Hadamard Flattening and Gaussian Pooling Sketch for Least Squares with Coordinate-wise Guarantee

Zhao Song* Lichen Zhang†

Aug 7, 2026

Abstract

Randomized sketch-and-solve algorithms accelerate overconstrained ℓ_2 regression by replacing the input with a much smaller, randomly projected problem. Standard subspace embeddings guarantee that the cost of the regression is nearly preserved, but coordinate-wise accuracy of the solution is more delicate: instead of preserving the objective value, we want the solution vector itself to be close to the optimal solution in ℓ_∞ norm. In particular, we want to find a vector $x' \in \mathbb{R}^d$ such that $\|x' - x^\star\|_\infty \leq \frac{\epsilon}{\sqrt{d}} \cdot \|Ax^\star - b\|_2 \cdot \|A^\dagger\|_{\mathrm{op}}$, where $A \in \mathbb{R}^{n \times d}$ is the design matrix, $b \in \mathbb{R}^n$ is the label vector, $x^\star \in \mathbb{R}^d$ is the optimal solution and $A^\dagger$ is the pseudo-inverse of A. Price, Song and Woodruff initiated the study of this problem and showed that the subsampled randomized Hadamard transform (SRHT) with $O(\epsilon^{-2} d^{1+\Theta(\sqrt{\log\log n / \log d})})$ rows achieves this guarantee. A subsequent work of Song, Ye, Yin and Zhang claimed to improve the row count to $O(\epsilon^{-2} d \log^3 n)$. Unfortunately, their proof relies on an independence assumption that does not hold in general, and we exhibit an explicit instance on which it fails.

To achieve a truly nearly-linear-in-d row count, we introduce a new fast, dense randomized transform, which combines a randomized Hadamard flattening, a random permutation, and balanced, disjoint Gaussian pooling. Conditioned on the Hadamard-and-permutation stage, the sketched problem becomes an exact Gaussian regression in which the noise is independent of the entire sketched design; this conditional independence is exactly what the earlier argument was missing. Our sketch yields the ℓ_∞ guarantee with $m = O(\epsilon^{-2} d \log d)$ rows, uses one Hadamard pass with a padded internal dimension $N = \widetilde{O}(n + \epsilon^{-2} d^3)$, and is efficient to apply: the sketched pair (SA, Sb) can be computed in $O(Nd \log N) = \widetilde{O}(nd + \epsilon^{-2} d^4)$ time.

1 Introduction

Least-squares regression is a workhorse of scientific computing, numerical linear algebra, and modern machine learning pipelines. In the overconstrained setting, a data matrix $A \in \mathbb{R}^{n \times d}$ with $n \gg d$ and a label vector $b \in \mathbb{R}^n$ are given, and one wishes to (approximately) minimize $\|Ax - b\|_2$ over $x \in \mathbb{R}^d$. A natural way to speed this up is to compress the problem before solving it: the sketch-and-solve paradigm draws an oblivious random matrix $S \in \mathbb{R}^{m \times n}$ with $m \ll n$, forms SA and Sb, and solves the m-row problem $\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_2$ instead. Randomized compression of this kind traces back to the Johnson–Lindenstrauss lemma [JL84]: any N points in Euclidean space can be mapped into $O(\epsilon^{-2} \log N)$ dimensions so that every pairwise distance is preserved up to a $1 \pm \epsilon$ factor, and Larsen and Nelson [LN17] proved that no embedding, linear or not, can do better in the worst case. Regression asks for more than pairwise distances, and the right tool is an oblivious

*magic.linuxkde@gmail.com.

†lichenz@mit.edu. MIT.

subspace embedding [Sar06, DMMS11]: a random matrix that simultaneously preserves the norm of every vector in a fixed low-dimensional subspace, which suffices for relative-error guarantees on the objective. Such embeddings admit fast implementations of two kinds: dense ones built from randomized Hadamard transforms, and sparse ones built from hashing, including CountSketch, which originated in streaming frequency estimation [CCFC02] and later enabled regression in input-sparsity time [CW17], and OSNAP [NN13], which refined the tradeoff between sparsity and embedding dimension.

Guarantees of this type, however, primarily control the residual or a global norm of the solution error. When S is an oblivious subspace embedding with distortion ϵ, the minimizer x' of the sketched problem satisfies $\|x' - x^\star\|_2 \le O(\epsilon) \cdot \|Ax^\star - b\|_2 \cdot \|A^\dagger\|_{\mathrm{op}}$, where $x^\star$ denotes the optimal solution. For a direction $a \in \mathbb{R}^d$ fixed independently of the sketch, Cauchy–Schwarz then gives

$$|\langle a, x' - x^\star\rangle| \le O(\epsilon) \cdot \|a\|_2 \cdot \|Ax^\star - b\|_2 \cdot \|A^\dagger\|_{\mathrm{op}}.$$

This estimate, however, leaves a factor of $\sqrt{d}$ on the table: for a generic direction a, the deviation should scale like a $1/\sqrt{d}$ fraction of the ℓ_2 error, and a bound of this refined form, applied simultaneously to $a = e_1, \ldots, e_d$, yields an ℓ_∞ guarantee on $x' - x^\star$. Price, Song and Woodruff [PSW17] confirmed this prediction for the subsampled randomized Hadamard transform (SRHT), albeit with a row count that is superlinear in d. [SYYZ23] later claimed a nearly-linear-in-d row count by combining a constant-distortion oblivious subspace embedding with an oblivious coordinate-wise embedding for a fixed pair of vectors. Unfortunately, their argument applies the fixed-vector guarantee to a direction that contains the inverse of the random sketched Gram matrix and therefore depends on the same sketch. The stated guarantee does not cover this adaptive choice, and as we show in Section 2, the gap is genuine rather than merely formal.

In this work, we introduce a new sketch, which we call the *balanced Gaussian-pooled transform*, and prove that it achieves the ℓ_∞ guarantee with a truly nearly-linear-in-d row count. Since this sketch is the central object of the paper, let us first state its ingredients explicitly. The transform factors as

$$S := BD_\gamma \Pi F_N D_\sigma J,$$

applied from right to left: J pads the input from dimension n to a larger internal dimension N by appending zeros, D_σ is a diagonal matrix of random signs, F_N is one normalized Hadamard transform, Π is a uniformly random permutation, D_γ is a diagonal matrix of independent standard Gaussians, and B partitions the N coordinates into m consecutive blocks of equal size and sums each block. We call the block-summing step *Gaussian pooling*, since each row of the sketch pools an entire block of Gaussian-weighted coordinates; the blocks are *balanced* in the sense that, with high probability, the Hadamard transform and the random permutation spread the subspace relevant to the regression so evenly that every block carries a nearly equal share of its energy.

The intuition behind the construction is to expose the randomness in two stages. The first stage, the randomized Hadamard transform followed by the random permutation, flattens and shuffles the (padded) column space of A together with the residual direction, so that every block becomes a nearly isotropic copy of this fixed $(d+1)$-dimensional subspace. Conditional on the first stage, the second stage takes over: because the pooling blocks are disjoint, the rows of the sketched problem are independent Gaussian vectors, and the sketched least squares becomes an exact Gaussian regression whose noise is independent of the entire design. The inverse sketched Gram matrix may then depend on the design in an arbitrary fashion, and no fixed-vector guarantee is ever applied to a sketch-dependent direction. With probability at least $1 - \delta$, the resulting estimator satisfies the fixed-direction guarantee with $m = O(\epsilon^{-2} d \log(1/\delta))$ rows, and the simultaneous guarantee over

all d coordinates with $m = O(\epsilon^{-2} d \log(d/\delta))$ rows. The transform uses a single Hadamard pass, but we emphasize that it is not the canonical SRHT: the analysis works with a padded internal dimension $N = \widetilde{O}(n + \epsilon^{-2} d^3)$.

1.1 Main Result

Let $A \in \mathbb{R}^{n \times d}$ have full column rank, let $b \in \mathbb{R}^n$, and write $x^\star := \arg\min_x \|Ax - b\|_2$, $\widehat{x} := \arg\min_x \|SAx - Sb\|_2$. For a fixed direction $a \in \mathbb{R}^d$, the target is

$$|a^\top(\widehat{x} - x^\star)| \le \frac{\epsilon}{\sqrt{d}} \|a\|_2 \, \|Ax^\star - b\|_2 \, \|A^\dagger\|_{\mathrm{op}}. \tag{1}$$

Taking $a = e_j$ and union bounding over $j \in [d]$ gives the corresponding ℓ_∞ estimate.

Theorem 1.1 of [SYYZ23] claims Eq. (1) for the canonical SRHT with $m = O(\epsilon^{-2} d \log^3(n/\delta))$ rows. As we explain in Section 2, the proof of that claim is incomplete, and the issue cannot be repaired within the same argument. What we prove instead is the following.

Theorem 1.1 (Main result)**.** *Fix $0 < \epsilon \le 1$ and $0 < \delta < 1/2$. There is a distribution over oblivious linear maps $S \in \mathbb{R}^{m \times n}$, each of which can be applied with a single randomized Hadamard transform, such that for every fixed A, b, a, a single draw of S satisfies Eq. (1) with probability at least $1 - \delta$ when $m = O(\epsilon^{-2} d \log(1/\delta))$. The guarantee holds for all d coordinates simultaneously using $m = O(\epsilon^{-2} d \log(d/\delta))$ rows, with the failure parameter δ/d used throughout the construction. The map is the balanced Gaussian-pooled transform defined in Definition 4.1. Its padded internal dimension N and total classical runtime T are*

$$N = \widetilde{O}(n + \epsilon^{-2} d^3), \qquad T = O(Nd \log N + \mathcal{T}_{\mathrm{mat}}(d, m, d) + d^\omega) = \widetilde{O}(nd + \epsilon^{-2} d^4).$$

Here $\mathcal{T}_{\mathrm{mat}}(a, b, c)$ denotes the arithmetic time required to multiply an $a \times b$ matrix by a $b \times c$ matrix, and $\widetilde{O}$ suppresses logarithms in $n, d, 1/\epsilon, 1/\delta$.

Remark 1.2. *For the canonical one-shot SRHT, Theorem 10 of Price, Song, and Woodruff [PSW17] is the rigorous dependency-safe baseline. Under their assumptions it gives the same fixed-direction regression conclusion with*

$$m = O(\epsilon^{-2} d^{1+\Theta(\sqrt{\log\log n / \log d})})$$

and polynomially small failure probability. Its Neumann-expansion proof handles reuse of the same sketch, but the row count is $d^{1+o(1)}$, not $d \,\mathrm{polylog}(n)$. Obtaining the latter for the canonical SRHT without the padded block construction would require an argument beyond this paper, for example, a suitable anisotropic inverse-Gram or fluctuation-averaging estimate; we leave closing this gap as an open problem.

Roadmap. In Section 2, we describe the dependency issue in the argument of [SYYZ23] and construct an explicit instance showing that the gap is genuine. In Section 3, we show that the natural fix of replacing the Rademacher diagonal in the SRHT with a Gaussian diagonal fails as well. In Section 4, we define the balanced Gaussian-pooled transform and explain the conditional-regression viewpoint that repairs the argument. In Section 5, we show that, conditional on the Hadamard-and-permutation stage, the pooled rows are independent Gaussian vectors with explicit covariances. In Sections 6 and 7, we develop the block geometry supplied by the randomized Hadamard transform and the exact conditional Gaussian regression representation. In Section 8, we prove concentration for the conditional design and bias, and record that our transform is an oblivious subspace embedding. In Sections 9 and 10, we prove the corrected core lemma and transfer it to least squares. Finally, in Section 11, we bound the running time of our sketch.

Acknowledgement & AI Disclosure

The sketch construction is developed by GPT Codex 5.6 Sol and Claude Code Fable 5. Proofs are generated part by GPT Codex 5.6 Sol and authors, and are verified independently by Claude Code Fable 5 and authors. The authors take full responsibility for verifying all claims and for the paper's content.

2 The original dependency bug

Take a thin singular value decomposition $A = U\Sigma V^\top$ and define the least-squares residual $r := b - Ax^\star$. The normal equations give $U^\top r = 0$. Whenever SU has full column rank,

$$\widehat{x} - x^\star = V\Sigma^{-1}(U^\top S^\top SU)^{-1}U^\top S^\top Sr.$$

Left-multiplying the preceding identity by $a^\top$ and using the notation defined below gives

$$a^\top(\widehat{x} - x^\star) = c^\top G^{-1}h, \tag{2}$$

where

$$c := \Sigma^{-1}V^\top a, \qquad G := U^\top S^\top SU, \qquad h := U^\top S^\top Sr.$$

Define $u(S) := UG^{-1}c = U(U^\top S^\top SU)^{-1}c$. Since G is symmetric, $c^\top G^{-1}h = u(S)^\top S^\top Sr$, and since $U^\top r = 0$, we have $u(S)^\top r = 0$. Hence Eq. (2) can be written in the exact OCE form $c^\top G^{-1}h = u(S)^\top S^\top Sr = u(S)^\top(S^\top S - I)r$.

Definition 2.1 (Oblivious coordinate-wise embedding, [SYYZ23])**.** *Fix $\beta > 0$, $0 < \delta < 1$, and positive integers m, n. A distribution $\mathcal{D}$ over matrices $S \in \mathbb{R}^{m\times n}$ is a (β, δ, n)-oblivious coordinate-wise embedding (OCE) if, for every fixed pair $g, h \in \mathbb{R}^n$ chosen independently of S, a draw $S \sim \mathcal{D}$ satisfies*

$$|g^\top(S^\top S - I_n)h| \le \frac{\beta}{\sqrt{m}}\|g\|_2\|h\|_2$$

with probability at least $1 - \delta$.

The original proof of [SYYZ23] applies Definition 2.1 to the pair $(u(S), r)$. This is not allowed because $u(S)$ depends on the same random sketch S. The definition has the quantifiers

$$\text{for every fixed } g, h, \qquad \Pr_S[\text{failure for } g, h] \le \delta,$$

not

$$\Pr_S[\text{failure for an adaptively chosen } u(S), r] \le \delta.$$

Conditioning first on the event that S embeds the column space of A (Definition 8.2) does not help: it changes the law of S, while $u(S)$ still depends on the remaining randomness. A norm bound on $u(S)$ does not make it fixed, independent, or covered by a uniform OCE statement.

2.1 The gap is genuine, not merely formal

The following compact example shows that OSE plus fixed-vector OCE does not imply the adaptive inverse-Gram conclusion.

Proposition 2.2 (OSE and fixed-vector OCE are insufficient)**.** *For every sufficiently large d, there is a distribution $\mathcal{D}_d$ on $S \in \mathbb{R}^{(d+1)\times(d+1)}$ such that all singular values of S lie in $[0.9, 1.1]$ deterministically and, for every fixed $g, h \in \mathbb{R}^{d+1}$ and every $0 < \delta < 1$,*

$$\Pr_{S\sim\mathcal{D}_d}[|g^\top(S^\top S - I_{d+1})h| > C\sqrt{\frac{\log(2/\delta)}{d}}\,\|g\|_2\|h\|_2] \le \delta.$$

Equivalently, after changing C by an absolute factor, $\mathcal{D}_d$ is a $(C\sqrt{\log(2/\delta)}, \delta, d+1)$-OCE according to Definition 2.1. Nevertheless, a fixed one-dimensional regression contrast is bounded below by an absolute constant with probability one.

Proof. Let $U := [e_1, \dots, e_d]$, draw v uniformly from the unit sphere in $\mathrm{span}(e_2, \dots, e_d)$, and fix $\rho := 1/20$ and $\alpha := 1/20$. Define

$$M(v) := \begin{bmatrix} B(v) & \alpha v \\ \alpha v^\top & 1 \end{bmatrix}, \qquad B(v) := I_d - \rho(e_1 v^\top + v e_1^\top), \qquad S := M(v)^{1/2}.$$

Let $\mathcal{D}_d$ denote the resulting distribution of S. On $\mathrm{span}(e_1, v, e_{d+1})$, the perturbation $M - I$ has eigenvalues 0 and $\pm\sqrt{\rho^2+\alpha^2}$; it vanishes on the orthogonal complement. Since $\sqrt{\rho^2+\alpha^2} < 1/10$, the matrix M is positive definite. Moreover, $S^\top S = M$, so the singular values of S are

$$1, \qquad \sqrt{1-\sqrt{\rho^2+\alpha^2}}, \qquad \sqrt{1+\sqrt{\rho^2+\alpha^2}},$$

together with additional copies of 1. They all lie in $[0.9, 1.1]$.

Let $\mathcal{H} := \mathrm{span}(e_2, \dots, e_d)$ and let $P_{\mathcal{H}}$ be the orthogonal projector onto $\mathcal{H}$. For fixed $g, h \in \mathbb{R}^{d+1}$, direct expansion gives

$$g^\top(M - I)h = v^\top w,$$

where the fixed vector

$$w := P_{\mathcal{H}}[-\rho(g_1 h_{1:d} + h_1 g_{1:d}) + \alpha(h_{d+1} g_{1:d} + g_{d+1} h_{1:d})]$$

satisfies

$$\|w\|_2 \le 2(\rho + \alpha)\|g\|_2\|h\|_2.$$

Consequently, spherical concentration [DG03, Lemma 2.2] gives a universal constant $C > 0$ such that, for every $0 < \delta < 1$,

$$\Pr[|g^\top(S^\top S - I_{d+1})h| > C\sqrt{\frac{\log(2/\delta)}{d}}\,\|g\|_2\|h\|_2] \le \delta.$$

Since the sketch has $m = d + 1$ rows, increasing C by an absolute factor shows that $\mathcal{D}_d$ is a $(C\sqrt{\log(2/\delta)}, \delta, d+1)$-OCE according to Definition 2.1.

Finally, take $A := U$, $b := e_{d+1}$, $r := e_{d+1}$, and $a := e_1$. Then $U^\top r = 0$, so the least-squares minimizer is $x^\star = 0$. Since $S^\top S = M$,

$$(SA)^\dagger Sb = (U^\top M U)^{-1} U^\top M r = B(v)^{-1}(\alpha v).$$

On the ordered basis (e_1, v) of $\mathrm{span}(e_1, v)$,

$$B(v) = \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}, \qquad B(v)^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Consequently,

$$|e_1^\top (SA)^\dagger Sb| = \frac{\alpha\rho}{1-\rho^2}.$$

This is independent of d, whereas the claimed fixed-OCE conclusion would tend to zero as $d \to \infty$. □

3 Why a Gaussian diagonal in the usual position also fails

One might replace the Rademacher diagonal in $m^{-1/2}PHD$ by a Gaussian diagonal. This does not repair the argument. Let H be an unnormalized Hadamard matrix, let P sample m rows independently and uniformly, and put

$$S := m^{-1/2}PHD_\gamma, \qquad K := m^{-1}H^\top P^\top PH.$$

Here $D_\gamma := \mathrm{diag}(\gamma_1, \ldots)$, with $\gamma_j \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, independently of P. For two distinct Hadamard columns, $K_{11} = K_{22} = 1$ and

$$\rho := K_{12} = \frac{1}{m}\sum_{k=1}^{m} \zeta_k$$

for independent Rademachers ζ_k. With $u := \frac{e_1+e_2}{\sqrt{2}}, r := \frac{e_1-e_2}{\sqrt{2}}$, we have $u^\top r = 0$, but the one-dimensional sketched regression coefficient is

$$\frac{u^\top S^\top Sr}{u^\top S^\top Su} = \frac{\gamma_1^2 - \gamma_2^2}{\gamma_1^2 + \gamma_2^2 + 2\rho\gamma_1\gamma_2}. \tag{3}$$

Hoeffding gives $\Pr[|\rho| > 1/2] \le 2e^{-m/8}$. Independently, the ratio $R_\gamma := \gamma_1/\gamma_2$ is standard Cauchy. The two disjoint events $|R_\gamma| \ge 2$ and $|R_\gamma| \le 1/2$ therefore give

$$\Pr[\max(\gamma_1^2, \gamma_2^2) \ge 4\min(\gamma_1^2, \gamma_2^2)] = \Pr[|R_\gamma| \ge 2] + \Pr[|R_\gamma| \le 1/2] = 2 - \frac{4}{\pi}\arctan 2.$$

To verify the claimed constant, put $a := \max(\gamma_1^2, \gamma_2^2)$ and $b := \min(\gamma_1^2, \gamma_2^2)$. On the imbalance event, $a \ge 4b$. If also $|\rho| \le 1/2$, then

$$|\gamma_1^2 - \gamma_2^2| = a - b \ge \frac{3a}{4},$$

whereas

$$|\gamma_1^2 + \gamma_2^2 + 2\rho\gamma_1\gamma_2| \le a + b + \sqrt{ab} \le \frac{7a}{4}.$$

Thus the absolute value in Eq. (3) is at least $3/7$. The two events are independent, so this happens with probability at least

$$(2 - \frac{4}{\pi}\arctan 2)(1 - 2e^{-m/8})$$

for arbitrarily large m. Our construction instead places the Gaussian weights *after* norm flattening; disjoint pooling supplies the independence that drives our analysis.

4 The balanced Gaussian-pooled transform

We begin with the formal definition of our sketch; the remainder of this section explains the ideas behind its analysis.

Definition 4.1 (Balanced Gaussian-pooled sketch)**.** *Round the desired output row count m up to a power of two, fix $\kappa \in (0, 1/2)$, and define $\Lambda := \log(Cnmd/(\kappa\delta))$. Choose s to be the smallest power of two satisfying*

$$s \ge \max\{\lceil \frac{n}{m} \rceil, C\kappa^{-2}(d+1)\Lambda^2\}, \qquad N := ms.$$

Then N is a power of two and $N \ge n$. Let $J : \mathbb{R}^n \to \mathbb{R}^N$ append zeros and let $F_N := H_N/\sqrt{N}$ be the normalized orthogonal Hadamard matrix.

Draw independently:

- *a Rademacher diagonal* $D_\sigma \in \mathbb{R}^{N\times N}$*;*
- *a uniform permutation matrix* $\Pi \in \mathbb{R}^{N\times N}$*;*
- *a Gaussian diagonal* $D_\gamma := \mathrm{diag}(\gamma_1, \ldots, \gamma_N)$*, with* $\gamma_j \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$*.*

Partition $[N]$ *into* m *consecutive blocks* $\mathcal{B}_1, \ldots, \mathcal{B}_m$*, each of size* s*. Let* $B \in \{0,1\}^{m\times N}$ *sum the coordinates in each block, i.e.,* $B_{i,j} := 1$ *if* $j \in \mathcal{B}_i$ *and* 0 *otherwise. Finally, define the balanced Gaussian-pooled sketch* $S \in \mathbb{R}^{m\times n}$ *by*

$$\boxed{S := BD_\gamma \Pi F_N D_\sigma J.}$$

The regression solver sees only SA *and* Sb*; there is no preliminary regression and no post-compression of another regression solution.*

The transform is dense in the original coordinates almost surely, but it is applied in factored form: pad, flip signs, take one Hadamard transform, permute, multiply by Gaussians, and sum blocks.

As we saw in Sections 2 and 3, the main obstacle is not subspace preservation alone, but the dependence created by the inverse sketched Gram matrix. Write $A = U\Sigma V^\top$, let $r := b - Ax^\star$, and set $X := SU$, $y := Sr$, $M := X^\top X$, and $c := \Sigma^{-1}V^\top a$. Whenever M is invertible, the normal equations give $a^\top(\widehat{x} - x^\star) = c^\top M^{-1} X^\top y$. A fixed-vector OCE bound cannot be applied with the direction $UM^{-1}c$, because this direction depends on the same sketch; this is precisely the failure mode described in Section 2. Our proof instead exposes the randomness in stages so that M^{-1} may depend on the design while, after fixing the first-stage randomness, the remaining noise is independent of the design.

Stage 1: balanced block geometry. The sketch in Definition 4.1 factors as $S = BD_\gamma QJ$, where $Q := \Pi F_N D_\sigma$. We first expose Q. For $r \neq 0$, Lemma 6.1 applies Q to the padded augmented orthonormal system $[JU, Jr/\|r\|_2]$ and shows that every balanced block is nearly isotropic simultaneously; the case $r = 0$ is handled separately in the core lemma. Claim 6.2 translates this event into a nearly uniform block design covariance C_i, a small design–residual covariance h_i, and controlled residual variance v_i. Orthogonality and $U^\top r = 0$ also give the exact identities $\sum_{i=1}^m C_i = I_d$ and $\sum_{i=1}^m h_i = 0$ in Claim 6.3.

Stage 2: exact conditional Gaussian regression. Fix a good realization of Q and then expose the Gaussian diagonal D_γ. Because the pooling blocks are disjoint, Claim 5.1 gives independent, centered, jointly Gaussian row pairs (x_i, y_i) across i, together with their exact conditional covariances. Definition 7.1 assembles these rows into X, y, and M. Lemma 7.2 residualizes each pair as $y_i = x_i^\top \theta_i + \xi_i$ and proves that, conditional on Q, the entire noise vector ξ is independent of the entire design X. Its cancellation and score-decomposition parts give $X^\top y = Z + X^\top \xi$, where $Z := \sum_{i=1}^m (x_i x_i^\top - C_i)\theta_i$. Thus Z is a conditionally centered, design-dependent fluctuation, whereas $X^\top \xi$ is driven by a Gaussian noise vector ξ that is independent of X conditional on Q. This exact separation is the key repair.

Stage 3: concentration and assembly. Lemma 8.1 shows, uniformly for every good Q, that M is well conditioned and Z is small. We then condition further on (Q, X): the quantities M^{-1} and Z are fixed, while $c^\top M^{-1} X^\top \xi$ is a centered Gaussian with controlled variance. The choice $\kappa = c_0/\sqrt{d}$ supplies the $d^{-1/2}$ factor for the bias, while the row count controls the Gaussian noise. Lemma 9.1

combines the block, design-and-bias, and noise events by the tower property, and Theorem 10.1 transfers the resulting core estimate to least squares and obtains the simultaneous coordinate bound by a union bound. Figure 1 records the order of conditioning and the associated failure budget.

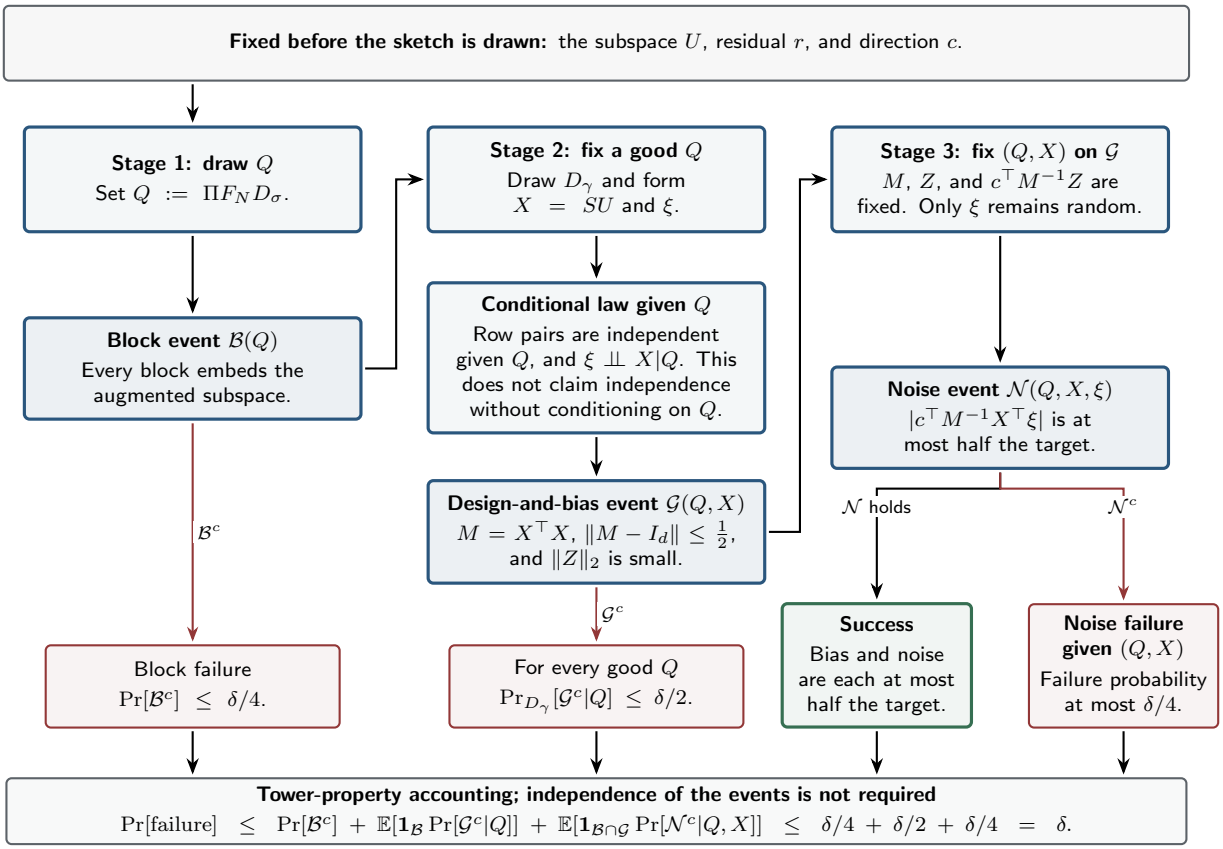

Figure 1: Conditioning hierarchy in the proof. After Q is fixed, D_γ generates (X, ξ), with $\xi \perp\!\!\!\perp X$ conditional on Q. The design-and-bias event is measurable with respect to (Q, X); after conditioning on (Q, X), only ξ remains random for the noise bound. The failure probabilities are combined by the tower property, not by independence of the events.

5 Conditional Gaussian pooling

The first property we extract from the construction is distributional: conditional on the Hadamard-and-permutation stage, the pooled rows are independent Gaussian vectors with explicit covariances.

Claim 5.1 (Conditional Gaussian pooling and exact block covariances)**.** *Put $Q := \Pi F_N D_\sigma$. Conditional on Q, for every $p \ge 1$ and fixed $T \in \mathbb{R}^{N\times p}$, define*

$$Y_T := BD_\gamma QT, \qquad \bar{T}_i := (QT)_{\mathcal{B}_i,*} \in \mathbb{R}^{s\times p}.$$

Conditional on Q, the rows of Y_T are mutually independent centered Gaussian vectors, and, for every $i \in [m]$,

$$\mathrm{Cov}_\gamma((Y_T)_{i,*}^\top | Q) = \bar{T}_i^\top \bar{T}_i.$$

Moreover,

$$\mathbb{E}_\gamma[Y_T^\top Y_T|Q] = T^\top T.$$

In particular, taking $T = J$ gives $Y_J = S$; hence, conditional on Q, the rows of S are mutually independent centered Gaussian vectors and $\mathbb{E}_\gamma[S^\top S|Q] = I_n$.

Proof. For each $i \in [m]$, put $g_i := (\gamma_j)_{j\in\mathcal{B}_i}$. Since B sums the coordinates in each block, $(Y_T)_{i,*}^\top = \bar{T}_i^\top g_i$. Conditional on Q, the vectors g_i are mutually independent standard Gaussian vectors because the blocks $\mathcal{B}_1, \dots, \mathcal{B}_m$ are disjoint. This proves the asserted conditional Gaussian law, independence, and covariance formula. Since the blocks partition $[N]$ and Q is orthogonal,

$$\mathbb{E}_\gamma[Y_T^\top Y_T|Q] = \sum_{i=1}^m \bar{T}_i^\top \bar{T}_i = T^\top Q^\top Q T = T^\top T.$$

Taking $T = J$ and using $J^\top J = I_n$ proves the final assertion. □

The isotropy identity alone is not a concentration statement; the conditional row independence is the stronger property used in the regression decomposition below.

6 Block geometry supplied by the randomized Hadamard transform

The role of the Hadamard stage is to make every block a very accurate embedding of the one fixed augmented subspace relevant to the regression. The parameter choice in Definition 4.1 implies Eq. (4), after increasing the universal constant C.

Lemma 6.1 (Simultaneous block embedding)**.** Let $0 < \delta < 1/2$, and let $W \in \mathbb{R}^{N\times p}$ be fixed independently of (D_σ, Π) and have orthonormal columns, with $1 \le p \le d+1$. Put $Q := \Pi F_N D_\sigma$. Let $W_i := (QW)_{\mathcal{B}_i,*}$. If

$$s \ge C\kappa^{-2} p \log^2(Npm/\delta), \tag{4}$$

then, with probability at least $1-\delta$ over (D_σ, Π), the event

$$\mathcal{E}_{\text{block}}(W) := \{\|mW_i^\top W_i - I_p\| \le \kappa \text{ for every } i \in [m]\}$$

occurs. Moreover, if $n \ge d$ and s, N, Λ are chosen as in Definition 4.1, then Eq. (4) holds after increasing the universal constant C, and

$$N = O(n + m\kappa^{-2}(d+1)\Lambda^2).$$

Proof. Put $Z_0 := F_N D_\sigma W$ and let $z_j := (Z_0)_{j,*}^\top$. Since $F_N D_\sigma$ is orthogonal and W has orthonormal columns, $\sum_{j=1}^N z_j z_j^\top = I_p$. Randomized-Hadamard flattening [Tro11, Lemma 3.3] gives, except with probability $\delta/2$,

$$\max_{j\in[N]} \|z_j\|_2^2 \le \frac{C(p + \log(2N/\delta))}{N}. \tag{5}$$

Fix D_σ for which Eq. (5) holds. For a fixed block i, the set $T_i := \Pi^{-1}\mathcal{B}_i$ is a uniform size-s subset of $[N]$, and

$$W_i^\top W_i = \sum_{j\in T_i} z_j z_j^\top.$$

Then we have $\mathbb{E}_\Pi[W_i^\top W_i | D_\sigma] = \frac{s}{N}\sum_{j=1}^N z_j z_j^\top = \frac{1}{m} I_p$. Moreover, Eq. (5) bounds every positive-semidefinite summand by

$$\|z_j z_j^\top\| \le \frac{C(p + \log(2N/\delta))}{N}.$$

Matrix Chernoff for sampling without replacement [Tro11, Theorem 2.2], based on the comparison principle of Gross and Nesme [GN10], now gives

$$\Pr_\Pi[\|mW_i^\top W_i - I_p\| > \kappa \,|\, D_\sigma] \le 2p \exp(-\frac{c\kappa^2 s}{p + \log(2N/\delta)}).$$

For completeness, let $L_0 := \log(Npm/\delta)$. In the present parameter range $L_0 \ge \log 2$, and

$$\frac{pL_0^2}{p + \log(2N/\delta)} \ge cL_0.$$

Thus Eq. (4), with a sufficiently large universal constant, makes the last probability at most $\delta/(2m)$. A union bound over the m blocks and the flattening failure proves that $\mathcal{E}_{\text{block}}(W)$ occurs with the stated probability. The blocks need not be independent.

For the final assertion, put $K := C\kappa^{-2}(d+1)\Lambda^2$. After increasing the universal constant C, we have $K \ge 1$. The minimality of the power-of-two choice of s in Definition 4.1 gives

$$N = ms < 2m \max\{\lceil n/m \rceil, K\} \le 2n + 2m + 2mK = O(n + mK),$$

where the last step uses $m\lceil n/m \rceil \le n + m$ and $K \ge 1$. Since $n \ge d$, $p \le d + 1$, and $\Lambda = \log(Cnmd/(\kappa\delta))$, this bound gives $\log(Npm/\delta) = O(\Lambda)$. Therefore $s \ge C\kappa^{-2}(d+1)\Lambda^2$ implies Eq. (4), after increasing C once more. □

Assume for now that $r \ne 0$ and identify U, r with their padded images JU, Jr. Since $U^\top U = I_d$ and $U^\top r = 0$, the matrix $W := [U, r/\|r\|_2] \in \mathbb{R}^{N\times(d+1)}$ has orthonormal columns. Apply Lemma 6.1 to this W and abbreviate $\mathcal{E}_{\text{block}} := \mathcal{E}_{\text{block}}(W)$. Write $\bar{U} := QU$, $\bar{r} := Qr$, and, for every $i \in [m]$, define the block restrictions $\bar{U}_i := (\bar{U})_{\mathcal{B}_i,*} \in \mathbb{R}^{s\times d}$, $\bar{r}_i := (\bar{r})_{\mathcal{B}_i} \in \mathbb{R}^s$.

Claim 6.2 (Blockwise covariance bounds)**.** *Under the standing assumption $r \ne 0$, for every $i \in [m]$, define*

$$C_i := \bar{U}_i^\top \bar{U}_i, \qquad h_i := \bar{U}_i^\top \bar{r}_i, \qquad v_i := \|\bar{r}_i\|_2^2.$$

On $\mathcal{E}_{\text{block}}$, the following hold simultaneously for every $i \in [m]$:

(a) $\frac{1-\kappa}{m} I_d \preceq C_i \preceq \frac{1+\kappa}{m} I_d$.
(b) $\|h_i\|_2 \le \frac{\kappa}{m}\|r\|_2$.
(c) $v_i \le \frac{1+\kappa}{m}\|r\|_2^2$.

Proof. On $\mathcal{E}_{\text{block}}$, we have $\|mW_i^\top W_i - I_{d+1}\| \le \kappa$, and the block Gram matrix is

$$mW_i^\top W_i - I_{d+1} = \begin{bmatrix} mC_i - I_d & mh_i/\|r\|_2 \\ mh_i^\top/\|r\|_2 & mv_i/\|r\|_2^2 - 1 \end{bmatrix}.$$

The upper-left principal compression satisfies $\|mC_i - I_d\| \le \kappa$, which proves Claim 6.2-(a). The upper-right rectangular compression satisfies $\|mh_i/\|r\|_2\| \le \kappa$, which proves Claim 6.2-(b). The lower-right scalar compression satisfies $|mv_i/\|r\|_2^2 - 1| \le \kappa$; its upper bound proves Claim 6.2-(c). □

Claim 6.3 (Block-sum identities)**.** *The block quantities defined in Claim 6.2 satisfy, independently of* $\mathcal{E}_{\text{block}}$*,*

$$\sum_{i=1}^{m} C_i = I_d, \qquad \sum_{i=1}^{m} h_i = U^\top r = 0.$$

Proof. For the covariance sum,

$$\sum_{i=1}^{m} C_i = \bar{U}^\top \bar{U} = U^\top Q^\top Q U = U^\top U = I_d.$$

The first step follows from the definition of C_i in Claim 6.2 and the fact that the blocks form a partition of the row index set $[N]$. The second step substitutes $\bar{U} = QU$. The third step uses $Q^\top Q = I_N$. The last step uses $U^\top U = I_d$.

For the cross-covariance sum,

$$\sum_{i=1}^{m} h_i = \bar{U}^\top \bar{r} = U^\top Q^\top Q r = U^\top r = 0.$$

The first step follows from the definition of h_i in Claim 6.2 and the fact that the blocks form a partition of the row index set $[N]$. The second step substitutes $\bar{U} = QU$ and $\bar{r} = Qr$. The third step uses $Q^\top Q = I_N$. The last step uses $U^\top r = 0$. □

7 Conditional Gaussian regression: the adaptive-safe step

Fix a realized Q for which $\mathcal{E}_{\text{block}}$ occurs. All probabilities in this section are now only over D_γ.

Definition 7.1 (Conditional observations, design, and Gram matrix)**.** *For every* $i \in [m]$*, put* $g_i := (\gamma_k)_{k \in \mathcal{B}_i} \in \mathbb{R}^s$*. Conditional on the fixed* Q*, the vectors* $g_1, \dots, g_m$ *are mutually independent and each is distributed as* $\mathcal{N}(0, I_s)$*. Define the conditional observation pair by*

$$x_i := \bar{U}_i^\top g_i \in \mathbb{R}^d, \qquad y_i := \bar{r}_i^\top g_i \in \mathbb{R}.$$

Equivalently, $x_i^\top$ *is the* i*-th row of* SU *and* y_i *is the* i*-th entry of* Sr*. Define the conditional design matrix* $X \in \mathbb{R}^{m \times d}$ *by*

$$X_{i,*} := x_i^\top \qquad \text{for every } i \in [m],$$

and define the response vector and conditional Gram matrix by

$$y := (y_1, \dots, y_m)^\top \in \mathbb{R}^m, \qquad M := X^\top X = \sum_{i=1}^{m} x_i x_i^\top \in \mathbb{R}^{d \times d}.$$

Apply Claim 5.1 with $T := [U, r] \in \mathbb{R}^{N \times (d+1)}$, using the padded identification above. Since $(QT)_{\mathcal{B}_i,*} = [\bar{U}_i, \bar{r}_i]$, the pairs (x_i, y_i) are, conditional on this fixed Q, mutually independent across i, centered jointly Gaussian, and have covariance

$$\begin{bmatrix} C_i & h_i \\ h_i^\top & v_i \end{bmatrix}.$$

Lemma 7.2 (Exact conditional regression decomposition)**.** For the standing $r \neq 0$ construction and the fixed Q satisfying $\mathcal{E}_{\text{block}}$, Claim 6.2-(a) implies that every C_i is positive definite. Define

$$\theta_i := C_i^{-1} h_i, \qquad \tau_i^2 := v_i - h_i^\top C_i^{-1} h_i.$$

Define $\xi_i := y_i - x_i^\top \theta_i$ and $\xi := (\xi_1, \dots, \xi_m)^\top \in \mathbb{R}^m$. Conditional on this fixed Q, the variables $\xi_1, \dots, \xi_m$ are mutually independent Gaussians with $\xi_i \sim \mathcal{N}(0, \tau_i^2)$, and the entire vector ξ is independent of the entire design matrix X. The defining identity is

$$y_i = x_i^\top \theta_i + \xi_i. \tag{6}$$

Moreover, we have:

(a) $\|\theta_i\|_2 \le 2\kappa \|r\|_2$ for every $i \in [m]$.
(b) $\tau_i^2 \le \frac{2}{m} \|r\|_2^2$ for every $i \in [m]$.
(c) $\sum_{i=1}^m C_i \theta_i = 0$.
(d) $X^\top y = Z + X^\top \xi$, where $Z := \sum_{i=1}^m (x_i x_i^\top - C_i)\theta_i$.

Proof. Throughout the proof, condition on the fixed Q. Since Q depends only on (D_σ, Π) and is independent of D_γ, the vectors g_i remain independent standard Gaussian vectors after this conditioning. Once Q is fixed, the quantities $\bar{U}_i, \bar{r}_i, C_i, h_i, v_i, \theta_i$, and τ_i^2 are deterministic. By Claim 6.2-(a),

$$C_i \succeq \frac{1-\kappa}{m} I_d \succ 0.$$

The first step follows from Claim 6.2-(a), and the second step follows from $\kappa < 1/2$. Thus C_i is invertible and $\theta_i = C_i^{-1} h_i$ is well defined. By the definitions in the lemma statement,

$$\xi_i = y_i - x_i^\top \theta_i = y_i - x_i^\top C_i^{-1} h_i.$$

Conditional on Q, the pair (x_i, y_i) is centered and jointly Gaussian, so (x_i, ξ_i) is also centered and jointly Gaussian. Moreover,

$$\mathbb{E}_\gamma[x_i \xi_i | Q] = \mathbb{E}_\gamma[x_i y_i | Q] - \mathbb{E}_\gamma[x_i x_i^\top | Q]\theta_i = h_i - C_i \theta_i = 0.$$

The first step substitutes $\xi_i = y_i - x_i^\top \theta_i$, and the second step uses the covariance blocks $\mathbb{E}_\gamma[x_i y_i | Q] = h_i$ and $\mathbb{E}_\gamma[x_i x_i^\top | Q] = C_i$, and the last step uses $C_i \theta_i = h_i$. Hence ξ_i is independent of x_i, because zero covariance implies independence for jointly Gaussian variables.

Its conditional variance satisfies

$$\mathbb{E}_\gamma[\xi_i^2 | Q] = v_i - \theta_i^\top h_i = v_i - h_i^\top C_i^{-1} h_i = \tau_i^2.$$

For the first step, expanding $(y_i - x_i^\top \theta_i)^2$ and using the three covariance blocks gives $v_i - 2\theta_i^\top h_i + \theta_i^\top C_i \theta_i$, which reduces to $v_i - \theta_i^\top h_i$ because $C_i \theta_i = h_i$. The second step uses $\theta_i = C_i^{-1} h_i$, and the last step is the definition of τ_i^2. In particular, $\tau_i^2 \ge 0$, and conditional on Q, $\xi_i \sim \mathcal{N}(0, \tau_i^2)$.

Conditional on Q, for distinct indices i and j, (x_i, ξ_i) and (x_j, ξ_j) are functions of the disjoint Gaussian blocks g_i and g_j, so these pairs are independent; in particular, the ξ_i are mutually independent. For completeness, the stacked vector formed by all x_i and ξ_i is a linear image of the jointly Gaussian vector formed by $g_1, \dots, g_m$, and is therefore jointly Gaussian. If $i \neq j$, block independence and centering give $\mathbb{E}_\gamma[x_j \xi_i | Q] = 0$, while the same identity for $i = j$ was proved above. Thus the cross-covariance between ξ and the stacked design vector $(x_1, \dots, x_m)$ is zero.

Joint Gaussianity now implies that ξ is independent of the entire design matrix X. This proves Eq. (6) and the asserted independence.

Proof of Lemma 7.2-(a). Because inverting a positive-definite matrix reciprocates its eigenvalues, Claim 6.2-(a) gives

$$\|C_i^{-1}\| \le \frac{m}{1-\kappa}.$$

Consequently,

$$\|\theta_i\|_2 \le \|C_i^{-1}\|\|h_i\|_2 \le \frac{\kappa}{1-\kappa}\|r\|_2 \le 2\kappa\|r\|_2.$$

The first step uses operator-norm submultiplicativity applied to $\theta_i = C_i^{-1}h_i$. The second step combines the preceding inverse bound with Claim 6.2-(b), and the last step uses $\kappa < 1/2$.

Proof of Lemma 7.2-(b). We have

$$0 \le \tau_i^2 = v_i - h_i^\top C_i^{-1} h_i \le v_i \le \frac{1+\kappa}{m}\|r\|_2^2 \le \frac{2}{m}\|r\|_2^2.$$

The first step follows because τ_i^2 is the conditional variance computed above. The second step is its definition, the third step uses $C_i^{-1} \succeq 0$, the fourth step is Claim 6.2-(c), and the last step uses $\kappa < 1/2$.

Proof of Lemma 7.2-(c). By the definition of θ_i, we have $C_i\theta_i = h_i$. Therefore, Claim 6.3 gives

$$\sum_{i=1}^m C_i\theta_i = \sum_{i=1}^m h_i = 0.$$

The first step applies $C_i\theta_i = h_i$ term by term, and the second step uses the identity $\sum_{i=1}^m h_i = 0$ in Claim 6.3.

Proof of Lemma 7.2-(d). Using Eq. (6),

$$X^\top y = \sum_{i=1}^m x_i y_i = \sum_{i=1}^m x_i x_i^\top \theta_i + \sum_{i=1}^m x_i\xi_i = \sum_{i=1}^m (x_i x_i^\top - C_i)\theta_i + \sum_{i=1}^m C_i\theta_i + X^\top\xi = Z + X^\top\xi.$$

The first step follows from the row definitions of X and y. The second step substitutes $y_i = x_i^\top\theta_i + \xi_i$ from Eq. (6). The third step adds and subtracts $\sum_{i=1}^m C_i\theta_i$ and uses $\sum_{i=1}^m x_i\xi_i = X^\top\xi$. The last step uses the definition of Z and the cancellation in Lemma 7.2-(c).

□

This is the key repair: since the noise ξ is independent of the full design after conditioning on Q, the inverse M^{-1} may depend arbitrarily on the design X, and no fixed-vector OCE statement is ever applied to a sketch-dependent vector.

8 Concentration of the design and centered bias

We record the standard scalar tools used below. If $G_1, \dots, G_m \sim \mathcal{N}(0,1)$ are independent and $a_1, \dots, a_m \ge 0$, then the Laurent–Massart inequality [LM00, Lemma 1] gives

$$\Pr[|\sum_{i=1}^m a_i(G_i^2 - 1)| > t] \le 2\exp(-c\min\{\frac{t^2}{\sum_{i=1}^m a_i^2}, \frac{t}{\max_{1\le i\le m} a_i}\}). \tag{7}$$

For centered jointly Gaussian G, H, the standard Gaussian-product bound [Ver18, Lemmas 2.7.7 and 2.7.10] gives

$$\|GH - \mathbb{E}[GH]\|_{\psi_1} \le C\sqrt{\mathbb{E}[G^2]\,\mathbb{E}[H^2]}, \tag{8}$$

where $\|Z\|_{\psi_1} := \inf\{t > 0 : \mathbb{E}[e^{|Z|/t}] \le 2\}$. Scalar Bernstein [Ver18, Theorem 2.8.1] converts independent ψ_1 bounds into the usual subexponential tail: if the Y_i are independent and centered with $\|Y_i\|_{\psi_1} \le K$, then, for $t \ge 1$, $\Pr[|\sum_{i=1}^m Y_i| > CK(\sqrt{mt} + t)] \le 2e^{-t}$. Finally, a $(1/4)$-net of S^{d-1} has size at most 9^d and controls a symmetric matrix norm up to factor two; a $(1/2)$-net has size at most 5^d and controls a vector norm up to factor two.

Lemma 8.1 (Design and bias concentration)**.** Fix any Q for which $\mathcal{E}_{\text{block}}$ (see definition in Lemma 6.1) occurs and let $L := d + \log(8/\delta)$. Let M be the conditional Gram matrix defined in Definition 7.1, and let $Z := \sum_{i=1}^m (x_i x_i^\top - C_i)\theta_i$ be the quantity defined in Lemma 7.2-(d). If $m \ge CL$, then, with conditional probability at least $1 - \delta/2$ over D_γ, the following hold:

(a) $\|M - I_d\| \le \frac{1}{2}$.

(b) $\|Z\|_2 \le C\kappa\|r\|_2(\sqrt{\frac{L}{m}} + \frac{L}{m})$.

These bounds are uniform over every fixed Q for which $\mathcal{E}_{\text{block}}$ occurs.

Proof. **Proof of Lemma 8.1-(a).** Fix a unit vector u. Conditional on the fixed Q, Claim 5.1 and Definition 7.1 imply that the vectors $x_1, \dots, x_m$ are independent and $x_i \sim \mathcal{N}(0, C_i)$. We define $a_i := u^\top C_i u$, $u^\top x_i = \sqrt{a_i} G_i$, where $G_1, \dots, G_m$ are independent standard Gaussians. The first relation defines a_i. Claim 6.2-(a), together with the standing bound $\kappa < 1/2$, shows that a_i is strictly positive. The second relation then follows by standardizing the independent Gaussian variables $u^\top x_i \sim \mathcal{N}(0, a_i)$.

The coefficient bounds are

$$\sum_{i=1}^m a_i = u^\top(\sum_{i=1}^m C_i)u = u^\top u = 1, \qquad \max_{1\le i\le m} a_i \le \frac{1+\kappa}{m} \le \frac{2}{m}, \qquad \sum_{i=1}^m a_i^2 \le (\max_{1\le i\le m} a_i)\sum_{i=1}^m a_i \le \frac{2}{m}.$$

For the first chain, the first step substitutes the definition of a_i, the second step uses Claim 6.3, and the last step uses $\|u\|_2 = 1$. For the second chain, the first step uses Claim 6.2-(a) and the last step uses $\kappa < 1/2$. For the third chain, the first step uses $a_i \ge 0$ term by term, and the last step uses the preceding two bounds.

Since $M = X^\top X = \sum_{i=1}^m x_i x_i^\top$,

$$u^\top(M - I_d)u = \sum_{i=1}^m (u^\top x_i)^2 - u^\top u = \sum_{i=1}^m a_i G_i^2 - \sum_{i=1}^m a_i = \sum_{i=1}^m a_i(G_i^2 - 1).$$

The first step expands the definition of M. The second step substitutes $u^\top x_i = \sqrt{a_i} G_i$ and uses $u^\top u = 1 = \sum_{i=1}^m a_i$. The last step collects the two sums term by term.

Eq. (7), with weights a_i and threshold $1/4$, gives

$$\Pr_\gamma[|u^\top(M - I_d)u| > 1/4 \,|\, Q] \le 2\exp(-c\min\{\frac{1}{16\sum_{i=1}^m a_i^2}, \frac{1}{4\max_{1\le i\le m} a_i}\}) \le 2e^{-cm}.$$

The first step applies Eq. (7) to the preceding identity. For the last step, the coefficient bounds give $1/(16\sum_{i=1}^m a_i^2) \ge m/32$ and $1/(4\max_{1\le i\le m} a_i) \ge m/8$, and the absolute factor is absorbed into c.

This fixed-u bound is uniform over unit vectors u. Let $\mathcal{N}_{1/4}$ be a $1/4$-net of S^{d-1} with size at most 9^d. A union bound gives

$$\Pr_\gamma[\max_{u\in\mathcal{N}_{1/4}} |u^\top(M - I_d)u| > 1/4 \,|\, Q] \le 2 \cdot 9^d e^{-cm} \le \delta/4.$$

The first step combines the preceding fixed-u tail bound with $|\mathcal{N}_{1/4}| \le 9^d$. The last step uses $m \ge C(d + \log(8/\delta))$ with a sufficiently large universal constant C. On the complementary event, $M - I_d$ is symmetric, so the symmetric-matrix net bound yields

$$\|M - I_d\| \le 2 \max_{u \in \mathcal{N}_{1/4}} |u^\top (M - I_d) u| \le \frac{1}{2}.$$

The first step is the standard 1/4-net bound for a symmetric matrix, whose factor is $(1-2\cdot 1/4)^{-1} = 2$. The last step uses the defining inequality of the complementary event. This proves Lemma 8.1-(a) with conditional failure probability at most $\delta/4$.

Proof of Lemma 8.1-(b). Conditional on the fixed Q, the vectors $x_1, \dots, x_m$ are independent centered Gaussians, while C_i and θ_i are deterministic. For a unit vector u, define

$$Y_i := u^\top (x_i x_i^\top - C_i)\theta_i = (u^\top x_i)(x_i^\top \theta_i) - u^\top C_i \theta_i.$$

The displayed step expands the matrix product. Each Y_i depends only on x_i, so the variables $Y_1, \dots, Y_m$ are conditionally independent. Moreover,

$$\mathbb{E}_\gamma[Y_i | Q] = u^\top (\mathbb{E}_\gamma[x_i x_i^\top | Q] - C_i)\theta_i = 0.$$

The first step substitutes the definition of Y_i and uses linearity of conditional expectation. The last step uses $\mathbb{E}_\gamma[x_i x_i^\top | Q] = C_i$. Thus Y_i is a centered product of the jointly Gaussian linear forms $u^\top x_i$ and $x_i^\top \theta_i$.

The variance of the second linear form satisfies

$$\theta_i^\top C_i \theta_i = h_i^\top C_i^{-1} h_i \le \|C_i^{-1}\| \|h_i\|_2^2 \le \frac{m}{1-\kappa} \frac{\kappa^2}{m^2} \|r\|_2^2 \le \frac{2\kappa^2}{m} \|r\|_2^2.$$

The first step substitutes $\theta_i = C_i^{-1} h_i$ and uses the symmetry of C_i. The second step applies the operator-norm bound to the quadratic form. The third step uses Claim 6.2-(a) and Claim 6.2-(b). The last step uses $\kappa < 1/2$. Similarly,

$$u^\top C_i u \le \frac{1+\kappa}{m} \le \frac{2}{m}.$$

The first step uses Claim 6.2-(a) and $\|u\|_2 = 1$, and the last step uses $\kappa < 1/2$.

Eq. (8) now gives

$$\|Y_i\|_{\psi_1} \le C \sqrt{u^\top C_i u} \sqrt{\theta_i^\top C_i \theta_i} \le \frac{C\kappa}{m} \|r\|_2.$$

The first step applies Eq. (8) to the centered Gaussian product defining Y_i. The last step substitutes the preceding two variance bounds and absorbs absolute numerical factors into C.

By the definitions of Z and Y_i,

$$u^\top Z = \sum_{i=1}^m u^\top (x_i x_i^\top - C_i)\theta_i = \sum_{i=1}^m Y_i.$$

The first step substitutes the definition of Z, and the last step uses the definition of Y_i. Scalar Bernstein, applied conditionally on Q to these independent centered variables with $\|Y_i\|_{\psi_1} \le C\kappa \|r\|_2/m$, gives, for every $t \ge 1$,

$$\Pr_\gamma[|u^\top Z| > C\kappa \|r\|_2 (\sqrt{\frac{t}{m}} + \frac{t}{m}) \,|\, Q] \le 2e^{-t}.$$

This step is the scalar Bernstein inequality with $K := C\kappa\|r\|_2/m$. Substituting this value into the Bernstein threshold $K(\sqrt{mt}+t)$ gives the displayed scale, after absorbing absolute constants into C.

Take $t := CL$, where $L = d + \log(8/\delta)$, and let $\mathcal{N}_{1/2}$ be a 1/2-net of S^{d-1} with size at most 5^d. A union bound gives

$$\Pr_{\gamma}[\max_{u\in\mathcal{N}_{1/2}} |u^\top Z| > C\kappa\|r\|_2(\sqrt{\frac{L}{m}} + \frac{L}{m}) \mid Q] \le 2\cdot 5^d e^{-CL} \le \frac{\delta}{4}.$$

The first step combines the fixed-u Bernstein bound with $|\mathcal{N}_{1/2}| \le 5^d$ and absorbs absolute factors from $t = CL$ into C. The last step uses $L = d + \log(8/\delta)$ and a sufficiently large universal constant C. On the complementary event, the vector net bound gives

$$\|Z\|_2 \le 2\max_{u\in\mathcal{N}_{1/2}} |u^\top Z| \le C\kappa\|r\|_2(\sqrt{\frac{L}{m}} + \frac{L}{m}).$$

The first step is the standard 1/2-net bound for a vector. The last step uses the defining inequality of the complementary event and absorbs the factor two into C. This proves Lemma 8.1-(b) with conditional failure probability at most $\delta/4$.

Adding the two conditional failure probabilities gives $\delta/4+\delta/4 = \delta/2$ and proves the lemma. The estimates are uniform over every fixed good Q because the conditional Gaussian and independence structure holds for every such Q, while all numerical bounds use only the deterministic inequalities defining $\mathcal{E}_{\text{block}}$. □

With the block geometry and the conditional Gaussian representation in hand, we can already record a classical property of the balanced Gaussian-pooled transform: it is an oblivious subspace embedding. This fact is not needed for the core lemma in Section 9, but we state it here for completeness and for comparison with prior work.

Definition 8.2 (Oblivious subspace embedding, [Sar06])**.** *Fix $0 < \eta < 1$, $0 < \delta < 1$, and $1 \le d \le n$. A distribution $\mathcal{D}$ over matrices $S \in \mathbb{R}^{m\times n}$ is an (η, δ, d)-oblivious subspace embedding (OSE) if, for every fixed $U \in \mathbb{R}^{n\times d}$ with $U^\top U = I_d$ chosen independently of S, a draw $S \sim \mathcal{D}$ satisfies*

$$(1-\eta)\|z\|_2^2 \le \|SUz\|_2^2 \le (1+\eta)\|z\|_2^2 \qquad \text{for every } z \in \mathbb{R}^d$$

with probability at least $1-\delta$. Equivalently, with the same probability, $\|U^\top S^\top SU - I_d\| \le \eta$. The word oblivious *means that the distribution $\mathcal{D}$ does not depend on U.*

Proposition 8.3 (The transform is an OSE)**.** *Let S be the balanced Gaussian-pooled transform of Definition 4.1 with $m \ge C\epsilon^{-2}(d + \log(1/\delta))$ for a sufficiently large universal constant C. Then S is an (ϵ, δ, d)-OSE.*

Proof. Fix $U \in \mathbb{R}^{n\times d}$ with $U^\top U = I_d$, independently of S, and put $W := JU$. Apply Lemma 6.1 to W with $p = d$ and failure probability $\delta/2$, absorbing the replacement of δ by $\delta/2$ into the universal constant. Condition on a realized Q for which $\mathcal{E}_{\text{block}}(W)$ occurs, and define $C_i := W_i^\top W_i$. By Claim 5.1, the rows $x_i^\top$ of SU are independent and $x_i \sim \mathcal{N}(0, C_i)$. Since the blocks partition $[N]$, Q is orthogonal, and $\kappa < 1/2$, we have $\sum_{i=1}^m C_i = I_d$, $0 \preceq C_i \preceq \frac{1+\kappa}{m} I_d \preceq \frac{2}{m} I_d$. Put $M := U^\top S^\top SU = \sum_{i=1}^m x_i x_i^\top$. For a fixed unit vector u, let $a_i := u^\top C_i u$. Then $\sum_{i=1}^m a_i = 1$, $\max_{1\le i\le m} a_i \le 2/m$, and $\sum_{i=1}^m a_i^2 \le 2/m$. Hence, for independent standard Gaussians $G_1, \dots, G_m$, $u^\top(M - I_d)u = \sum_{i=1}^m a_i(G_i^2 - 1)$. Eq. (7), applied with threshold $\epsilon/2$, gives, since $0 < \epsilon \le 1$, $\Pr_\gamma[|u^\top(M - I_d)u| >$

$\epsilon/2|Q] \le 2e^{-cm\epsilon^2}$. Let $\mathcal{N}_{1/4}$ be a 1/4-net of the unit sphere with size at most 9^d. For a sufficiently large universal constant C, the assumed lower bound on m and a union bound imply that, with conditional probability at least $1-\delta/2$, $\|M - I_d\| \le 2\max_{u\in\mathcal{N}_{1/4}} |u^\top(M-I_d)u| \le \epsilon$. Adding the failure probability of the block event proves the claim. The law of S is independent of U, so the embedding is oblivious. □

9 The corrected core lemma

Lemma 9.1 (Adaptive-safe one-shot core lemma)**.** Let $0 < \epsilon \le 1$, $0 < \delta < 1/2$, $U \in \mathbb{R}^{n\times d}$ have orthonormal columns, and let fixed $r \in \mathbb{R}^n$ and $c \in \mathbb{R}^d$ satisfy $U^\top r = 0$. Use the sketch from Definition 4.1 with

$$\kappa := \frac{c_0}{\sqrt{d}}, \qquad m \ge C\epsilon^{-2} d\log(8/\delta), \tag{9}$$

where $c_0 > 0$ is a sufficiently small universal constant; use the corresponding choices of s, N in Definition 4.1. With probability at least $1-\delta$, the matrix SU has full column rank and

$$|c^\top (U^\top S^\top SU)^{-1} U^\top S^\top Sr| \le \frac{\epsilon}{\sqrt{d}} \|c\|_2 \|r\|_2. \tag{10}$$

Proof. First suppose $r \ne 0$ and write $L := d + \log(8/\delta)$. Since $d \ge 1$ and $\delta < 1/2$, we have $L \le Cd\log(8/\delta)$. Thus the row assumption in Eq. (9) implies $m \ge CL$, as required by Lemma 8.1.

Put $W_r := [JU, Jr/\|r\|_2]$. Apply Lemma 6.1 to W_r, with its failure parameter set to $\delta/4$. Replacing δ by $\delta/4$ only adds $\log 4$ inside the logarithm in Eq. (4), which is absorbed by increasing the universal constant C. Fix any realized Q for which $\mathcal{E}_{\text{block}}(W_r)$ occurs. Conditional on this Q, Lemma 8.1 fails with probability at most $\delta/2$.

On its success event, Lemma 8.1-(a) gives $\|M - I_d\| \le 1/2$, so $M = X^\top X$ is positive definite and $\|M^{-1}\| \le 2$. By Definition 7.1, we have $SU = X$ and $Sr = y$. Consequently,

$$U^\top S^\top SU = M, \qquad U^\top S^\top Sr = X^\top y.$$

Lemma 7.2-(d) therefore gives

$$c^\top M^{-1} X^\top y = c^\top M^{-1} Z + c^\top M^{-1} X^\top \xi.$$

For the first term, Lemma 8.1-(b) gives

$$|c^\top M^{-1} Z| \le \|c\|_2 \|M^{-1}\| \|Z\|_2 \le 2\|c\|_2 \|Z\|_2 \le C\kappa \|c\|_2 \|r\|_2 \left(\sqrt{\frac{L}{m}} + \frac{L}{m}\right).$$

The row assumption and the displayed bound on L imply

$$\frac{L}{m} \le C\epsilon^2, \qquad \sqrt{\frac{L}{m}} + \frac{L}{m} \le C\epsilon,$$

where we used $0 < \epsilon \le 1$. Substituting $\kappa = c_0/\sqrt{d}$ from Eq. (9), we obtain

$$|c^\top M^{-1} Z| \le Cc_0 \frac{\epsilon}{\sqrt{d}} \|c\|_2 \|r\|_2.$$

Choose c_0 small enough that $Cc_0 \le 1/2$. Then the bias term is at most $\epsilon\|c\|_2\|r\|_2/(2\sqrt{d})$.

For the second term, condition on the pair (Q, X). This is essential: conditional on Q, ξ is independent of X, whereas mixing over different values of Q need not preserve independence. The

design-and-bias success event above is measurable with respect to (Q, X). Given (Q, X), the second term is a centered Gaussian with variance

$$\mathrm{Var}_\gamma[c^\top M^{-1}X^\top \xi | Q, X] = \sum_{i=1}^m \tau_i^2 (x_i^\top M^{-1} c)^2 \le \max_{1\le i\le m} \tau_i^2 \|XM^{-1}c\|_2^2 \le \frac{2\|r\|_2^2}{m}\|XM^{-1}c\|_2^2.$$

The last step uses Lemma 7.2-(b). Since $M = X^\top X$,

$$\|XM^{-1}c\|_2^2 = c^\top M^{-1}X^\top XM^{-1}c = c^\top M^{-1}c \le 2\|c\|_2^2.$$

Thus the conditional variance is at most $4\|c\|_2^2\|r\|_2^2/m$. The Gaussian tail bound at the target threshold gives

$$\Pr_\gamma[|c^\top M^{-1}X^\top \xi| > \frac{\epsilon}{2\sqrt{d}}\|c\|_2\|r\|_2 \,|\, Q, X] \le 2\exp(-\frac{\epsilon^2 m}{32d}) \le \frac{\delta}{4}$$

after enlarging C. This estimate is uniform over the measurable success event. Integrating the conditional estimates and adding the block-embedding, design-and-bias, and noise failures gives $\frac{\delta}{4} + \frac{\delta}{2} + \frac{\delta}{4} = \delta$. This proves Eq. (10), and $M \succ 0$ gives the asserted full column rank of SU.

If $r = 0$, apply Lemma 6.1 to JU with $p = d$ and failure probability $\delta/4$. Conditional on a good Q, repeat only the proof of Lemma 8.1-(a); its conditional failure probability is at most $\delta/4$. Hence $M \succ 0$ with probability at least $1 - \delta/2 \ge 1 - \delta$. Since $Sr = 0$, the numerator in Eq. (10) is identically zero. This proves both assertions in the remaining case. □

10 Regression consequence and desired row count

Theorem 10.1 (One-shot coordinate-wise regression)**.** *Let $0 < \epsilon \le 1$, $0 < \delta < 1/2$, let $A \in \mathbb{R}^{n\times d}$ have full column rank, let $b \in \mathbb{R}^n$, and fix $a \in \mathbb{R}^d$. Set $\kappa := c_0/\sqrt{d}$, choose m to be the smallest power of two satisfying $m \ge C\epsilon^{-2}d\log(8/\delta)$, and draw S according to Definition 4.1. Then, with probability at least $1 - \delta$, SA has full column rank, the sketched minimizer is unique, and*

$$|a^\top(\widehat{x} - x^\star)| \le \frac{\epsilon}{\sqrt{d}}\,\|a\|_2\,\|Ax^\star - b\|_2\,\|A^\dagger\|_{\mathrm{op}}. \tag{11}$$

For all coordinates simultaneously, it suffices to take

$$\boxed{m = O(\epsilon^{-2}d\log(d/\delta)).} \tag{12}$$

For this simultaneous statement, replace δ by δ/d in the entire parameter selection, including m, Λ, s, N.

Proof. Let $A = U\Sigma V^\top$ and $r := b - Ax^\star$. Then $U^\top r = 0$ and $b = Ax^\star + r$. On the good event in Lemma 9.1, SU has full column rank. Since $SA = (SU)\Sigma V^\top$, the matrix SA also has full column rank, so the sketched minimizer is unique.

The normal equations for the sketched problem give $(A^\top S^\top SA)(\widehat{x} - x^\star) = A^\top S^\top Sr$. Substituting the thin singular value decomposition and inverting the nonsingular factors yields $\widehat{x} - x^\star = V\Sigma^{-1}(U^\top S^\top SU)^{-1}U^\top S^\top Sr$. Apply Lemma 9.1 with $c := \Sigma^{-1}V^\top a$. Since $\|c\|_2 \le \|\Sigma^{-1}\|\|a\|_2 = \|A^\dagger\|\|a\|_2$, we obtain Eq. (11). For simultaneous coordinates, replace δ by δ/d in all parameters, including m, Λ, s, N. Apply the one-coordinate result to each fixed direction $a = e_j$ with failure probability δ/d, and union bound over $j \in [d]$. □

11 Running time

Lemma 11.1 (Sketch-and-solve running time)**.** Under the assumptions and parameter choices of Theorem 10.1, on its success event the one-shot estimator $\widehat{x}$ can be computed in the exact-arithmetic model using $O(Nd\log N+\mathcal{T}_{\mathrm{mat}}(d,m,d)+d^{\omega})$ arithmetic operations, where $\mathcal{T}_{\mathrm{mat}}(a,b,c)$ denotes the cost of multiplying an $a\times b$ matrix by a $b\times c$ matrix. Here $N=\widetilde{O}(n+\epsilon^{-2}d^{3})$. In particular, the general bound is $\widetilde{O}(nd+\epsilon^{-2}d^{4})$.

Proof. With $\kappa=c_0/\sqrt{d}$ and $p=d+1$, the second term in the block-size requirement of Definition 4.1 is $\widetilde{O}(d^2)$. Combining this with Eq. (12) and the choice of s,N in Definition 4.1 gives $N=O(n+md^2\Lambda^2)=\widetilde{O}(n+\epsilon^{-2}d^3)$. Applying S to one vector costs $O(N\log N)$. Applying it to all d columns of A and to b therefore costs $O(N(d+1)\log N)=O(Nd\log N)$. The first step sums the cost over the $d+1$ inputs, and the last step uses $d\geq 1$. The resulting dense least-squares problem has size $m\times d$. In the exact-arithmetic model, write $Y:=SA$ and $z:=Sb$. On the success event in Theorem 10.1, Y has full column rank, so $\widehat{x}$ is the unique solution of $(Y^{\top}Y)\widehat{x}=Y^{\top}z$. This is the normal equation for the full-column-rank least-squares problem. The two products can be formed together by multiplying $Y^{\top}$ by the $m\times(d+1)$ matrix whose columns are those of Y followed by z. After constant-factor padding in the last dimension, this costs $O(\mathcal{T}_{\mathrm{mat}}(d,m,d))$. Solving the resulting nonsingular $d\times d$ system costs $O(d^{\omega})$ [BH74]. This is an exact-arithmetic normal-equations calculation; it is not a claim that a backward-stable QR factorization has the same cost. Adding the sketching, product-formation, and linear-system costs gives $T=O(Nd\log N+\mathcal{T}_{\mathrm{mat}}(d,m,d)+d^{\omega})$. This step adds the three costs established above. Definition 4.1 and the parameter choices of Theorem 10.1 give $N=\widetilde{O}(n+\epsilon^{-2}d^3)$ and $m=\widetilde{O}(\epsilon^{-2}d)$. Consequently, $\mathcal{T}_{\mathrm{mat}}(d,m,d)=O(md^2)=\widetilde{O}(\epsilon^{-2}d^3)$ by classical matrix multiplication, while $d^{\omega}=O(d^3)$. Substituting these bounds gives $T=\widetilde{O}(nd+\epsilon^{-2}d^4)$. This step substitutes the bounds for N, $\mathcal{T}_{\mathrm{mat}}(d,m,d)$, and d^{ω} into the preceding running-time expression.

□

References

[BH74] James R. Bunch and John E. Hopcroft. Triangular factorization and inversion by fast matrix multiplication. *Mathematics of Computation*, 28(125):231–236, 1974.

[CCFC02] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *Automata, Languages and Programming*, volume 2380 of *Lecture Notes in Computer Science*, pages 693–703. Springer, 2002.

[CW17] Kenneth L. Clarkson and David P. Woodruff. Low-rank approximation and regression in input sparsity time. *Journal of the ACM*, 63(6):54:1–54:45, 2017.

[DG03] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

[DMMS11] Petros Drineas, Michael W. Mahoney, S. Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2011.

[GN10] David Gross and Vincent Nesme. Note on sampling without replacing from a finite collection of matrices. arXiv:1001.2738, 2010.

[JL84] William B. Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. In *Conference in Modern Analysis and Probability*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society, 1984.

[LM00] Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

[LN17] Kasper Green Larsen and Jelani Nelson. Optimality of the johnson–lindenstrauss lemma. In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science*, pages 633–638, 2017.

[NN13] Jelani Nelson and Huy L. Nguyen. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, pages 117–126, 2013.

[PSW17] Eric Price, Zhao Song, and David P. Woodruff. Fast regression with an ℓ_∞ guarantee. In *ICALP*, 2017. Theorem 10.

[Sar06] Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 143–152, 2006.

[SYYZ23] Zhao Song, Mingquan Ye, Junze Yin, and Lichen Zhang. A nearly-optimal bound for fast regression with an ℓ_∞ guarantee. arXiv:2302.00248v1, 2023.

[Tro11] Joel A. Tropp. Improved analysis of the subsampled randomized hadamard transform. *Advances in Adaptive Data Analysis*, 3(1–2), 2011.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.