymbol{B}_{11}^{-1}(\boldsymbol{I} - \lambda_{n-r+j}(\boldsymbol{B})\, \boldsymbol{B}_{11}^{-1})^{-1}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}.$$

Proof. Abbreviate $\tilde{\lambda}_{n-r+j} \equiv \lambda_{n-r+j}(\boldsymbol{B})$, $1 \leq j \leq r$. As in the proof of Theorem 2, apply the Sherman-Morrison formula [GV13, Section 2.1.4]

$$(\boldsymbol{B}_{11} - \tilde{\lambda}_{n-r+j}\, \boldsymbol{I})^{-1} = \boldsymbol{B}_{11}^{-1} + \tilde{\lambda}_{n-r+j}\, \boldsymbol{B}_{11}^{-1}(\boldsymbol{I} - \tilde{\lambda}_{n-r+j}\, \boldsymbol{B}_{11}^{-1})^{-1}\boldsymbol{B}_{11}^{-1},$$

and substitute the above into the expressions for

$$\tilde{\lambda}_{n-r+j} = \lambda_j(\boldsymbol{M}_j), \qquad 1 \leq j \leq r \tag{22}$$

from Lemma 3 where

$$\begin{aligned} \boldsymbol{M}_j &\equiv \boldsymbol{B}_{22} - \boldsymbol{B}_{12}^T(\boldsymbol{B}_{11} - \tilde{\lambda}_{n-r+j}\, \boldsymbol{I})^{-1}\boldsymbol{B}_{12} \\ &= \boldsymbol{B}_{22} - \boldsymbol{B}_{12}^T\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12} - \tilde{\lambda}_{n-r+j}\, \boldsymbol{B}_{12}^T\boldsymbol{B}_{11}^{-1}(\boldsymbol{I} - \tilde{\lambda}_{n-r+j}\, \boldsymbol{B}_{11}^{-1})^{-1}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}. \end{aligned}$$

From (14) follows the Loewner bound

$$\boldsymbol{M}_j \succeq \boldsymbol{Z}_j \equiv \boldsymbol{B}_{22} - \boldsymbol{B}_{12}^T\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12} - \|\boldsymbol{B}_{22}\|_2\, \boldsymbol{B}_{12}^T\boldsymbol{B}_{11}^{-1}(\boldsymbol{I} - \tilde{\lambda}_{r+j}\, \boldsymbol{B}_{11}^{-1})^{-1}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}.$$

This and (22) imply $\tilde{\lambda}_{n-r+j} = \lambda_j(\boldsymbol{M}_j) \geq \lambda_j(\boldsymbol{Z}_j)$, $1 \leq j \leq r$ [HJ13, Corollary 7.7.4]. □

The slightly weaker bound below extends Theorem 3 and focusses on a 'dominant' part of $\boldsymbol{B}_{11}$. This establishes the connection to Theorem 7, where $\boldsymbol{B}$ represents the perturbed matrix and the low order terms in $\boldsymbol{B}_{11}$ are captured by $\boldsymbol{C}_{11}$.

Theorem 6 (Second lower bound)**.** *Let $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ be symmetric positive semi-definite with* $\operatorname{rank}(\boldsymbol{B}) \geq n-r$ *for some $r \geq 1$, and partition*

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{12}^T & \boldsymbol{B}_{22} \end{bmatrix} \qquad \text{where} \quad \boldsymbol{B}_{11} \in \mathbb{R}^{(n-r)\times(n-r)}, \quad \boldsymbol{B}_{22} \in \mathbb{R}^{r\times r}.$$

If $\boldsymbol{B}_{11} = \boldsymbol{C}_{11} + \boldsymbol{C}_{12}$ where $\boldsymbol{C}_{11}$ is symmetric positive definite with $\lambda_{min}(\boldsymbol{C}_{11}) > \|\boldsymbol{B}_{22}\|_2$, and $\boldsymbol{C}_{12}$ is symmetric positive semi-definite, then

$$\lambda_{n-r+j}(\boldsymbol{B}) \geq \lambda_j\left(\boldsymbol{B}_{22} - \boldsymbol{B}_{12}^T\boldsymbol{C}_{11}^{-1}\boldsymbol{B}_{12}\right) - \frac{\|\boldsymbol{B}_{22}\|_2\, \|\boldsymbol{C}_{11}^{-1}\boldsymbol{B}_{12}\|_2^2}{1 - \|\boldsymbol{B}_{22}\|_2\, \|\boldsymbol{C}_{11}^{-1}\|_2}, \qquad 1 \leq j \leq r.$$

Proof. Define

$$\widehat{\boldsymbol{B}} \equiv \begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{12}^T & \boldsymbol{B}_{22} \end{bmatrix},$$

and abbreviate $\widehat{\lambda}_{n-r+j} \equiv \lambda_{n-r+j}(\widehat{\boldsymbol{B}})$, $1 \leq j \leq r$. From (18) in Lemma 4 follows

$$\lambda_{n-r+j}(\boldsymbol{B}) \geq \widehat{\lambda}_{n-r+j} = \lambda_j\left(\boldsymbol{B}_{22} - \boldsymbol{B}_{12}^T(\boldsymbol{C}_{11} - \widehat{\lambda}_{n-r+j}\, \boldsymbol{I})^{-1}\boldsymbol{B}_{12}\right), \quad 1 \leq j \leq r,$$

We proceed as in the proof of Theorem 5, and apply the Sherman-Morrison formula [GV13, Section 2.1.4],

$$(\boldsymbol{C}_{11} - \widehat{\lambda}_{n-r+j}\,\boldsymbol{I})^{-1} = \boldsymbol{C}_{11}^{-1} + \widehat{\lambda}_{n-r+j}\,\boldsymbol{C}_{11}^{-1}(\boldsymbol{I} - \widehat{\lambda}_{n-r+j}\,\boldsymbol{C}_{11}^{-1})^{-1}\boldsymbol{C}_{11}^{-1},$$

and (19) to the expression for

$$\widehat{\lambda}_{n-r+j} = \lambda_j(\boldsymbol{M}_j), \qquad 1 \le j \le r, \tag{23}$$

from Lemma 4, where

$$\begin{aligned}\boldsymbol{M}_j &\equiv \boldsymbol{B}_{22} - \boldsymbol{B}_{12}^T(\boldsymbol{C}_{11} - \widehat{\lambda}_{n-r+j}\,\boldsymbol{I})^{-1}\boldsymbol{B}_{12}, \qquad 1 \le j \le r \\ &= \boldsymbol{B}_{22} - \boldsymbol{B}_{12}^T\boldsymbol{C}_{11}^{-1}\boldsymbol{B}_{12} - \tilde{\lambda}_{n-r+j}\,\boldsymbol{B}_{12}^T\boldsymbol{C}_{11}^{-1}(\boldsymbol{I} - \tilde{\lambda}_{n-r+j}\,\boldsymbol{C}_{11}^{-1})^{-1}\boldsymbol{C}_{11}^{-1}\boldsymbol{B}_{12}.\end{aligned}$$

From (19) and (20) follows the lower bound

$$\begin{aligned}\boldsymbol{M}_j &\succeq \boldsymbol{B}_{22} - \boldsymbol{B}_{12}^T\boldsymbol{C}_{11}^{-1}\boldsymbol{B}_{12} - \|\boldsymbol{B}_{22}\|_2\,\boldsymbol{B}_{12}^T\boldsymbol{C}_{11}^{-1}(\boldsymbol{I} - \tilde{\lambda}_{r+j}\,\boldsymbol{C}_{11}^{-1})^{-1}\boldsymbol{C}_{11}^{-1}\boldsymbol{B}_{12} \\ &\succeq \boldsymbol{Z} \equiv \boldsymbol{B}_{22} - \boldsymbol{B}_{12}^T\boldsymbol{C}_{11}^{-1}\boldsymbol{B}_{12} - \underbrace{\frac{\|\boldsymbol{B}_{22}\|_2\,\|\boldsymbol{C}_{11}^{-1}\boldsymbol{B}_{12}\|_2^2}{1 - \|\boldsymbol{B}_{22}\|_2\,\|\boldsymbol{C}_{11}^{-1}\|_2}}_{\gamma}\boldsymbol{I}, \qquad 1 \le j \le r.\end{aligned}$$

Thus, $\boldsymbol{M}_j \succeq \boldsymbol{Z}$, $1 \le j \le r$. The Loewner properties [HJ13, Corollary 7.7.4] imply the same for the eigenvalues, $\lambda_j(\boldsymbol{M}_j) \ge \lambda_j(\boldsymbol{Z})$, $1 \le j \le r$. Combine this with the well conditioning of eigenvalues [GV13, Theorem 8.1.5],

$$\tilde{\lambda}_{n-r+j} = \lambda_j(\boldsymbol{M}_j) \ge \lambda_j(\boldsymbol{Z}) \ge \lambda_j(\boldsymbol{B}_{22} - \boldsymbol{B}_{12}^T\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}) - \gamma, \qquad 1 \le j \le r.$$

□

Theorem 6 reduces to Theorem 3 for $r = 1$, and to Theorem 5 for $\boldsymbol{C}_{12} = \boldsymbol{0}$.

3.2 A lower bound for a cluster of smallest singular values

We extend the bound for a single smallest singular value in section 2.2 to a cluster of smallest singular values. The resulting lower bound for the cluster of perturbed smallest singular values (Theorem 7) is based on the eigenvalue bounds in section 3.1, and expressed in terms of normwise absolute perturbations. We start with a summary of all assumptions (Assumptions 3.1), and end with a discussion of their generality (Remark 3.1).

Assumptions 3.1. *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with $m \ge n$ have $\operatorname{rank}(\boldsymbol{A}) \ge n - r$ for some $r \ge 1$. Let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be a full singular value decomposition, where $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times n}$ is diagonal, and $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ are orthogonal matrices. Partition commensurately,*

$$\boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_2 \\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}, \qquad \boldsymbol{E} = \boldsymbol{U}\begin{bmatrix}\boldsymbol{E}_{11} & \boldsymbol{E}_{12} \\ \boldsymbol{E}_{21} & \boldsymbol{E}_{22} \\ \boldsymbol{E}_{31} & \boldsymbol{E}_{32}\end{bmatrix}\boldsymbol{V}^T,$$

where $\boldsymbol{\Sigma}_1 \in \mathbb{R}^{(n-r)\times(n-r)}$ is nonsingular diagonal, and $\boldsymbol{\Sigma}_2 \in \mathbb{R}^{r\times r}$ is diagonal.

This bound below extends Theorem 4, and reduces to it for $r = 1$.

Theorem 7. *Let $\boldsymbol{A}, \boldsymbol{E} \in \mathbb{R}^{m\times n}$ satisfy Assumptions 3.1. If $1/\|\boldsymbol{\Sigma}_1^{-1}\|_2 > 4\|\boldsymbol{E}\|_2$ and $\|\boldsymbol{\Sigma}_2\|_2 < \|\boldsymbol{E}\|_2$, then*

$$\sigma_{n-r+j}(\boldsymbol{A} + \boldsymbol{E})^2 \ge \lambda_j(\boldsymbol{E}_{32}^T\boldsymbol{E}_{32} + (\boldsymbol{\Sigma}_2 + \boldsymbol{E}_{22})^T(\boldsymbol{\Sigma}_2 + \boldsymbol{E}_{22}) - \boldsymbol{R}_3) - r_4, \quad 1 \le j \le r,$$

where $\boldsymbol{R}_3$ contains terms of order 3

$$\boldsymbol{R}_3 \equiv \boldsymbol{E}_{12}^T\boldsymbol{W} + \boldsymbol{W}^T\boldsymbol{E}_{12}, \qquad \boldsymbol{W} \equiv (\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})^{-T}\begin{bmatrix}\boldsymbol{E}_{21} & \boldsymbol{E}_{31}^T\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_2 + \boldsymbol{E}_{22} \\ \boldsymbol{E}_{32}\end{bmatrix}$$

and r_4 contains terms of order 4 and higher,

$$r_4 \equiv \|\boldsymbol{W}\|_2^2 + 4\frac{\|\boldsymbol{E}\|_2^2\,\|(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})^{-1}(\boldsymbol{E}_{12} + \boldsymbol{W})\|_2^2}{1 - 4\|\boldsymbol{E}\|_2^2\|(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})^{-1}\|_2^2}.$$

Proof. We square the singular values of $\boldsymbol{A}+\boldsymbol{E}$ and consider the eigenvalues of the perturbed matrix

$$\boldsymbol{B} \equiv (\boldsymbol{A}+\boldsymbol{E})^T(\boldsymbol{A}+\boldsymbol{E}) = \boldsymbol{V}\begin{bmatrix}\boldsymbol{B}_{11} & \boldsymbol{B}_{12}\\ \boldsymbol{B}_{12}^T & \boldsymbol{B}_{22}\end{bmatrix}\boldsymbol{V}^T$$

where

$$\boldsymbol{B}_{11} = \underbrace{(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^T(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})}_{\boldsymbol{C}_{11}} + \underbrace{\boldsymbol{E}_{21}^T\boldsymbol{E}_{21}+\boldsymbol{E}_{31}^T\boldsymbol{E}_{31}}_{\boldsymbol{C}_{12}} \tag{24}$$

$$\begin{aligned}\boldsymbol{B}_{22} &= \boldsymbol{E}_{12}^T\boldsymbol{E}_{12} + (\boldsymbol{\Sigma}_2+\boldsymbol{E}_{22})^T(\boldsymbol{\Sigma}_2+\boldsymbol{E}_{22}) + \boldsymbol{E}_{32}^T\boldsymbol{E}_{32}\\ \boldsymbol{B}_{12} &= (\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^T\boldsymbol{E}_{12} + \boldsymbol{E}_{21}^T(\boldsymbol{\Sigma}_2+\boldsymbol{E}_{22}) + \boldsymbol{E}_{31}^T\boldsymbol{E}_{32}.\end{aligned} \tag{25}$$

From $\sigma_{\min}(\boldsymbol{\Sigma}_1) > 4\|\boldsymbol{E}\|_2$ follows that $\boldsymbol{C}_{11}$ is symmetric positive definite, while $\boldsymbol{C}_{12}$ is symmetric positive semi-definite and contains only second order terms. Abbreviate $\tilde{\lambda}_{n-r+j} \equiv \lambda_{n-r+j}(\boldsymbol{B}) = \sigma_{n-r+j}(\boldsymbol{A}+\boldsymbol{E})^2$, $1 \le j \le r$.

The proof proceeds in two steps:

1. Confirming that $\boldsymbol{C}_{11}$ satisfies the assumptions of Theorem 6.
2. Deriving the lower bounds for $\tilde{\lambda}_{n-r+j}$ from Theorem 6.

1. Confirm that $\boldsymbol{C}_{11}$ satisfies the assumptions of Theorem 6 We show that $\lambda_{\min}(\boldsymbol{C}_{11}) > \|\boldsymbol{B}_{22}\|_2$, by bounding $\|\boldsymbol{B}_{22}\|_2$ from above and $\lambda_{\min}(\boldsymbol{C}_{11})$ from below.

Regarding the upper bound for $\|\boldsymbol{B}_{22}\|_2$, the expression for $\boldsymbol{B}_{22}$ in (25) and the assumption $\|\boldsymbol{\Sigma}_2\|_2 < \|\boldsymbol{E}\|_2$ imply

$$\|\boldsymbol{B}_{22}\|_2 = \left\| \begin{bmatrix}\boldsymbol{E}_{12}^T & (\boldsymbol{E}_{22}+\boldsymbol{\Sigma}_2)^T & \boldsymbol{E}_{32}^T\end{bmatrix}^T \right\|_2^2 \le (\|\boldsymbol{\Sigma}_2\|_2 + \|\boldsymbol{E}\|_2)^2 \le 4\|\boldsymbol{E}\|_2^2. \tag{26}$$

Regarding the lower bound for $\lambda_{\min}(\boldsymbol{C}_{11})$, view $\boldsymbol{C}_{11} = (\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^T(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})$ as a singular value problem, so that $\lambda_{\min}(\boldsymbol{C}_{11}) = \sigma_{\min}(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^2$. The well-conditioning of singular values [GV13, Corollary 8.6.2] implies

$$|\,\sigma_{\min}(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11}) - \sigma_{\min}(\boldsymbol{\Sigma}_1)| \le \|\boldsymbol{E}_{11}\|_2 \le \|\boldsymbol{E}\|_2.$$

Adding the assumption $\sigma_{\min}(\boldsymbol{\Sigma}_1) = 1/\|\boldsymbol{\Sigma}_1^{-1}\|_2 > 4\|\boldsymbol{E}\|_2$ gives

$$\sigma_{\min}(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11}) \ge \sigma_{\min}(\boldsymbol{\Sigma}_1) - \|\boldsymbol{E}\|_2 > 4\,\|\boldsymbol{E}\|_2 - \|\boldsymbol{E}\|_2 = 3\|\boldsymbol{E}\|_2.$$

Now combine this lower bound for $\lambda_{\min}(\boldsymbol{C}_{11})$ with (26),

$$\lambda_{\min}(\boldsymbol{C}_{11}) = \sigma_{\min}(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^2 > 9\|\boldsymbol{E}\|_2^2 > 4\|\boldsymbol{E}\|_2^2 \ge \|\boldsymbol{B}_{22}\|_2.$$

Hence $\lambda_{\min}(\boldsymbol{C}_{11}) > \|\boldsymbol{B}_{22}\|_2$, and $\boldsymbol{C}_{11}$ satisfies the assumptions of Theorem 3.

2. Derive the lower bounds for $\tilde{\lambda}_{n-r+j}$ from Theorem 6 In these bounds,

$$\tilde{\lambda}_{n-r+j} \ge \lambda_j(\boldsymbol{S}) - \frac{\|\boldsymbol{B}_{22}\|_2\,\|\boldsymbol{C}_{11}^{-1}\boldsymbol{B}_{12}\|_2^2}{1-\|\boldsymbol{B}_{22}\|_2\,\|\boldsymbol{C}_{11}^{-1}\|_2}, \qquad \boldsymbol{S} \equiv \boldsymbol{B}_{22} - \boldsymbol{B}_{12}^T\boldsymbol{C}_{11}^{-1}\boldsymbol{B}_{12}, \tag{27}$$

the key term is $\boldsymbol{C}_{11}^{-1}\boldsymbol{B}_{12}$. Insert the expression for $\boldsymbol{B}_{12}$ from (25),

$$\begin{aligned}(\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^{-T}\boldsymbol{B}_{12} &= \boldsymbol{E}_{12} + (\boldsymbol{\Sigma}_1+\boldsymbol{E}_{11})^{-T}\left(\boldsymbol{E}_{21}^T(\boldsymbol{\Sigma}_2+\boldsymbol{E}_{22}) + \boldsymbol{E}_{31}^T\boldsymbol{E}_{32}\right)\\ &= \boldsymbol{E}_{12} + \boldsymbol{W}.\end{aligned} \tag{28}$$

Combine the expression for $\boldsymbol{C}_{11}$ from (24) with the above,

$$\boldsymbol{C}_{11}^{-1}\boldsymbol{B}_{12} = (\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})^{-1} \underbrace{(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})^{-T}\boldsymbol{B}_{12}}_{\boldsymbol{E}_{12}+\boldsymbol{W}} = (\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})^{-1}(\boldsymbol{E}_{12} + \boldsymbol{W}) \tag{29}$$

Multiply the above by $\boldsymbol{B}_{12}^T$ on the left, and use (28),

$$\begin{aligned}\boldsymbol{B}_{12}^T\boldsymbol{C}_{11}^{-1}\boldsymbol{B}_{12} &= \boldsymbol{B}_{12}^T(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})^{-1}(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})^{-T}\boldsymbol{B}_{12} \\ &= (\boldsymbol{E}_{12} + \boldsymbol{W})^T(\boldsymbol{E}_{12} + \boldsymbol{W}) = \boldsymbol{E}_{12}^T\boldsymbol{E}_{12} + \boldsymbol{E}_{12}^T\boldsymbol{W} + \boldsymbol{W}^T\boldsymbol{E}_{12} + \boldsymbol{W}^T\boldsymbol{W}.\end{aligned}$$

Substitute the above, and $\boldsymbol{B}_{22}$ from (25) into $\boldsymbol{S}$ from (27),

$$\begin{aligned}\boldsymbol{S} &= \boldsymbol{E}_{12}^T\boldsymbol{E}_{12} + (\boldsymbol{\Sigma}_2 + \boldsymbol{E}_{22})^T(\boldsymbol{\Sigma}_2 + \boldsymbol{E}_{22}) + \boldsymbol{E}_{32}^T\boldsymbol{E}_{32} \\ &\qquad - (\boldsymbol{E}_{12}^T\boldsymbol{E}_{12} + \boldsymbol{E}_{12}^T\boldsymbol{W} + \boldsymbol{W}^T\boldsymbol{E}_{12} + \boldsymbol{W}^T\boldsymbol{W}) \\ &= \boldsymbol{E}_{32}^T\boldsymbol{E}_{32} + (\boldsymbol{\Sigma}_2 + \boldsymbol{E}_{22})^T(\boldsymbol{\Sigma}_2 + \boldsymbol{E}_{22}) - \boldsymbol{R}_3 - \boldsymbol{W}^T\boldsymbol{W}.\end{aligned}$$

The well conditioning of eigenvalues [GV13, Theorem 8.1.5] implies

$$\lambda_j(\boldsymbol{S}) \geq \lambda_j(\boldsymbol{E}_{32}^T\boldsymbol{E}_{32} + (\boldsymbol{\Sigma}_2 + \boldsymbol{E}_{22})^T(\boldsymbol{\Sigma}_2 + \boldsymbol{E}_{22}) - \boldsymbol{R}_3) - \|\boldsymbol{W}\|_2^2. \tag{30}$$

Substitute the bound for $\|\boldsymbol{B}_{22}\|_2$ from (26), and (28) into the second summand of (27),

$$\frac{\|\boldsymbol{B}_{22}\|_2\,\|\boldsymbol{C}_{11}^{-1}\boldsymbol{B}_{12}\|_2^2}{1 - \|\boldsymbol{B}_{22}\|_2\,\|\boldsymbol{C}_{11}^{-1}\|_2} \leq 4\frac{\|\boldsymbol{E}\|_2^2\,\|(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})^{-1}(\boldsymbol{E}_{12} + \boldsymbol{W})\|_2^2}{1 - 4\|\boldsymbol{E}\|_2^2\|(\boldsymbol{\Sigma}_1 + \boldsymbol{E}_{11})^{-1}\|_2^2}. \tag{31}$$

At last insert (30) and (31) into (27). □

Remark 3.1. *The assumptions in Theorem 7 are not restrictive. Only a small gap of* $3\|\boldsymbol{E}\|_2$ *is required to separate the small singular value cluster of* $\boldsymbol{A}$ *from the remaining singular values,*

$$\|\boldsymbol{\Sigma}_2\|_2 < \|\boldsymbol{E}\|_2 < 4\|\boldsymbol{E}\|_2 \leq 1/\|\boldsymbol{\Sigma}_1^{-1}\|_2.$$

4 Numerical experiments

We present numerical experiments to illustrate that downcasting to lower precision can increase small singular values, thus confirming that our bounds in sections 2 and 3 are informative models for the effects of reduced arithmetic precision.

After describing the algorithms for computing the singular values (Section 4.1), we present the numerical experiments (Section 4.2).

4.1 Generation and computation of singular values

The code for the numerical experiments consists of two algorithms: Algorithm 2 in Appendix A generates the diagonal matrix $\boldsymbol{\Sigma}$ containing the exact singular values, while Algorithm 1 generates the matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ from $\boldsymbol{\Sigma}$ in double precision, and then computes the singular values of $\boldsymbol{A}$ and those of its lower precision versions `single`($\boldsymbol{A}$) and `half`($\boldsymbol{A}$). We use `Julia` programming language for our computations. The scripts for reproducing the numerical experiment are published in our git repository[4].

The n singular values in the diagonal matrix $\boldsymbol{\Sigma}$ generated by Algorithm 2 consist of two clusters: a cluster $\boldsymbol{\Sigma}_1$ of large singular values, and a cluster $\boldsymbol{\Sigma}_2$ of small singular values. Each cluster is defined by the following input parameters: the number of singular values; the smallest and largest singular value; and the gap between the two clusters. Specifically, cluster $\boldsymbol{\Sigma}_1$ consists of $k_1 > 0$ singular values, the largest one being 10^{s_1} and the smallest one being $10^{s_1-d_1}$. Here $d_1 \geq 0$ controls the distance between smallest and largest singular value. If $d_1 > 0$ and $k_1 > 2$, then the interior singular values of $\boldsymbol{\Sigma}_1$ are sampled uniformly at random in the interval $[10^{s_1-d_1}, 10^{s_1}]$.

[4] `https://github.com/cboutsikas/small_sigmas_increase.git`

The parameter g controls the gap between the clusters, which is set to 10^g. Cluster $\mathbf{\Sigma}_2$ consists of $k_2 \equiv n - k_1 \geq 0$ singular values, the largest one being $10^{s_1-d_1-g}$, and the smallest one being $10^{s_1-d_1-g-d_2}$, where $d_2 \geq 0$ controls the distance between the smallest and largest singular value in cluster $\mathbf{\Sigma}_2$. If $d_2 > 0$ and $k_2 > 2$, then the interior singular values of $\mathbf{\Sigma}_2$ are sampled uniformly at random in the interval $[10^{s_1-d_1-g-d_2}, 10^{s_1-d_1-g}]$.

4.2 Numerical results and discussion

We present numerical experiments that corroborate bounds in sections 2 and 3. Our experiments are performed on matrices $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with $\text{rank}(\boldsymbol{A}) = n$, $m = 4,096$ and $n = 256$. We emphasize that changing the matrix dimensions while keeping the aspect ratio m/n fixed does not change our conclusions.

Figures 1–4 show the exact singular values as well as the singular values computed in double precision, which turn out to be identical in all cases. In addition, Figures 1 and 3 show the singular values computed in *single precision*, while Figures 2 and 4 show the singular values computed in *half precision*.

To guarantee that our empirical evaluations satisfy Assumptions 2.1 and 3.1, we compute the singular values with the Golub-Kahan Bi-Diagonalization[5] algorithm [GK65] in the respective precision. As in [GV13, Algorithm 8.6.2] we assume that

$$\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{V} = \mathbf{\Sigma} + \boldsymbol{E},$$

satisfies $\|\boldsymbol{E}\|_2 \approx u\|\boldsymbol{A}\|_2$, where u is the unit roundoff [GV13, section], which depends on the underlying precision[6]. Thus, we can express the assumption in Remark 3.1 as follows:

$$\sigma_{n-r+1} = \|\mathbf{\Sigma}_2\|_2 \lesssim \sigma_{\max}\, u \lesssim 4\,\sigma_{\max} u \lesssim 1/\|\mathbf{\Sigma}_1^{-1}\|_2 = \sigma_{n-r}, \tag{32}$$

for some $r \geq 1$. Table 1 shows the increase in the smallest singular value for $r = 1$; and the average increase in the r smallest singular values $r > 1$. Since we describe a qualitative model, an increase can be observed even when the assumptions are not satisfied. We provide demonstrations of two such cases in Appendix B.

	$\min(\mathbf{\Sigma})$	$\min(\mathbf{\Sigma}^d)$	$\min(\mathbf{\Sigma}^s)$	$\texttt{avg}(\mathbf{\Sigma}_2^s)$	$\min(\mathbf{\Sigma}^h)$	$\texttt{avg}(\mathbf{\Sigma}_2^h)$
Fig. 1	10^{-7}	10^{-7}	6×10^{-7}	5×10^{-7}	N/A	N/A
Fig. 2	10^{-3}	10^{-3}	N/A	N/A	4×10^{-3}	3×10^{-3}
Fig. 3	10^{-5}	10^{-5}	6×10^{-4}	4×10^{-4}	N/A	N/A
Fig. 4	10^{-4}	10^{-4}	N/A	N/A	8×10^{-3}	6×10^{-3}

Table 1: Smallest singular values in Figures 1–4: exact ($\mathbf{\Sigma}$); double precision ($\mathbf{\Sigma}^d$); single precision ($\mathbf{\Sigma}^s$); and half precision ($\mathbf{\Sigma}^h$). The quantities $\texttt{avg}(\mathbf{\Sigma}_2^s)$ and $\texttt{avg}(\mathbf{\Sigma}_2^h)$ represent the average increase of the r smallest singular values for single and half precision.

4.2.1 A single smallest singular value

We illustrate that downcasting to lower precision can increase the smallest singular value, thus confirming that Theorem 4 represents a proper qualitative model for the effects of reduced precision. In Figures 1 and 2, the small singular value cluster $\mathbf{\Sigma}_2$ consists of a single singular value, while the large singular value cluster $\mathbf{\Sigma}_1$ contains 255 singular values.

Figure 1 The cluster $\mathbf{\Sigma}_1$ contains 255 distinct singular values in the interval $[10^{-4}, 10^2]$, while $\mathbf{\Sigma}_2$ contains the single singular value 10^{-7}. The values in Assumption 2.1 and (32) are

$$\underbrace{\sigma_n}_{10^{-7}} \lesssim \underbrace{\sigma_{\max}\, u}_{6\times10^{-6}} \lesssim \underbrace{4\,\sigma_{\max}\, u}_{2.4\times10^{-5}} \lesssim \underbrace{\sigma_{n-1}}_{10^{-4}}.$$

In single precision, the smallest singular value has increased by almost 5×10^{-7}.

[5] `Julia` computes the SVD with the LAPACK routine `dgesvd()`, which employs the Bi-Diagonalization method.

[6] In double precision (binary64), $u = 2^{-53} \approx 1.11\times10^{-16}$; in single precision (binary32), $u = 2^{-24} \approx 5.96\times10^{-8}$; and in half precision (binary16) $u = 2^{-11} \approx 4.88\times10^{-4}$.

Figure 2 The cluster $\mathbf{\Sigma}_1$ contains 255 distinct singular values in the interval $[10^{-1}, 10^{1}]$, while $\mathbf{\Sigma}_2$ contains the single singular value 10^{-3}. The values in Assumption 2.1 and (32)) are

$$\underbrace{\sigma_n}_{10^{-3}} \lesssim \underbrace{\sigma_{\max}\, u}_{5\times10^{-3}} \lesssim \underbrace{4\,\sigma_{\max}\, u}_{2\times10^{-2}} \lesssim \underbrace{\sigma_{n-1}}_{10^{-1}}.$$

In half precision, the smallest singular value has increased to 4×10^{-3}.

4.2.2 A cluster of small singular values

We illustrate that downcasting to lower precision can increase the values of the cluster of small singular values, thus confirming that Theorem 7 represents a proper qualitative model for the effects of reduced precision. In Figures 3 and 4, the small singular value cluster $\mathbf{\Sigma}_2$ contains 28 singular values, while the large singular value cluster $\mathbf{\Sigma}_1$ contains 228 singular values.

Figure 3 The cluster $\mathbf{\Sigma}_1$ contains 228 distinct singular values in the interval $[10^{-1}, 10^{5}]$, while $\mathbf{\Sigma}_2$ contains 28 singular values in the interval $[10^{-5}, 10^{-3}]$. The values in Assumption 2.1 and (32) are

$$\underbrace{\sigma_{n-r}}_{10^{-3}} \lesssim \underbrace{\sigma_{\max}\, u}_{6\times10^{-3}} \lesssim \underbrace{4\,\sigma_{\max}\, u}_{2.4\times10^{-2}} \lesssim \underbrace{\sigma_{n-r+1}}_{10^{-1}}.$$

In single precision, the smallest singular value of $\mathbf{\Sigma}_2$ has increased to 6×10^{-4}, with an average increase of 4×10^{-4} for the r smallest singular values.

Figure 4 The cluster $\mathbf{\Sigma}_1$ contains 228 distinct singular values in the interval $[10^{0}, 10^{2}]$, while $\mathbf{\Sigma}_2$ contains 28 singular values in the interval $[10^{-4}, 10^{-2}]$. The values in Assumption 2.1 and (32) are

$$\underbrace{\sigma_{n-r}}_{10^{-2}} \lesssim \underbrace{\sigma_{\max}\, u}_{5\times10^{-2}} \lesssim \underbrace{4\,\sigma_{\max}\, u}_{10^{-1}} \lesssim \underbrace{\sigma_{n-r+1}}_{10^{0}}.$$

In half precision, the smallest singular value of $\mathbf{\Sigma}_2$ has increased to 8×10^{-3}, with an average increase of 7×10^{-3} for the r smallest singular values.

5 Future Work

We investigated the change in the computed singular values of a full column-rank matrix $\boldsymbol{A}$ after it has been is downcast to a lower arithmetic precision. Our lower bounds in Theorem 1 represent a *qualitative* model for the increase in the smallest singular values of the perturbed matrix $\boldsymbol{A} + \boldsymbol{E}$, which is confirmed by the experiments in section 4.

Future work will consist of a *quantitative* analysis to determine the exact order of magnitude of the increase in the small singular values and the structural properties of $\boldsymbol{A}$ that can contribute to it, including specifically the size of the gap that separates the small singular values from the larger singular values; and the condition number of $\boldsymbol{A}$ with respect to left inversion.

In addition, the influence of the third order perturbation terms needs to be investigated, as they might possibly become dominant for ill-conditioned matrices $\boldsymbol{A}$.

A Algorithms

We present pseudo codes for two algorithms: The function `create_sigmas` in Algorithm 2 computes the singular values $\mathbf{\Sigma}$ according to the specifications in the input parameters `params`. Algorithm 1 constructs $\boldsymbol{A}$ from $\mathbf{\Sigma}$ in double precision, and then computes the singular values $\mathbf{\Sigma}^d$ of $\boldsymbol{A}$, $\mathbf{\Sigma}^s$ of $\texttt{single}(\boldsymbol{A})$, and $\mathbf{\Sigma}^h$ of $\texttt{half}(\boldsymbol{A})$. If $d_1 = 0$ or $d_2 = 0$ in Algorithm 2, then the cluster $\mathbf{\Sigma}_1$ or $\mathbf{\Sigma}_2$ consists of a single singular value of multiplicity k_1 or k_2, respectively.

Algorithm 1 Singular values of $\boldsymbol{A}$, $\texttt{single}(\boldsymbol{A})$ and $\texttt{half}(\boldsymbol{A})$

Input: Large matrix dimension m, $\texttt{params}$
Output: Singular values of $\boldsymbol{A}$ in double, single, half precision

$\boldsymbol{\Sigma} \leftarrow \texttt{create\_sigmas}(\texttt{params})$ {Exact singular values}
$n \leftarrow \texttt{length}(\boldsymbol{\Sigma})$ {Small dimension of $\boldsymbol{A}$}
$[\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{V}] \leftarrow \texttt{SVD}(\texttt{randn}(m, n))$ {Left and right singular vectors for $\boldsymbol{A}$}
$\boldsymbol{A} \leftarrow \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ {Compute $\boldsymbol{A}$ in double precision}
$\boldsymbol{\Sigma}^d \leftarrow \texttt{SVD}(\boldsymbol{A})$ {Singular values of double precision $\boldsymbol{A}$}
$\boldsymbol{\Sigma}^s \leftarrow \texttt{SVD}(\texttt{double}(\texttt{single}(\boldsymbol{A}))$ {Singular values of single precision $\boldsymbol{A}$}
$\boldsymbol{\Sigma}^h \leftarrow \texttt{SVD}(\texttt{double}(\texttt{half}(\boldsymbol{A}))$ {Singular values of half precision $\boldsymbol{A}$}
return $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^d$, $\boldsymbol{\Sigma}^s$, $\boldsymbol{\Sigma}^h$

Algorithm 2 Exact singular values: function $\texttt{create\_sigmas}$

Input: $\texttt{params} = \{s_1, g, k_1, k_2, d_1, d_2\}$
Output: Exact singular values $\boldsymbol{\Sigma}$

$\boldsymbol{\Sigma} \leftarrow \texttt{zeros}(k_1 + k_2, 1)$ {Initialize vector of all singular values}
$\boldsymbol{\Sigma}_1 \leftarrow \texttt{zeros}(k_1, 1)$ {Initialize cluster of large singular values}
$\boldsymbol{\Sigma}_2 \leftarrow \texttt{zeros}(k_2, 1)$ {Initialize cluster of small singular values}
$\boldsymbol{\Sigma}_1(1) \leftarrow 10^{s_1}$ {Largest singular value}
if $k_1 > 1$ **then**
  $\boldsymbol{\Sigma}_1(k_1) \leftarrow 10^{s_1 - d_1}$ {Smallest singular value in $\boldsymbol{\Sigma}_1$}
end if
{Uniform sampling of interior singular values in cluster $\boldsymbol{\Sigma}_1$}
for $j = 2 : k_1 - 1$ **do**
  $\boldsymbol{\Sigma}_1(j) \leftarrow \texttt{Uniform}([\boldsymbol{\Sigma}_1(\texttt{k}_1), \boldsymbol{\Sigma}_1(1)])$
end for
$\boldsymbol{\Sigma}_2(1) \leftarrow 10^{s_1 - d_1 - g}$ {Largest singular value in $\boldsymbol{\Sigma}_2$}
if $k_2 > 1$ **then**
  $\boldsymbol{\Sigma}_2(k_2) \leftarrow 10^{s_1 - d_1 - g - d_2}$ {Smallest singular value in $\boldsymbol{\Sigma}_2$}
end if
{Uniform sampling of interior singular values in cluster $\boldsymbol{\Sigma}_2$}
for $j = 2 : k_2 - 1$ **do**
  $\boldsymbol{\Sigma}_2(j) \leftarrow \texttt{Uniform}([\boldsymbol{\Sigma}_2(\texttt{k}_2), \boldsymbol{\Sigma}_2(1)])$
end for
$\boldsymbol{\Sigma} \leftarrow [\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2]$ {Concatenate the two singular value clusters}
return $\texttt{sort}(\boldsymbol{\Sigma})$ {Return sorted singular values in non-ascending order}

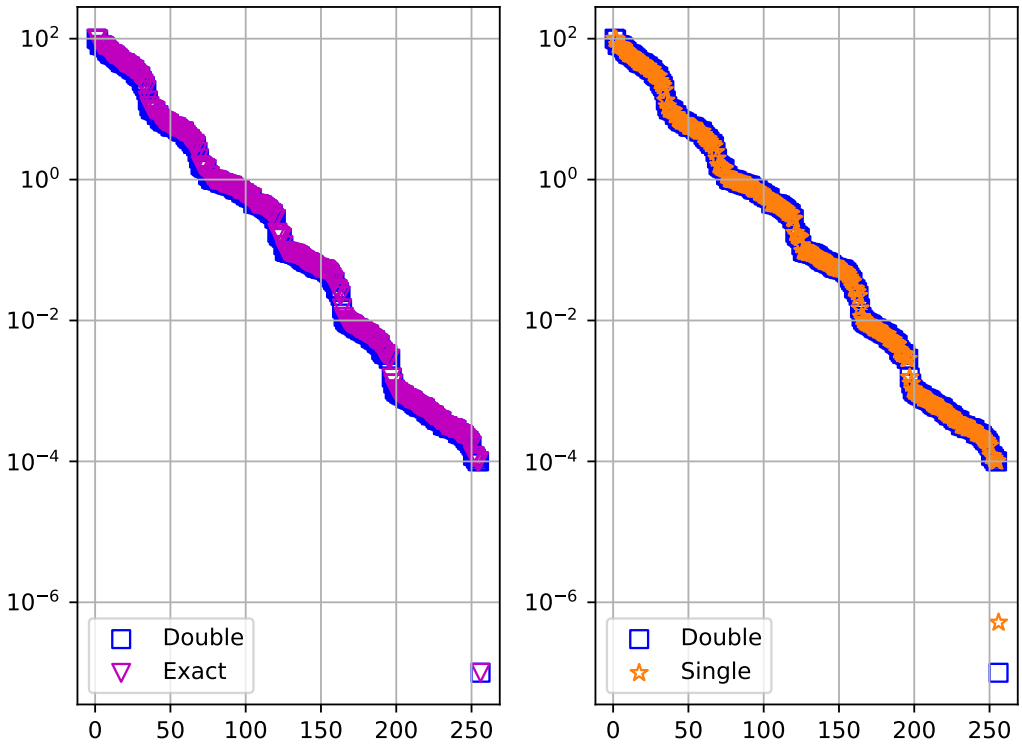

Figure 1: The matrix $\boldsymbol{A} \in \mathbb{R}^{4096\times 256}$ has 255 distinct singular values in $[10^{-4}, 10^{2}]$, and a single small singular value 10^{-7}. All panels: Double precision singular values (squares). Left: Exact singular values (triangles). Right: Single precision singular values (stars).

B Supplementary material

We illustrate that downcasting to lower precision can the increase the set of the smallest singular values, even when the Assumptions 2.1, 3.1 are not satisfied. We present two additional plots, and specifically Figure 5 shows the increase of the computed smallest singular values in single ($r = 28$) and Figure 6 shows the increase of the computed smallest singular value in half ($r = 1$).

	$\min(\boldsymbol{\Sigma})$	$\min(\boldsymbol{\Sigma}^d)$	$\min(\boldsymbol{\Sigma}^s)$	$\texttt{avg}(\boldsymbol{\Sigma}_2^s)$	$\min(\boldsymbol{\Sigma}^h)$	$\texttt{avg}(\boldsymbol{\Sigma}_2^h)$
Fig. 5	10^{-7}	10^{-7}	5×10^{-5}	4×10^{-5}	N/A	N/A
Fig. 6	10^{-3}	10^{-3}	N/A	N/A	6×10^{-3}	5×10^{-3}

Table 2: Smallest singular values in Figures 5–6: exact ($\boldsymbol{\Sigma}$); double precision ($\boldsymbol{\Sigma}^d$); single precision ($\boldsymbol{\Sigma}^s$); and half precision ($\boldsymbol{\Sigma}^h$). The quantities $\texttt{avg}(\boldsymbol{\Sigma}_2^s)$ and $\texttt{avg}(\boldsymbol{\Sigma}_2^h)$ represent the average increase of the r smallest singular values for single and half precision.

Figure 5 The cluster $\boldsymbol{\Sigma}_1$ contains 228 distinct singular values in the interval $[10^{-3}, 10^{4}]$, while $\boldsymbol{\Sigma}_2$ contains the smallest 28 singular values in the interval $[10^{-7}, 10^{-4}]$. The values in Assumption 3.1 and (32) are not

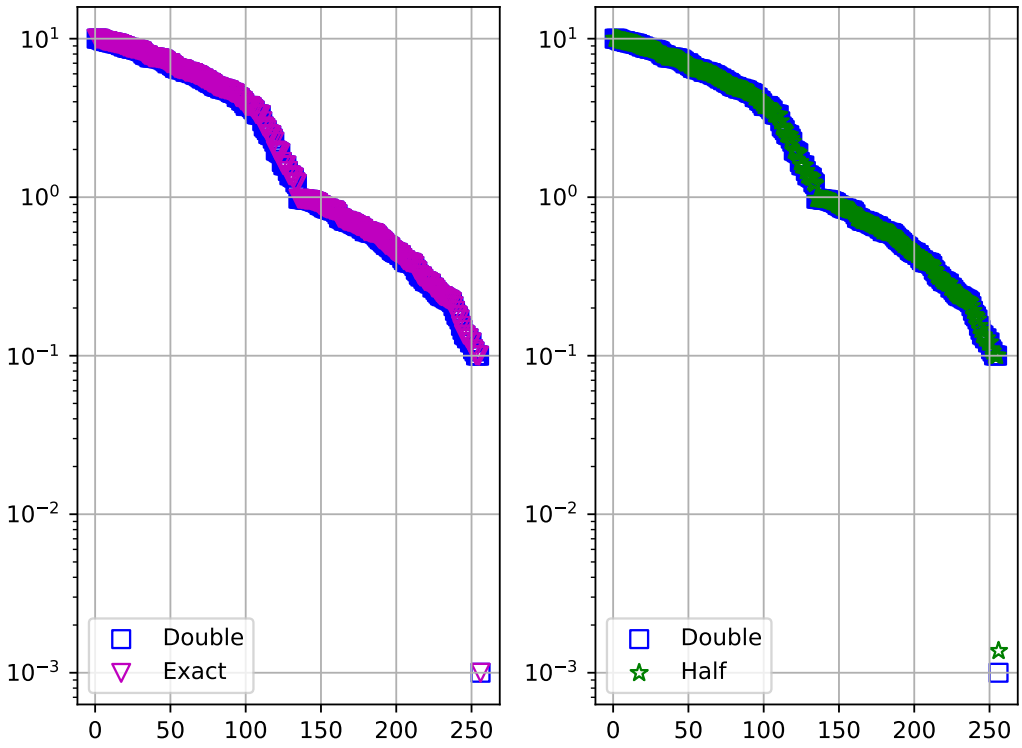

Figure 2: The matrix $\boldsymbol{A} \in \mathbb{R}^{4096\times 256}$ has 255 distinct singular values in $[10^{-1}, 10^{1}]$, and a single small singular value 10^{-3}. All panels: Double precision singular values (squares). Left: Exact singular values (triangles). Right: Half precision singular values (stars).

satisfied since

$$\underbrace{\sigma_{n-r}}_{10^{-4}} \lesssim \underbrace{\sigma_{\max}\, u}_{6\times 10^{-2}},$$
$$\underbrace{4\,\sigma_{\max}\, u}_{2.4\times 10^{-1}} > \underbrace{\sigma_{n-r+1}}_{10^{-3}}.$$

However, the smallest singular value of $\boldsymbol{\Sigma}_2$ has increased to 5×10^{-6}, with an average increase of 4×10^{-6} for the r smallest singular values.

Figure 6 The cluster $\boldsymbol{\Sigma}_1$ contains 255 distinct singular values in the interval $[10^{-2}, 10^{2}]$, while $\boldsymbol{\Sigma}_2$ contains the single singular value 10^{-3}. The values in Assumption 2.1 and (32) are not satisfied since

$$\underbrace{\sigma_{n}}_{10^{-3}} \lesssim \underbrace{\sigma_{\max}\, u}_{5\times 10^{-2}},$$
$$\underbrace{4\,\sigma_{\max}\, u}_{2\times 10^{-1}} > \underbrace{\sigma_{n-1}}_{10^{-2}}.$$

However, in half precision, the smallest singular value has increased to 6×10^{-3}.

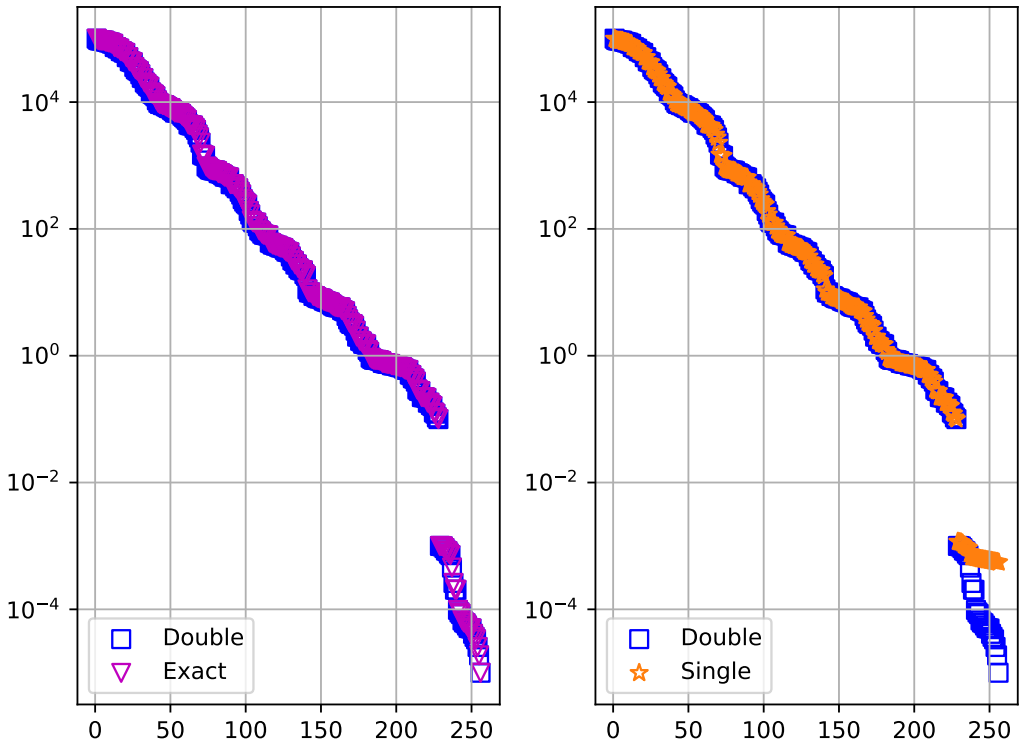

Figure 3: The matrix $\boldsymbol{A} \in \mathbb{R}^{4096\times256}$ has 228 distinct singular values in $[10^{-1}, 10^{5}]$, and 28 distinct singular values in $[10^{-5}, 10^{-3}]$. All panels: Double precision singular values (squares). Left: Exact singular values (triangles). Right: Single precision singular values (stars).

References

[GK65] Gene Golub and William Kahan. "Calculating the singular values and pseudo-inverse of a matrix". In: *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis* 2.2 (1965), pp. 205–224.

[GV13] G. H. Golub and C. F. Van Loan. *Matrix computations*. Fourth. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, 2013.

[HJ13] R. A. Horn and C. R. Johnson. *Matrix analysis*. Second. Cambridge University Press, Cambridge, 2013.

[HP92] Y. P. Hong and C.-T. Pan. "A lower bound for the smallest singular value". In: *Linear Algebra Appl.* 172 (1992), pp. 27–32. ISSN: 0024-3795. DOI: 10.1016/0024-3795(92)90016-4. URL: https://doi.org/10.1016/0024-3795(92)90016-4.

[Hua08] T.-Z. Huang. "Estimation of $\|A^{-1}\|_\infty$ and the smallest singular value". In: *Comput. Math. Appl.* 55.6 (2008), pp. 1075–1080. ISSN: 0898-1221. DOI: 10.1016/j.camwa.2007.04.036. URL: https://doi.org/10.1016/j.camwa.2007.04.036.

[Joh89] C. R. Johnson. "A Gersgorin-type lower bound for the smallest singular value". In: *Linear Algebra Appl.* 112 (1989), pp. 1–7. ISSN: 0024-3795. DOI: 10.1016/0024-3795(89)90583-1. URL: https://doi.org/10.1016/0024-3795(89)90583-1.

[JS98] C. R. Johnson and T. Szulc. "Further lower bounds for the smallest singular value". In: *Linear Algebra Appl.* 272 (1998), pp. 169–179. ISSN: 0024-3795. DOI: 10.1016/S0024-3795(97)00330-3. URL: https://doi.org/10.1016/S0024-3795(97)00330-3.

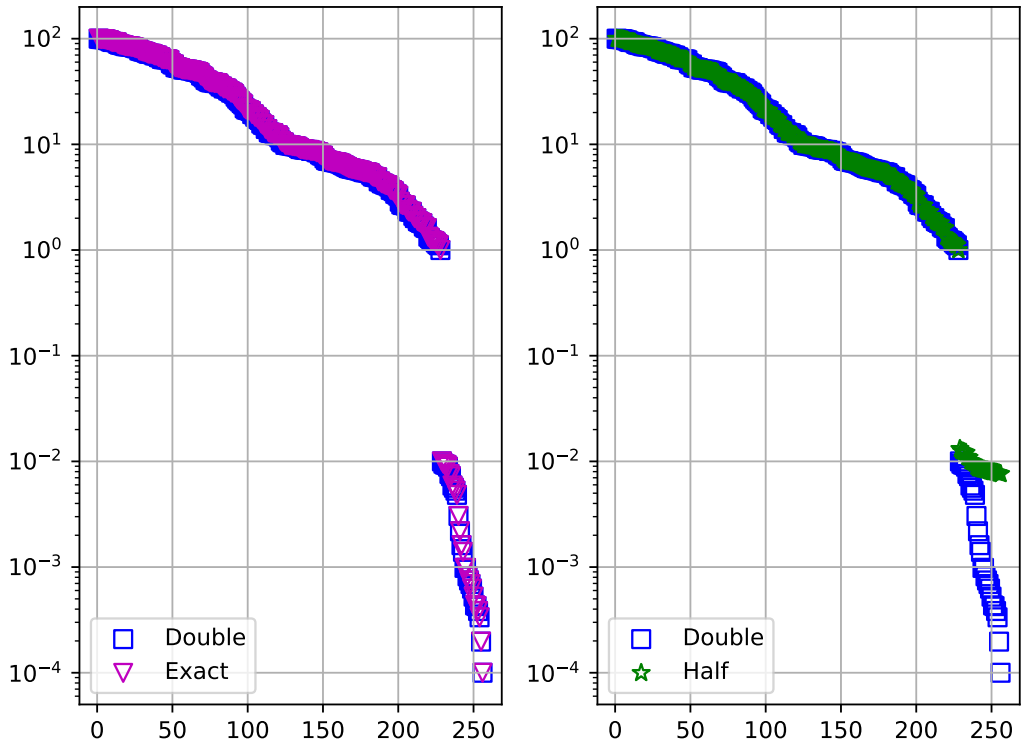

Figure 4: The matrix $\boldsymbol{A} \in \mathbb{R}^{4096\times 256}$ has 228 distinct singular values in $[10^0, 10^2]$, and 28 distinct singular values in $[10^{-4}, 10^{-2}]$. All panels: Double precision singular values (squares). Left: Exact singular values (triangles). Right: Half precision singular values (stars).

[Li20] C. Li. "Schur complement-based infinity norm bounds for the inverse of SDD matrices". In: *Bull. Malays. Math. Sci. Soc.* 43.5 (2020), pp. 3829–3845. ISSN: 0126-6705. DOI: 10.1007/s40840-020-00895-x. URL: https://doi.org/10.1007/s40840-020-00895-x.

[LX21] M. Lin and M. Xie. "On some lower bounds for smallest singular value of matrices". In: *Appl. Math. Lett.* 121 (2021), Paper No. 107411, 7. ISSN: 0893-9659. DOI: 10.1016/j.aml.2021.107411. URL: https://doi.org/10.1016/j.aml.2021.107411.

[Ois23] S. Oishi. "Lower bounds for the smallest singular values of generalized asymptotic diagonal dominant matrices". In: *Jpn. J. Ind. Appl. Math.* (July 2023). DOI: 10.1007/s13160-023-00596-5.

[Par80] B. N. Parlett. *The Symmetric Eigenvalue Problem*. Englewood Cliffs: Prentice Hall, 1980.

[Rum09] S. M. Rump. "Inversion of extremely ill-conditioned matrices in floating-point". In: *Japan J. Indust. Appl. Math.* 26.2-3 (2009), pp. 249–277. ISSN: 0916-7005. URL: http://projecteuclid.org/euclid.jjiam/1265033781.

[San21] C. Sang. "Schur complement-based infinity norm bounds for the inverse of *DSDD* matrices". In: *Bull. Iranian Math. Soc.* 47.5 (2021), pp. 1379–1398. ISSN: 1017-060X. DOI: 10.1007/s41980-020-00447-w. URL: https://doi.org/10.1007/s41980-020-00447-w.

[Shu22] X. Shun. "Two new lower bounds for the smallest singular value". In: *J. Math. Inequal.* 16.1 (2022), pp. 63–68. DOI: 10.7153/jmi-2022-16-05. URL: https://doi.org/10.7153/jmi-2022-16-05.

[SS90] G. W. Stewart and J. G. Sun. *Matrix perturbation theory*. Computer Science and Scientific Computing. Academic Press, Inc., Boston, MA, 1990.

[Ste06] Michael Stewart. "Perturbation of the SVD in the presence of small singular values". In: *Linear Algebra Appl.* 419.1 (2006), pp. 53–77. ISSN: 0024-3795. DOI: 10.1016/j.laa.2006.04.013.

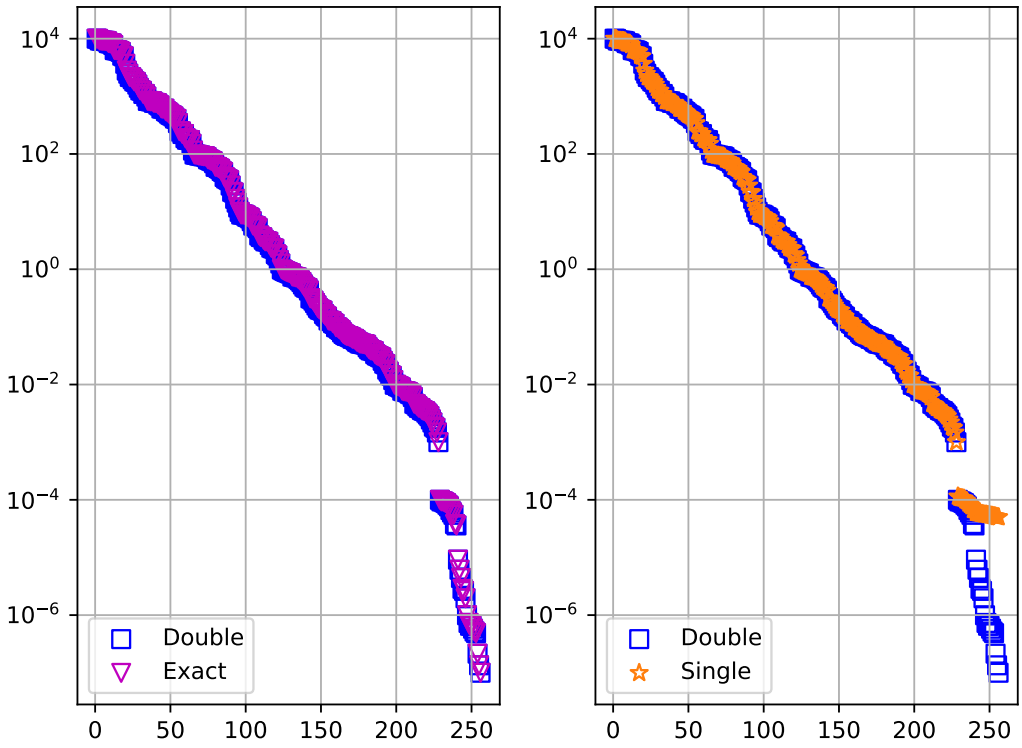

Figure 5: The matrix $\boldsymbol{A} \in \mathbb{R}^{4096\times256}$ has 228 distinct singular values in $[10^{-3}, 10^{4}]$, and 28 distinct singular values in $[10^{-7}, 10^{-4}]$. All panels: Double precision singular values (squares). Left: Exact singular values (triangles). Right: Single precision singular values (stars).

[Ste84] G. W. Stewart. "A second order perturbation expansion for small singular values". In: *Linear Algebra Appl.* 56 (1984), pp. 231–235.

[Var75] J. M. Varah. "A lower bound for the smallest singular value of a matrix". In: *Linear Algebra Appl.* 11 (1975), pp. 3–5. ISSN: 0024-3795. DOI: 10.1016/0024-3795(75)90112-3. URL: https://doi.org/10.1016/0024-3795(75)90112-3.

[YG97] Y. Yu and D. Gu. "A note on a lower bound for the smallest singular value". In: *Linear Algebra Appl.* 253 (1997), pp. 25–38. ISSN: 0024-3795. DOI: 10.1016/0024-3795(95)00784-9. URL: https://doi.org/10.1016/0024-3795(95)00784-9.

[Zou12] L. Zou. "A lower bound for the smallest singular value". In: *J. Math. Inequal.* 6.4 (2012), pp. 625–629. ISSN: 1846-579X. DOI: 10.7153/jmi-06-60. URL: https://doi.org/10.7153/jmi-06-60.

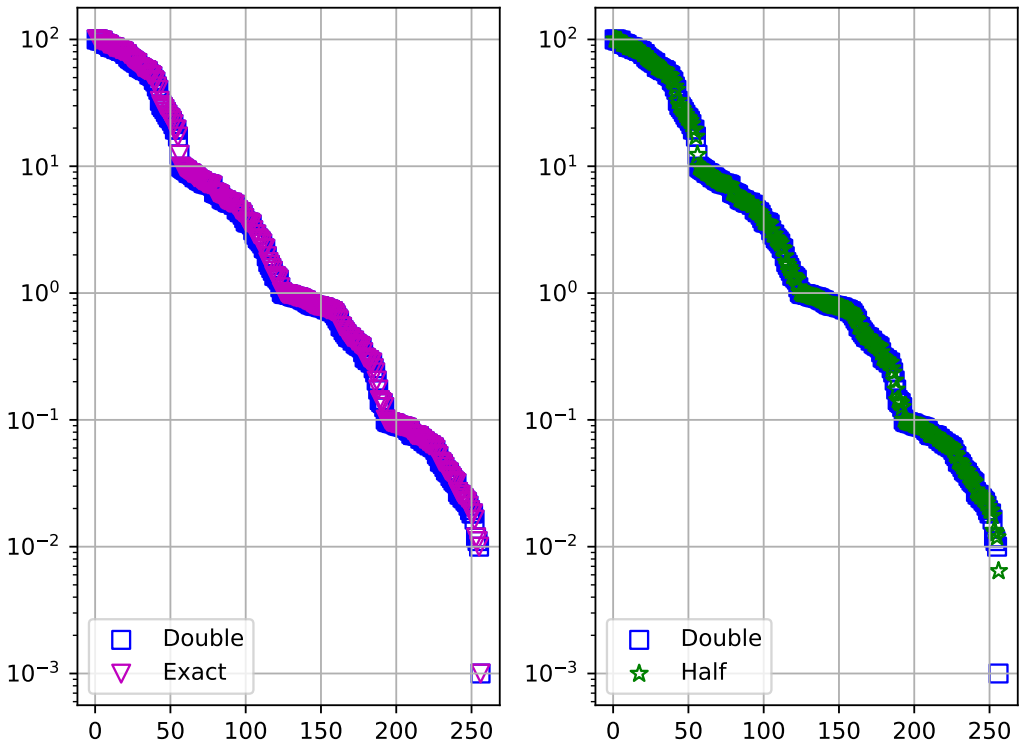

Figure 6: The matrix $\boldsymbol{A} \in \mathbb{R}^{4096\times 256}$ has 228 distinct singular values in $[10^{-2}, 10^{2}]$, and a single small singular value 10^{-3}. All panels: Double precision singular values (squares). Left: Exact singular values (triangles). Right: Half precision singular values (stars).