Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Stochastic rounding turns rounding errors into a residual: randomized Kaczmarz in low precision

Liuwen Chang

clwww406@outlook.com

Xuchang University

Research Article

Keywords: randomized Kaczmarz, stochastic rounding, low precision, mixed precision, rounding error analysis, tail averaging

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11576594/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Stochastic rounding turns rounding errors into a residual: randomized Kaczmarz in low precision

Liuwen Chang
School of Science, Xuchang University, Xuchang 461000, Henan, China
clwww406@outlook.com

Abstract

Randomized Kaczmarz with iterates stored in a low-precision format stalls at an error floor. For round-to-nearest, the known bound on this floor grows like $\tilde{\kappa}^2 u$, where $\tilde{\kappa} = \|A\|_F \|A^+\|$ is the Demmel condition number and u the unit roundoff, and iterative refinement guarantees $\tilde{\kappa} u$. With stochastic rounding, the rounding errors form a martingale difference sequence and act exactly like an additional inconsistency of the linear system: the long-run mean of $\|A(\hat{x}_k - x_\star)\|^2$ equals $\|r_\star\|^2 + \|A\|_F^2 \eta$, where $r_\star$ is the least-squares residual and η the mean squared rounding error per step. We call $\|A\|_F \sqrt{\eta}$ the rounding residual. This identity yields a floor of order $\tilde{\kappa} u$ without refinement, an exact trace formula for the stationary second moment, an additive split of the error into an exact-arithmetic part and a rounding part, a $T^{-1/2}$ rate for tail averages, and a bias bound for stochastic rounding with few random bits; at most $\log_2 \tilde{\kappa} + 4$ bits were needed in our tests. For random singular vectors the trace formula is well approximated by the floor $(\eta(\kappa_F^2 + n^2)/(2n))^{1/2}$ in terms of the measured η, with $\kappa_F = \|A\|_F \|A^+\|_F$. In experiments with binary16, bfloat16 and four singular value profiles, this prediction matches the median measured floor to within 4% in all 43 configurations, with n from 50 to 500 and $\tilde{\kappa}$ from 10 to 640, and at $\tilde{\kappa} = 640$ in binary16 round-to-nearest is 12 times less accurate than stochastic rounding for the iterates and 37 times for their tail averages.

Keywords: randomized Kaczmarz; stochastic rounding; low precision; mixed precision; rounding error analysis; tail averaging

MSC: 65F10, 65G50, 65F20, 60G42

1 Introduction

The randomized Kaczmarz method [1,2] solves $Ax = b$ by projecting the iterate onto the hyperplane of one randomly chosen equation at a time. Each step touches a single row and the current iterate, so the cost of the method is dominated by moving these vectors through memory. Storing the matrix and the iterate in a 16-bit format halves this traffic compared with single precision, mixed precision is by now a standard tool in numerical linear algebra [16,17], and current accelerators make such formats cheap: binary16 and bfloat16 are standard on GPUs, and NVIDIA Blackwell, AMD CDNA4, Graphcore and Tesla Dojo provide instructions that convert single-precision results to 16-bit or 8-bit formats with stochastic rounding [9].

Low-precision storage leaves an error floor. With round-to-nearest, Dereźiński, Epperly, Needell and Xue [5] proved that randomized Kaczmarz in a uniform working precision reaches a forward error of order $\tilde{\kappa}^2 u \|x_\star\|$, where $\tilde{\kappa} = \|A\|_F \|A^+\|$, while a backward stable direct solver reaches $\tilde{\kappa} u \|x_\star\|$; they restore forward stability by iterative refinement. Rounding errors differ from noise in the data, whose propagation through Kaczmarz and other stationary iterations was analysed by Elfving [15]: they are injected at every step and their size depends on the iterate. The bound of Dereźiński et al. is a worst case; in our experiments the floor of round-to-nearest grows overall like $\tilde{\kappa}^{1.3}$ to $\tilde{\kappa}^{1.5}$ with local slopes up to 2.4, the error of its tail average varies by up to a factor of ten between matrices with the same $\tilde{\kappa}$, and in small examples the iterate stops at a grid point away from the solution (Figure 1). The floor is therefore not predictable from $\tilde{\kappa}$ and is only partly removed by averaging.

Stochastic rounding rounds a real number up or down with probabilities proportional to its distances from the two neighbouring floating-point numbers [7]. Its rounding errors have zero conditional mean and are mean independent [6], so for summation, inner products and other finite algorithms the constant nu in

worst-case bounds can be replaced by $\sqrt{n}\,u$ with high probability [6,7]. When only r random bits are used, a bias of size 2^{-r} ulp appears and $r \geq \frac{1}{2}\log_2 n$ bits suffice for sums of length n [8,10]. For iterative methods, the analyses of gradient descent with stochastic rounding [11,12] concern convergence rates and the threshold below which progress stops. Neither line of work quantifies the stationary error of a projection method, its dependence on the matrix, or the effect of averaging, and the 2026 survey of stochastic rounding [9] lists no work on Kaczmarz or other stationary iterations.

The closest results come from adaptive filtering. Randomized Kaczmarz is the normalized least-mean-squares algorithm with unit step applied to randomly sampled rows, and the effect of quantizing the weights of the least-mean-squares algorithm was analysed in the 1970s and 1980s [18–20]. These analyses assume that the quantization errors are white noise independent of all other quantities and that the input is a stationary signal with independent regressors, and they give closed-form steady-state errors. The noise model is known to fail when the updates are shorter than half a quantization step, where the adaptation stalls [18,21,22]. Stochastic rounding coincides with truncation after one unit in the last place of nonsubtractive rectangular dither [24]. It makes the zero-mean part of the noise model exact for every step size and removes the stall. Random row sampling makes the sampled rows independent of the past, which the classical analyses assume for the regressors; the rounding errors still depend on the iterate through their variance, and the analysis below uses their conditional mean and a bound on their variance, not independence. The results below are therefore rigorous counterparts of the classical roundoff formulas, for fixed matrices instead of stationary signals and with the noise variance measured rather than assumed. How large the floor of low-precision Kaczmarz is, what determines it, and what the rounding mode contributes had not been established without such modelling assumptions.

We answer these questions for iterates stored with stochastic rounding, abbreviated SR below, and compare with round-to-nearest, abbreviated RN. The results are the following.

1. SR makes the rounding errors act exactly like an inconsistency of the system (Theorem 1):

$$\lim_{K\to\infty} \frac{1}{K} \sum_{k<K} \mathbb{E}\,\|Ae_k\|^2 = \|r_\star\|^2 + \|A\|_F^2 \bar{\eta}.$$

   We call $\|A\|_F\sqrt{\bar{\eta}}$ the rounding residual. As a consequence the forward error floor is at most $\tilde{\kappa}u\|x_\star\|(1 + O(\tilde{\kappa}u))$, without refinement (Theorem 2).

2. The stationary second moment has an exact trace formula, bracketed by two computable bounds (Theorem 3, Corollary 1). If the leverage scores are proportional to the squared row norms and the noise variances are uncorrelated with the diagonal of $(A^TA)^{-1}$, it gives the floor $(\frac{\eta}{2n}(\kappa_F^2 + n^2))^{1/2}$ with $\kappa_F = \|A\|_F\|A^+\|_F$ and the measured noise level η; both conditions hold approximately for random singular vectors and can fail for singular vectors aligned with the coordinate axes. The squared error is the squared error in exact arithmetic plus a rounding term determined by the noise covariance of the same run, in the stationary regime and at every finite step (Corollary 2).

3. Tail averages of SR iterates converge at the rate $T^{-1/2}$, with the same order of dependence on $\tilde{\kappa}$ as the bound of Epperly, Goldshlager and Webber [4] for inconsistent systems and the squared residual replaced by the squared residual plus a bound on the squared rounding residual (Theorem 4). RN averages can stop at a bias floor, as they do in our tests.

4. With r random bits the bias accumulates coherently and is amplified by up to $\tilde{\kappa}^2$, while the unbiased part grows like $\tilde{\kappa}$ (Theorem 5). Comparing the two terms of the bound gives a sufficient number of bits $\log_2\tilde{\kappa} + O(1)$, the analogue for iterative methods of the rule $r \geq \frac{1}{2}\log_2 n$ for sums, with the relaxation time $\tilde{\kappa}^2$ in place of n; in our experiments between $\log_2\tilde{\kappa} + 2.7$ and $\log_2\tilde{\kappa} + 3.7$ bits reach the floor of ideal SR within 5%.

We also show that the null-space component of the error is a martingale for rank-deficient systems (Proposition 1) and that SR as the inner solver of iterative refinement contracts by the SR floor in every refinement step (Corollary 3). Table 1 places these results next to the closest work.

Table 1. Closest work. "Floor" means a statement about the error that persists as the iteration count grows.

Work	Rounding model	Method	Result on the error
Dereźiński et al. [5]	deterministic, $\lvert\delta\rvert \le u$	RK, accelerated RK	floor $\lesssim \tilde{\kappa}^2 u$; refinement gives $\tilde{\kappa} u$
Epperly et al. [4]	exact arithmetic	RK, tail average	rate T^{-1} in mean square for inconsistent systems
Connolly et al. [6], Croci et al. [7]	SR	finite algorithms	$\sqrt{n}\,u$ bounds with high probability
El Arar et al. [8]	SR with r bits	summation	$O(\sqrt{n}\,u + n2^{-r}u)$, $r \ge \frac{1}{2}\log_2 n$
Xia et al. [11,12]	SR, biased SR	gradient descent	convergence rate above a gradient threshold
Caraiscos, Liu [19], Ghanassi et al. [22]	additive white noise model	LMS, stationary input	steady-state error formula; model fails in the stall regime
This paper	SR, SR with r bits	RK, tail average, refinement	identity, floor $O(\tilde{\kappa}u)$, exact second moment, $T^{-1/2}$, $r = \log_2\tilde{\kappa} + O(1)$

Section 2 fixes the model. Sections 3–7 contain the analysis, Section 8 the application to refinement, Section 9 the experiments, Section 10 practical recommendations, and Section 11 the conclusions. Longer proofs are in Appendix A.

2 Setting

Let $A \in \mathbb{R}^{m\times n}$ have full column rank and rows $a_1^T, \dots, a_m^T$, and let $b \in \mathbb{R}^m$. Put $x_\star = A^+ b$ and $r_\star = b - Ax_\star$, so that $A^T r_\star = 0$. Let $S = A^T A/\|A\|_F^2$, so that $\lambda_{\min}(S) = \tilde{\kappa}^{-2}$ with $\tilde{\kappa} = \|A\|_F\|A^+\|$. At step k a row $i = i_k$ is drawn independently of the past with probability $p_i = \|a_i\|^2/\|A\|_F^2$, and with $P_i = a_i a_i^T/\|a_i\|^2$ we form

$$v_k = \hat{x}_k + \frac{b_i - a_i^T\hat{x}_k}{\|a_i\|^2}\,a_i, \qquad \hat{x}_{k+1} = \mathrm{fl}(v_k). \qquad (1)$$

The vector v_k is formed in a higher precision, which we idealize as exact, and fl rounds every component to a binary format with p significant bits, unit roundoff $u = 2^{-p}$ and smallest normal exponent $e_{\min}$. We denote by $\mathrm{ulp}(t)$ the spacing of the format at t; for normalized t, $u|t| < \mathrm{ulp}(t) \le 2u|t|$, and for subnormal t the spacing is $\epsilon_s = 2^{e_{\min}-p+1}$, so that $\mathrm{ulp}(t) \le \max\{2u|t|, \epsilon_s\}$ for all t. For binary16, $\epsilon_s = 2^{-24}$. We assume that no overflow occurs, so that every real number in the range of the iteration lies between two floating-point numbers; the test problems are scaled so that all quantities stay far below the overflow threshold. This model describes an iterate held in memory in a 16-bit format and updated in single precision, which is how the stochastic rounding conversions of current hardware are used [9]. It is also how low precision is usually simulated [13]. In the tests of Section 9, forming v_k in single precision changes the median errors of SR by at most 3%.

Write $e_k = \hat{x}_k - x_\star$, $\delta_k = \hat{x}_{k+1} - v_k$ and $g_i = r_{\star,i}\,a_i/\|a_i\|^2$. Since $b_i - a_i^T\hat{x}_k = r_{\star,i} - a_i^T e_k$,

$$e_{k+1} = (I - P_i)e_k + g_i + \delta_k. \qquad (2)$$

Let $\mathcal{F}_k$ be generated by the rows and rounding decisions before step k and $\mathcal{G}_k = \sigma(\mathcal{F}_k, i_k)$. Then $\hat{x}_k$ is $\mathcal{F}_k$-measurable and v_k is $\mathcal{G}_k$-measurable.

SR rounds each component t of v_k up to $\lceil t \rceil$ with probability $(t - \lfloor t \rfloor)/(\lceil t \rceil - \lfloor t \rfloor)$ and down otherwise, with independent random numbers for every component and step [7]. Equivalently, a uniform random number in $[0, \operatorname{ulp}(t))$ is added to t and the result is rounded toward $-\infty$, which is nonsubtractive rectangular dither of one unit in the last place [24]. Such dither makes the rounding error mean zero for every input but leaves its variance dependent on the input, which is the content of (S2) below.

Lemma 1. *Under SR,*

(S1) $\mathbb{E}[\delta_k \mid \mathcal{G}_k] = 0$*;*

(S2) the components of δ_k *are conditionally independent given* $\mathcal{G}_k$*, and*

$$\mathbb{E}[\delta_{k,j}^2 \mid \mathcal{G}_k] = (v_{k,j} - \lfloor v_{k,j} \rfloor)(\lceil v_{k,j} \rceil - v_{k,j}) \le \min\{1/4 \operatorname{ulp}(v_{k,j})^2,\ \operatorname{ulp}(v_{k,j})\, |v_{k,j} - \hat{x}_{k,j}|\}.$$

In particular $\mathbb{E}[\|\delta_k\|^2 \mid \mathcal{G}_k] \le u^2 \|v_k\|^2 + n\epsilon_s^2/4$*, and the last term is absent if no component of* v_k *is subnormal.*

Proof. Property (S1) and the variance formula are the defining properties of SR [7, Thm. 6.3]. With $\alpha = t - \lfloor t \rfloor$ and $\beta = \lceil t \rceil - t$ we have $\alpha + \beta = \operatorname{ulp}(t)$, so $\alpha\beta \le \operatorname{ulp}(t)^2/4$, and also $\alpha\beta \le \min\{\alpha, \beta\}\operatorname{ulp}(t)$. The distance $\min\{\alpha, \beta\}$ from t to the grid is at most $|t - \hat{x}_{k,j}|$ because $\hat{x}_{k,j}$ is a floating-point number. Finally $\operatorname{ulp}(t)^2/4 \le \max\{u^2 t^2, \epsilon_s^2/4\} \le u^2 t^2 + \epsilon_s^2/4$. □

By the second bound in (S2), the noise level adapts to the step size: an update shorter than one ulp is rounded away with high probability and leaves a small rounding error. In the random test problems of Figure 2 the measured noise level $\mathbb{E}\|\delta_k\|^2$ is 35% to 54% of the uniform model $\sum_j \operatorname{ulp}(x_{\star,j})^2/6$, with median 47%, and in the tomography problem of Section 9, where the updates are much shorter, it is 4% to 6%. For this reason all predictions below use the measured noise level η.

3 The rounding residual

Figure 1 compares the two rounding modes on a consistent 3×2 system and a grid with $p = 6$ bits. There the RN iterate reaches a grid point from which every row update is shorter than half an ulp, and it stays there; the error of 9.8×10^{-3} is the distance from $x_\star$ to that point, and averaging cannot remove it. The SR iterate keeps moving among the grid points around $x_\star$ and its average is 70 times more accurate. The mean of $\|Ae_k\|^2$ along its path is exactly that of the same iteration in exact arithmetic on an inconsistent system whose residual has the energy of the rounding errors, as the next theorem shows.

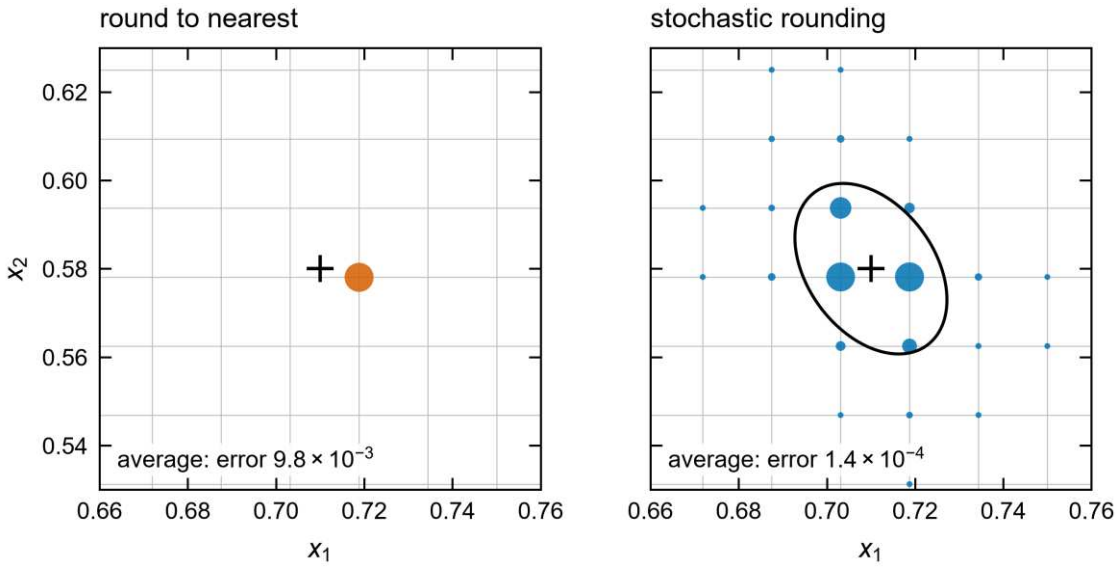


Figure 1. RN gets stuck, SR fluctuates around $x_\star$ with a predictable spread. Randomized Kaczmarz for $A = [1,\ 0.3;\ 0.95,\ 0.45;\ 0.2,\ 1]$, $x_\star = (0.71, 0.58)^T$, $b = Ax_\star$, with the iterate stored with $p = 6$ bits; grey lines are the floating-point grid, marker area is proportional to the number of visits in the last 2×10^4 of 6×10^4 steps, and + marks $x_\star$. The ellipse is the two-standard-deviation ellipse of the second moment predicted by Theorem 3 with the measured rounding noise; the predicted root-mean-square error is 1.411×10^{-2} and the measured one 1.409×10^{-2}.

Theorem 1 (residual identity). *Assume (S1). For every* $k \geq 0$,

$$\mathbb{E}\|e_{k+1}\|^2 = \mathbb{E}\|e_k\|^2 - \frac{\mathbb{E}\|Ae_k\|^2}{\|A\|_F^2} + \frac{\|r_\star\|^2}{\|A\|_F^2} + \mathbb{E}\|\delta_k\|^2. \qquad (3)$$

If $\sup_k \mathbb{E}\|e_k\|^2 < \infty$ *and* $\bar{\eta} = \lim_{K\to\infty} K^{-1} \sum_{k<K} \mathbb{E}\,\|\delta_k\|^2$ *exists, then*

$$\lim_{K\to\infty} \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\,\|Ae_k\|^2 = \|r_\star\|^2 + \|A\|_F^2\,\bar{\eta}. \qquad (4)$$

Proof. Expand the square of (2). Since $(I - P_i)^T(I - P_i) = I - P_i$ and $\mathbb{E}[P_{i_k} \mid \mathcal{F}_k] = S$, the first term contributes $\mathbb{E}\|e_k\|^2 - \mathbb{E}[e_k^T S e_k]$, and $\mathbb{E}[e_k^T S e_k] = \mathbb{E}\|Ae_k\|^2/\|A\|_F^2$. Next, $\mathbb{E}\left\|g_{i_k}\right\|^2 = \sum_i p_i\, r_{\star,i}^2/\|a_i\|^2 = \|r_\star\|^2/\|A\|_F^2$. The cross term of the first two terms vanishes identically because $a_i^T(I - P_i) = 0$, and the cross terms with δ_k have zero expectation by (S1), since $(I - P_i)e_k$ and g_i are $\mathcal{G}_k$-measurable. Summing (3) over $k < K$, dividing by K and letting $K \to \infty$ gives (4). □

Identity (4) says that SR adds $\|A\|_F^2\bar{\eta}$ to the mean of $\|Ae_k\|^2$, as an inconsistency of that size would. Since $r_\star$ is orthogonal to the range of A, the computable residual satisfies $\|b - A\hat{x}_k\|^2 = \|r_\star\|^2 + \|Ae_k\|^2$. Arguments for inconsistent systems that use the residual only through the one-step relation (3), such as the proof of the tail-averaging bound of Epperly, Goldshlager and Webber [4], therefore carry over with $\|r_\star\|^2$ replaced by $\|r_\star\|^2 + \|A\|_F^2\bar{\eta}$; Theorem 4 below is an example. Rounding noise and residual noise still cause different forward errors, because residual noise enters along the sampled row and rounding noise in all coordinates; Section 4 quantifies the difference. Averaged over the 64 chains of a configuration, the two sides of (4) agree to within 1.4% for SR in all our experiments, while for RN on consistent systems their ratio ranges from 1.7 to 50 and grows with $\tilde{\kappa}$. For a consistent system the left side is the mean squared residual and η is available from δ_k during the run, so the ratio is a cheap diagnostic of rounding bias. A single chain observed over $8\tilde{\kappa}^2$ steps gives SR ratios between 0.76 and 1.41; for $\tilde{\kappa} \geq 40$ they stay below 1.3, while single RN chains exceed 1.6.

Theorem 2 (forward error floor). *Assume (S1)–(S2) and* $\tilde{\kappa}u < \frac{1}{2}$*, and let* $\rho = 1 - \tilde{\kappa}^{-2}(1 - \tilde{\kappa}u - \tilde{\kappa}^2u^2)$. *Then*

$$\mathbb{E}\|e_k\|^2 \le \rho^k\|e_0\|^2 + \frac{\tilde{\kappa}^2(\|r_\star\|^2/\|A\|_F^2 + u^2(1+\tilde{\kappa}u)c_\star^2 + n\epsilon_s^2/4)}{1 - \tilde{\kappa}u - \tilde{\kappa}^2u^2}, \qquad c_\star^2 = \|x_\star\|^2 + \frac{\|r_\star\|^2}{\|A\|_F^2}.$$

For a consistent system without underflow, $\operatorname{limsup}_k \mathbb{E}\|e_k\|^2 \le \tilde{\kappa}^2u^2\|x_\star\|^2(1 + O(\tilde{\kappa}u))$.

Proof. By (3) and $\mathbb{E}\|Ae_k\|^2 \ge \|A\|_F^2\tilde{\kappa}^{-2}\mathbb{E}\|e_k\|^2$, $\mathbb{E}\|e_{k+1}\|^2 \le (1 - \tilde{\kappa}^{-2})\mathbb{E}\|e_k\|^2 + \|r_\star\|^2/\|A\|_F^2 + \mathbb{E}\|\delta_k\|^2$. By Lemma 1, $\mathbb{E}\|\delta_k\|^2 \le u^2\mathbb{E}\|v_k\|^2 + n\epsilon_s^2/4$, and $v_k = (I - P_i)e_k + (x_\star + g_i)$. The cross term of these two vectors need not vanish, so we use $(a+b)^2 \le (1+\epsilon^{-1})a^2 + (1+\epsilon)b^2$ with $\epsilon = \tilde{\kappa}u$, together with $\mathbb{E}\|x_\star + g_i\|^2 = c_\star^2$, which holds because $\mathbb{E}g_i = A^Tr_\star/\|A\|_F^2 = 0$. This gives

$$\mathbb{E}\|e_{k+1}\|^2 \le \left(1 - \tilde{\kappa}^{-2} + u^2 + \frac{u}{\tilde{\kappa}}\right)\mathbb{E}\|e_k\|^2 + \frac{\|r_\star\|^2}{\|A\|_F^2} + u^2(1+\tilde{\kappa}u)c_\star^2 + \frac{n\epsilon_s^2}{4},$$

and the first factor equals $\rho < 1$. The subnormal term is negligible unless $\|x_\star\| \lesssim \sqrt{n}\,\epsilon_s/(2u)$, which for binary16 means $\|x_\star\| \lesssim 6 \times 10^{-5}\sqrt{n}$. Summing the geometric series, $1/(1-\rho) = \tilde{\kappa}^2/(1 - \tilde{\kappa}u - \tilde{\kappa}^2u^2)$. □

Up to the factor $1 + O(\tilde{\kappa}u)$, Theorem 2 reaches the accuracy $\tilde{\kappa}u\|x_\star\|$ of a backward stable direct solver by changing only the rounding mode; for binary16 and $\tilde{\kappa} = 640$, where $\tilde{\kappa}u = 0.31$, the factor $((1+\tilde{\kappa}u)/(1 - \tilde{\kappa}u - \tilde{\kappa}^2u^2))^{1/2}$ in the bound is 1.49. The comparison with [5] is made within one model: their analysis rounds every operation, but its only input is the per-step bound $\|\hat{x}_{k+1} - v_k\| \lesssim u(\|x_\star\| + \|\hat{x}_k\|)$, which RN satisfies in the model of Section 2, so their floor $\tilde{\kappa}^2u$ applies to RN storage as well. In the proof of Theorem 2, property (S1) lets the rounding errors enter only through their squared norms, so the contraction factor $1 - \tilde{\kappa}^{-2}$ is paid once in the second moment, giving $\tilde{\kappa}^2u^2$; a recursion for norms, which is what a deterministic rounding model allows, pays it in the first moment and gives $\tilde{\kappa}^2u$ [5]. For binary16 and $\tilde{\kappa} = 40$, the factors $\tilde{\kappa}u$ and $\tilde{\kappa}^2u$ are 2% and 78%. The theorem assumes $\tilde{\kappa}u < \frac{1}{2}$, which admits $\tilde{\kappa}$ up to about 1000 in binary16 and 128 in bfloat16; the deterministic analysis requires $\tilde{\kappa}^2u \ll 1$, that is $\tilde{\kappa} \ll 45$ and $\tilde{\kappa} \ll 16$. In the experiments of Section 9 the RN error reaches $\tilde{\kappa}u$ at $\tilde{\kappa} = 640$ in binary16, while the SR error stays a factor eleven below it.

4 The stationary second moment

The bound $\tilde{\kappa}u\|x_\star\|$ is attained only if the error concentrates on the smallest singular direction. The exact floor follows from the second moment. Let $m_k = \mathbb{E}e_k$; by (S1) and $\mathbb{E}g_i = 0$, $m_{k+1} = (I - S)m_k$, so $m_k \to 0$. Put $\Sigma_k = \mathbb{E}[e_ke_k^T]$, $G = \mathbb{E}[g_ig_i^T] = \|A\|_F^{-2}\sum_i r_{\star,i}^2 P_i$, $D_k = \mathbb{E}[\delta_k\delta_k^T]$, which is diagonal by (S2), and let $\ell_i = a_i^T(A^TA)^{-1}a_i$ be the leverage scores.

Theorem 3 (second moment). *Assume (S1)–(S2) and* $\tilde{\kappa}u < \frac{1}{2}$*, and let* $(\bar{\Sigma}, \bar{D})$ *be a limit point of the Cesàro averages of* (Σ_k, D_k). *Then*

$$S\bar{\Sigma} + \bar{\Sigma}S = G + \bar{D} + \mathbb{E}[P_i\bar{\Sigma}P_i], \qquad (5)$$

$$\operatorname{tr}\bar{\Sigma} = 1/2\operatorname{tr}(S^{-1}(G + \bar{D})) + 1/2\sum_{i=1}^{m}\ell_i\frac{a_i^T\bar{\Sigma}a_i}{\|a_i\|^2}, \qquad (6)$$

and, with $\mu(A) = \|A\|_F^2\max_i\ell_i/\|a_i\|^2$,

$$1/2\operatorname{tr}(S^{-1}(G + \bar{D})) \le \operatorname{tr}\bar{\Sigma} \le 1/2\operatorname{tr}(S^{-1}(G + \bar{D})) + 1/2\,\mu(A)\operatorname{tr}(G + \bar{D}). \qquad (7)$$

The proof is in Appendix A. It multiplies (5) by S^{-1}, takes the trace, and bounds the last term with Theorem 1.

Corollary 1 (floor law). *Consider a consistent system and let* $\kappa_F = \|A\|_F \|A^+\|_F$.

(a) If $d_- I \preccurlyeq \bar{D} \preccurlyeq d_+ I$, *then* $\frac{1}{2} d_- \kappa_F^2 \le \mathrm{tr}\bar{\Sigma} \le \frac{1}{2} d_+ (\kappa_F^2 + n\mu(A))$.

(b) If the leverage scores are proportional to the sampling probabilities, $\ell_i = np_i$ *for all* i, *then* $\mu(A) = n$ *and*

$$\mathrm{tr}\bar{\Sigma} = 1/2\,\mathrm{tr}(S^{-1}\bar{D}) + n/2\,\mathrm{tr}\bar{D};$$

for $\bar{D} = (\eta/n)I$ *this is* $\mathrm{tr}\bar{\Sigma} = \frac{\eta}{2n}(\kappa_F^2 + n^2)$.

Proof. (a) $\mathrm{tr}S^{-1} = \|A\|_F^2 \mathrm{tr}((A^TA)^{-1}) = \kappa_F^2$ and $\mathrm{tr}\bar{D} \le nd_+$. (b) If $\ell_i = np_i$, the last term of (6) equals $\frac{n}{2}\sum_i p_i\, a_i^T \bar{\Sigma} a_i / \|a_i\|^2 = \frac{n}{2}\mathrm{tr}(S\bar{\Sigma})$, and $\mathrm{tr}(S\bar{\Sigma}) = \mathrm{tr}\bar{D}$ by Theorem 1. □

The condition in (b) holds for square matrices with rows of equal length, since then $\ell_i = 1$ and $p_i = 1/n$, and approximately for matrices with Haar random singular vectors, whose row norms and leverage scores both concentrate. The term $\frac{n}{2}\mathrm{tr}\bar{D}$ explains why the floor exceeds $\frac{1}{2}\mathrm{tr}(S^{-1}\bar{D})$ for moderate $\tilde{\kappa}$. For isotropic noise the two terms are $\frac{\eta}{2n}\kappa_F^2$ and $\frac{n}{2}\eta$, so since $\kappa_F^2 \ge n^2$ the second is at most as large as the first and becomes negligible once $\kappa_F \gg n$; for anisotropic noise this comparison can fail. The leading term depends on how the noise is distributed over the coordinates. With $s_j = (S^{-1})_{jj}$, whose mean over j is κ_F^2/n, and $d_j = \bar{D}_{jj}$, whose mean is η/n with $\eta = \mathrm{tr}\bar{D}$,

$$\mathrm{tr}(S^{-1}\bar{D}) = \frac{\eta}{n}\kappa_F^2 + \sum_{j=1}^{n}\left(s_j - \frac{\kappa_F^2}{n}\right)\left(d_j - \frac{\eta}{n}\right), \qquad (8)$$

so $\frac{1}{2}\mathrm{tr}(S^{-1}\bar{D}) = \frac{\eta}{2n}\kappa_F^2$ exactly when the weights s_j and the noise variances d_j are uncorrelated over the coordinates. The variances d_j scale with the squared magnitudes of the components of $x_\star$ and with the lengths of the updates in each coordinate, while for right singular vectors that are not aligned with the coordinate axes every weight s_j mixes all singular values, so that the correlation is small. If it is negligible and $\ell_i \approx np_i$, the root-mean-square error is, in terms of the measured noise level and without fitted constants,

$$\left(\frac{\eta}{2n}(\kappa_F^2 + n^2)\right)^{1/2}. \qquad (9)$$

The noise level is available during the run, since $\delta_k = \hat{x}_{k+1} - v_k$ is formed at every step, whereas the forward error is not; when κ_F is known or estimated and the two conditions above are plausible, (9) turns the one into an estimate of the other. For the 176 matrices of Figure 2, which have Haar random singular vectors, the ratio of $\mathrm{tr}(S^{-1}\bar{D})$ to $\frac{\eta}{n}\kappa_F^2$ for $\bar{D} \propto \mathrm{diag}(\mathrm{ulp}(x_{\star,j})^2)$ lies between 0.83 and 1.17 with median 1.00; for the tomography matrix of Section 9, where the background pixels of the image are small, it is 0.78. For singular vectors aligned with the coordinate axes the correlation in (8) can be large; Section 9 gives examples in which the ratio of the measured floor to (9) ranges from 0.10 to 1.36, while the solution of (5) with the measured $\bar{D}$ stays within 6%. Because $(\kappa_F^2 + n^2)/n \le \tilde{\kappa}^2 + n \le 2\tilde{\kappa}^2$, the law refines Theorem 2; when only a few singular values are small the floor is far below $\tilde{\kappa}u\|x_\star\|$. For $n = 50$, geometric singular values, $p = 14$ and $\bar{D} = \mathrm{diag}(\mathrm{ulp}(x_{\star,j})^2/6)$, the trace of the exact solution of (5) exceeds $\frac{1}{2}\mathrm{tr}(S^{-1}\bar{D}) + \frac{n}{2}\mathrm{tr}\bar{D}$ by 0.7%, 0.6%, 0.2% and 0.1% at $\tilde{\kappa} \approx 15$, 40, 100 and 250, whereas it exceeds the lower bound $\frac{1}{2}\mathrm{tr}(S^{-1}\bar{D})$ by 60%, 16%, 4% and 1%.

Corollary 2 (additivity).

(a) Under the assumptions of Theorem 3, let Σ_{ex} *solve (5) with* $\bar{D} = 0$*, which is the stationary second moment in exact arithmetic, and let* Σ_D *solve it with* $G = 0$ *and the noise covariance* $\bar{D}$ *of the same run. Then* $\bar{\Sigma} = \Sigma_{\mathrm{ex}} + \Sigma_D$.

(b) Assume (S1). Let x_t' *be the iterates of the same iteration in exact arithmetic, with the same rows and the same starting point, and let* $z \in \mathbb{R}^n$ *be fixed. Then for every* t

$$\mathbb{E}\|\hat{x}_t - z\|^2 = \mathbb{E}\|x_t' - z\|^2 + \mathbb{E}\|\hat{x}_t - x_t'\|^2,$$
$$\mathbb{E}[(\hat{x}_t - x_t')(\hat{x}_t - x_t')^T] = \sum_{k<t} \mathcal{M}^{t-1-k}(D_k),$$

where $\mathcal{M}(X) = \mathbb{E}[(I - P_i)X(I - P_i)]$*, and* $\mathbb{E}\|\hat{x}_t - x_t'\|^2 \le \sum_{k<t} \mathbb{E}\,\|\delta_k\|^2$.

Proof. (a) Write (5) as $\bar{\Sigma} - \mathcal{M}(\bar{\Sigma}) = G + \bar{D}$. The map $\mathcal{M}$ preserves the positive semidefinite cone and $\mathrm{tr}\mathcal{M}(X) = \mathrm{tr}X - \mathrm{tr}(SX) \le (1 - \tilde{\kappa}^{-2})\mathrm{tr}X$ for $X \succcurlyeq 0$, so the Neumann series $\sum_k \mathcal{M}^k$ converges on symmetric matrices and the solution is linear in $G + \bar{D}$. (b) The difference $d_t = \hat{x}_t - x_t'$ satisfies $d_0 = 0$ and $d_{t+1} = (I - P_i)d_t + \delta_t$, and $(I - P_i)d_t$ is $\mathcal{G}_t$-measurable, so (S1) gives $\mathbb{E}[d_{t+1}d_{t+1}^T] = \mathcal{M}(\mathbb{E}[d_t d_t^T]) + D_t$. The error $x_t' - z$ is a function of the rows $i_0, \dots, i_{t-1}$, and d_t is a sum of terms $\Pi_k \delta_k$ in which Π_k depends only on the rows after step k. These rows are drawn independently of $\mathcal{G}_k$ and of δ_k, so $\mathbb{E}[\delta_k \mid \mathcal{G}_k, i_{k+1}, \dots, i_{t-1}] = 0$ and the cross term $\mathbb{E}[(x_t' - z)^T d_t]$ vanishes. The last inequality follows from $\mathrm{tr}\mathcal{M}(X) \le \mathrm{tr}X$. □

Corollary 2 separates the two contributions within one run. The rounding term depends on the residual through $\bar{D}$, because the noise variance depends on the length of the updates. For $A = (1,1)^T$ and $b = (1,1 + h)^T$, with h the spacing of the format at 1, every projection lands on 1 or $1 + h$ and $\bar{D} = 0$, while the consistent system with the same solution $x_\star = 1 + h/2$ has $\bar{D} = h^2/4$. In the experiment of Figure 3(c) the noise level instead grows with the relative residual $\gamma = \|r_\star\|/(\|A\|_F\|x_\star\|)$, from 54% of the uniform model at $\gamma/u = 0.07$ to 82% at $\gamma/u = 1$ and 97% at $\gamma/u = 10$, because longer updates spread the positions within one ulp more evenly. The floor of the consistent problem is therefore neither an upper nor a lower bound for the rounding term. Low-precision storage adds little error when the rounding term is small compared with the error of randomized Kaczmarz in exact arithmetic, which grows with γ. In Figure 3(c), binary16 storage increases the root-mean-square error by factors 19.7, 1.95 and 1.02 at $\gamma/u = 0.07$, 1 and 10, and Corollary 2(a) with $\mathrm{tr}\Sigma_D$ from (9) and the η measured in each run predicts 20.1, 1.93 and 1.02; with the η of the nearly consistent system the prediction at $\gamma/u = 1$ would be 1.68. The cruder estimate $\sqrt{1 + \eta/\gamma^2}$, which treats rounding noise like residual noise, predicts 6.4 instead of 19.7, because residual noise enters along the rows while rounding noise is isotropic and is amplified by κ_F^2.

Part (b) applies to iterations stopped after t steps, as in semiconvergent use for ill-posed problems, where the noise covariance D_k changes with k. To evaluate it, let $S = \sum_j \lambda_j w_j w_j^T$ with $\lambda_j = \sigma_j^2/\|A\|_F^2$ and replace $\mathbb{E}[P_i X P_i]$ by $\mathrm{tr}(SX)\,S$. This keeps $\mathrm{tr}\mathcal{M}(X)$ exact and, if $\ell_i = np_i$, also the weighted trace $\mathrm{tr}(S^{-1}\mathcal{M}(X))$ used in Theorem 3, and on the diagonal y of X in the basis (w_j) it acts as the matrix $B = I - 2\Lambda + \lambda\lambda^T$ with $\Lambda = \mathrm{diag}(\lambda)$. Then

$$\mathbb{E}\|\hat{x}_t - x_t'\|^2 \approx \mathbf{1}^T \sum_{k<t} B^{t-1-k} c_k, \qquad c_{k,j} = w_j^T D_k w_j. \qquad (10)$$

For constant $D_k = \bar{D}$ the limit is $\mathbf{1}^T(I - B)^{-1}c = \frac{1}{2}\mathrm{tr}(S^{-1}\bar{D}) + \frac{n}{2}\mathrm{tr}\bar{D}$, the expression of Corollary 1(b), by the Sherman–Morrison formula. Directions that have not converged after t steps have not accumulated rounding noise either.

5 Tail averaging

Let $\rho_0 = 1 - \tilde{\kappa}^{-2}$, $a_k = \mathbb{E}\|e_k\|^2$, $\nu_k = \|r_\star\|^2/\|A\|_F^2 + \mathbb{E}\|\delta_k\|^2$ and $\nu^* = \sup_k \nu_k$, which is finite by Theorem 2. For $2 \le t_b < t$ let $T = t - t_b$ and $\bar{x}_t = T^{-1}\sum_{s=t_b}^{t-1} \hat{x}_s$.

Theorem 4 (tail averaging). *Assume (S1) and $\nu^* < \infty$; the latter holds under (S1)–(S2) and $\tilde{\kappa}u < \frac{1}{2}$ by Theorem 2. Then*

$$\mathbb{E}\|\bar{x}_t - x_\star\|^2 \le \rho_0^{2t_b}\min\left\{1, \frac{\tilde{\kappa}^4}{T^2}\right\}\|e_0\|^2 + \frac{2\tilde{\kappa}^4\nu^*}{T} + \frac{\tilde{\kappa}^4}{T^2}\left(4\tilde{\kappa}^2\nu^* + (\tilde{\kappa}^2+2)\rho_0^{t_b-2}\|e_0\|^2\right).$$

The proof, in Appendix A, writes $e_{k+1} = (I-S)e_k + \zeta_k$ with martingale differences ζ_k and bounds $\mathbb{E}\|\zeta_k\|^2$ by a telescoping quantity obtained from (3). The leading term has the form of the bound of Epperly, Goldshlager and Webber for inconsistent systems [4, Thm. 2], with $\|r_\star\|^2$ replaced by $\|A\|_F^2\nu^*$, the squared residual plus the squared rounding residual, as Theorem 1 suggests. For a consistent system ν^* is of order $u^2\|x_\star\|^2$ by Lemma 1, so by the bound the error of the average decreases like $\tilde{\kappa}^2 u\|x_\star\|/\sqrt{T}$ up to a constant, and comparison with Theorem 2 suggests that the average beats the iterates once T is of order $\tilde{\kappa}^2$; in Figure 3(a) this happens already at $T \approx \tilde{\kappa}^2/4$. The constant is not sharp: for $T \to \infty$, $\mathbb{E}\|\bar{x}_t - x_\star\|^2 \sim T^{-1}\mathrm{tr}(S^{-1}QS^{-1})$ with Q the Cesàro limit of $\mathbb{E}[\zeta_k\zeta_k^T]$. For a format with $p = 14$ and $n = 50$ this asymptotic value is within 3% and 13% of the measured error at $\tilde{\kappa} = 40$ and 100, while the bound of Theorem 4 is about five times larger.

Averaging need not help RN in the same way. Its rounding errors need not have zero conditional mean, so the average can converge to a biased point; in Figure 3(a) the RN average stops near 10^{-3} while the SR average continues to decrease at the rate $T^{-0.46}$.

6 Stochastic rounding with few random bits

Hardware implementations of SR use a fixed number r of random bits [9,10]. Then (S1) fails and we assume instead

(S1$_r$) $\beta_k := \mathbb{E}[\delta_k \mid \mathcal{G}_k]$ satisfies $|\beta_{k,j}| \le 2^{-r}\mathrm{ulp}(v_{k,j}) \le 2^{-r}\max\{2u|v_{k,j}|, \epsilon_s\}$,

while $\tilde{\delta}_k = \delta_k - \beta_k$ satisfies $\mathbb{E}[\tilde{\delta}_k \mid \mathcal{G}_k] = 0$ and $\mathbb{E}[\tilde{\delta}_{k,j}^2 \mid \mathcal{G}_k] \le \frac{1}{4}\mathrm{ulp}(v_{k,j})^2$. Rounding first to $p+r$ bits by truncation and then stochastically to p bits [8] satisfies (S1$_r$).

Theorem 5 (random bits). *Assume (S1$_r$), and let $\beta^* = \sup_k(\mathbb{E}\|\beta_k\|^2)^{1/2}$ and $\tilde{\eta}^* = \sup_k \mathbb{E}\left\|\tilde{\delta}_k\right\|^2$.*

(a) $\mathbb{E}e_{k+1} = (I-S)\mathbb{E}e_k + \mathbb{E}\beta_k$, hence $\mathrm{limsup}_k\|\mathbb{E}e_k\| \le \tilde{\kappa}^2\sup_k\|\mathbb{E}\beta_k\| \le \tilde{\kappa}^2\beta^*$.

(b) For every $\theta \in (0,1)$,

$$\limsup_k \mathbb{E}\|e_k\|^2 \le \frac{\tilde{\kappa}^2}{1-\theta}\left[(1+\theta\tilde{\kappa}^{-2})\frac{\|r_\star\|^2}{\|A\|_F^2} + \tilde{\eta}^* + \left(1+\frac{\tilde{\kappa}^2}{\theta}\right)\beta^{*2}\right].$$

Without subnormal components, $\beta^* \le 2^{1-r}uV$ and $\tilde{\eta}^* \le u^2V^2$ with $V^2 = \sup_k\mathbb{E}\|v_k\|^2$, and $V \approx \|x_\star\|$ for a consistent system near its solution; subnormal components add at most $2^{-r}\sqrt{n}\,\epsilon_s$ to β^*. The bias then contributes $O(\tilde{\kappa}^{2\,2^{-r}}u\|x_\star\|)$ to the root-mean-square error while the unbiased part contributes $O(\tilde{\kappa}u\|x_\star\|)$, and the bound on the bias term of (b) does not exceed that on the unbiased term once $r \ge 1 + \frac{1}{2}\log_2(1 + \tilde{\kappa}^2/\theta)$, which is $\log_2\tilde{\kappa} + \frac{3}{2} + o(1)$ for $\theta = \frac{1}{2}$. This is the analogue of the rule $r \ge \frac{1}{2}\log_2 n$ for sums of length n [8]: unbiased errors accumulate like a random walk over the relaxation time $\tilde{\kappa}^2$ of the iteration, the number of steps over which the contraction factor $1 - \tilde{\kappa}^{-2}$ forgets an error, and the bias accumulates coherently over the same time. The condition compares bounds and is not necessary for the error; the

measured threshold is 1.2 to 2.2 bits larger, see below. Part (a) also predicts the bias itself. If the iterates stay close to $x_\star$ and their positions within one ulp are spread uniformly, then $\bar{\beta}_j \approx -\mathrm{sgn}(x_{\star,j})2^{-r-1}\mathrm{ulp}(x_{\star,j})$ for truncation toward zero, and the stationary mean error is $S^{-1}\bar{\beta}$. In Figure 3(b) the resulting prediction $(\|S^{-1}\bar{\beta}\|^2 + N^2)^{1/2}$, with N the floor of ideal SR, agrees with the measurement to within 14% in the median over the matrices wherever the predicted error is below 5%. On each matrix we divide the root-mean-square error with r bits by that of ideal SR; the smallest number of bits for which the median of this ratio over the matrices is at most 1.05 is 8, 9 and 11 for $\tilde{\kappa} = 20$, 80 and 320, that is r between $\log_2\tilde{\kappa} + 2.7$ and $\log_2\tilde{\kappa} + 3.7$.

7 Rank-deficient systems

Without full column rank, let $x_\star = A^+ b$ and let $P_{\mathcal{N}}$ be the orthogonal projector onto the null space of A. All P_i and g_i live in the range of A^T, so applying $P_{\mathcal{N}}$ to (2) gives the following statement.

Proposition 1 (null-space drift). *Assume (S1). Then $P_{\mathcal{N}} e_{k+1} = P_{\mathcal{N}} e_k + P_{\mathcal{N}}\delta_k$, so $(P_{\mathcal{N}} e_k)$ is a martingale and*

$$\mathbb{E}\|P_{\mathcal{N}} e_k\|^2 = \|P_{\mathcal{N}} e_0\|^2 + \sum_{j<k} \mathbb{E}\left\|P_{\mathcal{N}}\delta_j\right\|^2.$$

With $P_{\mathcal{R}} = I - P_{\mathcal{N}}$, the component $P_{\mathcal{R}} e_k$ satisfies (2) with δ_k replaced by $P_{\mathcal{R}}\delta_k$. This noise still satisfies (S1), but its covariance $P_{\mathcal{R}} D_k P_{\mathcal{R}}$ is in general not diagonal; for $A = (1,1)$ the two projected components are perfectly correlated. Theorems 1, 3 and 4 use (S1) and the second moments of the noise but not the independence in (S2), so they hold for the range component with $\tilde{\kappa} = \|A\|_F\|A^+\|$ and $\bar{D}$ replaced by $P_{\mathcal{R}}\bar{D}P_{\mathcal{R}}$, provided the noise moments stay bounded. Theorem 2 bounds these moments through $\|v_k\|$, which contains the null-space component, so it applies as long as $\|P_{\mathcal{N}} e_k\|$ is small compared with $\|x_\star\|$. With r random bits the bias $P_{\mathcal{N}}\beta_k$ accumulates linearly in the null space. If $\mathbb{E}\|P_{\mathcal{N}}\delta_k\|^2$ stays near a nonzero value, the null-space error grows like $\sqrt{k}$ times its root, which is of order $u\|x_\star\|$; it does not grow at all if the iterates reach a point that no update moves. For a deterministic rounding model it can grow linearly [14]. In Figure 3(d), with five zero singular values and $n = 50$, the SR null-space error grows by a factor 4.3 when k grows by a factor 21, against $\sqrt{21} = 4.6$, while the RN error grows by 6.7. The drift does not affect $A\hat{x}_k$ and is removed by one projection onto the range of A^T at the end.

8 Stochastic rounding inside iterative refinement

Iterative refinement with a randomized Kaczmarz inner solver [5] keeps the outer iterate x_j and solves $Ae = b - Ax_j$ approximately by K inner steps started at $e = 0$, with the residual $b_i - a_i^T(x_j + e)$ formed on the fly. In the model of Section 2, both x_j and e are stored in the low-precision format and $x_{j+1} = \mathrm{fl}(x_j + e)$.

Corollary 3 (refinement with an SR inner solver). *Let the system be consistent, let the inner iterate be rounded by SR, the outer update by RN, assume $\tilde{\kappa}u < \frac{1}{2}$ and that no underflow occurs. Then*

$$(\mathbb{E}\left\|x_{j+1} - x_\star\right\|^2)^{1/2} \le (1+u)\, q\, (\mathbb{E}\left\|x_j - x_\star\right\|^2)^{1/2} + u\|x_\star\|,$$

where

$$q = \rho^{K/2} + \tilde{\kappa}u\left(\frac{1+\tilde{\kappa}u}{1-\tilde{\kappa}u-\tilde{\kappa}^2u^2}\right)^{1/2}.$$

Proof. Conditionally on x_j, the inner iteration is randomized Kaczmarz with SR for the consistent system $Ae = A(x_\star - x_j)$, started at 0. Theorem 2 with $x_\star$ replaced by $x_\star - x_j$ and $r_\star = 0$ bounds the mean

squared inner error after K steps by $(\rho^K + \tilde{\kappa}^2 u^2(1+\tilde{\kappa}u)/(1-\tilde{\kappa}u-\tilde{\kappa}^2u^2))\|x_\star - x_j\|^2$. With $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$ this gives $(\mathbb{E}\|x_j + e - x_\star\|^2)^{1/2} \le q(\mathbb{E}\|x_j - x_\star\|^2)^{1/2}$. The outer RN rounding of $x_j + e$ changes it by at most $u\|x_j + e\| \le u(\|x_j + e - x_\star\| + \|x_\star\|)$ in norm, and the triangle inequality in L^2 completes the proof. □

The first term of q is a worst case, attained when the initial error lies in the slowest direction; for errors spread over many directions the transient decays much faster, and each step then contracts by about the floor of the inner solver. If $(1+u)q < 1$, which for large K holds for $\tilde{\kappa}u < 0.46$, the bound contracts at the rate $(1+u)q$ toward $u\|x_\star\|/(1-(1+u)q)$, which is of the order of the storage limit $u\|x_\star\|$ when q stays away from one; it gets there after about $\log(1/u)/\log(1/q)$ steps. With an RN inner solver the same argument, based on the deterministic floor $\tilde{\kappa}^2 u$, needs $\tilde{\kappa}^2 u$ to be small. Near the storage limit the correction e is small and its components can become subnormal in binary16; the term $n\epsilon_s^2/4$ of Theorem 2 then enters the inner floor, and by Theorem 2 its contribution is at most $\tilde{\kappa}\sqrt{n}\,\epsilon_s/\big(2(1-\tilde{\kappa}u-\tilde{\kappa}^2u^2)^{1/2}\big)$ in norm, which is 1.8×10^{-4} for $n = 50$ and $\tilde{\kappa} = 640$, of the order of the storage limit. In our tests the SR version nevertheless reaches the storage limit for every $\tilde{\kappa} \le 640$. In Table 3 both inner solvers reach the storage limit for $\tilde{\kappa} \le 320$ in binary16. At $\tilde{\kappa} = 320$ the SR version needs three refinement steps and the RN version five, and the first step reduces the error to 1.4×10^{-2} for SR and 1.0×10^{-1} for RN, close to the floors 1.5×10^{-2} and 9.9×10^{-2} of the plain methods. At $\tilde{\kappa} = 640$, where $\tilde{\kappa}u = 0.31$, the SR version still reaches the storage limit in three steps, while the RN version contracts by only about 0.33 per step and after six steps is still 5.5 times above the limit.

9 Numerical experiments

Setup. Test matrices have Haar random singular vectors and the four singular value profiles of [5, Table 1]: exponential, polynomial, "highrank" with linear decay followed by a constant tail, and harmonic, with the free parameter chosen to give a prescribed $\tilde{\kappa}$. Unless stated otherwise $m = n = 50$, $x_\star$ is a random unit vector, $b = Ax_\star$, and A and b are stored in the target format with RN, $x_\star$ being the exact solution of the stored system. Rounding to binary16, bfloat16 and a format with $p = 20$ is simulated in double precision with NumPy, including subnormal numbers; RN agrees bit for bit with the IEEE binary16 conversion. Errors are measured against the solution of the stored system, so they exclude the effect of rounding A and b, which is the same for RN and SR. Each configuration uses eight matrices with eight independent chains each, the experiments with $n = 200$ and 500 four matrices with four chains each, and the refinement experiment eight matrices with four chains each. Every chain starts from $x_0 = 0$, discards $\tilde{\kappa}^2(\ln(1/u) + 4)$ steps and then records $8\tilde{\kappa}^2$ steps. We report the root-mean-square error of the iterates over this window and the error of their average. Consecutive iterates are correlated over about $\tilde{\kappa}^2$ steps, so uncertainties are given as 95% percentile intervals of a bootstrap with 4000 resamples that draws matrices with replacement and, within each drawn matrix, chains with replacement; RN and SR chains are drawn independently. Random number generators are seeded per matrix and per run, so every number in this section is reproducible from the code.

Floors. Figure 2 shows the floor of the iterates and of their tail average against $\tilde{\kappa}$. The SR floors follow the prediction (9) with the measured noise level in all three formats and at every $\tilde{\kappa}$; the median ratio of measurement to prediction lies between 0.97 and 1.02, its 95% bootstrap interval contains one in every configuration, and for single matrices the ratio lies between 0.91 and 1.07. Without the term n^2 the prediction would underestimate the floor by a factor of about 1.4 at $\tilde{\kappa} = 10$. For small $\tilde{\kappa}$ the RN iterates are slightly more accurate, by up to a factor two, because the SR error per step is bounded by $2u$ rather than u and keeps moving; the curves cross near $\tilde{\kappa} \approx 30$ in bfloat16 and $\tilde{\kappa} \approx 40$–$80$ in the other formats. Beyond the crossing the RN floor grows faster and spreads over a factor of up to three between matrices. At $\tilde{\kappa} = 640$ in binary16, where $\tilde{\kappa}^2 u = 200$, the median RN error is 32%, as large as the accuracy $\tilde{\kappa}u = 31\%$ of a backward stable direct solver, while the SR error is 2.8% and follows the prediction. The tail

averages separate earlier: for $\tilde{\kappa} \geq 20$ the SR average is at least as accurate as the RN average on every matrix, and the median advantage grows to 13 at $\tilde{\kappa} = 320$ and 37 at $\tilde{\kappa} = 640$ in binary16, with 95% bootstrap intervals from 9.9 to 23 and from 30 to 59, and to 32 at $\tilde{\kappa} = 640$ for $p = 20$. For the iterates the median advantage at $\tilde{\kappa} = 640$ in binary16 is 12, with interval 9.9 to 23. Table 2 repeats the comparison for the four singular value profiles and for $n = 200$ and 500. The prediction holds equally well for the larger matrices, and the advantage of SR averages over RN averages grows with $\tilde{\kappa}$ for every profile and size.

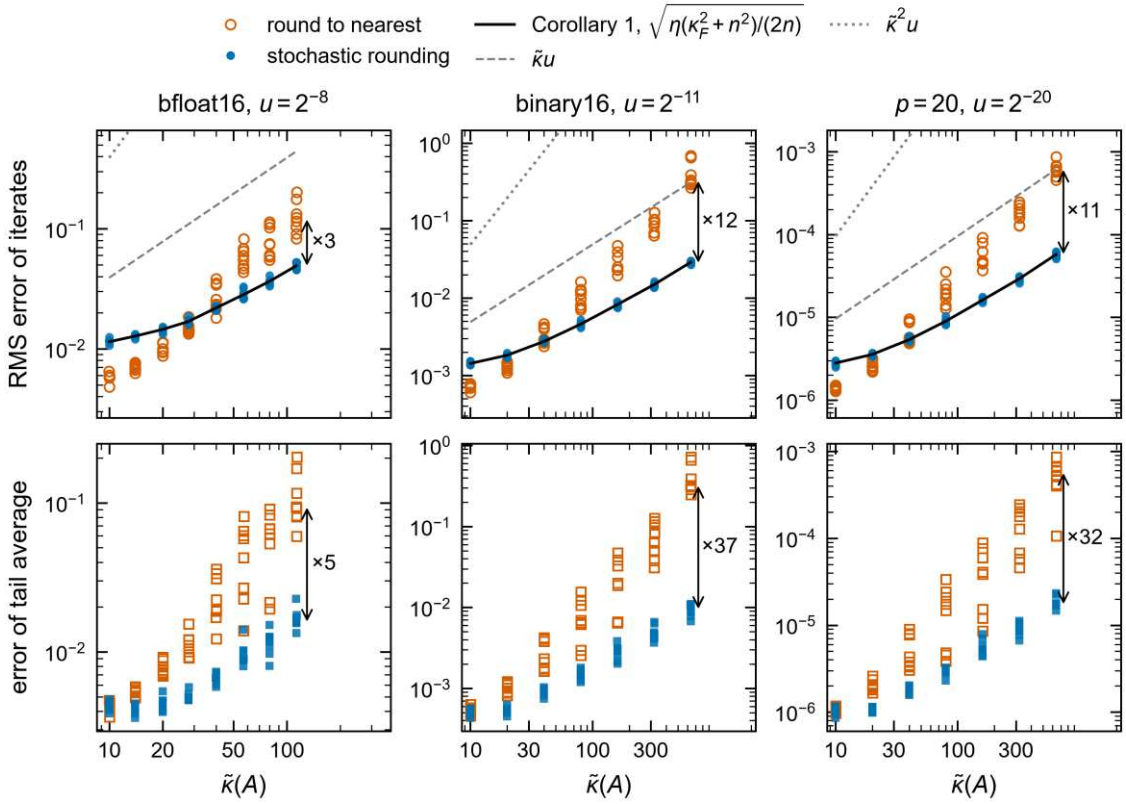


Figure 2. The SR floor follows the law (9) with the measured noise level, and SR averages are more accurate than RN averages for $\tilde{\kappa} \geq 20$. Errors are relative to $\|x_\star\|$. Top: root-mean-square error of the iterates; bottom: error of the tail average over $8\tilde{\kappa}^2$ steps; one marker per matrix, exponential profile, $n = 50$. Arrows give the median ratio of RN to SR errors at the largest $\tilde{\kappa}$; at $\tilde{\kappa} = 640$ in binary16 it is 12 for the iterates and 37 for the averages.

Table 2. Measured SR floor divided by the prediction (9), median over the matrices, and median ratio of RN to SR tail-average errors; binary16. For $n = 50$ there are eight matrices per entry and $\tilde{\kappa} = 20{,}80{,}320$; for $n = 200$ and 500 there are four matrices and $\tilde{\kappa} = 40{,}100{,}200$. All entries of the left block lie within 4% of one.

		floor / prediction			RN average / SR average		
profile	n	increasing $\tilde{\kappa}$			increasing $\tilde{\kappa}$		
exponential	50	0.999	1.008	0.967	1.8	5.2	13.4
polynomial	50	1.021	0.997	1.002	1.6	4.2	25.1
highrank	50	0.999	1.022	0.991	1.7	4.5	20.8
harmonic	50	1.004	0.988	0.981	1.7	5.0	13.3
exponential	200	1.000	1.001	1.005	2.0	3.8	8.1

polynomial	200	1.002	1.006	1.005	1.7	3.9	6.6
exponential	500	1.013	0.997	0.991	1.3	2.6	4.3
polynomial	500	0.996	1.004	0.997	1.3	2.6	4.6

Aligned singular vectors. To test the conditions behind (9), we repeat the binary16 experiment at $\tilde{\kappa} = 80$ and 320 with right singular vectors equal to the coordinate axes, once with a random unit vector $x_\star$ and once with $|x_{\star,j}|$ graded over two orders of magnitude, the large components lying on the smallest or on the largest singular values; Haar random singular vectors serve as control. For random $x_\star$, (9) overestimates the floor by median factors 1.7 and 3.2 at $\tilde{\kappa} = 80$ and 320: coordinates with small column norms receive short updates, and on the five coordinates with the largest weights s_j the median noise is 11% and 3% of the uniform model, against 50% and 32% over all coordinates. When the large components lie on the large singular values the factor grows to 3.7 and 10, and when they lie on the small singular values the measured floor is 1.36 and 1.09 times (9). In all cases the solution of (5) with the measured diagonal $\bar{D}$ matches the measured floor to within 6%. For the Haar control, (9) with the noise level measured on the first half of the window predicts the error on the second half with median ratios 1.01 and 0.98.

Single-precision updates. The analysis idealizes v_k as exact. Forming the residual, the step and v_k in single precision before rounding to binary16, for the matrices of Figure 2 at $\tilde{\kappa} = 320$ and 640 and with the same rows, changes the median over the matrices of the SR iterate error by at most 1% and that of the tail-average error by at most 3%; for single matrices the ratio lies between 0.93 and 1.05 for the iterates and between 0.75 and 1.26 for the tail average, whose value on one matrix is a median over eight chains. Accumulating the average in single precision once per sweep, as Algorithm 1 allows, has no visible effect, whereas accumulating it at every step increases its error by 30% at $\tilde{\kappa} = 320$ and by a factor 2.2 at $\tilde{\kappa} = 640$, because the running sum of up to $8\tilde{\kappa}^2$ iterates loses digits.

Averaging, random bits, inconsistency and rank deficiency. Figure 3 collects the remaining predictions in binary16. Panel (a) shows the decrease of tail averages with the window length at $\tilde{\kappa} = 80$: SR follows $T^{-1/2}$, RN levels off near 10^{-3}, and the gap at $T = 128\tilde{\kappa}^2$ is a factor 2.8. Panel (b) shows, for truncation toward zero, the ratio used for the bit thresholds of Section 6 against the number of random bits; the predictions of Theorem 5(a) are drawn solid where the predicted error is below 5% and dashed elsewhere. Rounding toward $-\infty$, as in the implementation $\lfloor v/\mathrm{ulp} + U \rfloor$ with U uniform on r-bit numbers, gives the same bit counts except at $\tilde{\kappa} = 320$, where it needs 12 bits. Panel (c) uses overdetermined systems with $m = 200$, $n = 50$, $\tilde{\kappa} = 80$ and residuals orthogonal to the range of A, and compares the error with low-precision storage with the error of the same iteration in double precision. Panel (d) shows the null-space drift of Proposition 1; there A and b are kept in double precision, because rounding A to binary16 would lift its zero singular values to values between 10^{-7} and 10^{-4}.

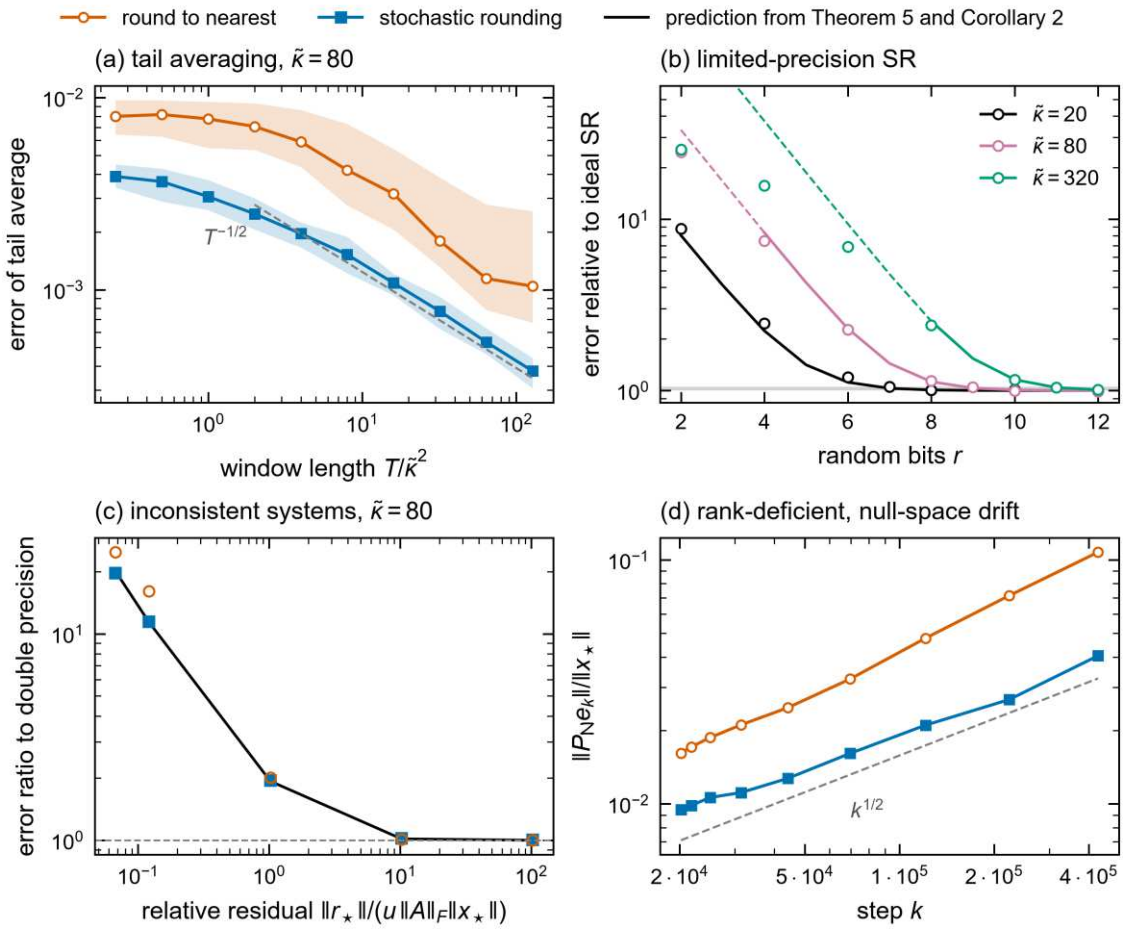


Figure 3. The measurements follow the predictions of Theorems 4 and 5, Corollary 2 and Proposition 1 in binary16; errors are relative to $\|x_\star\|$. (a) SR averages decrease like $T^{-1/2}$ while RN averages stop near 10^{-3}. (b) Error with r random bits divided by the error of ideal SR on the same matrix, median over matrices; lines are the predictions of Theorem 5(a), dashed where the linearization is not valid, and the grey band marks 5%. (c) Error relative to double precision against the relative residual; the line is Corollary 2(a) with (9) and the noise level measured in each run. (d) Null-space component for a matrix with five zero singular values. Shaded bands in (a) are interquartile ranges over 64 chains.

Refinement. Table 3 lists the error after each refinement step of Section 8, with $4\tilde{\kappa}^2$ inner steps per refinement.

Table 3. Iterative refinement with inner rounding RN or SR, binary16, median relative error after each refinement over eight matrices with four chains each; the storage limit $\|\mathrm{fl}(x_\star) - x_\star\|/\|x_\star\|$ is about 2×10^{-4}. The SR inner solver needs fewer refinements as $\tilde{\kappa}$ grows.

$\tilde{\kappa}$	inner	step 1	step 2	step 3	step 4	steps to limit
80	RN	1.1×10^{-2}	2.9×10^{-4}	2.2×10^{-4}	2.2×10^{-4}	3
80	SR	4.6×10^{-3}	2.2×10^{-4}	2.2×10^{-4}	2.2×10^{-4}	2
160	RN	2.8×10^{-2}	1.1×10^{-3}	2.1×10^{-4}	2.0×10^{-4}	3
160	SR	9.2×10^{-3}	2.2×10^{-4}	2.0×10^{-4}	2.0×10^{-4}	2
320	RN	1.0×10^{-1}	1.2×10^{-2}	1.4×10^{-3}	2.9×10^{-4}	5
320	SR	1.4×10^{-2}	3.4×10^{-4}	2.0×10^{-4}	2.0×10^{-4}	3

$\tilde{\kappa}$	inner	step 1	step 2	step 3	step 4	steps to limit
640	RN	3.2×10^{-1}	1.1×10^{-1}	3.4×10^{-2}	1.1×10^{-2}	> 6
640	SR	2.9×10^{-2}	8.3×10^{-4}	2.2×10^{-4}	2.2×10^{-4}	3

Computed tomography. We reconstruct a 32×32 Shepp–Logan phantom from 90 parallel-beam projections with 4050 rays, so $n = 1024$ and $\tilde{\kappa} = 307$. The data are perturbed by Gaussian noise whose standard deviation is 1% of the mean absolute value of Ax_{true}, with a fixed seed; A and b are kept in double precision and only the iterate is stored in low precision. Each of eight low-precision chains runs next to a double-precision chain with the same rows, so that Corollary 2(b) can be checked directly. After 40 sweeps, binary16 storage changes the root-mean-square error by 0.3% with SR and 0.1% with RN, and bfloat16 storage by 12% with SR and 5% with RN (Table 4). The relative residual is $\gamma = 9.6 \times 10^{-4}$, below the unit roundoff of bfloat16, so the stationary SR floor would exceed the error in double precision; two effects keep the cost small. After 40 sweeps, about $1.7\tilde{\kappa}^2$ steps, the slowest directions have accumulated little rounding noise, and the updates are so short that the measured noise level is only 4% of the uniform model for bfloat16. The cross term of Corollary 2(b) is below 0.3% of $\mathbb{E}\|\hat{x}_t - x_{\text{true}}\|^2$. The noise level changes by up to a third over the sweeps; driven by the measured diagonal noise covariance of every sweep, (10) predicts $\mathbb{E}\|\hat{x}_t - x'_t\|^2$ after 5, 10, 20 and 40 sweeps to within 13% for both formats, about twice the sampling error of the measurement, whereas, with the same noise input, isotropic noise overestimates it by up to 23% and replacing $\mathbb{E}[P_i X P_i]$ by SXS underestimates it by up to 34%. Added to the error in double precision, the prediction gives final errors 0.0479 and 0.0531 against the measured 0.0479 and 0.0537. The rounding mode matters little in this regime, and RN is more accurate than SR for bfloat16.

Table 4. Relative image error for the CT problem, root mean square over eight chains; the least-squares solution has error 0.0373. After 40 sweeps, binary16 storage changes the error by at most 0.3%, bfloat16 storage by 5% with RN and 12% with SR.

storage	sweep 5	sweep 40	minimum	tail average
double	0.062	0.048	0.047	0.037
binary16, RN	0.062	0.048	0.047	0.037
binary16, SR	0.062	0.048	0.047	0.037
bfloat16, RN	0.061	0.050	0.049	0.038
bfloat16, SR	0.066	0.054	0.052	0.038

10 Practical recommendations

Algorithm 1 collects the choices suggested by the analysis and the experiments.

Algorithm 1. Randomized Kaczmarz with stochastically rounded storage

Require: A, b, storage format with unit roundoff u, number r of random bits, burn-in t_b, window T, stride s

1. $\hat{x} \leftarrow 0$ in the storage format; $z \leftarrow 0$ and $c \leftarrow 0$ in working precision
2. **for** $k = 0,1,\dots,t_b + T - 1$ **do**
3. draw i with probability $\|a_i\|^2 / \|A\|_F^2$
4. $v \leftarrow \hat{x} + \big((b_i - a_i^T \hat{x})/\|a_i\|^2\big) a_i$ in working precision
5. $\hat{x} \leftarrow \mathrm{fl}(v)$ with stochastic rounding using r random bits

6. **if** $k \geq t_b$ and s divides $k - t_b$ **then** $z \leftarrow z + \hat{x}$, $c \leftarrow c + 1$
7. **end for**
8. $\bar{x} \leftarrow z/c$; if A is rank deficient, project $\hat{x}$ and $\bar{x}$ onto the range of A^T
9. **return** $\hat{x}$ and $\bar{x}$

Theorem 2 covers storage formats with $\tilde{\kappa}u < \frac{1}{2}$ and guarantees a floor of order $\tilde{\kappa}u\|x_\star\|$. For a consistent system, Lemma 1 gives $\eta \leq u^2\mathbb{E}\|v_k\|^2 + n\epsilon_s^2/4$, which is about $u^2\|x_\star\|^2$ near the solution, and (9) then gives the estimate $u\|x_\star\|((\kappa_F^2 + n^2)/(2n))^{1/2}$ of the floor, two to three times the measured floor in our random tests. For an inconsistent system the rounding term of Corollary 2 is to be compared with the error of the method in exact arithmetic, with the noise level measured on that system, since η grows with the residual; if it is much smaller, low-precision storage adds little error. During the run, η costs one subtraction per component and step. Turning it into an estimate of the forward error with (9) also needs κ_F, known or estimated, and the conditions of Section 4; for singular vectors aligned with the coordinate axes the estimate can be off by a factor ten, and the general formula (5) needs the noise variances by coordinate. With few random bits, $r \geq \log_2\tilde{\kappa} + 4$ reaches the floor of ideal SR within 5% in our tests (Section 6). The burn-in $t_b \approx \tilde{\kappa}^2\ln(1/u)$ makes the transient term of Theorem 4 comparable with the floor, and the window T is then chosen from the $T^{-1/2}$ rate. An estimate of $\tilde{\kappa}$ that is off by a factor of two changes r by one bit and t_b by a factor of four.

The average is kept in working precision, which would cost the memory traffic that low-precision storage saves if it were updated at every step. Consecutive iterates are correlated over about $\tilde{\kappa}^2$ steps, so the average may be accumulated only every s steps with $s \ll \tilde{\kappa}^2$. In a test with binary16, $n = 50$, $\tilde{\kappa} = 80$ and a window of $32\tilde{\kappa}^2$ steps, the median error of the average is 7.15×10^{-4} for $s = 1$, 7.14×10^{-4} for $s = 50$, which is one sweep over the rows, 7.16×10^{-4} for $s = 400$ and 7.54×10^{-4} for $s = 3200 = \tilde{\kappa}^2/2$. Accumulating once per sweep therefore adds n working-precision updates per m steps and, in this test, changed the error by less than 0.3%. With a single-precision accumulator the stride also keeps the running sum accurate, see Section 9.

Some accelerators generate the random bits of SR in hardware [9], while the conversion instructions of NVIDIA GPUs take them as an operand supplied by the program [23]; in software, one r-bit random integer per stored component and step is needed, which a counter-based generator produces from the step and component indices at a few integer operations per component and without memory traffic. When refinement is used, the SR iterate is the better inner solver (Section 8), and the outer iterate can be rounded to nearest.

11 Conclusions

With stochastic rounding, the rounding errors of randomized Kaczmarz act exactly as an additional residual of energy $\|A\|_F^2\eta$. The mechanism is that unbiased errors add in mean square, so the contraction of the iteration is paid once; this lowers the guaranteed floor from order $\tilde{\kappa}^2u$ to order $\tilde{\kappa}u$, lets tail averages converge like $T^{-1/2}$, and limits the amplification of the bias of r-bit rounding to $\tilde{\kappa}^2$, so that at most $\log_2\tilde{\kappa} + 4$ random bits were needed in our tests. The stationary second moment has an exact trace formula. For random singular vectors it is well approximated by $(\frac{\eta}{2n}(\kappa_F^2 + n^2))^{1/2}$ in terms of the measured noise level, which matched the median floor to within 4% in all our random tests; for singular vectors aligned with the coordinate axes the noise covariance has to be resolved by coordinate, and the solution of (5) with the measured covariance stayed accurate. Stochastic rounding is advantageous for ill-conditioned problems, for averaged iterates and as the inner solver of refinement, while for well-conditioned problems and for iterations stopped early, as in tomography, round-to-nearest is as good or slightly better. The analysis assumes that the update is formed in higher precision and only the stored iterate is rounded; updates formed in single precision changed the median SR errors by at most 3% in our tests, and rounding

inside low-precision inner products is not covered. The noise level η has to be measured, because it depends on the length of the updates and grows with the residual.

Declarations

Funding. Not applicable. **Competing interests.** The author declares no competing interests. **Data and code availability.** The code and data that support the findings of this study are available from the corresponding author on request.

References

[1] S. Kaczmarz, Angenäherte Auflösung von Systemen linearer Gleichungen, Bull. Int. Acad. Polon. Sci. Lett. A 35 (1937) 355–357.

[2] T. Strohmer, R. Vershynin, A randomized Kaczmarz algorithm with exponential convergence, J. Fourier Anal. Appl. 15 (2009) 262–278.

[3] D. Needell, Randomized Kaczmarz solver for noisy linear systems, BIT Numer. Math. 50 (2010) 395–403.

[4] E. N. Epperly, G. Goldshlager, R. J. Webber, Randomized Kaczmarz with tail averaging, Appl. Comput. Harmon. Anal. 80 (2026) 101812.

[5] M. Dereźiński, E. N. Epperly, D. Needell, A. Xue, Numerical instabilities in the Kaczmarz method and stabilization by iterative refinement, arXiv:2605.17185, 2026.

[6] M. P. Connolly, N. J. Higham, T. Mary, Stochastic rounding and its probabilistic backward error analysis, SIAM J. Sci. Comput. 43 (2021) A566–A585.

[7] M. Croci, M. Fasi, N. J. Higham, T. Mary, M. Mikaitis, Stochastic rounding: implementation, error analysis and applications, R. Soc. Open Sci. 9 (2022) 211631.

[8] E.-M. El Arar, M. Fasi, S.-I. Filip, M. Mikaitis, Probabilistic error analysis of limited-precision stochastic rounding, arXiv:2408.03069, 2024.

[9] E.-M. El Arar, M. Fasi, S.-I. Filip, M. Mikaitis, What is new in stochastic rounding: a survey on theory, hardware, and applications, arXiv:2603.06060, 2026.

[10] A. Fitzgibbon, S. Felix, On stochastic rounding with few random bits, arXiv:2504.20634, 2025.

[11] L. Xia, S. Massei, M. E. Hochstenbach, B. Koren, On stochastic roundoff errors in gradient descent with low-precision computation, J. Optim. Theory Appl. 200 (2024) 634–668.

[12] L. Xia, S. Massei, M. E. Hochstenbach, On the convergence of the gradient descent method with stochastic fixed-point rounding errors under the Polyak–Łojasiewicz inequality, Comput. Optim. Appl. 90 (2025) 753–799.

[13] N. J. Higham, S. Pranesh, Simulating low precision floating-point arithmetic, SIAM J. Sci. Comput. 41 (2019) C585–C602.

[14] N. J. Higham, P. A. Knight, Finite precision behavior of stationary iteration for solving singular systems, Linear Algebra Appl. 192 (1993) 165–186.

[15] T. Elfving, Noise propagation in linear stationary iterations, Numer. Algorithms 99 (2025) 603–625.

[16] E. Carson, N. J. Higham, Accelerating the solution of linear systems by iterative refinement in three precisions, SIAM J. Sci. Comput. 40 (2018) A817–A847.

[17] N. J. Higham, T. Mary, Mixed precision algorithms in numerical linear algebra, Acta Numer. 31 (2022) 347–414.

[18] R. D. Gitlin, J. E. Mazo, M. G. Taylor, On the design of gradient algorithms for digitally implemented adaptive filters, IEEE Trans. Circuit Theory 20 (1973) 125–136.

[19] C. Caraiscos, B. Liu, A roundoff error analysis of the LMS adaptive algorithm, IEEE Trans. Acoust. Speech Signal Process. 32 (1984) 34–41.

[20] S. T. Alexander, Transient weight misadjustment properties for the finite precision LMS algorithm, IEEE Trans. Acoust. Speech Signal Process. 35 (1987) 1250–1258.

[21] J. C. M. Bermudez, N. J. Bershad, A nonlinear analytical model for the quantized LMS algorithm: the arbitrary step size case, IEEE Trans. Signal Process. 44 (1996) 1175–1183.

[22] M. Ghanassi, B. Champagne, P. Kabal, On the steady-state mean squared error of the fixed-point LMS algorithm, Signal Process. 87 (2007) 3226–3233.

[23] NVIDIA Corporation, Parallel Thread Execution ISA, data movement and conversion instructions: cvt, https://docs.nvidia.com/cuda/parallel-thread-execution/, accessed 31 October 2026.

[24] R. A. Wannamaker, S. P. Lipshitz, J. Vanderkooy, J. N. Wright, A theory of nonsubtractive dither, IEEE Trans. Signal Process. 48 (2000) 499–516.

Appendix A. Proofs of Theorems 3, 4 and 5

Proof of Theorem 3. Squaring (2) as a matrix and conditioning on $\mathcal{G}_k$,

$$\Sigma_{k+1} = \mathbb{E}[(I-P_i)\Sigma_k(I-P_i)] + G + D_k + \sum_i p_i\left((I-P_i)m_k g_i^T + g_i m_k^T(I-P_i)\right),$$

since the terms containing δ_k once vanish by (S1). The last sum tends to zero with m_k. The Cesàro averages are bounded by Theorem 2; passing to a limit along a subsequence and using $\mathbb{E}[(I-P_i)X(I-P_i)] = X - SX - XS + \mathbb{E}[P_i X P_i]$ gives (5). Multiplying (5) by S^{-1} and taking traces, $\mathrm{tr}(S^{-1}(S\bar{\Sigma}+\bar{\Sigma}S)) = 2\mathrm{tr}\bar{\Sigma}$; moreover $P_i\bar{\Sigma}P_i = (a_i^T\bar{\Sigma}a_i/\|a_i\|^2)P_i$, $\mathrm{tr}(S^{-1}P_i) = \|A\|_F^2\ell_i/\|a_i\|^2$ and $p_i\|A\|_F^2 = \|a_i\|^2$, which gives (6). The last term in (6) is nonnegative. Since $\ell_i \le \mu(A)p_i$, it is at most $\frac{1}{2}\mu(A)\sum_i p_i\, a_i^T\bar{\Sigma}a_i/\|a_i\|^2 = \frac{1}{2}\mu(A)\mathrm{tr}(S\bar{\Sigma})$, and $\mathrm{tr}(S\bar{\Sigma}) = \mathrm{tr}(G+\bar{D})$ by Theorem 1. ▫

Proof of Theorem 4. Write $e_{k+1} = (I-S)e_k + \zeta_k$ with $\zeta_k = (S-P_i)e_k + g_i + \delta_k$. By (S1) and $\mathbb{E}g_i = 0$, $\mathbb{E}[\zeta_k \mid \mathcal{F}_k] = 0$. Hence $e_s = m_s + n_s$ with $m_s = (I-S)^s e_0$ and $n_s = \sum_{k<s}(I-S)^{s-1-k}\zeta_k$, and

$$\bar{x}_t - x_\star = \bar{m} + \bar{n}, \qquad \bar{n} = \frac{1}{T}\sum_{k=0}^{t-2} W_k\,\zeta_k, \qquad W_k = \sum_{s=\max(t_b,k+1)}^{t-1}(I-S)^{s-1-k}.$$

Since $\bar{m}$ is deterministic and $\mathbb{E}\bar{n} = 0$, $\mathbb{E}\|\bar{x}_t - x_\star\|^2 = \|\bar{m}\|^2 + \mathbb{E}\|\bar{n}\|^2$. From $0 \preccurlyeq I - S \preccurlyeq \rho_0 I$ and $\sum_{s\ge t_b}\rho_0^s = \tilde{\kappa}^2\rho_0^{t_b}$ we get $\|\bar{m}\| \le \rho_0^{t_b}\min\{1, \tilde{\kappa}^2/T\}\|e_0\|$, $\|W_k\| \le \tilde{\kappa}^2$ for $k \ge t_b - 1$, and $\|W_k\| \le \tilde{\kappa}^2\rho_0^{t_b-1-k}$ for $k < t_b - 1$. By orthogonality of martingale differences, $\mathbb{E}\|\bar{n}\|^2 = T^{-2}\sum_k \mathbb{E}\,\|W_k\zeta_k\|^2$. Using $P_i^2 = P_i$, $P_i g_i = g_i$, $\mathbb{E}g_i = 0$ and (S1), $\mathbb{E}\|\zeta_k\|^2 = \mathbb{E}[e_k^T S e_k] - \mathbb{E}[e_k^T S^2 e_k] + \nu_k$, and by (3), $\mathbb{E}[e_k^T S e_k] = a_k - a_{k+1} + \nu_k$; therefore $\mathbb{E}\|\zeta_k\|^2 \le a_k - a_{k+1} + 2\nu_k$. Also $a_{k+1} \le \rho_0 a_k + \nu^*$, so $a_k \le \rho_0^k a_0 + \tilde{\kappa}^2\nu^*$.

For the T indices $t_b - 1 \le k \le t-2$, telescoping gives $\sum\mathbb{E}\|W_k\zeta_k\|^2 \le \tilde{\kappa}^4(a_{t_b-1} + 2T\nu^*) \le \tilde{\kappa}^4(\rho_0^{t_b-1}a_0 + \tilde{\kappa}^2\nu^* + 2T\nu^*)$. For $k < t_b - 1$ let $w_k = \rho_0^{2(t_b-1-k)}$, which increases with k. Summation by parts gives

$$\sum_{k<t_b-1} w_k\,(a_k - a_{k+1}) \le w_0 a_0 + \sum_{k=1}^{t_b-2}(w_k - w_{k-1})a_k \le \rho_0^{2(t_b-1)}a_0 + \tilde{\kappa}^2\rho_0^{t_b}a_0 + \tilde{\kappa}^2\nu^*,$$

where we used $a_k \le \rho_0^k a_0 + \tilde{\kappa}^2 \nu^*$, $\sum_k (w_k - w_{k-1}) \le 1$ and $\sum_k w_k \rho_0^k \le \rho_0^{t_b}/(1-\rho_0)$. With $2\nu^* \sum_k w_k \le 2\tilde{\kappa}^2 \nu^*$ this part is at most $\tilde{\kappa}^4\big((\tilde{\kappa}^2+1)\rho_0^{t_b-2} a_0 + 3\tilde{\kappa}^2\nu^*\big)$. Adding both parts and dividing by T^2 proves the theorem. ▫

Proof of Theorem 5. (a) Take expectations in $e_{k+1} = (I - P_i)e_k + g_i + \beta_k + \tilde{\delta}_k$ and iterate, using $\|I - S\| \le \rho_0$ and $\sum_j \rho_0^j = \tilde{\kappa}^2$. (b) The cross terms with $\tilde{\delta}_k$ vanish, so $\mathbb{E}\|e_{k+1}\|^2 = \mathbb{E}\|y_k + \beta_k\|^2 + \mathbb{E}\big\|\tilde{\delta}_k\big\|^2$ with $y_k = (I - P_i)e_k + g_i$. Apply $\|y+\beta\|^2 \le (1+\epsilon)\|y\|^2 + (1+\epsilon^{-1})\|\beta\|^2$ with $\epsilon = \theta\tilde{\kappa}^{-2}$ and $\mathbb{E}\|y_k\|^2 \le \rho_0 \mathbb{E}\|e_k\|^2 + \|r_\star\|^2/\|A\|_F^2$, which follows from the proof of Theorem 1. Since $(1+\theta\tilde{\kappa}^{-2})(1-\tilde{\kappa}^{-2}) \le 1 - (1-\theta)\tilde{\kappa}^{-2}$, summing the geometric series proves (b). ▫