

STOCHASTIC ROUNDING IMPLICITLY REGULARIZES TALL-AND-THIN MATRICES*

GREGORY DEXTER[†], CHRISTOS BOUTSIKAS[†], LINKAI MA[†], ILSE C. F. IPSEN[‡], AND
PETROS DRINEAS[†]

Abstract. Motivated by the popularity of stochastic rounding in the context of machine learning and the training of large-scale deep neural network models, we consider stochastic nearness rounding of real matrices \mathbf{A} with many more rows than columns. We provide novel theoretical evidence, supported by extensive experimental evaluation, that with high probability, the smallest singular value of a stochastically rounded matrix is well bounded away from zero—regardless of how close \mathbf{A} is to being rank-deficient and even if \mathbf{A} is rank-deficient. In other words, stochastic rounding *implicitly regularizes* tall-*and-thin* matrices \mathbf{A} so that the rounded version has full column rank. Our proofs leverage powerful results in random matrix theory, and the idea that stochastic rounding errors do not concentrate in low-dimensional column spaces.

Key words. stochastic rounding, singular values, Weyl’s inequality, rank, arithmetic precision, random matrix theory, variance, Hoeffding’s inequality, union bound

MSC codes. 68W20, 65F15, 65F22, 65G50, 15A18, 15A42

DOI. 10.1137/24M1647679



See reproducibility of
computational results
at end of the article.

1. Introduction. Stochastic rounding (SR), proposed over 70 years ago, is a probabilistic approach to rounding. According to [12], the earliest proposal for SR appeared in a one-paragraph abstract of a communication presented by Forsythe [19] in 1950 at the 52nd meeting of the American Mathematical Society, in the context of reducing the accumulation of round-off errors in solving systems of ordinary differential equations. Relatedly, the idea of modeling rounding errors as random variables to handle imprecise data in exact arithmetic goes back to 1949 and the work of von Neumann and Goldstine [37].

Despite its illustrious beginnings, SR has been largely overlooked by the numerical analysis community. Over the past few years, SR has enjoyed a resurgence in popularity, mainly due to the increasing interest for low-precision floating point arithmetic in the context of machine learning applications and the training of large-scale deep neural network models. Currently, major chip designers own numerous SR-related patents, which seems to indicate that we might soon reach an inflection point for a wider adoption of SR in hardware and software. A nonexhaustive list includes Graph-Core IPU that support SR in binary32 and binary16 [1]; the Loihi chip [14]; AMD [31]; NVIDIA [3]; IBM [8, 9]; VIA Technologies [23]; DensBits Technologies [29]; and

*Received by the editors March 19, 2024; accepted for publication (in revised form) November 19, 2024; published electronically February 12, 2025.

<https://doi.org/10.1137/24M1647679>

Funding: The work of the first three authors was partially supported by NSF grants CCF-2209509, CCF-1814041, and DMS-1760353. The work of the fourth author was supported in part by NSF grants DMS-1760374 and CCF-2209510 and by DOE grant DE-SC0022085. The work of the fifth author was partially supported by NSF grants CCF-2209509, CCF-1814041, and DMS-1760353 and by DOE grant DE-SC0022085.

[†]Department of Computer Science, Purdue University, West Lafayette, IN 47907 USA (gdexter@purdue.edu, cboutsik@purdue.edu, ma856@purdue.edu, pdrineas@purdue.edu).

[‡]Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 USA (ipsen@ncsu.edu).

GSI Technology [30]. A detailed discussion of the history of SR and probabilistic error analysis as well as devices and patents for SR can be found in [12].

Recall that, given a number $x \in \mathbb{R}$ and a finite set of numbers $\mathcal{F} \subset \mathbb{R}$, rounding refers to the process of matching x to a number $\tilde{x} \in \mathcal{F}$. This can be done deterministically or *stochastically*: a common modality for SR is to round with a probability that depends on the distance of x from the two points that enclose it in \mathcal{F} . For example, if $\mathcal{F} = \{0, 1\}$, the value $x = 0.7$ is rounded to $\tilde{x} = 1$ with probability .7 and to $\tilde{x} = 0$ with probability .3.

This version of SR is sometimes called *SR-nearness* or *mode-1 SR* and is an unbiased estimator. Of particular interest is the case where \mathcal{F} is the set of *normalized floating point* numbers.

1.1. Our results. Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ be a tall-and-thin matrix with $n \gg d$. We present novel theoretical evidence, supported by extensive experimental evaluation, that guarantees with high probability that, after SR, the smallest singular value of the rounded matrix *is bounded away from zero*. This holds regardless of how close to rank-deficient \mathbf{A} might be, or even if \mathbf{A} is rank-deficient, assuming that the rounding process has access to enough randomness (see (1.1) for a precise definition). If the stochastically rounded \mathbf{A} were to be used in a downstream regression or classification problem, this is akin to saying that SR *implicitly regularizes* \mathbf{A} . Such regularization effects are often beneficial in downstream machine learning algorithms and, in particular, in training deep neural network models and large language models [21, 38]. Thus, SR could serve as an implicit regularizer in modern machine learning applications and might bypass the need for explicit regularization. The references in [12, section 8.2] point to a long list of machine learning applications that could benefit from properties of SR.

To give a taste of our results, let \mathcal{F} be the set of normalized floating point numbers. Applying SR entrywise to $\mathbf{A} \in \mathbb{R}^{n \times d}$ gives the stochastically rounded version $\tilde{\mathbf{A}} \in \mathcal{F}^{n \times d}$.

Our main result, Theorem 4.1, combined with (4.7) proves a lower bound for the smallest singular value of $\tilde{\mathbf{A}}$. Formally, let $\sigma_d(\cdot)$ denote the smallest singular value of an $n \times d$ matrix where $n \geq d$; let β be the basis and p be the working precision of the floating point representation; and assume for simplicity of exposition that all entries of \mathbf{A} are in the interval $[-1, 1]$. (See section 4.2 for the general case where the entries of \mathbf{A} are arbitrary.) We prove that SR guarantees, with high probability, the following absolute bound for the smallest singular value of the stochastically rounded \mathbf{A} :

$$(1.1) \quad \sigma_d(\tilde{\mathbf{A}}) \geq \beta^{1-p} \sqrt{n} (\sqrt{\nu} - \epsilon_{n,d}).$$

Here $0 \leq \nu \leq 1$ is the *minimum normalized variance* of the SR process over all columns of \mathbf{A} (defined in section 4.2), and $\epsilon_{n,d}$ captures *lower-order* terms that depend only on the dimensions n and d of \mathbf{A} .

We discuss the parameters to interpret the above bound.

1. As the “tall” dimension n of the matrices \mathbf{A} and $\tilde{\mathbf{A}}$ grows, the smallest singular value of the rounded matrix $\tilde{\mathbf{A}}$ increases. This is because the columns of $\tilde{\mathbf{A}}$ have more opportunity to be linearly independent, as they become longer.
2. The parameter ν , formally defined in section 4.2, captures the stochasticity affecting the rounding of the entries of \mathbf{A} .

To intuitively understand the importance of ν , consider the special case when \mathbf{A} consists of two identical columns whose entries are elements of \mathcal{F} . Any rounding process, including SR, keeps the matrix intact, so that $\tilde{\mathbf{A}} = \mathbf{A}$ has

two identical columns. Therefore, the smallest singular values of \mathbf{A} and $\tilde{\mathbf{A}}$ are equal to zero, since both matrices are rank-deficient. In that case, $\nu = 0$, since there is no flexibility in the rounding process.

Indeed, SR is most powerful when the two points in \mathcal{F} that enclose the entry to be rounded have meaningful probabilities associated with them.

3. The parameter $\epsilon_{n,d}$ captures *lower-order* terms that depend only on the dimensions n and d of the input matrix. Corollary 4.2 states that if $\mathbf{A} \in \mathbb{R}^{n \times d}$ is sufficiently tall and thin, that is, $d = o((n/\log n)^{1/4})$, then

$$\lim_{n \rightarrow \infty} \epsilon_{n,d} = 0.$$

Thus, we can drop $\epsilon_{n,d}$ from the bound (1.1), which gives

$$(1.2) \quad \sigma_d(\tilde{\mathbf{A}}) \gtrsim \beta^{1-p} \sqrt{n\nu}.$$

This estimate is strongly supported by our empirical evaluations in section 5 and essentially matches our lower bound in section 4.6. We conjecture that (1.2) characterizes the true behavior of SR on essentially all tall-and-thin matrices.

An interesting aspect of (1.1) and (1.2) is that they *do not* depend on the closeness of \mathbf{A} to rank-deficiency. Thus, SR guarantees that the stochastically rounded matrix $\tilde{\mathbf{A}}$ invariably has its smallest singular value bounded away from zero and thus has full column rank.

Our proof techniques build upon results from random matrix theory (RMT). Of particular importance is Theorem 4.3, which first appeared in [16] and bounds the smallest singular value of matrices whose entries are independent but not identically distributed random variables. Our proof first decomposes the matrix of rounding errors into two components, and then bounds the smallest singular value of the first component via RMT, and the norm of the second component with a scalar concentration inequality. The idea is that there is no concentration of error in low-dimensional subspaces.

The paper is organized as follows. Section 2 presents notation and basic background, including SR; section 3 discusses prior work; section 4 presents and discusses our bounds; section 5 presents our experimental evaluations; section 6 discusses future research directions; and Appendix B presents a singular value bound for Gaussian perturbations.

2. Background. We define notation in section 2.1, review SR in section 2.2, bound the deviation of a sum random variables from its expectation in section 2.4, and recall the union bound for probabilities in section 2.5.

2.1. Notation. We use bold uppercase letters to denote matrices and bold lowercase letters to denote vectors. The singular values of a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ with $n \geq d$ are denoted by $\sigma_1(\mathbf{A}) \geq \dots \geq \sigma_d(\mathbf{A}) \geq 0$. We use standard notation for matrix and vector norms and denote the natural logarithm of n by $\log n$.

The expectation of a random variable X is denoted by $\mathbb{E}[X]$ and its variance by $\text{Var}[X]$. The probability of an event \mathcal{E} is $0 \leq \Pr[\mathcal{E}] \leq 1$. The overbar in $\bar{\mathcal{E}}$ represents the complement of the event \mathcal{E} . Recall that $\Pr[\mathcal{E}] \leq p$ is equivalent to $\Pr[\bar{\mathcal{E}}] \geq 1 - p$.

The statement $f(n) = o(g(n))$ means

$$\lim_{n \rightarrow \infty} f(n)/g(n) = 0.$$

2.2. Stochastic rounding and its properties. We review the SR model in [4].

Let $\mathcal{F} \subset \mathbb{R}$ be a fixed, finite set of numbers. For a number real $x \in [\min \mathcal{F}, \max \mathcal{F}]$, we represent the enclosing numbers in \mathcal{F} by

$$(2.1) \quad \lceil x \rceil = \min\{y \in \mathcal{F} : y \geq x\} \quad \text{and} \quad \lfloor x \rfloor = \max\{y \in \mathcal{F} : y \leq x\}.$$

For $\lceil x \rceil \neq \lfloor x \rfloor$, *SR-nearness* of $x \in \mathbb{R}$ is defined as

$$\text{StochasticRound}(x) = \begin{cases} \lceil x \rceil & \text{with probability } \frac{x - \lfloor x \rfloor}{\lceil x \rceil - \lfloor x \rfloor}, \\ \lfloor x \rfloor & \text{otherwise.} \end{cases}$$

If $\lfloor x \rfloor = \lceil x \rceil$, then $\text{StochasticRound}(x) = x$. SR-nearness produces an unbiased estimator of x , namely

$$(2.2) \quad \mathbb{E}[\text{StochasticRound}(x)] = x.$$

The generalization of SR-nearness to matrices is immediate. If $\mathbf{A} \in \mathbb{R}^{n \times d}$, then SR-nearness produces the random matrix $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times d}$ with elements

$$(2.3) \quad \tilde{\mathbf{A}}_{ij} = \text{StochasticRound}(\mathbf{A}_{ij}), \quad 1 \leq i \leq n, 1 \leq j \leq d.$$

The random matrix of absolute SR rounding errors is

$$(2.4) \quad \mathbf{E} \equiv \tilde{\mathbf{A}} - \mathbf{A}.$$

The entries of \mathbf{E} are independent random variables, and the matrix-valued expectation equals $\mathbb{E}[\mathbf{E}] = \mathbf{0}$.

2.3. Normalized floating point numbers. In the context of a basis β and working precision p , a *normalized* floating point number x can be uniquely represented as [2]

$$x = s \cdot m \cdot \beta^{e-p},$$

where $s = \pm 1$ is the sign, e is the exponent, and the *significand* m is an integer in the interval

$$\beta^{p-1} \leq m < \beta^p.$$

With \mathcal{F} representing the set of normalized floating point numbers, the rounding model in section 2.2 is exactly the SR-nearness model from [4], and as in [4], we ignore numerical overflow and underflow.

Suppose the real number $x \in [\min \mathcal{F}, \max \mathcal{F}]$ is not an element of \mathcal{F} . According to (2.1), x is enclosed by the successive floating point numbers $\lfloor x \rfloor$ and $\lceil x \rceil$. As in deterministic floating point arithmetic, the SR floating point version is either $\lfloor x \rfloor$ or $\lceil x \rceil$, with an error of¹

$$(2.5) \quad \max\{|x - \lfloor x \rfloor|, |x - \lceil x \rceil|\} \leq \beta^{1-p}|x|.$$

The absolute distance between the two enclosing floating numbers equals [4, section II.A, Figure 1]

$$(2.6) \quad \lceil x \rceil - \lfloor x \rfloor = \beta^{e-p}.$$

¹Compared to SR-nearness, the bound for the IEEE-754 RN mode (round-to-nearest, ties to even) is tighter by a factor of $1/2$.

2.4. Bounding the sum of random variables. The following well-known inequality by Hoeffding bounds the deviation of a sum of n independent random variables from its expectation. The slight restatement below is adapted to our context.

THEOREM 2.1 (Theorem 2 in [25]). *Let X_1, \dots, X_n be independent random variables with $m_i \leq X_i \leq M_i$, $1 \leq i \leq n$. Then, for any $t > 0$,*

$$\mathbb{P} \left(\left| \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \right| \geq t \right) \leq 2 \exp \left(\frac{-2t^2}{\sum_{i=1}^n (M_i - m_i)^2} \right).$$

2.5. The union bound. The following union bound states that the probability of the union of k events \mathcal{E}_i , $1 \leq i \leq k$, is bounded above by the sum of the individual probabilities, i.e.,

$$\Pr[\mathcal{E}_1 \cup \dots \cup \mathcal{E}_k] \leq \sum_{i=1}^k \Pr[\mathcal{E}_i].$$

3. Prior work. There exists a large body of prior work discussing the implementation and error analysis of SR in the context of floating point arithmetic. In the 1950s, Forsythe [20] modeled round-off errors as random variables. A few years later, Hull and Swenson [27] presented probabilistic models for round-off errors and concluded that such models are, in general, very good in theory and in practice. The focus of our work is the SR-nearness mode, which first appeared in [32], while the SR-up-or-down mode was analyzed in [36]. Other work analyzed SR for the heat equation [13], proved that SR prevents stagnation [11], analyzed SR for floating point summation [22], and proposed alternative frameworks to characterize SR errors for sequential summation and Horner's method for evaluating polynomials based on the computation of the variance and Chebyshev's inequality instead of martingales [5]. Software emulators of SR include Verificarlo [15], Verrou [18], and Cadna [28]. Finally, [12] presents a survey of error analysis and applications of SR including a more general analysis of SR errors that are not necessarily independent, but only weakly independent.

Our contribution is to demonstrate that SR rounding tends to increase the smallest singular value of tall-and-thin matrices, thus performing implicit regularization when these matrices are used in downstream machine learning and data analysis applications. Our work was partially motivated by [34], where Gaussian elimination without pivoting is analyzed in the smoothed complexity model, and the smallest singular value of $\mathbf{A} + \mathbf{E}$ is bounded from below, for a matrix \mathbf{E} whose entries are independent, identically distributed Gaussian random variables. While the proof techniques of [34] are not directly portable to our setting, the motivation is somewhat similar.

Finally, our own prior work [7] demonstrates both theoretically and experimentally that perturbing a real matrix \mathbf{A} of full column rank can potentially increase the smallest singular values under certain assumptions involving singular value gaps, thus establishing a qualitative model for the increase in small singular values after a matrix has been downcast to a lower arithmetic precision. However, the bounds in [7] have a different flavor and are not directly comparable to the bounds here.

4. Bounding the smallest singular value of the perturbed matrix. We present a simple example for the effect of SR-nearness on the smallest singular value in section 4.1, derive a singular value bound in section 4.2 for general random errors and in section 4.3 for SR-nearness errors, review the RMT result for our proof in section 4.4, and finally present the proof of Theorem 4.1 in section 4.5 and its tightness in section 4.6.

The goal is to quantify how SR-nearness applied to a tall and skinny matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ with $n \gg d$ increases its smallest singular value. This is in contrast to deterministic rounding, which can drive the smallest singular value to zero.

A first approach for bounding the smallest singular value $\sigma_d(\mathbf{A} + \mathbf{E})$ of the rounded matrix $\mathbf{A} + \mathbf{E}$ might rely on Weyl's inequality,

$$(4.1) \quad \sigma_d(\mathbf{A} + \mathbf{E}) \geq \sigma_d(\mathbf{A}) - \|\mathbf{E}\|_2.$$

However, (4.1) is not informative if $\sigma_d(\mathbf{A}) = 0$. More generally, (4.1) produces positive bounds only if the smallest singular value of \mathbf{A} is larger than $\|\mathbf{E}\|_2$.

This is the reason why we are exploring lower bounds that *do not* depend on the smallest singular value of \mathbf{A} . We argue that, under appropriate assumptions, the smallest singular value of the rounded matrix $\mathbf{A} + \mathbf{E}$ cannot be too small, and that its value does not depend on the singular values of the original matrix \mathbf{A} .

To this end, we employ RMT to derive bounds that *only depend on the distribution* of the entries of \mathbf{E} . Our lower bounds for $\sigma_d(\mathbf{A} + \mathbf{E})$ can be positive even if $\sigma_d(\mathbf{A}) = 0$. Therefore, SR-nearness can *increase* the smallest singular value even in the extreme case of rank-deficient matrices

4.1. A simple example. We illustrate the effect of SR on the smallest singular value of \mathbf{A} , in the special case where $\mathbf{A} \in \mathbb{R}^{n \times 2}$ is a rank-one matrix, all of whose entries are equal to $1/2$. Suppose we want to round \mathbf{A} so as to represent each entry in terms of a single bit, i.e., $\mathcal{F} = \{0, 1\}$.

A deterministic model that rounds $1/2$ to one produces a rounded matrix that is rank-deficient as well. In contrast, SR-nearness sets each entry of $\tilde{\mathbf{A}}$ to zero or one with equal probability. Hence $\text{rank}(\tilde{\mathbf{A}}) = 1$ only if the two columns of $\tilde{\mathbf{A}}$ are identical—an event whose probability becomes exponentially small as n increases. Here is an example of what SR-nearness may look like for an 8×2 matrix:

$$\mathbf{A} = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \\ 1/2 & 1/2 \\ 1/2 & 1/2 \\ 1/2 & 1/2 \\ 1/2 & 1/2 \\ 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}, \quad \tilde{\mathbf{A}} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

In Appendix A we prove that the smallest singular value of the $n \times 2$ matrix $\tilde{\mathbf{A}}$ satisfies the following lower bound, with probability at least 0.997:

$$(4.2) \quad \sigma_2^2(\tilde{\mathbf{A}}) \geq 0.25 \cdot n - 8\sqrt{n} = \Omega(n) \Rightarrow \sigma_2(\tilde{\mathbf{A}}) \geq \Omega(\sqrt{n}).$$

This illustrates that SR-nearness is highly likely to produce a significant increase in the smallest singular value.

4.2. General random perturbation. Our main result in Theorem 4.1 bounds the smallest singular value of the rounded matrix $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{E} \in \mathbb{R}^{n \times d}$ away from zero, provided (i) there is enough randomness in the perturbation \mathbf{E} , and (ii) \mathbf{A} is sufficiently tall and thin.

We define the minimum normalized columnwise variance ν of \mathbf{E} as follows:

$$(4.3) \quad \nu \equiv \frac{1}{n\mathcal{R}^2} \min_{1 \leq j \leq d} \sum_{i=1}^n \text{Var}(\mathbf{E}_{ij}) \quad \text{with} \quad \max_{i,j} |\mathbf{E}_{ij}| \leq \mathcal{R}.$$

Clearly, $\text{Var}(\mathbf{E}_{ij}) = \text{Var}(\tilde{\mathbf{A}}_{ij})$, which implies that $\text{Var}(\mathbf{E}_{ij}) \leq \mathcal{R}^2$. Therefore,

$$0 \leq \nu \leq 1.$$

Intuitively, ν characterizes the amount of randomness in SR-nearness. Our main theorem below depends on the minimal columnwise variance of the perturbation \mathbf{E} .

THEOREM 4.1. *Let \mathbf{A} and $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$ be real $n \times d$ matrices with $n \gg d$. Here \mathbf{E} models random perturbations with minimal normalized column variance ν and \mathcal{R} from (4.3).*

If $n \geq 836$, then with probability at least $1 - \frac{1}{n^c} - \frac{2d^2}{n^2}$,

$$\sigma_d(\tilde{\mathbf{A}}) \geq \mathcal{R}\sqrt{n}(\sqrt{\nu} - \varepsilon_{n,d}),$$

where

$$(4.4) \quad \varepsilon_{n,d} \equiv \sqrt{\frac{d}{n}} + 2d^2 \sqrt{\frac{\log n}{n}} + \frac{C(\log n)^{2/3}}{n^{1/30}} \cdot \left(\frac{d}{n}\right)^{1/54},$$

and c and C are absolute constants.²

The comments below provide intuition for Theorem 4.1.

1. The lower bound for $\sigma_d(\tilde{\mathbf{A}})$ is very general; it holds for any random matrix \mathbf{E} , regardless of whether it models SR-nearness errors or not. However, Theorem 4.1 requires that \mathbf{E} merely change entries of \mathbf{A} by a small amount, quantified by \mathcal{R} . In other words, a small value of \mathcal{R} prevents large changes in individual entries of \mathbf{A} , thereby preventing them from exerting disproportionate influence on the smallest singular value of $\tilde{\mathbf{A}}$.
2. For the bound to be positive, we need $\varepsilon_{n,d} \leq \sqrt{\nu} \leq 1$. If

$$(4.5) \quad d = o((n/\log n)^{1/4}),$$

then the first two terms of $\varepsilon_{n,d}$ must approach 0 as $n \rightarrow \infty$. This is because $d = o((n/\log n)^{1/4})$ is equivalent to $\lim_{n \rightarrow \infty} \frac{d^4 \log n}{n} = 0$, which in turn implies

$$\lim_{n \rightarrow \infty} 2d^2 \sqrt{\frac{\log n}{n}} = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} \sqrt{\frac{d}{n}} = 0.$$

The third term also approaches 0 as $n \rightarrow \infty$, because

$$(\log n)^{2/3} = o(n^{1/30}) \iff \lim_{n \rightarrow \infty} (\log n)^{2/3}/n^{1/30} = 0.$$

The ratio $(\log n)^{2/3}/n^{1/30}$ goes to zero slowly as n grows. For example, n needs to be larger than 10^{50} for this ratio to drop to $1/2$. An important question for future research is the strengthening of Theorem 4.1 to reduce $\varepsilon_{n,d}$.

In the appendix, we prove Theorem B.1 for the special case where the elements \mathbf{E}_{ij} are independent, identically distributed Gaussian normal random variables and show that the smallest singular value of $\tilde{\mathbf{A}}$ is again bounded away from zero with high probability.

However, Theorem B.1 is much sharper than Theorem 4.1. For example, even if $n = 900$ and $d = 25$, the bound $\sigma_d(\tilde{\mathbf{A}}) \geq 1$ holds with probability at least

²These constants are unspecified in [16, Theorem 2.10 and Remark 2.11].

0.98 and does not need a bound on $\max_{i,j} |\mathbf{E}_{ij}|$; it suffices that the variance of \mathbf{E}_{ij} is bounded. Hence it might be possible to significantly reduce $\epsilon_{n,d}$ in Theorem 4.1 or even eliminate it, via RMT work on non-Gaussian, nonidentically distributed perturbations. The experiments in section 5 illustrate that $\epsilon_{n,d}$ is an artifact of our analysis and not a real concern in practice.

3. The success probability $1 - \frac{1}{n^c} - \frac{2d^2}{n^2}$ approaches 1 as $n \rightarrow \infty$, in spite of the unspecified absolute constant c . This is because the assumption (4.5) on d implies

$$\lim_{n \rightarrow \infty} 1/n^c = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} 2d^2/n^2 = 0.$$

The above discussion implies the following bound for sufficiently tall and skinny matrices.

COROLLARY 4.2. *Under the assumptions of Theorem 4.1 if also*

$$d = o((n/\log n)^{1/4}),$$

then with probability approaching one,

$$(4.6) \quad \sigma_d(\tilde{\mathbf{A}}) \gtrsim \mathcal{R} \sqrt{n \cdot \nu}.$$

4.3. Perturbations from SR-nearness. Applying Theorem 4.1 to SR-nearness with normalized floating point numbers \mathcal{F} is straightforward.

First, we focus on the case where \mathbf{A}_{ij} round to normalized numbers in $[-1, 1]$ closest to zero, $1 \leq i \leq n$, $1 \leq j \leq d$. From (2.5) follows that the error in the elements of $\tilde{\mathbf{A}} \in \mathcal{F}^{n \times d}$ is at most β^{1-p} . Thus

$$(4.7) \quad \max_{i,j} |\mathbf{A}_{ij} - \tilde{\mathbf{A}}_{ij}| = \max_{i,j} |\mathbf{E}_{ij}| \leq \mathcal{R} \equiv \beta^{1-p} |\mathbf{A}_{ij}|.$$

Theorem 4.1 and Corollary 4.2 apply immediately with \mathcal{R} from (4.7).

Now assume that $\mathbf{A} \in \mathbb{R}^{n \times d}$ is a general matrix. From (2.6) follows that the error in $\tilde{\mathbf{A}}_{ij}$ is at most $\beta^{e_{ij}-p}$, where e_{ij} is the exponent for \mathbf{A}_{ij} , $1 \leq i \leq n$, $1 \leq j \leq d$. Thus

$$(4.8) \quad \max_{i,j} |\mathbf{A}_{ij} - \tilde{\mathbf{A}}_{ij}| = \max_{i,j} |\mathbf{E}_{ij}| \leq \max_{i,j} \beta^{e_{ij}-p} \leq \mathcal{R} \equiv \beta^{e_{\max}-p},$$

where $e_{\max} = \max_{i,j} \{e_{ij}\}$. Theorem 4.1 and Corollary 4.2 immediately apply with \mathcal{R} from (4.8).

4.4. A random matrix theory bound. We present the basis for our proof, which is a lower bound on the minimum singular value from [16, Theorem 2.10 and Remark 2.11] for matrices whose elements are independent zero-mean random numbers that are not necessarily identically distributed. This latter fact is crucial, since the elements of $\mathbf{E} = \tilde{\mathbf{A}} - \mathbf{A}$ are independent but not identically distributed.³

THEOREM 4.3 (Theorem 2.10 and Remark 2.11 in [16]). *Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ with $n \geq d$ have independent zero-mean entries, so that $\mathbb{E}[\mathbf{X}] = \mathbf{0}$. Suppose there are $q > 0$, $\kappa \geq 1$, and $0 < \gamma \leq 1$ such that*

³To be precise, we adapt results from the arXiv version [16], specifically in the regime where $\gamma \rightarrow 0$, as the published version [17] provides results only for the case when $\gamma = O(1)$. See also [10] for an overview and results on matrix concentration inequalities that have a similar flavor to the bounds used in our paper.

$$(4.9) \quad \max_{1 \leq i \leq n, 1 \leq j \leq d} |\mathbf{X}_{ij}| \leq 1/q,$$

$$(4.10) \quad \max_{1 \leq i \leq n, 1 \leq j \leq d} \mathbb{E}|\mathbf{X}_{ij}|^2 \leq \kappa/n,$$

$$(4.11) \quad \max_{1 \leq j \leq d} \sum_{i=1}^n \mathbb{E}|\mathbf{X}_{ij}|^2 \leq 1,$$

$$(4.12) \quad \max_{1 \leq i \leq n} \sum_{j=1}^d \mathbb{E}|\mathbf{X}_{ij}|^2 \leq \gamma,$$

$$(4.13) \quad \min_{1 \leq j \leq d} \sum_{i=1}^n \mathbb{E}|\mathbf{X}_{ij}|^2 \geq \tilde{\rho}_{\min} \geq \sqrt{\gamma}.$$

If also

$$(4.14) \quad \sqrt{\log n} \leq q \leq n^{\frac{1}{10}} \kappa^{-\frac{1}{9}} \gamma^{-\frac{1}{18}},$$

then, with probability at least $1 - n^{-c}$,

$$\sigma_d(\mathbf{X}) \geq \sqrt{\tilde{\rho}_{\min}} - \sqrt{\gamma} - Cq^{-1/3}(\log n)^{2/3},$$

where c and C are absolute constants.

Assumptions (4.9)–(4.13) in Theorem 4.3 require the following quantities associated with \mathbf{X} to be sufficiently small: the elements, the variances of the elements, the maximal variance of the columns, and the maximal variance of the rows. However, to guarantee enough randomness, the minimal variance of the columns is not allowed to be too small.

Assumption (4.14) implies $q \geq \sqrt{\log n} \geq 3$, so that in (4.9) all elements in \mathbf{X} must be bounded by one in magnitude,

$$\max_{1 \leq i \leq n, 1 \leq j \leq d} |\mathbf{X}_{ij}| \leq 1/q \leq 1.$$

The following example supplies intuition for Theorem 4.3 and previews its proof in section 4.5. It illustrates that for sufficiently tall and skinny matrices, the lower bound for the smallest singular value is meaningful and holds with a success probability close to one.

Example 1. Consider Theorem 4.3 in the special case when all elements of \mathbf{X} have the same variance. We show that this variance equals $1/n$, thus decreases with increasing row dimension. Since $\gamma = d/n$, the lower bound holds with probability close to 1 for sufficiently tall and skinny matrices.

Suppose that in assumption (4.10)

$$\mathbb{E}|\mathbf{X}_{ij}|^2 = \kappa/n, \quad 1 \leq i \leq n, \quad 1 \leq j \leq d,$$

for some $\kappa \geq 1$. Inserting this into the column variances (4.11) gives

$$\kappa = \max_{1 \leq j \leq d} \sum_{i=1}^n \mathbb{E}|\mathbf{X}_{ij}|^2 \leq 1.$$

This, together with $\kappa \geq 1$, implies $\kappa = 1$, so that all elements have variance

$$\mathbb{E}|\mathbf{X}_{ij}|^2 = 1/n, \quad 1 \leq i \leq n, \quad 1 \leq j \leq d.$$

Thus, the variance decreases with increasing row dimension number. Inserting $\mathbb{E}|\mathbf{X}_{ij}|^2 = 1/n$ into the row variances (4.12) gives

$$\frac{d}{n} = \max_{1 \leq i \leq n} \sum_{j=1}^d \mathbb{E}|\mathbf{X}_{ij}|^2 \leq \gamma.$$

This together with the assumption $n \geq d$ implies $\gamma = d/n \leq 1$, as required. Inserting $\mathbb{E}|\mathbf{X}_{ij}|^2 = 1/n$ into the minimal column variances (4.13) gives

$$1 = \max_{1 \leq j \leq d} \sum_{i=1}^n \mathbb{E}|\mathbf{X}_{ij}|^2 = \tilde{\rho}_{\min}$$

so that $1 = \tilde{\rho}_{\min} \geq \sqrt{\gamma} = \sqrt{d/n}$ in (4.13) is automatically fulfilled. The lower bound for $\sigma_d(\mathbf{X})$ becomes

$$\sigma_d(\mathbf{X}) \geq 1 - \sqrt{d/n} - Cq^{-1/3}(\log n)^{2/3}.$$

Thus, the smallest singular value approaches 1 with increasing probability as \mathbf{X} becomes taller and skinnier.

An open problem for future work is a better understanding for which assumptions (4.9)–(4.14) are truly necessary for Theorem 4.3 to hold. We conjecture that many can be relaxed or even eliminated. Indeed, any improvement of Theorem 4.3 could result in significant improvements for and generalizations of our Theorem 4.1.

4.5. Proof of Theorem 4.1. First we present a brief outline of the proof, and then the proof proper.

Outline of proof. The proof consists of four main steps.

1. We introduce the orthogonal projector $\mathbf{P}_{\mathbf{A}}$ onto the column space of \mathbf{A} . This allows us to focus on $\mathbf{P}_{\mathbf{A}}\mathbf{E}$. Weyl's inequality yields a lower bound on the smallest singular value of $(\mathbf{I} - \mathbf{P}_{\mathbf{A}})\mathbf{E}$ by lower bounding the smallest singular value of \mathbf{E} and upper bounding the largest singular value of $\mathbf{P}_{\mathbf{A}}\mathbf{E}$.
2. Application of Theorem 4.3 shows that the smallest singular value of \mathbf{E} is sufficiently large.
3. The largest singular value of the projection $\mathbf{P}_{\mathbf{A}}\mathbf{E}$ is small, because $\mathbf{P}_{\mathbf{A}}$ projects \mathbf{E} on the low-dimensional subspace of dimension d , and application of Hoeffding's inequality shows that \mathbf{E} does not concentrate well in any low-dimensional subspace.
4. At last, we combine the bounds for the smallest singular value of \mathbf{E} and the largest singular value of $\mathbf{P}_{\mathbf{A}}\mathbf{E}$ via a union bound on their probabilities.

Proof. We follow the four steps outlined above.

1. We decompose the task of lower bounding $\sigma_d(\mathbf{A} + \mathbf{E})$ into two parts with the help of an orthogonal projector $\mathbf{P} \in \mathbb{R}^{n \times n}$. The effect of the orthogonal projector can be quantified by the singular value product inequalities [26, Theorem 3.3.16], which together with $\|\mathbf{P}\|_2 = 1$ imply

$$(4.15) \quad \sigma_d(\mathbf{P}(\mathbf{A} + \mathbf{E})) \leq \sigma_d(\mathbf{A} + \mathbf{E})\|\mathbf{P}\|_2 \leq \sigma_d(\mathbf{A} + \mathbf{E}).$$

Let $\mathbf{P}_{\mathbf{A}} \in \mathbb{R}^{n \times n}$ be the orthogonal projector on the column space of \mathbf{A} and $\mathbf{P}_{\mathbf{A},\perp} \equiv \mathbf{I}_n - \mathbf{P}_{\mathbf{A}}$ be the orthogonal projector onto the left nullspace of \mathbf{A} .

From (4.15) follows

$$(4.16) \quad \begin{aligned} \sigma_d(\mathbf{A} + \mathbf{E}) &\geq \sigma_d(\mathbf{P}_{\mathbf{A},\perp}(\mathbf{A} + \mathbf{E})) = \sigma_d(\mathbf{P}_{\mathbf{A},\perp}\mathbf{A} + \mathbf{P}_{\mathbf{A},\perp}\mathbf{E}) \\ &= \sigma_d(\mathbf{P}_{\mathbf{A},\perp}\mathbf{E}). \end{aligned}$$

Weyl's inequality [6, Theorem III.2.1] implies

$$(4.17) \quad \sigma_d(\mathbf{P}_{\mathbf{A},\perp}\mathbf{E}) = \sigma_d(\mathbf{E} - \mathbf{P}_{\mathbf{A}}\mathbf{E}) \geq \sigma_d(\mathbf{E}) - \|\mathbf{P}_{\mathbf{A}}\mathbf{E}\|_2.$$

We have now broken our task into two parts. First, we must make sure that the smallest singular value of the random matrix \mathbf{E} is large enough. Second, the projection of \mathbf{E} onto the d -dimensional column space of \mathbf{A} must be small, that is, \mathbf{E} should not concentrate in any d -dimensional space.

2. To bound $\sigma_d(\mathbf{E})$ from below, set $\mathbf{X} = \frac{1}{\mathcal{R}\sqrt{n}}\mathbf{E}$, where $\mathbb{E}[\mathbf{X}] = \mathbf{0}$. We show that \mathbf{X} satisfies the assumptions of Theorem 4.3 by setting $\kappa = 1$, $\gamma = d/n$, $\tilde{\rho}_{\min} = \nu$, and $q = n^{1/10} \cdot \gamma^{-1/18}$,
From $\max_{i,j} |\mathbf{E}_{ij}| \leq \mathcal{R}$ and $q = n^{1/10} \cdot \gamma^{-1/18}$ follows (4.9),

$$\max_{1 \leq i \leq n, 1 \leq j \leq d} |\mathbf{X}_{ij}| \leq \frac{1}{\sqrt{n}} \leq \frac{1}{q}.$$

This implies (4.10),

$$\max_{1 \leq i \leq n, 1 \leq j \leq d} \mathbb{E}|\mathbf{X}_{ij}|^2 \leq \frac{1}{n} = \frac{\kappa}{n},$$

which in turn implies (4.11),

$$\max_{1 \leq j \leq d} \sum_{i=1}^n \mathbb{E}|\mathbf{X}_{ij}|^2 \leq \sum_{i=1}^n \frac{1}{n} = 1,$$

as well as (4.12),

$$\max_{1 \leq i \leq n} \sum_{j=1}^d \mathbb{E}|\mathbf{X}_{ij}|^2 \leq \sum_{j=1}^d \frac{1}{n} = \frac{d}{n} = \gamma.$$

From

$$\mathbb{E}[\mathbf{X}_{ij}^2] = \mathbb{E}[\mathbf{X}_{ij}]^2 + \text{Var} \left[\frac{1}{\mathcal{R}\sqrt{n}}\mathbf{E}_{ij} \right] = \frac{1}{n \cdot \mathcal{R}^2} \text{Var}[\mathbf{E}_{ij}]$$

and (4.3) follows (4.13)

$$\min_{1 \leq j \leq d} \sum_{i=1}^n \mathbb{E}|\mathbf{X}_{ij}|^2 = \min_{1 \leq j \leq d} \sum_{i=1}^n \frac{1}{\mathcal{R}^2 \cdot n} \cdot \text{Var}[\mathbf{E}_{ij}] = \nu = \tilde{\rho}_{\min}.$$

We now focus on the assumption (4.14), $\sqrt{\log n} \leq q \leq n^{1/10} \kappa^{-1/9} \gamma^{-1/18}$. The second inequality is immediately satisfied, because $\kappa = 1$ and $q = n^{1/10} \cdot \gamma^{-1/18}$. The first inequality is equivalent to

$$\log n \leq n^{14/45} \cdot d^{-1/9},$$

hence

$$(4.18) \quad d \leq \frac{n^{14/5}}{(\log n)^9}.$$

Let the assumption $n \geq 836$ hold, and suppose $d^4 \geq n$. Then $2d^2 \sqrt{\log n/n} > 2$ and Theorem 4.1 vacuously holds, since the right-hand side of the bound is negative.

Now suppose that $d^4 < n$. Then for $n \geq 836$ (4.18) holds,

$$d < n^{1/4} \leq n^{14/5}/(\log n)^9.$$

Thus, \mathbf{X} satisfies the assumptions of Theorem 4.3, so that with probability at least $1 - n^{-c}$,

$$\sigma_d(\mathbf{X}) \geq \sqrt{\nu} - \sqrt{\frac{d}{n}} - \frac{C(\log n)^{2/3}}{n^{1/30}} \cdot \left(\frac{d}{n}\right)^{1/54},$$

where c and C are absolute constants. Recall the definition of \mathbf{X} and multiply both sides by $\mathcal{R} \cdot \sqrt{n}$,

$$(4.19) \quad \sigma_d(\mathbf{E}) \geq \mathcal{R}\sqrt{\nu \cdot n} - \mathcal{R}\sqrt{d} - \frac{C \cdot \mathcal{R} \cdot (\log n)^{2/3}}{n^{1/30}} \cdot \left(\frac{d}{n}\right)^{1/54} \cdot \sqrt{n}.$$

3. To bound $\|\mathbf{P}_A \mathbf{E}\|_2$ from above, first consider the case where \mathbf{P}_A projects onto a one-dimensional space, so that $\mathbf{P}_A = \mathbf{u}\mathbf{u}^T$ for some unit vector $\mathbf{u} \in \mathbb{R}^n$, and

$$\|\mathbf{P}_A \mathbf{E}\|_2 = \|\mathbf{u}\mathbf{u}^T \mathbf{E}\|_2 = \|\mathbf{E}^T \mathbf{u}\|_2.$$

We now bound each entry of $\mathbf{E}^T \mathbf{u}$ via Theorem 2.1. First check the assumptions. From $\max_{i,j} |\mathbf{E}_{ij}| \leq \mathcal{R}$ follows $|\mathbf{E}_{ij} \mathbf{u}_i| \leq |\mathbf{u}_i| \cdot \mathcal{R}$ for a fixed unit vector $\mathbf{u} \in \mathbb{R}^n$.

Hence, the values of the random variable $\mathbf{E}_{ij} \mathbf{u}_i$ are in the interval $[m_{ij}, M_{ij}]$, with $(M_{ij} - m_{ij})^2 \leq 4\mathbf{u}_i^2 \mathcal{R}^2$. From $\|\mathbf{u}\|_2 = 1$ follows

$$\sum_{i=1}^n (M_{ij} - m_{ij})^2 \leq 4\mathcal{R}^2.$$

Applying Theorem 2.1 gives

$$\begin{aligned} \mathbb{P} \left[\left| \sum_{i=1}^n \mathbf{E}_{ij} \mathbf{u}_i \right| \geq t \right] &\leq 2 \exp \left(\frac{-2t^2}{\sum_{i=1}^n (M_{ij} - m_{ij})^2} \right), \quad 1 \leq j \leq d, \\ &\leq 2 \exp \left(\frac{-t^2}{2\mathcal{R}^2} \right). \end{aligned}$$

The vector norm relations

$$\|\mathbf{E}^T \mathbf{u}\|_2 \leq \|\mathbf{E}^T \mathbf{u}\|_1 = \sum_{j=1}^d \left| \sum_{i=1}^n \mathbf{E}_{ij} \mathbf{u}_i \right|$$

followed by application of the above inequality with $t = 2\mathcal{R}\sqrt{\log n}$ and application of the union bound from section 2.5 over all $j = 1 \dots d$ gives

$$(4.20) \quad \mathbb{P} \left(\|\mathbf{E}^T \mathbf{u}\|_2 \geq 2d \cdot \mathcal{R}\sqrt{\log n} \right) \leq d \cdot 2 \exp \left(\frac{-4 \cdot \mathcal{R}^2 \log n}{2\mathcal{R}^2} \right) = \frac{2d}{n^2}.$$

We extend this to the case where \mathbf{P}_A projects on a d -dimensional subspace, so that $\mathbf{P}_A = \sum_{j=1}^d \mathbf{u}_j \mathbf{u}_j^T$ for orthonormal vectors $\mathbf{u}_j \in \mathbb{R}^n$. The triangle inequality implies

$$\|\mathbf{P}_A \mathbf{E}\|_2 = \left\| \sum_{j=1}^d \mathbf{u}_j \mathbf{u}_j^T \mathbf{E} \right\|_2 \leq \sum_{j=1}^d \|\mathbf{u}_j \mathbf{u}_j^T \mathbf{E}\|_2.$$

Application of (4.20) and application of the union bound from section 2.5 over all d vectors \mathbf{u}_j gives

$$(4.21) \quad \mathbb{P} \left[\|\mathbf{P}_A \mathbf{E}\|_2 \geq 2d^2 \cdot \mathcal{R} \sqrt{\log n} \right] \leq \frac{2d^2}{n^2}.$$

4. Finally, use $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$ and combine (4.16), (4.17), (4.19), and (4.21) to get

$$\begin{aligned} \sigma_d(\tilde{\mathbf{A}}) &\geq \mathcal{R} \sqrt{\nu \cdot n} - \mathcal{R} \sqrt{d} - \frac{C \cdot \mathcal{R} \cdot \log^{2/3}(n)}{n^{1/30}} \cdot \left(\frac{d}{n}\right)^{1/54} \cdot \sqrt{n} \\ &\quad - 2d^2 \cdot \mathcal{R} \sqrt{\log n}. \end{aligned}$$

Application of a union bound shows that the above holds with probability at least $1 - 1/n^c - 2d^2/n^2$. At last, factor out $\mathcal{R} \sqrt{n}$ and use the abbreviation $\varepsilon_{n,d}$.

4.6. Theorem 4.1 is tight. To show that Theorem 4.1 is asymptotically tight, we exhibit $n \times d$ matrices \mathbf{A} whose stochastically rounded version $\tilde{\mathbf{A}}$ has a smallest singular value no larger than $\mathcal{R} \sqrt{n\nu}$ times a small constant, slightly larger than one, with ν from (4.3).

THEOREM 4.4. *For every $n \geq d$ and $1 \leq \ell \leq n$, there exists a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ with normalized minimum column variance $\nu = \ell/n$, whose stochastically rounded version $\tilde{\mathbf{A}}$ has a smallest singular value*

$$\sigma_d(\tilde{\mathbf{A}}) \leq \left(1 + \sqrt{1/(d-1)}\right) \mathcal{R} \sqrt{\nu n}.$$

Proof. Let $\mathcal{F} = \{0, 1\}$ and set $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$. Fix $\nu \in (0, 1)$ such that $\nu n = \ell$ for some integer $1 \leq \ell \leq d$. Then, $\nu = \ell/n$. Let the entries in the first ℓ rows of \mathbf{A} be equal to $1/2$ and let the remaining $n - \ell$ rows be zero.

Then, $\text{Var}(\mathbf{E}_{ij}) = 1/4$ for $1 \leq i \leq \ell, 1 \leq j \leq d$; $\text{Var}(\mathbf{E}_{ij}) = 0$ for $\ell + 1 \leq i \leq n, 1 \leq j \leq d$; $\mathcal{R} = \max_{i,j} |\mathbf{E}_{ij}| = 1/2$; and

$$\frac{1}{n\mathcal{R}^2} \sum_{i=1}^n \text{Var}(\mathbf{E}_{ij}) = \frac{4}{n} \sum_{i=1}^{\ell} \frac{1}{4} = \frac{\ell}{n} = \nu, \quad 1 \leq j \leq d.$$

By construction, \mathbf{A} has a single nonzero singular value, and $\sigma_j(\mathbf{A}) = 0, 2 \leq j \leq d$. Weyl's inequality [6, Theorem III.2.1] implies

$$(4.22) \quad \sigma_d(\tilde{\mathbf{A}}) = \sigma_d(\mathbf{A} + \mathbf{E}) \leq \sigma_2(\mathbf{A}) + \sigma_{d-1}(\mathbf{E}) = \sigma_{d-1}(\mathbf{E}).$$

Since $\mathbf{E}_{ij} = \pm 1/2$ for $1 \leq i \leq \ell, 1 \leq j \leq d$ and zero otherwise, we have $\|\mathbf{E}\|_F^2 = \ell d/4$. Inserting this into the Frobenius norm

$$\|\mathbf{E}\|_F^2 = \sum_{i=1}^d \sigma_i^2(\mathbf{E}) \geq \sum_{i=1}^{d-1} \sigma_i^2(\mathbf{E}) \geq (d-1) \sigma_{d-1}^2(\mathbf{E})$$

gives

$$\sigma_{d-1}^2(\mathbf{E}) \leq \frac{1}{d-1} \|\mathbf{E}\|_F^2 = \frac{d}{d-1} \cdot \frac{1}{4} \ell = \frac{d}{4(d-1)} n \nu.$$

At last, take square roots, combine with (4.22), and abbreviate $\mathcal{R} = 1/2$. \square

5. Experiments. The numerical experiments illustrate the behavior of the smallest singular value under SR-nearness rounding and support our results from section 4. The experiments are designed so that the effects of fine-grained changes in precision can be easily discerned. The scripts for reproducing the numerical experiment are available in our git repository.⁴

Design of experiments. We investigate the dependence of $\sigma_d(\tilde{\mathbf{A}})$ on various factors, such as the aspect ratio n/d , the smallest singular value $\sigma_d(\mathbf{A})$, and the minimal column variance ν . To provide statistical significance, for each \mathbf{A} , we generate 100 stochastically rounded matrices $\tilde{\mathbf{A}}$ and compute their singular values $\sigma_d(\tilde{\mathbf{A}})$.

Furthermore, we examine both fixed-point arithmetic with base $\beta = 10$ and floating point arithmetic, where we demote the matrix in lower precision using SR-nearness. Experiments in fixed-point arithmetic are presented for rank-deficient matrices (section 5.1.1) and full-rank matrices (section 5.1.2), and in floating point arithmetic for rank-deficient matrices (section 5.2.1) and matrices with controlled ν (section 5.2.2).

Matrices. The matrices have $n = 10^4$ rows and $d = 10, 100, \text{ and } 1000$ columns. Their elements are drawn from different distributions, including skewed and nonsymmetric ones. We start with matrices $\mathbf{A} \in \mathbb{R}^{n \times d}$ whose elements are independent, identically distributed random variables from the standard normal distribution $\mathcal{N}(0, 1)$ or the $\text{Lognormal}(0, 3)$ distribution. Subsequently, we adjust the smallest singular value to fit the desired setting, e.g., for singular \mathbf{A} we force $\sigma_d(\mathbf{A}) = 0$. In the experiments where we control ν , we modify the elements of \mathbf{A} directly and force singularity by setting two columns equal to each other.

We present our main experimental findings in Tables 5.1–5.6 and provide additional plots in Appendix C.

5.1. Experiments with fixed-point arithmetic. SR-nearness rounds the elements of \mathbf{A} to elements of the set $\mathcal{F}^{\{p\}}$, $1 \leq p \leq 3$, where

$$(5.1) \quad \mathcal{F}^{\{p\}} = \left\{ \pm m/10^p \quad \text{for all integers } m = \underbrace{0, 1, 2, \dots, 10^p - 1}_{\leq p \text{ digits}} \right\} \cup \{\pm 1\}.$$

This is equivalent to rounding to a signed base-10 fixed-point precision with at most p digits in the fractional part.

5.1.1. Tables 5.1, 5.2 and Figures C.1, C.2: Rank-deficient matrices.

The rank-deficient matrices \mathbf{A} have a smallest singular value $\sigma_d(\mathbf{A}) = 0$. The goal is to understand how the aspect ratio n/d affects the behavior of $\sigma_d(\tilde{\mathbf{A}})$ on matrices drawn from different distributions.

Remark 5.1. We again emphasize that our bound predicts the behavior of the smallest singular value of the stochastically rounded matrix *asymptotically*. Modifying our bound by just a small constant to be $0.9 \cdot \mathcal{R} \sqrt{n\nu}$ instead of $\mathcal{R} \sqrt{n\nu}$ results in zero

⁴https://github.com/cboutisikas/stoch_rounding_iplicit_reg.

TABLE 5.1

The percentage of matrices (recall that we perform 100 stochastic roundings for each parameter setting; see also Figure C.1) violating the estimate of (1.2). The elements of \mathbf{A} are in $\mathcal{N}(0, 1)$, while the elements $\tilde{\mathbf{A}}_{ij}$ belong to $\mathcal{F}^{\{p\}}$, thus $\|\tilde{\mathbf{A}}_{ij}\| - \|\mathbf{A}_{ij}\| \leq 10^{-p}$, for $p = 1, \dots, 3$. Notice that when the number of columns is $d = 10$ or 100, the bound $\mathcal{R}\sqrt{n\nu}$ is violated in approximately 25% to 50% of the test cases; the bound is always violated for $d = 1,000$ (squarish matrix). However, a very mild relaxation of the bound from $\mathcal{R}\sqrt{n\nu}$ to $0.9 \cdot \mathcal{R}\sqrt{n\nu}$ immediately fixes this issue, resulting in zero violations in all settings. Additionally, the relative error between the estimate provided by our bound and the minimum observed increase in the smallest singular value of the rounded matrix remains consistently below 6%. More precisely, we compute $s_{\min} = \min\{\sigma_{\min}(\tilde{\mathbf{A}})\}$ over all 100 roundings for a specific parameter setting and report the relative error $1 - s_{\min}/\mathcal{R}\sqrt{n\nu}$. In light of the small relative error, it is clear that our bound provides a useful estimate for the magnitude of regularization, even for very modest aspect ratios (see Remark 5.1).

| d | $\%(\sigma_{\min}(\tilde{\mathbf{A}}) < c \cdot \mathcal{R}\sqrt{n\nu})$ | | | | | | $1 - s_{\min}/\mathcal{R}\sqrt{n\nu}$ | | |
|-------|--|--------|-------|--------|-------|--------|---------------------------------------|-------|-------|
| | p = 1 | | p = 2 | | p = 3 | | p = 1 | p = 2 | p = 3 |
| | c = 1 | c = .9 | c = 1 | c = .9 | c = 1 | c = .9 | | | |
| 10 | 26% | 0% | 46% | 0% | 30% | 0% | .01 | .01 | .01 |
| 100 | 48% | 0% | 37% | 0% | 51% | 0% | .02 | .01 | .02 |
| 1,000 | 100% | 0% | 100% | 0% | 100% | 0% | .06 | .06 | .06 |

TABLE 5.2

The percentage of matrices (recall that we perform 100 stochastic roundings for each parameter setting; see also Figure C.2) violating the estimate of (1.2). The elements of \mathbf{A} are in $\text{Lognormal}(0, 3)$, while the elements $\tilde{\mathbf{A}}_{ij}$ belong to $\mathcal{F}^{\{p\}}$, thus $\|\tilde{\mathbf{A}}_{ij}\| - \|\mathbf{A}_{ij}\| \leq 10^{-p}$, for $p = 1, \dots, 3$. Notice that when the number of columns is $d = 10$ or 100, the bound $\mathcal{R}\sqrt{n\nu}$ is violated in approximately 0% to 35% of the test cases; the bound is always violated for $d = 1,000$ (squarish matrix). However, a very mild relaxation of the bound from $\mathcal{R}\sqrt{n\nu}$ to $0.9 \cdot \mathcal{R}\sqrt{n\nu}$ immediately fixes this issue, resulting in zero violations in all settings. Additionally, the relative error between the estimate provided by our bound and the minimum observed increase in the smallest singular value of the rounded matrix remains consistently below 6%. More precisely, we compute $s_{\min} = \min\{\sigma_{\min}(\tilde{\mathbf{A}})\}$ over all 100 roundings for a specific parameter setting and report the relative error $1 - s_{\min}/\mathcal{R}\sqrt{n\nu}$. If $s_{\min} > \mathcal{R}\sqrt{n\nu}$, we mark the respective entry as N/A.

| d | $\%(\sigma_{\min}(\tilde{\mathbf{A}}) < c \cdot \mathcal{R}\sqrt{n\nu})$ | | | | | | $1 - s_{\min}/\mathcal{R}\sqrt{n\nu}$ | | |
|-------|--|--------|-------|--------|-------|--------|---------------------------------------|-------|-------|
| | p = 1 | | p = 2 | | p = 3 | | p = 1 | p = 2 | p = 3 |
| | c = 1 | c = .9 | c = 1 | c = .9 | c = 1 | c = .9 | | | |
| 10 | 0% | 0% | 36% | 0% | 22% | 0% | N/A | .01 | .01 |
| 100 | 0% | 0% | 15% | 0% | 34% | 0% | N/A | .01 | .01 |
| 1,000 | 100% | 0% | 100% | 0% | 100% | 0% | .04 | .05 | .06 |

violations. In this case, the relative error on the right-hand side of the table is not relevant, since even the smallest observed singular value s_{\min} of the rounded matrices would exceed the estimate $0.9 \cdot \mathcal{R}\sqrt{n\nu}$. This highlights that our estimate is very accurate as a lower bound for the regularization even for almost square matrices.

5.1.2. Tables 5.3, 5.4 and Figures C.3, C.4: Full rank. The full-rank matrices \mathbf{A} have the smallest singular value $\sigma_d(\mathbf{A}) = 10^{-2}$. The goal is to understand how the aspect ratio n/d affects the behavior of $\sigma_d(\tilde{\mathbf{A}})$.

5.2. Experiments with floating point numbers. Given a matrix \mathbf{A} in double precision, we form $\tilde{\mathbf{A}}$ by demoting \mathbf{A} to single precision using SR-nearness. To achieve this, we emulate the computations with the chop library [24].

5.2.1. Table 5.5 and Figure C.5: Rank-deficient matrices. The rank-deficient matrices \mathbf{A} have a smallest singular value $\sigma_d(\mathbf{A}) = 0$. The goal is to

TABLE 5.3

The percentage of matrices (recall that we perform 100 stochastic roundings for each parameter setting; see also Figure C.3) violating the estimate of (1.2). The elements of \mathbf{A} are in $\mathcal{N}(0, 1)$, while the elements $\tilde{\mathbf{A}}_{ij}$ belong to $\mathcal{F}^{(p)}$, thus $\|\tilde{\mathbf{A}}_{ij}\| - \|\mathbf{A}_{ij}\| \leq 10^{-p}$, for $p = 1, \dots, 3$. Notice that when the number of columns is $d = 10$ or 100 , the bound $\mathcal{R}\sqrt{n\nu}$ is violated in approximately 0% to 40% of the test cases; the bound is almost always violated for $d = 1,000$ (suarish matrix). However, a very mild relaxation of the bound from $\mathcal{R}\sqrt{n\nu}$ to $0.9 \cdot \mathcal{R}\sqrt{n\nu}$ immediately fixes this issue, resulting in zero violations in all settings. Additionally, the relative error between the estimate provided by our bound and the minimum observed increase in the smallest singular value of the rounded matrix remains consistently below 6%. More precisely, we compute $s_{\min} = \min\{\sigma_{\min}(\tilde{\mathbf{A}})\}$ over all 100 roundings for a specific parameter setting and report the relative error $1 - s_{\min}/\mathcal{R}\sqrt{n\nu}$. If $s_{\min} > \mathcal{R}\sqrt{n\nu}$, we mark the respective entry as N/A.

| d | $\%(\sigma_{\min}(\tilde{\mathbf{A}}) < c \cdot \mathcal{R}\sqrt{n\nu})$ | | | | | | $1 - s_{\min}/\mathcal{R}\sqrt{n\nu}$ | | |
|-------|--|--------|-------|--------|-------|--------|---------------------------------------|-------|-------|
| | p = 1 | | p = 2 | | p = 3 | | p = 1 | p = 2 | p = 3 |
| | c = 1 | c = .9 | c = 1 | c = .9 | c = 1 | c = .9 | | | |
| 10 | 37% | 0% | 29% | 0% | 0% | 0% | .02 | .02 | N/A |
| 100 | 39% | 0% | 36% | 0% | 0% | 0% | .01 | .02 | N/A |
| 1,000 | 100% | 0% | 100% | 0% | 96% | 0% | .06 | .06 | .03 |

TABLE 5.4

The percentage of matrices (recall that we perform 100 stochastic roundings for each parameter setting; see also Figure C.4) violating the estimate of (1.2). The elements of \mathbf{A} are in $\text{Lognormal}(0, 3)$, while the elements $\tilde{\mathbf{A}}_{ij}$ belong to $\mathcal{F}^{(p)}$, thus $\|\tilde{\mathbf{A}}_{ij}\| - \|\mathbf{A}_{ij}\| \leq 10^{-p}$, for $p = 1, \dots, 3$. Notice that when the number of columns is $d = 10$ or 100 , the bound $\mathcal{R}\sqrt{n\nu}$ is violated in approximately 0% to 30% of the test cases; the bound is almost always violated for $d = 1,000$ (suarish matrix). However, a very mild relaxation of the bound from $\mathcal{R}\sqrt{n\nu}$ to $0.9 \cdot \mathcal{R}\sqrt{n\nu}$ immediately fixes this issue, resulting in zero violations in all settings. Additionally, the relative error between the estimate provided by our bound and the minimum observed increase in the smallest singular value of the rounded matrix remains consistently below 6%. More precisely, we compute $s_{\min} = \min\{\sigma_{\min}(\tilde{\mathbf{A}})\}$ over all 100 roundings for a specific parameter setting and report the relative error $1 - s_{\min}/\mathcal{R}\sqrt{n\nu}$. If $s_{\min} > \mathcal{R}\sqrt{n\nu}$, we mark the respective entry as N/A.

| d | $\%(\sigma_{\min}(\tilde{\mathbf{A}}) < c \cdot \mathcal{R}\sqrt{n\nu})$ | | | | | | $1 - s_{\min}/\mathcal{R}\sqrt{n\nu}$ | | |
|-------|--|--------|-------|--------|-------|--------|---------------------------------------|-------|-------|
| | p = 1 | | p = 2 | | p = 3 | | p = 1 | p = 2 | p = 3 |
| | c = 1 | c = .9 | c = 1 | c = .9 | c = 1 | c = .9 | | | |
| 10 | 0% | 0% | 8% | 0% | 0% | 0% | N/A | .01 | N/A |
| 100 | 2% | 0% | 27% | 0% | 0% | 0% | .001 | .01 | N/A |
| 1,000 | 100% | 0% | 100% | 0% | 95% | 0% | .05 | .06 | .03 |

understand how the aspect ratio n/d affects the behavior of $\sigma_d(\tilde{\mathbf{A}})$ on matrices drawn from different distributions when \mathbf{A} is being demoted to a lower precision via SR-nearness.

5.2.2. Table 5.6 and Figure C.6: Matrices with controlled ν . We start with a $10^4 \times d$ matrix whose elements are independent, identically distributed random variables $\mathcal{N}(0, 1)$, and enforce rank-deficiency by setting two columns equal to each other. For each column dimension d , we create two matrices: \mathbf{A}^h with a “high” value of ν and \mathbf{A}^l with a “low” value of ν such as $\nu(\mathbf{A}^h)/\nu(\mathbf{A}^l) \approx 100$. The goal is to also understand how ν might affect the behavior of $\sigma_d(\tilde{\mathbf{A}})$.

Conclusions from our experimental evaluations. We want to emphasize that the theoretical bound used in our experiments is not identical to the one presented in Theorem 4.1, but rather our conjecture on what the true lower bound should be (see discussion after (1.2)). The rationale behind this choice is that the precise bound is overly pessimistic for modest values of the aspect ratio n/d . This is probably due

TABLE 5.5

The percentage of matrices (recall that we perform 100 stochastic roundings for each parameter setting; see also Figure C.5) violating the estimate of (1.2). The elements of \mathbf{A} are drawn from either $\mathcal{N}(0, 1)$ or $\text{Lognormal}(0, 3)$, while the elements of \mathbf{A}_{ij} are obtained by stochastically rounding the corresponding elements \mathbf{A}_{ij} to single precision. We present results for both tested distributions. Notice that when the number of columns is $d = 10$ or 100 , the bound $\mathcal{R}\sqrt{n\nu}$ is violated in approximately 0% to 30% of the test cases; the bound is always violated for $d = 1,000$ (squarish matrix) when elements drawn from $\mathcal{N}(0, 1)$, whereas there are no violations when the elements are drawn from $\text{Lognormal}(0, 3)$. Still, a very mild relaxation of the bound from $\mathcal{R}\sqrt{n\nu}$ to $0.8 \cdot \mathcal{R}\sqrt{n\nu}$ effectively eliminates almost every violation across all settings. Furthermore, we compute $s_{\min} = \min\{\sigma_{\min}(\tilde{\mathbf{A}})\}$ over all 100 roundings for a specific parameter setting and report the relative error $1 - s_{\min}/\mathcal{R}\sqrt{n\nu}$. If $s_{\min} > \mathcal{R}\sqrt{n\nu}$, we mark the respective entry as N/A.

| d | %($\sigma_{\min}(\tilde{\mathbf{A}}) < \mathcal{R}\sqrt{n\nu}$) | | | | $1 - s_{\min}/\mathcal{R}\sqrt{n\nu}$ | |
|-------|---|----------|--------------------------|----------|---------------------------------------|--------------------------|
| | $\mathcal{N}(0, 1)$ | | $\text{Lognormal}(0, 3)$ | | $\mathcal{N}(0, 1)$ | $\text{Lognormal}(0, 3)$ |
| | $c = 1$ | $c = .8$ | $c = 1$ | $c = .8$ | | |
| 10 | 0% | 0% | 2% | 0% | N/A | .07 |
| 100 | 3% | 0% | 28% | 3% | .001 | .2 |
| 1,000 | 100% | 0% | 0% | 0% | .04 | N/A |

TABLE 5.6

The percentage of matrices (recall that we perform 100 stochastic roundings for each parameter setting; see also Figure C.6) violating the estimate of (1.2). The elements of \mathbf{A} are drawn from $\mathcal{N}(0, 1)$ and subsequently we construct two matrices $\mathbf{A}^h, \mathbf{A}^l$ corresponding to a “high” and a “low” value of ν , respectively, meaning that $\nu(\mathbf{A}^h)/\nu(\mathbf{A}^l) \approx 100$. Notice that when the number of columns is $d = 10$ or 100 , the bound $\mathcal{R}\sqrt{n\nu}$ is violated in approximately 0% to 55% of the test cases; the bound is almost always violated for $d = 1,000$ (squarish matrix). As indicated by our theory, matrices with higher value of ν lead to fewer violations (excluding the ill-advised case of $d = 1,000$). Again, a very mild relaxation of the bound from $\mathcal{R}\sqrt{n\nu}$ to $0.8 \cdot \mathcal{R}\sqrt{n\nu}$ effectively eliminates almost every violation across all settings. Furthermore, we compute $s_{\min} = \min\{\sigma_{\min}(\tilde{\mathbf{A}})\}$ over all 100 roundings for a specific parameter setting and report the relative error $1 - s_{\min}/\mathcal{R}\sqrt{n\nu}$. If $s_{\min} > \mathcal{R}\sqrt{n\nu}$, we mark the respective entry as N/A.

| d | %($\sigma_{\min}(\tilde{\mathbf{A}}) < \mathcal{R}\sqrt{n\nu}$) | | | | $1 - s_{\min}/\mathcal{R}\sqrt{n\nu}$ | |
|-------|---|----------|------------------------|----------|---------------------------------------|------------------------|
| | $\tilde{\mathbf{A}}^h$ | | $\tilde{\mathbf{A}}^l$ | | $\tilde{\mathbf{A}}^h$ | $\tilde{\mathbf{A}}^l$ |
| | $c = 1$ | $c = .8$ | $c = 1$ | $c = .8$ | | |
| 10 | 46% | 0% | 54% | 0% | .02 | .1 |
| 100 | 0% | 0% | 35% | 0% | N/A | .1 |
| 1,000 | 100% | 0% | 75% | 2% | .05 | .2 |

to state-of-the-art RMT bounds, which typically require much larger and impractical values of n . We show that, in practice, modifying the bound by a small constant (i.e., reducing $\mathcal{R}\sqrt{n\nu}$ to $c \cdot \mathcal{R}\sqrt{n\nu}$, with $c \geq .8$) results in excellent behavior in all our experimental settings.

6. Future work. First, we need to relax the assumptions for the singular value bound in Theorem 4.1, so they resemble the assumptions of Theorem B.1, where the perturbations are Gaussian. While we don’t expect the gap between bounds for SR-nearness and Gaussian perturbations to be completely bridged, we need to understand how the former bounds can be improved. This will require novel and more powerful RMT results along the lines of Theorem 4.3.

Second, we conjecture that for all sufficiently tall-and-thin matrices \mathbf{A} , SR-nearness produces a matrix whose smallest singular value is bounded away from zero as in (1.2). Although removal of $\epsilon_{n,d}$ from (1.2) seems infeasible with state-of-the-art RMT bounds, our numerical experiments strongly support this conjecture.

Appendix A. Proof for section 4.1. Distinguish the columns of the rounded matrix,

$$\tilde{\mathbf{A}} = \begin{bmatrix} \tilde{\mathbf{A}}_1 & \tilde{\mathbf{A}}_2 \end{bmatrix}.$$

Since, in expectation, half of the entries of $\tilde{\mathbf{A}}$ are equal to one and the other half are equal to zero, the columns of $\tilde{\mathbf{A}}$ have expected squared norms equal to

$$(A.1) \quad \mathbb{E} \|\tilde{\mathbf{A}}_j\|_2^2 = \frac{n}{2}, \quad j = 1, 2.$$

In the inner product between the two columns,

$$\tilde{\mathbf{A}}_1^T \tilde{\mathbf{A}}_2 = \sum_{i=1}^n \tilde{\mathbf{A}}_{i1} \tilde{\mathbf{A}}_{i2},$$

the i th summand $\tilde{\mathbf{A}}_{i1} \tilde{\mathbf{A}}_{i2} = 1$ only if $\mathbf{A}_{i,1} = \mathbf{A}_{i,2} = 1$, which occurs with probability $1/4$. Hence the expected inner product between the two columns equals

$$(A.2) \quad \mathbb{E}(\mathbf{A}_1^T \mathbf{A}_2) = \frac{n}{4}.$$

The 2×2 Gram matrix $\tilde{\mathbf{A}}^T \tilde{\mathbf{A}}$ has expectation

$$\mathbb{E}[\tilde{\mathbf{A}}^T \tilde{\mathbf{A}}] = \begin{bmatrix} \frac{n}{2} & \frac{n}{4} \\ \frac{n}{4} & \frac{n}{2} \end{bmatrix} = \frac{n}{4} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix},$$

which immediately implies that $\sigma_2(\mathbb{E}[\tilde{\mathbf{A}}^T \tilde{\mathbf{A}}]) = n/4$.

Denote by $\mathbf{B} = \tilde{\mathbf{A}}^T \tilde{\mathbf{A}} - \mathbb{E}[\tilde{\mathbf{A}}^T \tilde{\mathbf{A}}]$ the deviation of the Gram matrix from its expectation. Weyl's inequality and the bound $\|\mathbf{B}\|_2 \leq \|\mathbf{B}\|_F$ give

$$(A.3) \quad \begin{aligned} \sigma_2(\tilde{\mathbf{A}})^2 &= \sigma_2(\tilde{\mathbf{A}}^T \tilde{\mathbf{A}}) = \sigma_2(\mathbb{E}[\tilde{\mathbf{A}}^T \tilde{\mathbf{A}}] - (\mathbb{E}[\tilde{\mathbf{A}}^T \tilde{\mathbf{A}}] - \tilde{\mathbf{A}}^T \tilde{\mathbf{A}})) \\ &\geq \sigma_2(\mathbb{E}[\tilde{\mathbf{A}}^T \tilde{\mathbf{A}}]) - \|\mathbf{B}\|_2 \\ &\geq \sigma_2(\mathbb{E}[\tilde{\mathbf{A}}^T \tilde{\mathbf{A}}]) - \|\mathbf{B}\|_F \\ &= \frac{n}{4} - \|\mathbf{B}\|_F. \end{aligned}$$

It remains to bound $\|\mathbf{B}\|_F$. Since $(\tilde{\mathbf{A}}^T \tilde{\mathbf{A}})_{ij} = \sum_{k=1}^n \tilde{\mathbf{A}}_{ki} \tilde{\mathbf{A}}_{kj}$ is the sum of n random variables that are either zero or one, we can invoke Theorem 2.1,

$$\mathbb{P}[|\mathbf{B}_{ij}| \geq t] = \mathbb{P}\left[|(\tilde{\mathbf{A}}^T \tilde{\mathbf{A}})_{ij} - \mathbb{E}[(\tilde{\mathbf{A}}^T \tilde{\mathbf{A}})_{ij}]| \geq t\right] \leq 2 \exp\left(\frac{-2t^2}{n}\right).$$

Applying a union bound over the four events that represent the entries of \mathbf{B} being less than t gives the failure probability

$$\mathbb{P}\left[\sum_{i,j=1}^2 |\mathbf{B}_{ij}| \geq 4t\right] \leq 8 \exp\left(\frac{-2t^2}{n}\right)$$

and the success probability of the complementary event,

$$\mathbb{P} \left[\sum_{i,j=1}^2 |\mathbf{B}_{ij}| \leq 4t \right] \geq 1 - 8 \exp \left(\frac{-2t^2}{n} \right).$$

Vector p -norm relations imply

$$\|\mathbf{B}\|_F = \|\text{vec}(\mathbf{B})\|_2 \leq \|\text{vec}(\mathbf{B})\|_1 = \sum_{i,j=1}^2 |\mathbf{B}_{ij}|.$$

Insert this into the success probability,

$$\mathbb{P} (\|\mathbf{B}\|_F \leq 4t) \geq 1 - 8 \exp \left(\frac{-2t^2}{n} \right),$$

and combine with (A.3) to obtain a lower bound for $\sigma_2^2(\tilde{\mathbf{A}})$,

$$\mathbb{P} \left(\sigma_2^2(\tilde{\mathbf{A}}) \geq \frac{n}{4} - 4t \right) \geq 1 - 8 \exp \left(\frac{-2t^2}{n} \right).$$

Setting $t = 2\sqrt{n}$ gives

$$(A.4) \quad \sigma_2^2(\tilde{\mathbf{A}}) \geq 0.25 \cdot n - 8\sqrt{n}$$

with probability at least 0.997.

Appendix B. Gaussian perturbations. We consider perturbations $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$, where⁵ $\mathbf{E}_{i,j} = \mathcal{N}(0, 1)$, and show that the smallest singular value of $\tilde{\mathbf{A}}$ is bounded away from zero with high probability. While this perturbation model is not relevant for SR, we do note that Theorem B.1 is much sharper than Theorem 4.1 and Corollary 4.2.

For example, if $n = 900$ and $d = 25$, then Theorem B.1 with $t = 4$ shows that $\sigma_d(\tilde{\mathbf{A}}) \geq 1$ with probability at least 0.98, *without any additional assumptions*. This provides evidence that the assumptions of Theorem 4.1 and Corollary 4.2 could be significantly relaxed for non-Gaussian, nonidentically distributed perturbations. The sharper bounds in this section are derived from the strong concentration properties of Gaussian distributions, which allow the use of more powerful measure concentration inequalities. Additionally, the invariance of Gaussian distributions under the unitary transformations further simplifies the analysis.

The proof follows the same high-level proof structure as in section 4.5. We start by stating our main result.

THEOREM B.1. *Let $\mathbf{A} \in \mathbb{R}^{n \times d}$, and let $\mathbf{E} \in \mathbb{R}^{n \times d}$ be a random matrix with independent, identically distributed Gaussian entries, i.e., $\mathbf{E}_{ij} = \mathcal{N}(0, 1)$. Then, for any $t > 0$,*

$$\mathbb{P} \left(\sigma_d(\mathbf{A} + \mathbf{E}) \geq \sqrt{n} - (1+t)\sqrt{d} - t \right) \geq 1 - (2d+1)e^{-t^2/2}.$$

Proof. As in section 4.5, we decompose the task of lower bounding $\sigma_d(\mathbf{A} + \mathbf{E})$ into two parts. Again, for any orthogonal projector \mathbf{P} ,

$$\sigma_d(\mathbf{A} + \mathbf{E}) \geq \sigma_d(\mathbf{P}(\mathbf{A} + \mathbf{E})).$$

⁵By rescaling the bound by $\sigma > 0$, we can extend the bound to any \mathbf{E} with $\mathbf{E}_{ij} = \mathcal{N}(0, \sigma)$.

Let $\mathbf{P}_{\mathbf{A}} \in \mathbb{R}^{n \times n}$ be the orthogonal projector onto the d dimensional column space of \mathbf{A} , and let $\mathbf{P}_{\mathbf{A}, \perp} = \mathbf{I} - \mathbf{P}_{\mathbf{A}}$ be the orthogonal projector onto the $n - d$ dimensional left null space of \mathbf{A} . Then

$$\sigma_d(\mathbf{A} + \mathbf{E}) \geq \sigma_d(\mathbf{P}_{\mathbf{A}, \perp}(\mathbf{A} + \mathbf{E})) = \sigma_d(\mathbf{P}_{\mathbf{A}, \perp}\mathbf{E}).$$

Weyl's inequality implies

$$(B.1) \quad \sigma_d(\mathbf{P}_{\mathbf{A}, \perp}\mathbf{E}) = \sigma_d(\mathbf{E} - (\mathbf{I} - \mathbf{P}_{\mathbf{A}, \perp})\mathbf{E}) \geq \sigma_d(\mathbf{E}) - \|\mathbf{P}_{\mathbf{A}}\mathbf{E}\|_2.$$

We have now broken our task into two conceptual components. First, we must make sure that the random matrix \mathbf{E} has a large minimum singular value. Second, the projection of \mathbf{E} must be small. In other words, the matrix \mathbf{E} should not concentrate in any d -dimensional space.

We start by lower bounding $\sigma_d(\mathbf{E})$. From [33, expression below (1.11)] follows that for any $t > 0$,

$$(B.2) \quad \mathbb{P}(\sigma_d(\mathbf{E}) \leq \sqrt{n} - \sqrt{d} - t) \leq e^{-t^2/2}.$$

Next, we bound $\|\mathbf{P}_{\mathbf{A}}\mathbf{E}\|_2$ by first leveraging the unitary invariance of the Gaussian distribution. If $\mathbf{g} \in \mathbb{R}^n$ is a vector with independent, identically distributed standard normal entries, then the distribution of $\mathbf{U}\mathbf{g}$ is the same as the distribution of \mathbf{g} for any orthogonal matrix $\mathbf{U} \in \mathbb{R}^{n \times n}$. Since the columns of \mathbf{E} are independent, we can apply this result columnwise to conclude that the distribution of \mathbf{E} is equal to the distribution of $\mathbf{U}\mathbf{E}$ for any orthogonal matrix \mathbf{U} . Let

$$(B.3) \quad \mathbf{P}_{\mathbf{A}} = \mathbf{U} \begin{bmatrix} \mathbf{I}_d & \mathbf{0}_{d \times (n-d)} \\ \mathbf{0}_{(n-d) \times d} & \mathbf{0}_{(n-d) \times (n-d)} \end{bmatrix} \mathbf{U}^T$$

be an eigenvalue decomposition where $\mathbf{U} \in \mathbb{R}^{n \times n}$ is an orthogonal matrix. From the rotational invariance of the Gaussian distribution follows that the entries of $\mathbf{U}^T\mathbf{E} \in \mathbb{R}^{n \times d}$ are also independent, identically distributed Gaussian normal random variables. The unitary invariance of the two-norm implies

$$(B.4) \quad \|\mathbf{P}_{\mathbf{A}}\mathbf{E}\|_2 = \|\mathbf{G}\|_2 \quad \text{where} \quad \mathbf{G} \equiv \begin{bmatrix} \mathbf{I}_d & \mathbf{0}_{d \times (n-d)} \end{bmatrix} \mathbf{U}^T\mathbf{E} \in \mathbb{R}^{d \times d}$$

is the matrix that contains the leading d rows of $\mathbf{U}^T\mathbf{E}$.

From Lemma B.3 and (B.4) follows

$$\mathbb{P}(\|\mathbf{P}_{\mathbf{A}}\mathbf{E}\|_2 = \mathbb{P}(\|\mathbf{G}\|_2 \geq t\sqrt{d}) \leq 2de^{-t^2/2}$$

for all $t > 0$. Combining this with (B.2) and applying a union bound to control the two failure probabilities gives

$$\mathbb{P}\left(\sigma_d(\mathbf{E}) - \|\mathbf{P}_{\mathbf{A}}\mathbf{E}\|_2 \leq \sqrt{n} - (1+t)\sqrt{d} - t\right) \leq (2d+1)e^{-t^2/2}$$

for all $t > 0$. The complement of the above event, for all $t > 0$, is

$$\mathbb{P}\left(\sigma_d(\mathbf{E}) - \|\mathbf{P}_{\mathbf{A}}\mathbf{E}\|_2 \geq \sqrt{n} - (1+t)\sqrt{d} - t\right) \geq 1 - (2d+1)e^{-t^2/2}.$$

At last, combine the above with (B.1). □

In order to bound the largest singular value of a Gaussian matrix, we will use the following concentration inequality from prior work.

THEOREM B.2 (Theorem 4.1.1 in [35]). *Consider a finite sequence of $\{\mathbf{B}_k\}$ of fixed real-valued matrices with dimension $d_1 \times d_2$, and let $\{\gamma_k\}$ be a finite sequence of independent standard normal variables. Introduce the Gaussian series:*

$$\mathbf{Z} = \sum_k \gamma_k \mathbf{B}_k.$$

Let $\mathcal{V}(\mathbf{Z})$ be the matrix variance statistic of the sum:

$$\begin{aligned} \mathcal{V}(\mathbf{Z}) &= \max\{\mathbb{E}\|\mathbf{Z}\mathbf{Z}^T\|_2, \mathbb{E}\|\mathbf{Z}^T\mathbf{Z}\|_2\} \\ &= \max\left\{\mathbb{E}\left\|\sum_k \mathbf{B}_k \mathbf{B}_k^T\right\|_2, \mathbb{E}\left\|\sum_k \mathbf{B}_k^T \mathbf{B}_k\right\|_2\right\}. \end{aligned}$$

Then,

$$\mathbb{E}\|\mathbf{Z}\|_2 \leq \sqrt{2 \cdot \mathcal{V}(\mathbf{Z}) \log(d_1 + d_2)}.$$

Furthermore, for all $t \geq 0$,

$$\mathbb{P}(\|\mathbf{Z}\|_2 \geq t) \leq (d_1 + d_2) \exp\left(\frac{-t^2}{2 \cdot \mathcal{V}(\mathbf{Z})}\right).$$

The following lemma bounds the largest singular value of Gaussian matrices, whose entries are independent, identically distributed $\mathcal{N}(0, 1)$ random variables.

LEMMA B.3. *Let $\mathbf{G} \in \mathbb{R}^{d \times d}$ be a random matrix with independent, identically distributed standard normal entries. Then, for all $t \geq 0$,*

$$\mathbb{P}(\|\mathbf{G}\|_2 \geq t\sqrt{d}) \leq 2de^{-t^2/2}.$$

Proof. We will apply Theorem B.2. First write \mathbf{G} as a sum of outer products

$$\mathbf{G} = \sum_{i,j=1}^d g_{ij} \mathbf{e}_i \mathbf{e}_j^T,$$

where g_{ij} are independent, identically distributed standard normal variables and $\mathbf{e}_i \in \mathbb{R}^d$, $1 \leq i \leq d$, are the columns of the identity matrix $\mathbf{I} \in \mathbb{R}^{d \times d}$. In order to apply the aforementioned theorem we need to bound

$$\left\| \sum_{i,j=1}^d \mathbf{e}_i \mathbf{e}_j^T (\mathbf{e}_i \mathbf{e}_j^T)^T \right\|_2 = \left\| \sum_{i,j=1}^d \mathbf{e}_i \mathbf{e}_i^T \right\|_2 = \|d \cdot \mathbf{I}_d\|_2 = d.$$

Similarly, $\|\sum_{i,j=1}^d (\mathbf{e}_i \mathbf{e}_j^T)^T \mathbf{e}_i \mathbf{e}_j^T\|_2 = d$. Hence in Theorem B.2 the parameter $\mathcal{V}(\mathbf{G})$ is equal to d , and so

$$\mathbb{P}[\|\mathbf{G}\|_2 \geq t] \leq 2de^{-t^2/2d}.$$

Rescaling the parameter t by a \sqrt{d} factor concludes the proof. □

Appendix C. Additional plots.

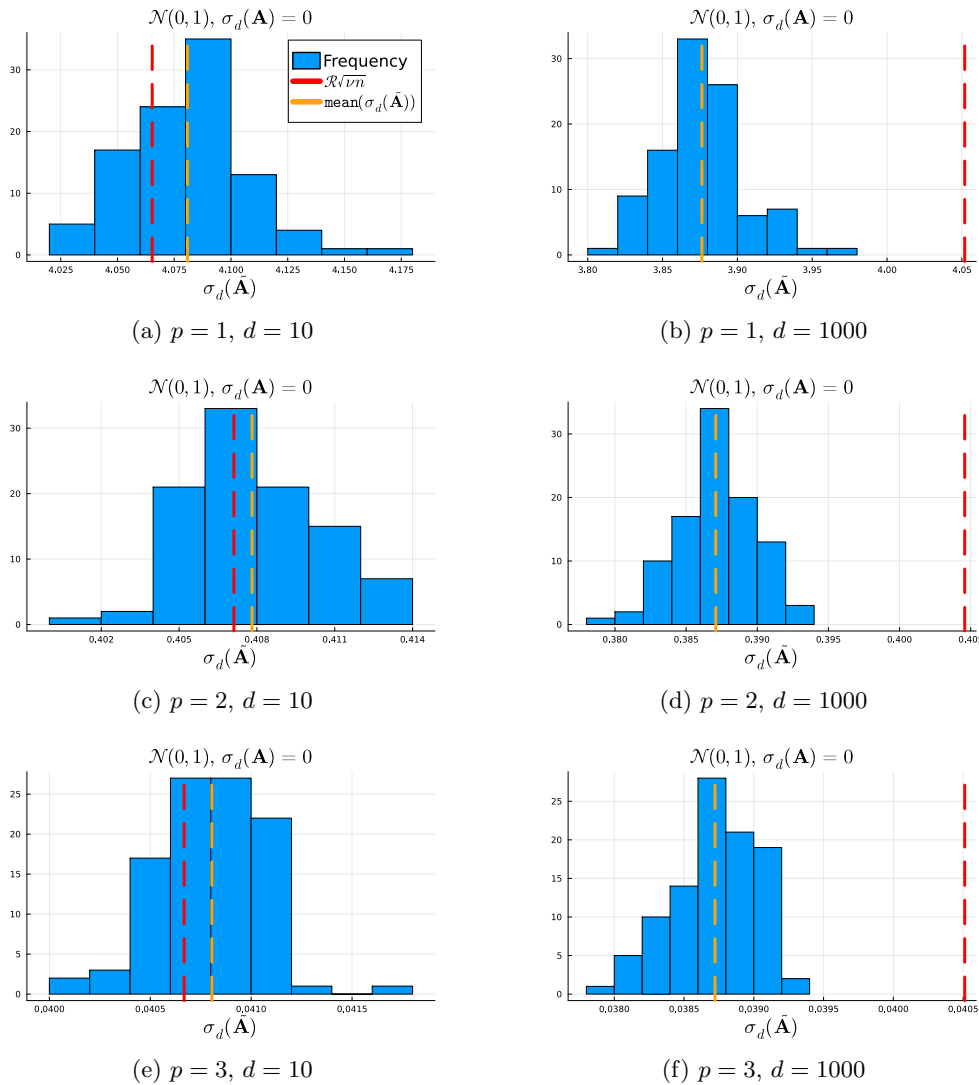


FIG. C.1. The elements of \mathbf{A} are random variables in $\mathcal{N}(0, 1)$ with $\sigma_d(\mathbf{A}) = 0$. The stochastically rounded $\tilde{\mathbf{A}}$ has elements in \mathcal{F}^p for $p = 1, 2, 3$. The horizontal axis represents the values of $\sigma_d(\tilde{\mathbf{A}})$ over 100 runs, grouped into at most 10 bins. The vertical axis represents the number of $\sigma_d(\tilde{\mathbf{A}})$ in each bin. The orange dashed vertical line represents the average value of $\sigma_d(\tilde{\mathbf{A}})$, while the red dashed vertical line represents the lower bound estimate (1.2). Each panel corresponds to a different combination of p and d . In each row, the precision p is fixed, while the column dimension d varies.

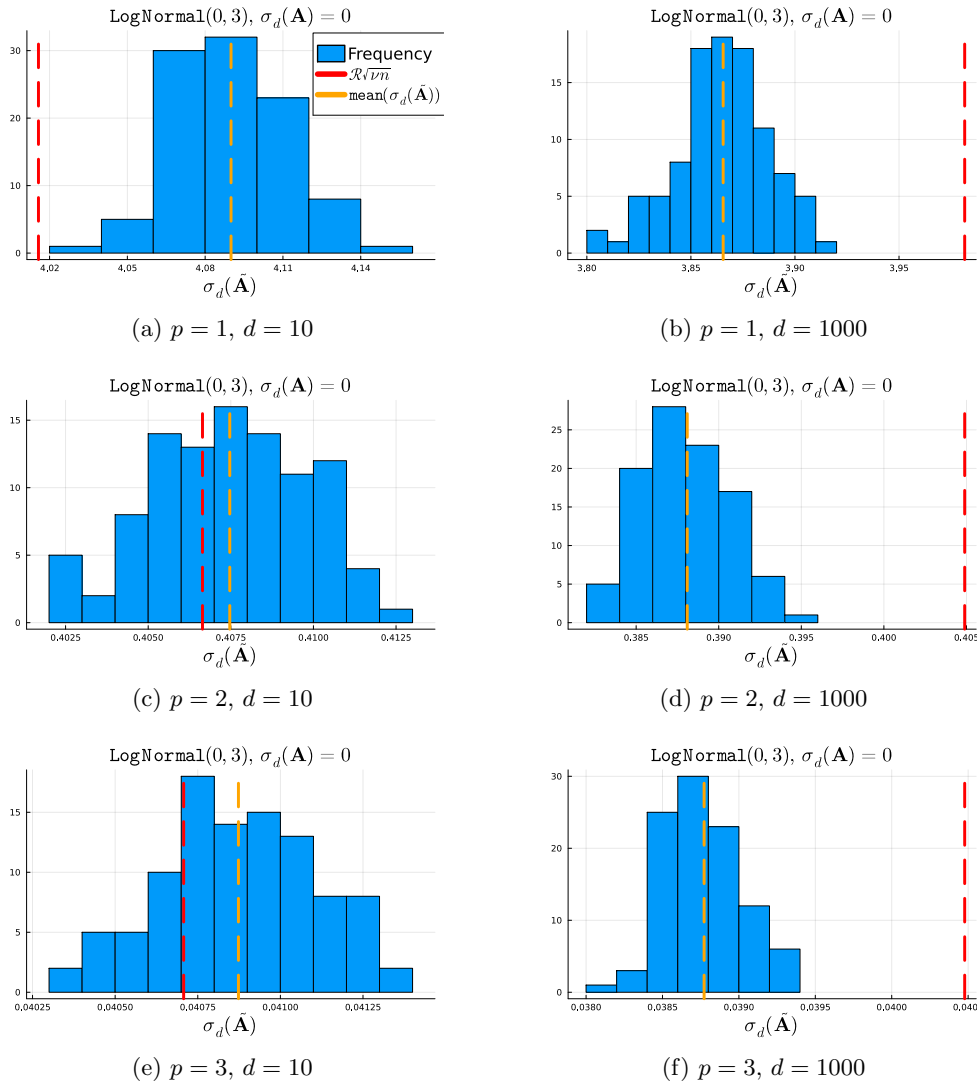


FIG. C.2. The matrices are initially drawn from a log-normal distribution with the smallest singular value set to 0, and stochastically rounded to \mathcal{F}^p , for $p = 1, \dots, 3$. The horizontal axis represents the distribution of $\sigma_d(\tilde{\mathbf{A}})$ over 100 repetitions, grouped into up to 10 bins. The vertical axis shows the frequency with which each $\sigma_d(\tilde{\mathbf{A}})$ appears in each bin. The orange dashed vertical line represents the average value of $\sigma_d(\tilde{\mathbf{A}})$, while the red dashed vertical line represents the lower bound estimate (1.2). Each panel corresponds to a different combination of p and d . In each row, the precision p is fixed, while the column dimension d varies.

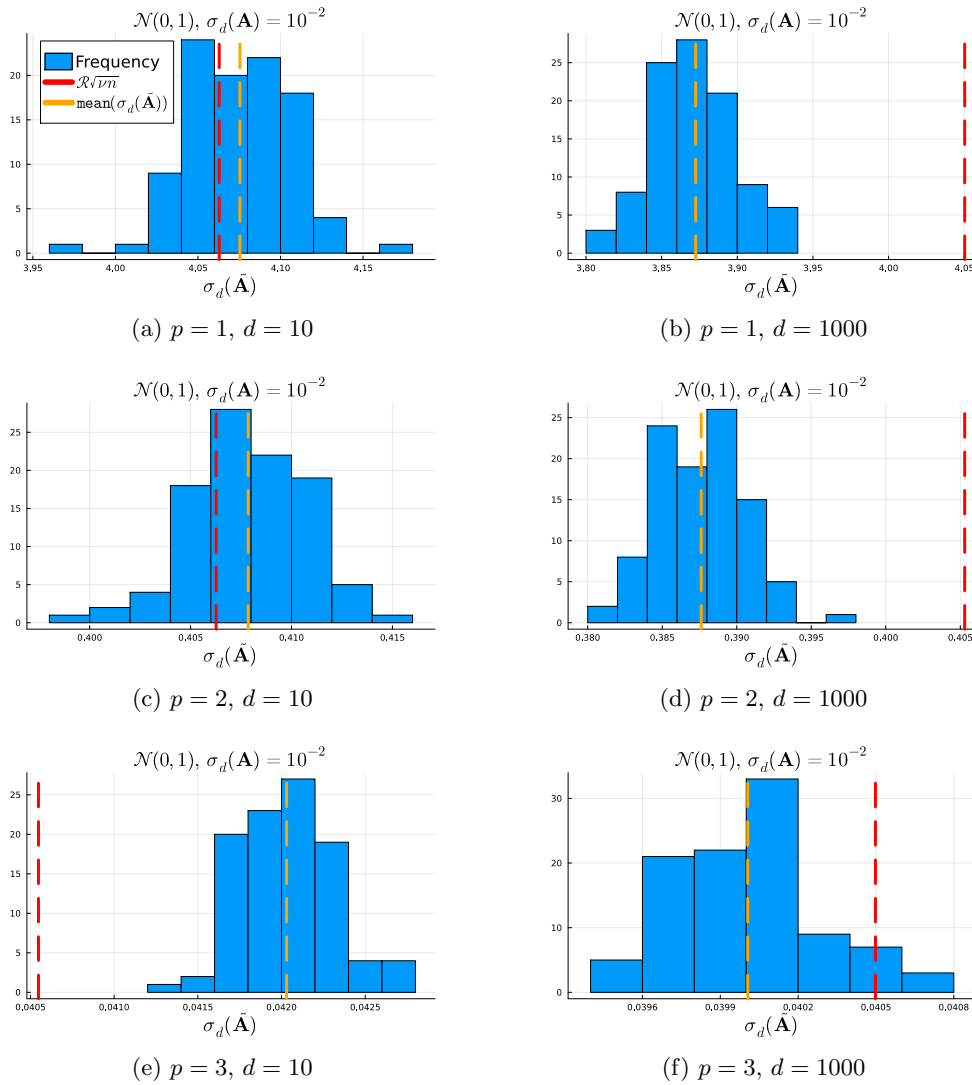


FIG. C.3. The matrices are initially drawn from a standard normal distribution with the smallest singular value set to 10^{-2} , and stochastically rounded to \mathcal{F}^p , for $p = 1, \dots, 3$. The horizontal axis represents the distribution of $\sigma_d(\tilde{\mathbf{A}})$ over 100 repetitions, grouped into up to 10 bins. The vertical axis shows the frequency with which each $\sigma_d(\tilde{\mathbf{A}})$ appears in each bin. The orange dashed vertical line represents the average value of $\sigma_d(\tilde{\mathbf{A}})$, while the red dashed vertical line represents the lower bound estimate (1.2). Each panel corresponds to a different combination of p and d . In each row, the precision p is fixed, while the column dimension d varies.

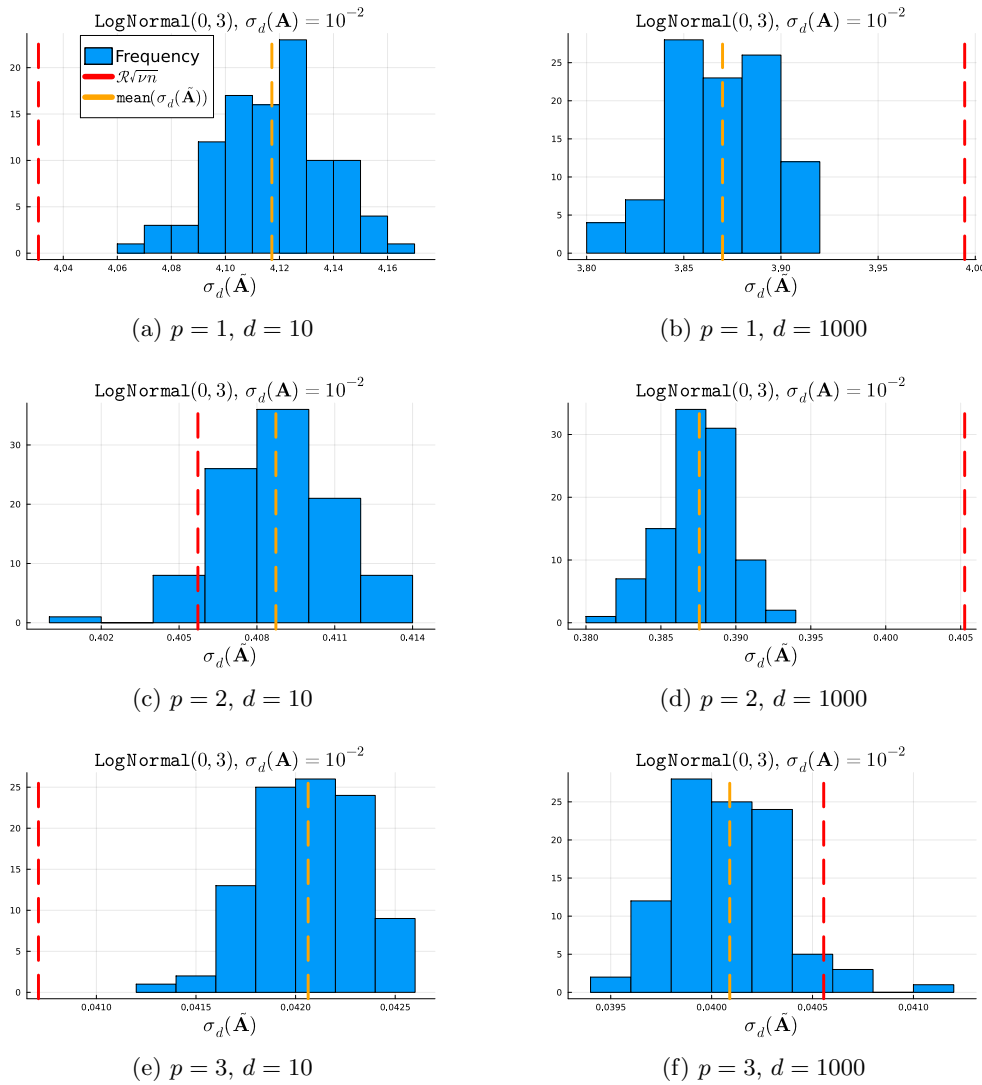


FIG. C.4. The matrices are initially drawn from a log-normal distribution with the smallest singular value set to 10^{-2} , and stochastically rounded to \mathcal{F}^p , for $p = 1, \dots, 3$. The horizontal axis represents the distribution of $\sigma_d(\tilde{\mathbf{A}})$ over 100 repetitions, grouped into up to 10 bins. The vertical axis shows the frequency with which each $\sigma_d(\tilde{\mathbf{A}})$ appears in each bin. The orange dashed vertical line represents the average value of $\sigma_d(\tilde{\mathbf{A}})$, while the red dashed vertical line represents the lower bound estimate (1.2). Each panel corresponds to a different combination of p and d . In each row, the precision p is fixed, while the column dimension d varies.

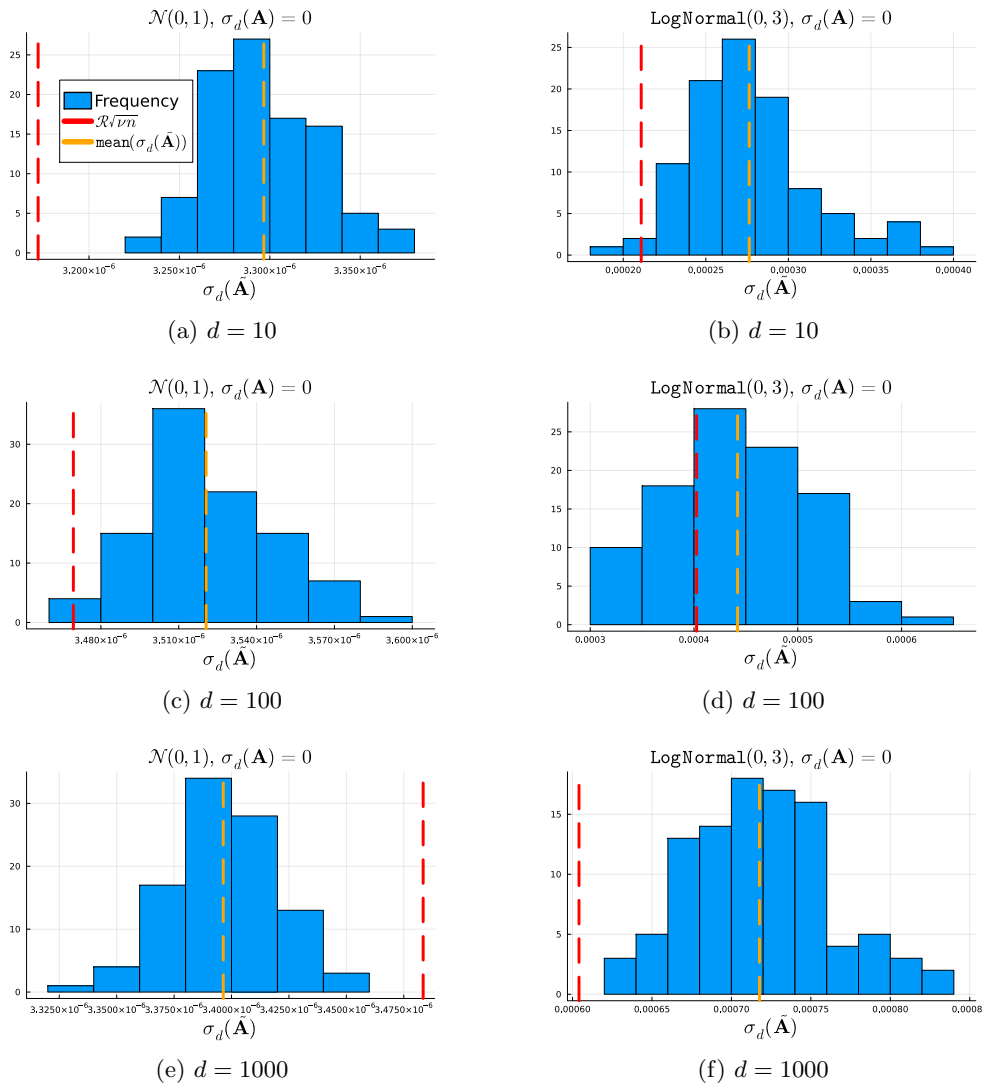


FIG. C.5. The matrices are initially drawn from both a standard normal and a log-normal distribution with the smallest singular value set to 0, and stochastically rounded to single precision. The horizontal axis represents the distribution of $\sigma_d(\tilde{\mathbf{A}})$ over 100 repetitions, grouped into up to 10 bins. The vertical axis shows the frequency with which each $\sigma_d(\tilde{\mathbf{A}})$ appears in each bin. The orange dashed vertical line represents the average value of $\sigma_d(\tilde{\mathbf{A}})$, while the red dashed vertical line represents the lower bound estimate (1.2). Each panel corresponds to a different combination of the initial distribution and d . In each row, column dimension d is fixed, while the distribution varies.

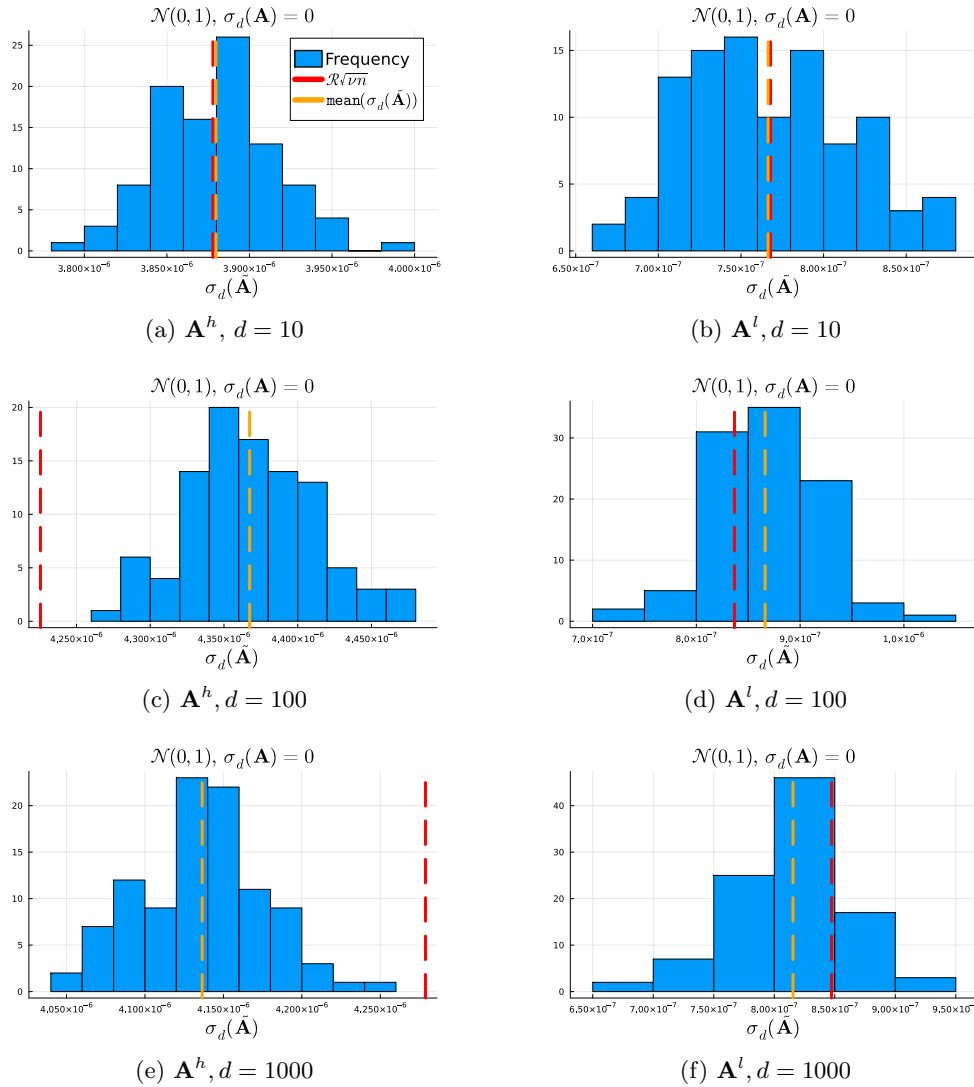


FIG. C.6. The matrices are initially drawn from a standard normal distribution with the smallest singular value equal to 0, and stochastically rounded to single precision. The horizontal axis represents the distribution of $\sigma_d(\tilde{\mathbf{A}})$ over 100 repetitions, grouped into up to 10 bins. The vertical axis shows the frequency with which each $\sigma_d(\tilde{\mathbf{A}})$ appears in each bin. The orange dashed vertical line represents the average value of $\sigma_d(\tilde{\mathbf{A}})$, while the red dashed vertical line represents the lower bound estimate (1.2). Each panel corresponds to a different combination of “high” or “low” ν and d . In each row, the column dimension d is fixed, while the value of ν varies.

Reproducibility of computational results. This paper has been awarded the “SIAM Reproducibility Badge: Code and data available” as a recognition that the authors have followed reproducibility principles valued by SIMAX and the scientific computing community. Code and data that allow readers to reproduce the results in this paper are available at https://github.com/cboutsikas/stoch_rounding_iplicit_reg.

REFERENCES

- [1] *Mixed-Precision Arithmetic for AI: A Hardware Perspective*, <https://docs.graphcore.ai/projects/ai-float-white-paper/en/latest/ai-float.html>.
- [2] *IEEE Standard for Floating-Point Arithmetic*, 754-2019 (revision of IEEE 754-2008), IEEE, 2019, <https://doi.org/10.1109/IEEESTD.2019.8766229>.
- [3] J. M. ALBEN, P. MICKEVICIUS, H. WU, AND M. Y. SIU, *Stochastic Rounding of Numerical Values*, 2019, <https://patents.google.com/patent/US10684824B2/en?q=US+10%2c684%2c824+B2>.
- [4] E.-M. E. ARAR, D. SOHIER, P. DE OLIVEIRA CASTRO, AND E. PETIT, *The positive effects of stochastic rounding in numerical algorithms*, in Proceedings of the 29th Symposium on Computer Arithmetic, IEEE, 2022, pp. 58–65, <https://doi.org/10.1109/ARITH54963.2022.00018>.
- [5] E.-M. E. ARAR, D. SOHIER, P. DE OLIVEIRA CASTRO, AND E. PETIT, *Stochastic rounding variance and probabilistic bounds: A new approach*, SIAM J. Sci. Comput., 45 (2023), pp. C255–C275, <https://doi.org/10.1137/22M1510819>.
- [6] R. BHATIA, *Matrix Analysis*, Grad. Texts in Math. 169, Springer, New York, 2013.
- [7] C. BOUTSIKAS, P. DRINEAS, AND I. C. F. IPSEN, *Small singular values can increase in lower precision*, SIAM J. Matrix Anal. Appl., 45 (2024), pp. 1518–1540, <https://doi.org/10.1137/23M1557209>.
- [8] J. D. BRADBURY, S. R. CARLOUGH, B. R. PRASKY, AND E. M. SCHWARZ, *Reproducible Stochastic Rounding for Out of Order Processors*, 2016, <https://patents.google.com/patent/US10083008B2/en?q=US+10083008+>.
- [9] J. D. BRADBURY, S. R. CARLOUGH, B. R. PRASKY, AND E. M. SCHWARZ, *Stochastic Rounding Floating-Point Multiply Instruction Using Entropy from a Register*, 2016, <https://patents.google.com/patent/US10445066B2/en?q=US+10445066>.
- [10] T. BRAILOVSKAYA AND R. VAN HANDEL, *Universality and sharp matrix concentration inequalities*, Geom. Funct. Anal., 34 (2024), pp. 1–105, <https://doi.org/10.1007/s00039-024-00692-9>.
- [11] M. P. CONNOLLY, N. J. HIGHAM, AND T. MARY, *Stochastic rounding and its probabilistic backward error analysis*, SIAM J. Sci. Comput., 43 (2021), pp. A566–A585, <https://doi.org/10.1137/20M1334796>.
- [12] M. CROCI, M. FASI, N. J. HIGHAM, T. MARY, AND M. MIKAITIS, *Stochastic rounding: Implementation, error analysis and applications*, Roy. Soc. Open Sci., 9 (2022), <https://doi.org/10.1098/rsos.211631>.
- [13] M. CROCI AND M. B. GILES, *Effects of round-to-nearest and stochastic rounding in the numerical solution of the heat equation in low precision*, IMA J. Numer. Anal., 43 (2023), pp. 1358–1390, <https://doi.org/10.1093/imanum/drac012>.
- [14] M. DAVIES, N. SRINIVASA, T.-H. LIN, G. CHINYA, Y. CAO, S. H. CHODAY, G. DIMOU, P. JOSHI, N. IMAM, S. JAIN, ET AL., *Loihi: A neuromorphic manycore processor with on-chip learning*, IEEE Micro, 38 (2018), pp. 82–99, <https://doi.org/10.1109/MM.2018.112130359>.
- [15] C. DENIS, P. D. O. CASTRO, AND E. PETIT, *Verificarlo: Checking Floating Point Accuracy through Monte Carlo Arithmetic*, in 2016 IEEE 23rd Symposium on Computer Arithmetic (ARITH), IEEE, 2016, pp. 55–62, <https://doi.org/10.1109/ARITH.2016.31>.
- [16] I. DUMITRIU AND Y. ZHU, *Extreme Singular Values of Inhomogeneous Sparse Random Rectangular Matrices*, preprint, arXiv:2209.12271, 2022.
- [17] I. DUMITRIU AND Y. ZHU, *Extreme singular values of inhomogeneous sparse random rectangular matrices*, Bernoulli, 30 (2024), pp. 2904–2931, <https://doi.org/10.3150/23-BEJ1699>.
- [18] F. FÉVOTTE AND B. LATHULIERE, *VERROU: A CESTAC evaluation without recompilation*, in Proceedings of the 17th International Symposium on Scientific Computing, Computer Arithmetic and Verified Numerics, 2016.
- [19] G. FORSYTHE, *Round-off errors in numerical integration on automatic machinery*, Bull. Amer. Math. Soc., 56 (1950), pp. 55–65, <https://doi.org/10.1090/S0002-9904-1950-09343-4>.

- [20] G. E. FORSYTHE, *Reprint of a note on rounding-off errors*, SIAM Rev., 1 (1959), pp. 66–67, <https://doi.org/10.1137/1001011>.
- [21] S. GUPTA, A. AGRAWAL, K. GOPALAKRISHNAN, AND P. NARAYANAN, *Deep learning with limited numerical precision*, in Proceedings of the International Conference on Machine Learning, PMLR, 2015, pp. 1737–1746.
- [22] E. HALLMAN AND I. C. F. IPSEN, *Precision-aware deterministic and probabilistic error bounds for floating point summation*, Numer. Math., 155 (2023), pp. 83–119, <https://doi.org/10.1007/s00211-023-01370-y>.
- [23] G. G. HENRY AND D. R. REED, *Processor with Memory Array Operable as Either Cache Memory or Neural Network Unit Memory*, 2016, <https://patents.google.com/patent/US10664751B2/en?q=us+10664751>.
- [24] N. J. HIGHAM AND S. PRANESH, *Simulating low precision floating-point arithmetic*, SIAM J. Sci. Comput., 41 (2019), pp. C585–C602, <https://doi.org/10.1137/19M1251308>.
- [25] W. HOEFFDING, *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc., 58 (1963), pp. 13–30, <https://doi.org/10.1080/01621459.1963.10500830> (accessed 2024-02-18).
- [26] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 2012, <https://doi.org/10.1017/CBO9781139020411>.
- [27] T. E. HULL AND J. R. SWENSON, *Tests of probabilistic models for propagation of roundoff errors*, Commun. ACM, 9 (1966), pp. 108–113, <https://doi.org/10.1145/365170.365212>.
- [28] F. JÉZÉQUEL AND J.-M. CHESNEAUX, *CADNA: A library for estimating round-off error propagation*, Comput. Phys. Commun., 178 (2008), pp. 933–955, <https://doi.org/10.1016/j.cpc.2008.02.003>.
- [29] O. A. KANTER AND I. BAR, *Apparatus and Methods for Hardware-Efficient Unbiased Rounding*, 2008, <https://patents.google.com/patent/US8972472B2/en?q=US+8%2c972.472+>.
- [30] S. LIFSCHES, *In-Memory Stochastic Rounder*, 2018, <https://patents.google.com/patent/US10803141B2/en?q=10%2c803%2c141>.
- [31] G. H. LOH, *Stochastic Rounding Logic*, 2018, <https://patents.google.com/patent/US10628124B2/en?q=US10628124B2>.
- [32] D. S. PARKER, *Monte Carlo Arithmetic: Exploiting Randomness in Floating-Point Arithmetic*, University of California, LA, 1997, <https://citeseerx.ist.psu.edu/document?doi=3b55cfb2325bc081f113612204733c2a3bd650fe>.
- [33] M. RUDELSON AND R. VERSHYNIN, *Smallest singular value of a random rectangular matrix*, Comm. Pure Appl. Math., 62 (2009), pp. 1707–1739, <https://doi.org/10.1002/cpa.20294>.
- [34] A. SANKAR, D. A. SPIELMAN, AND S. H. TENG, *Smoothed analysis of the condition numbers and growth factors of matrices*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 446–476, <https://doi.org/10.1137/S0895479803436202>.
- [35] J. A. TROPP, *An introduction to matrix concentration inequalities*, Found. Trends Mach. Learn., 8 (2015), pp. 1–230, <https://doi.org/10.1561/22000000048>.
- [36] J. VIGNES, *Discrete stochastic arithmetic for validating results of numerical software*, Numer. Algorithms, 37 (2004), pp. 377–390, <https://doi.org/10.1023/B:NUMA.0000049483.75679.ce>.
- [37] J. VON NEUMANN AND H. H. GOLDSTINE, *Numerical inverting of matrices of high order*, Bull. Amer. Math. Soc., 53 (1947), pp. 1021–1099, <https://doi.org/10.1090/S0002-9904-1947-08909-6>.
- [38] N. WANG, J. CHOI, D. BRAND, C.-Y. CHEN, AND K. GOPALAKRISHNAN, *Training deep neural networks with 8-bit floating point numbers*, Adv. Neural Inf. Process. Syst., 31 (2018), https://papers.nips.cc/paper_files/paper/2018/hash/335d3d1cd7ef05ec77714a215134914c-Abstract.html.