Algorithmic Gaussianization through Sketching: Converting Data into Sub-gaussian Random Designs

Michał Dereziński   DEREZIN@UMICH.EDU

University of Michigan

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Algorithmic Gaussianization is a phenomenon that can arise when using randomized sketching or sampling methods to produce smaller representations of large datasets: For certain tasks, these sketched representations have been observed to exhibit many robust performance characteristics that are known to occur when a data sample comes from a sub-gaussian random design, which is a powerful statistical model of data distributions. However, this phenomenon has only been studied for specific tasks and metrics, or by relying on computationally expensive methods. We address this by providing an algorithmic framework for gaussianizing data using sparse sketching operators, proving that it is possible to efficiently construct data sketches that are nearly indistinguishable (in terms of total variation distance) from sub-gaussian random designs. In particular, relying on a recently introduced sketching technique called Leverage Score Sparsified (LESS) embeddings, we show that one can construct an $n \times d$ sketch of an $N \times d$ matrix $\mathbf{A}$, where $n \ll N$, that is nearly indistinguishable from a sub-gaussian design, in time $O(\mathrm{nnz}(\mathbf{A}) \log N + nd^2)$, where $\mathrm{nnz}(\mathbf{A})$ is the number of non-zero entries in $\mathbf{A}$. As a consequence, strong statistical guarantees and precise asymptotics available for the estimators produced from sub-gaussian designs (e.g., for least squares and Lasso regression, covariance estimation, low-rank approximation, etc.) can be straightforwardly adapted to our sketching framework. We illustrate this with a new approximation guarantee for sketched least squares, among other examples. The key technique that enables our analysis is a novel variant of the Hanson-Wright inequality on the concentration of random quadratic forms, which we establish for random vectors that arise from sparse sketches.

Keywords: Sketching, Least squares, Randomized Numerical Linear Algebra, Sub-gaussianity

1. Introduction

In a standard statistical learning setup, we are given a sample of i.i.d. points $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$, and our goal is to perform an estimation or prediction task. For example, in linear regression, we aim to learn a linear model $\mathbf{w}^*$ from labels/responses $y_i = \mathbf{x}_i^\top \mathbf{w}^* + \xi_i$, where ξ_i represents the noise. Naturally, the performance of prediction models for linear regression, such as ordinary least squares (OLS) and Lasso, depends greatly on the properties of the sample distribution, as well as on the distribution of the noise ξ, e.g., whether they are heavy-tailed or not. To that end, there is extensive literature (see Section 5.2) which provides precise analysis of statistical and machine learning models under strong distributional assumptions on the data, one of the most common assumptions being Gaussianity and sub-gaussianity. When, the data and noise exhibit small sub-gaussian tails (e.g., by having a constant sub-gaussian Orlicz norm), then estimators such as OLS and Lasso, as well as model selection methods such as cross-validation, exhibit provably good, even optimal, performance. In modern learning tasks these distributional assumptions are rarely met. On the other hand, we can often benefit from a great abundance of cheap data available in many domains. To that

end, we ask: Can we leverage this data abundance to algorithmically introduce sub-gaussianity into a data distribution, and when is it practical?

A key motivation in this context is randomized sketching, which is useful in situations when running a learning/estimation algorithm directly on the entire dataset is computationally prohibitive. Instead, we produce a smaller sketch of the data, e.g., via importance sampling or random projections, and use it as a surrogate data sample. To illustrate this, suppose that our goal is to solve a least squares regression task $(\mathbf{A}, \mathbf{b})$:

$$\text{Find} \qquad \mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \, L(\mathbf{w}) \qquad \text{for} \quad L(\mathbf{w}) = \|\mathbf{A}\mathbf{w} - \mathbf{b}\|^2,$$

where $\mathbf{A}$ is a large tall $N \times d$ matrix and $\mathbf{b}$ is an N-dimensional vector. In the context of sketching, a classical way of computationally reducing this task to a much smaller and more tractable instance is via the so-called Gaussian embedding: We apply a randomized linear transformation to both $\mathbf{A}$ and $\mathbf{b}$, where the transformation is defined by an $n \times N$ sketching matrix $\mathbf{S}$ with i.i.d. Gaussian entries from $\mathcal{N}(0, 1/n)$ to produce a much smaller $n \times d$ sketched regression task $(\mathbf{X}, \mathbf{y}) = (\mathbf{S}\mathbf{A}, \mathbf{S}\mathbf{b})$, where $n \ll N$. Remarkably, for an arbitrary fixed input $(\mathbf{A}, \mathbf{b})$, the resulting algorithmically generated data sample $(\mathbf{X}, \mathbf{y})$ exactly matches the standard Gaussian random design: each row vector $\mathbf{x}_i^\top$ of $\mathbf{X}$ is a Gaussian vector with covariance $\mathbf{\Sigma} = \frac{1}{n}\mathbf{A}^\top\mathbf{A}$, whereas each entry of $\mathbf{y}$ is distributed according to a linear model $y_i = \mathbf{x}_i^\top\mathbf{w}^* + \xi_i$ with independent mean zero Gaussian noise.[1] Note that we did not impose any linear noise model on the original problem $(\mathbf{A}, \mathbf{b})$, but rather, it arises naturally through the sketching transformation. In particular, this allows us to derive the exact expected approximation error of the sketched ordinary least squares (OLS) estimator:

$$\mathbb{E}\big[L(\widehat{\mathbf{w}}) - L(\mathbf{w}^*)\big] = \frac{d}{n - d - 1} \cdot L(\mathbf{w}^*) \qquad \text{for} \qquad \widehat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \, \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2. \qquad (1)$$

This statistical model implies many other strong performance guarantees for a variety of tasks such as covariance estimation, OLS, Lasso, PCA etc. (e.g., Koltchinskii and Lounici, 2017; Dobriban and Wager, 2018; Miolane and Montanari, 2021), which directly apply to data transformed by Gaussian embeddings, regardless of the distribution of the matrix $\mathbf{A}$. We refer to this phenomenon as Algorithmic Gaussianization. Unfortunately, Gaussian embeddings carry a substantial preprocessing cost, compared to, say, uniformly down-sampling the data, and so, many other sketching techniques have been proposed, e.g., CountSketch, Sparse Johnson-Lindenstrauss Transforms (SJLT), Subsampled Randomized Hadamard Transforms (SRHT) and Leverage Score Sampling (e.g., Sarlos, 2006; Drineas et al., 2006; Clarkson and Woodruff, 2017), which can be viewed as computationally efficient algorithms for transforming a large dataset into a small data sample. Existing work has shown various approximation guarantees for these algorithms, via arguments based on subspace embeddings and the Johnson-Lindenstrauss property (Woodruff, 2014; Drineas and Mahoney, 2016; Dereziński and Mahoney, 2021). However, even though some level of gaussianization is implied by these guarantees, exact parallels with Gaussian random designs are not available.

In this work, we establish a framework for studying Algorithmic Gaussianization. Specifically, in our main result (Theorem 3), we characterize the total variation distance between a sketched data sample and the closest sub-gaussian random design. This approach can be used to show sub-gaussian properties for extremely sparse sketching transformations, including the so-called Leverage Score Sparsified (LESS) embeddings, whose time complexity matches that of other state-of-the-art techniques (Dereziński et al., 2021b). For these transformations, our result yields a sub-gaussian

---

1. This follows because $\mathbf{x}_i^\top\mathbf{w}^*$ and $y_i - \mathbf{x}_i^\top\mathbf{w}^*$ are uncorrelated and jointly Gaussian, and therefore, independent.

property known as the Hanson-Wright inequality (Rudelson and Vershynin, 2013), which is widely used to establish both asymptotic and non-asymptotic guarantees for statistical models. We also provide a matching lower bound (Theorem 5), which shows that some sketching methods, such as leverage score sampling, do not enjoy the same sub-gaussian properties.

We illustrate the strength of our framework on several examples, including least squares (Theorem 7), Lasso regression (Corollary 10), and Randomized SVD (Corollary 15 in Appendix A). In particular, as a second main result of independent interest, we show that the Gaussian approximation error formula (1) for the sketched OLS estimator can be non-asymptotically extended to efficient sparse sketches such as LESS embeddings. Namely, we show that for any $\mathbf{A}$ and $\mathbf{b}$, if $\mathbf{S}$ is a LESS embedding matrix, then the sketched OLS estimator satisfies:

$$\text{(Theorem 7)} \qquad \mathbb{E}\big[L(\widehat{\mathbf{w}}) - L(\mathbf{w}^*)\big] \approx_{1+\epsilon} \frac{d}{n-d-1} \cdot L(\mathbf{w}^*) \qquad \text{for} \qquad \epsilon = \tilde{O}(1/\sqrt{d}).$$

Here, $a \approx_{1+\epsilon} b$ denotes a $(1+\epsilon)$-approximation $b/(1+\epsilon) \le a \le (1+\epsilon)b$. This type of guarantee for sketched OLS is a first of its kind (outside of Gaussian embeddings), as prior work only showed upper bounds (with additional constant and logarithmic factors) of the form $\tilde{O}(d/n) \cdot L(\mathbf{w}^*)$ (e.g., see Woodruff, 2014), or asymptotic results (Dobriban and Liu, 2019). Remarkably, we show empirically (Appendix B) that for LESS embeddings the above approximation error estimate is extremely accurate, whereas for other popular sketching methods that are not covered by our theory (such as row sampling and SRHT), the approximation error can be much more problem dependent.

2. Main result: Hanson-Wright inequality for sparse sketches

In this section, we present our main result, showing that certain sparse sketching operators produce data sketches whose distributions are nearly indistinguishable from sub-gaussian random designs.

As there are a number of closely related notions of sub-gaussianity, we clarify this here, with a more detailed discussion in Section 5 (see Figure 3 for an illustration of different sub-gaussian properties). We say that a variable X is K-sub-gaussian if its corresponding sub-gaussian Orlicz norm, i.e., $\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}\exp(X^2/t^2) \le 2\}$, is bounded by K. For a random vector, perhaps the simplest and most popular model of sub-gaussianity is to assume that it has independent sub-gaussian entries. However, entry-wise independence is a fairly strong assumption. The multivariate sub-gaussian norm, i.e., $\|\mathbf{x}\|_{\psi_2} = \sup_{\mathbf{v}:\|\mathbf{v}\|=1}\|\mathbf{v}^\top\mathbf{x}\|_{\psi_2}$, is a popular relaxation that allows for sub-gaussian vectors with dependent entries. However, this notion is too weak for some applications, e.g., in high-dimensional statistics (Bai and Silverstein, 2010), or to obtain our least squares result (Theorem 7). These settings often rely on a stronger property, called the Hanson-Wright inequality, which has seen significant interest in the literature (e.g., Hsu et al., 2012; Rudelson and Vershynin, 2013; Adamczak, 2015; Vershynin, 2020; Bamberger et al., 2021). In the classical version given below (due to Rudelson and Vershynin, 2013), the inequality is established for vectors with independent sub-gaussian entries, however, in our main results, we demonstrate that it is applicable far beyond that setting. In the following statement, c denotes an absolute constant.

Lemma 1 (Hanson-Wright inequality) *Let $\mathbf{x}$ have independent K-sub-gaussian entries with mean zero and unit variance. Then, it satisfies the Hanson-Wright inequality with constant K:*

$$\Pr\Big\{|\mathbf{x}^\top\mathbf{B}\mathbf{x} - \text{tr}(\mathbf{B})| \ge t\Big\} \le 2\exp\bigg(-c\min\Big\{\frac{t^2}{K^4\|\mathbf{B}\|_F^2}, \frac{t}{K^2\|\mathbf{B}\|}\Big\}\bigg) \qquad \textit{for any } \mathbf{B}. \quad (2)$$
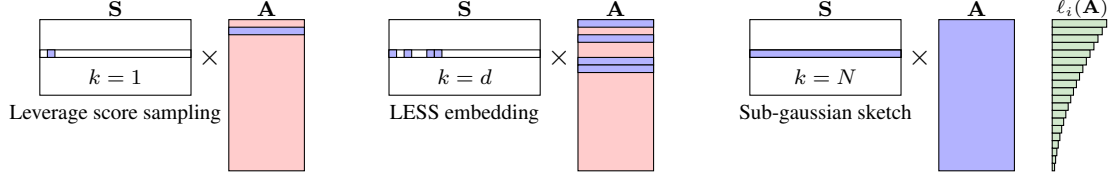
Figure 1: Illustration of p-sparsified sketches, highlighting the interaction of one row of $\mathbf{S}$ with $\mathbf{A}$.

Consider an $n \times d$ random matrix $\mathbf{X}$. We will say that $\mathbf{X}$ is a sub-gaussian random design satisfying Hanson-Wright inequality with constant K, if it can be decomposed so that $\mathbf{X} = \mathbf{Z}\mathbf{\Sigma}^{1/2}$, where $\mathbf{\Sigma}$ is the positive definite covariance matrix and $\mathbf{Z}$ consists of i.i.d. isotropic row vectors $\mathbf{z}_i^\top$ that satisfy (2). Note that the rows of $\mathbf{Z}$ may not have independent entries.

Before we present our main result for sparse sketches, let us start with a simpler example of dense sub-gaussian sketches. Consider an $N \times d$ data matrix $\mathbf{A}$ with full column rank and an $n \times N$ sketching matrix $\mathbf{S}$ consisting of independent random $\pm 1/\sqrt{n}$ entries (scaled random signs). The scaling in $\mathbf{S}$ is standard, and chosen so that the singular values of the resulting sketch $\mathbf{SA}$ are of the same order as the singular values of $\mathbf{A}$. Now, let $\mathbf{\Sigma} = \mathbf{A}^\top \mathbf{A}$ and $\mathbf{Z} = \sqrt{n}\mathbf{SU}$, where $\mathbf{U} = \mathbf{A}\mathbf{\Sigma}^{-1/2}$. Then, we have $\mathbf{SA} = \frac{1}{\sqrt{n}}\mathbf{Z}\mathbf{\Sigma}^{1/2}$, where the scaling by $1/\sqrt{n}$ could also be absorbed into $\mathbf{\Sigma}$. It is easy to see that the rows of $\mathbf{Z}$ are i.i.d. and isotropic. Moreover, since the rows of $\sqrt{n}\mathbf{S}$ have i.i.d. sub-gaussian entries with $K = O(1)$, and the Hanson-Wright inequality is preserved under the transformation from $\mathbf{x}$ to $\mathbf{U}^\top \mathbf{x}$, it follows from Lemma 1 that the rows of $\mathbf{Z}$ also satisfy (2) with the same constant K. Even in this example, the rows of $\mathbf{Z}$ do not have independent entries, which is why we use inequality (2) to define sub-gaussianity, rather than the assumptions of Lemma 1.

We next show that a similar reduction can be achieved for a class of sparse sketching matrices.

Definition 2 (p**-sparsified sub-gaussian sketch**) *Let* $p = (p_1, ..., p_N)$ *be a probability distribution. A p-sparsified sub-gaussian matrix $\mathbf{S}$ of size n with k non-zeros per row has n i.i.d. row vectors* $\frac{1}{\sqrt{n}}\sum_{i=1}^{k} \frac{r_i}{\sqrt{kp_{I_i}}}\mathbf{e}_{I_i}^\top$, *where $I_i \sim p$ and r_i are i.i.d. mean zero, unit variance and $O(1)$-sub-gaussian.*

In the following theorem, our main result, we show that many sparse sketches as defined above are very close to a sub-gaussian random design that satisfies the Hanson-Wright inequality with a small constant K. The closeness is measured using the total variation distance, denoted by d_{tv}, which allows transferring virtually any property from the design matrix to the sketch. The best guarantees are obtained when the sparsifying distribution p is close to the so-called leverage score distribution of matrix $\mathbf{A}$ (Drineas et al., 2006): the ith leverage score of a rank d matrix $\mathbf{A}$ is defined as $\ell_i(\mathbf{A}) = \mathbf{a}_i^\top (\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{a}_i$, where $\mathbf{a}_i^\top$ is the ith row of $\mathbf{A}$. Note that we have $\sum_{i=1}^{N}\ell_i(\mathbf{A}) = d$. Also, in the statement, we use $\mathbf{C} \approx_{1+\epsilon} \mathbf{D}$ to denote a $(1+\epsilon)$-approximation between two positive semidefinite (psd) matrices $\mathbf{C}$ and $\mathbf{D}$ in terms of the Loewner psd ordering.

Theorem 3 (Main result) *Consider an $N \times d$ matrix $\mathbf{A}$ with rank d, and a p-sparsified sub-gaussian matrix $\mathbf{S}$ of size n with k non-zeros per row. Define $\mu := \max_i \ell_i(\mathbf{A})/p_i \geq d$. For any $\delta > 0$, there is a $d \times d$ psd matrix $\widetilde{\mathbf{\Sigma}}$ and an $n \times d$ random matrix $\mathbf{Z}$ with i.i.d. mean zero isotropic row vectors, each satisfying the Hanson-Wright inequality (2) with $K = O\big(1 + \sqrt{\mu \log(\mu n/\delta)/k}\big)$, such that:*

$$d_{\mathrm{tv}}\big(\mathbf{SA}, \tfrac{1}{\sqrt{n}}\mathbf{Z}\widetilde{\mathbf{\Sigma}}^{1/2}\big) \leq \delta \qquad and \qquad \widetilde{\mathbf{\Sigma}} \approx_{1+\delta} \mathbf{A}^\top \mathbf{A}.$$
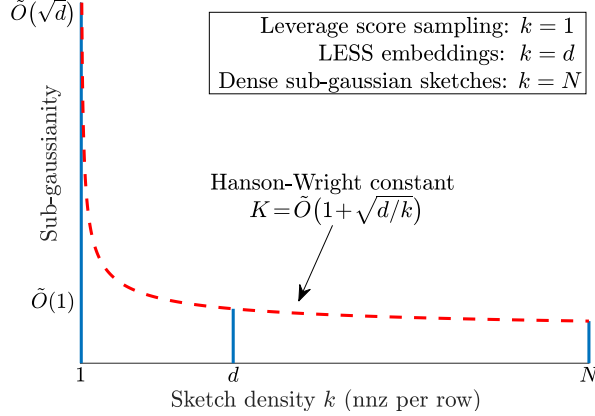
4

Figure 2: Illustration for Theorem 3. Matching lower bound is given in Theorem 5. Note that a p-sparsified sketch with $k = N$ is not strictly equivalent to a standard sub-gaussian sketch, due to with-replacement sampling of non-zero entries (chosen to simplify the analysis).

To illustrate this result, consider the scenario where $N \gg d$ and p is the approximate leverage score sampling distribution, i.e., $p_i \approx_{O(1)} \ell_i(\mathbf{A})/d$ for all i, in which case $\mu = \Theta(d)$ and so the Hanson-Wright constant satisfies $K = \tilde{O}(1 + \sqrt{d/k})$. In this case, as shown in Figure 1 (where the rows of $\mathbf{A}$ are sorted by leverage scores for the purpose of illustration), as we vary sketch density k, i.e., the number of non-zeros per row of $\mathbf{S}$, we recover: leverage score sampling for $k = 1$, LESS embeddings for $k = d$, and dense sub-gaussian sketches for $k = N$ (up to minor differences in the definitions, which do not affect the conclusions). For leverage score sampling, we can only get $K = \tilde{O}(\sqrt{d})$, which is essentially vacuous. However, to get the optimal $K = \tilde{O}(1)$, we only need $k = d \ll N$. Increasing k further has little effect on the Hanson-Wright constant, as we observe in Figure 2, which illustrates the dependence of K on sketch density. In Theorem 5 below, we provide a nearly-matching lower bound of $K = \tilde{\Omega}(1 + \sqrt{d/k})$, which in particular confirms that leverage score sampling does not enjoy a near-optimal Hanson-Wright guarantee.

Remark 4 *Below are the key examples of sparse sketches that fall under Theorem 3, for which the Hanson-Wright constant satisfies $K = \tilde{O}(1)$:*

1. **LESS embeddings.** *Here, p is an approximate leverage score sampling distribution, i.e., we have $p_i \approx_{O(1)} \ell_i(\mathbf{A})/d$ for all i. The number of non-zeros needs to satisfy $k = \Omega(d)$. Sketching takes $O(\mathrm{nnz}(\mathbf{A}) \log N + nd^2)$ time, where $\mathrm{nnz}(\mathbf{A})$ is the number of non-zero entries in $\mathbf{A}$.*

2. **Uniformly sparsified sketches.** *We use a uniform distribution p, i.e., $p_i = \frac{1}{N}$ for all i. The number of non-zeros needs to satisfy $k = \Omega(N \cdot \max_i \ell_i(\mathbf{A}))$, which can be anywhere between d and N, depending on $\mathbf{A}$. Here, sketching takes $O(ndk)$ time.*

3. **Preconditioned sparse sketches.** *Replacing $\mathbf{A}$ with $\tilde{\mathbf{A}} = \mathbf{HDA}$, where $\mathbf{H}$ is a fast Hadamard transform and $\mathbf{D}$ is diagonal with random ± 1 entries, we can ensure that all leverage scores are nearly uniform: $\ell_i(\tilde{\mathbf{A}}) \approx d/N$ (Tropp, 2011), while preserving the data covariance, $\tilde{\mathbf{A}}^\top \tilde{\mathbf{A}} = \mathbf{A}^\top \mathbf{A}$. Then, the uniformly sparsified sketch needs only $k = \Omega(d)$ non-zeros. This procedure takes $O(Nd \log N + nd^2)$ time for preconditioning and sketching.*

Based on the above result, LESS embeddings and preconditioned sparse sketches provide the most effective Gaussianizing property for arbitrary data matrices $\mathbf{A}$. However, simpler uniformly sparsified sketches can be equally effective for certain input matrices (depending on the leverage score distribution). We note that the nearly-linear time complexity of LESS embeddings is primarily based on the cost of approximating the leverage scores of $\mathbf{A}$ (Drineas et al., 2012), which can be done in $O(\mathrm{nnz}(\mathbf{A}) \log N + d^3 \log(d))$ time. Thus, when using LESS embeddings, our Hanson-Wright reduction can be achieved in time nearly linear in the input size (a.k.a. input sparsity time), with its complexity matching state-of-the-art sketching methods up to logarithmic factors (see Section 5.2).

Certain sparse sketches, such as CountSketch or SJLT, do not technically fit into our framework (we assume row-wise independence, while they have column-wise independence). Yet, these methods align closely with uniformly sparsified sketches (e.g., for CountSketch, with $k \approx N/n$ non-zeros per row). This suggests that the sparsity needed to recover sub-gaussian behavior with these methods depends both on the leverage score distribution and the dimensions of the data matrix.

Next, we demonstrate the sharpness of Theorem 3 by constructing examples of data matrices and p-sparsified sketches for which one cannot find a sufficiently good sub-gaussian random design that is close in total variation distance. Importantly, the result gives a lower bound of $K = \tilde{\Omega}(\sqrt{d})$ for leverage score sampling (i.e., $k = 1$), and shows that LESS embeddings require $k = \tilde{\Omega}(d)$ non-zeros per row to get $K = \tilde{O}(1)$, matching our upper bounds up to logarithmic factors (see Figure 2).

Theorem 5 (Lower bound) *For any $N \geq d$, there is a rank d matrix $\mathbf{A} \in \mathbb{R}^{N \times d}$ with the property: Given any $n, k \geq 1$, let $\mathbf{S}$ be a p-sparsified sub-gaussian matrix of size n with k non-zeros per row, where variables r_i are random ± 1 signs, and the distribution p satisfies $p_i \approx_{O(1)} \ell_i(\mathbf{A})/d$ for all i. Suppose there is $\mathbf{Z}$ having i.i.d. isotropic row vectors satisfying Hanson-Wright inequality (2) with K, such that $d_{\mathrm{tv}}(\mathbf{S}\mathbf{A}, \frac{1}{\sqrt{n}}\mathbf{Z}\widetilde{\Sigma}^{1/2}) \leq 1/2$ for some $\widetilde{\Sigma} \approx_{O(1)} \mathbf{A}^\top \mathbf{A}$. Then:*

$$K = \Omega\Big(1 + \sqrt{\tfrac{d/k}{\log d}}\Big).$$

Remark 6 *Recall that when $p_i \approx_{O(1)} \ell_i(\mathbf{A})/d$ (i.e., approximate leverage score sampling), then $\mu = \Theta(d)$ in Theorem 3, so the lower bound matches our upper bound up to logarithmic factors for leverage score sampling and LESS embeddings. In Appendix G we show an even stronger result, lower bounding the sub-gaussian norm $\| \cdot \|_{\psi_2}$ instead of the Hanson-Wright constant K, so that $\mathbf{Z}$ cannot be a sub-gaussian design even in this weaker sense (see Figure 3 for comparison).*

To demonstrate the benefits of our main results, recall that small total variation distance, namely $d_{\mathrm{tv}}(\mathbf{S}\mathbf{A}, \frac{1}{\sqrt{n}}\mathbf{Z}\widetilde{\Sigma}^{1/2}) \leq \delta$, means that we can couple the sketch $\mathbf{S}\mathbf{A}$ with the sub-gaussian design so that they are identical with probability $1 - \delta$. This effectively allows us to transfer any property that holds with high probability for the sub-gaussian design to a property that holds for the sketch. The only caveat is that the covariance matrix gets slightly distorted in the process, but both this distortion and the failure probability are controlled by δ which can be made negligibly small. We use this in Section 3 to show new guarantees for sparse sketches in least squares and Lasso regression.

3. Applications: Least squares and Lasso regression

We next present some of the implications of Theorem 3. In Section 3.1, we present our main application, giving a precise characterization of the expected approximation error of the sketched least

squares estimator with LESS embeddings. Here, this is not merely a corollary of our main result, but rather a result of independent interest, and to our knowledge, it was not previously known even for dense sub-gaussian sketches. Then, in Section 3.2, we demonstrate how our results imply new efficient algorithms for approximately solving Lasso regression and other constrained optimization tasks. Further applications to low-rank approximation, as well as corollaries for the subspace embedding and Johnson-Lindenstrauss properties of LESS embeddings, are given in Appendix A.

3.1. Sketched least squares with LESS embeddings

As a motivating application of our theory, we study the approximation properties of the sketched least squares estimator, showing that, when the sketch is a LESS embedding, then it achieves nearly the same expected approximation error as for a Gaussian embedding, even down to constant factors.

Consider an $N \times d$ data matrix $\mathbf{A}$ with full column rank, where $N \gg d$, and an N-dimensional vector $\mathbf{b}$. Recall that our goal is to approximately solve the least squares task $(\mathbf{A}, \mathbf{b})$, i.e., find:

$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} L(\mathbf{w}) = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{b}, \quad \text{where} \quad L(\mathbf{w}) = \|\mathbf{A}\mathbf{w} - \mathbf{b}\|^2.$$

Given an $n \times N$ sketching matrix $\mathbf{S}$, the sketched least squares estimator $\widehat{\mathbf{w}}$ for this model is the solution of the sketched version of the problem, i.e., $(\mathbf{SA}, \mathbf{Sb})$. If $\mathbf{S}$ is a Gaussian embedding, then via a reduction to a Gaussian random design, the expected error of this estimator is given by the exact formula (1), but the cost of constructing such a sketch is $O(Ndn)$, which is prohibitively expensive. If we let $\mathbf{S}$ be a LESS embedding matrix with d non-zeros per row (see Remark 4), then the cost of computing this estimator is $O(\mathrm{nnz}(\mathbf{A}) \log N + nd^2)$, which includes the cost of approximating the leverage scores, sketching the data, and computing $\widehat{\mathbf{w}}$ from the sketch. We show that the resulting estimator enjoys nearly the same expected error guarantee as the Gaussian embedding. To avoid numerical precision issues which may occur when computing the exact expectation (there is a small but non-zero probability that $\mathbf{SA}$ is very ill-conditioned), we assume that: the LESS embedding matrix is constructed so that the random variables r_i from Definition 2 are random ± 1 sign variables, and the probability distribution satisfies $p_i \approx_{O(1)} \ell_i(\mathbf{A})/d + 1/N$. We also restrict the entries of the sketched solution $\widehat{\mathbf{w}}$ to a finite range $\mathcal{R} = [-D, D]$, for a sufficiently large $D = \mathrm{poly}(N, \kappa(\mathbf{A}), \|\mathbf{A}\|, \|\mathbf{b}\|)$ which is absorbed by the logarithmic factors in the result (see Appendix E for details).

Theorem 7 *Fix $\mathbf{A}$ of rank d and let $\mathbf{S}$ be a LESS embedding as above, with $n \geq \tilde{O}(d)$ rows and $k = d$ non-zeros per row. For any $\mathbf{b}$, the estimator $\widehat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathcal{R}^d} \|\mathbf{S}(\mathbf{A}\mathbf{w} - \mathbf{b})\|^2$ satisfies:*

$$\mathbb{E}\big[L(\widehat{\mathbf{w}}) - L(\mathbf{w}^*)\big] \approx_{1+\epsilon} \frac{d}{n - d - 1} \cdot L(\mathbf{w}^*) \qquad \text{for} \qquad \epsilon = \tilde{O}(1/\sqrt{d}),$$

where $L(\mathbf{w}) = \|\mathbf{A}\mathbf{w} - \mathbf{b}\|^2$ and $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w}} L(\mathbf{w})$.

The result easily extends to any sketching method covered by Remark 4, as long as the entries of $\mathbf{S}$ are almost surely bounded, as well as to dense random ± 1 matrices, affecting only the polylogarithmic factors hidden in $\tilde{O}$.

Remark 8 *The obtained expression for the expected multiplicative approximation error, i.e., $\frac{d}{n-d-1}$, is completely independent of the data matrix $\mathbf{A}$ or the regression vector $\mathbf{b}$, and it matches the formula (1) for the Gaussian embedding. Comparable non-asymptotic guarantees for other fast sketching methods, such as SJLT, SRHT, or leverage score sampling, take the form of $\tilde{O}(d/n) \cdot L(\mathbf{w}^*)$,*

7

where the precise constant and logarithmic terms are problem-dependent (e.g., see Woodruff, 2014; Chen and Price, 2019). To our knowledge, this result was not previously known even for dense sub-gaussian sketches.

Proof sketch   The proof, given in Appendix E, proceeds differently from the existing bounds for sketched least squares. Our approach is inspired by the asymptotic analysis of the OLS estimator in high-dimensional statistics (e.g., Hastie et al., 2019), but we obtain a non-asymptotic bound and we avoid any statistical assumptions on the label noise. We start by rewriting the expected loss $\mathbb{E}[L(\widehat{\mathbf{w}})]$ as the expectation of the leave-one-out cross-validation estimator, $L_{\mathrm{CV}} = \sum_{i=1}^{n}(\tilde{\mathbf{a}}_i^\top \widehat{\mathbf{w}}^{-i} - \tilde{b}_i)^2$, where $(\tilde{\mathbf{a}}_i^\top, \tilde{b}_i)$ is the ith row of the sketched problem $(\mathbf{SA}, \mathbf{Sb})$ and $\widehat{\mathbf{w}}^{-i}$ is computed by excluding this row from the sketch. We then use the Hanson-Wright inequality via Theorem 3 to show that all of the leverage scores of the sketch are with high probability very close to uniform (Lemma 22):

$$\ell_i(\mathbf{SA}) \approx_{1+\epsilon} \frac{d}{n} \quad \text{with} \quad \epsilon = \tilde{O}(1/\sqrt{d}\,) \quad \text{for all } i.$$

Using a standard short-cut formula, this allows us to approximate the expectation of L_{CV} as follows:

$$\mathbb{E}[L_{\mathrm{CV}}] \approx_{1+\epsilon'} \left(1 - \tfrac{d}{n}\right)^{-2} \mathbb{E}\big[\|\mathbf{S}(\mathbf{A}\widehat{\mathbf{w}} - \mathbf{b})\|^2\big] \quad \text{with} \quad \epsilon' = \tilde{O}(\sqrt{d}/n).$$

Finally, our key contribution in the analysis is a new approximate expectation formula for the sketch of the so-called *hat matrix*, $\mathbf{H} = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}$ (i.e., the orthogonal projection onto the column-span of $\mathbf{A}$), which arises when deriving the optimum least squares loss: $L(\mathbf{w}^*) = \mathbf{b}^\top(\mathbf{I} - \mathbf{H})\mathbf{b}$. Specifically, in Lemma 23 we show that the hat matrix $\widehat{\mathbf{H}}$ of the sketch $\mathbf{SA}$ satisfies:

$$\mathbb{E}\big[\mathbf{S}^\top(\mathbf{I} - \widehat{\mathbf{H}})\mathbf{S}\big] \approx_{1+\epsilon'} \left(1 - \tfrac{d}{n}\right) \cdot (\mathbf{I} - \mathbf{H}) \quad \text{with} \quad \epsilon' = \tilde{O}(\sqrt{d}/n),$$

where $\approx_{1+\epsilon}$ is defined in terms of the psd matrix ordering, and the formula holds exactly when $\mathbf{S}$ is a Gaussian embedding. The result then follows because $\mathbb{E}[\|\mathbf{S}(\mathbf{A}\widehat{\mathbf{w}} - \mathbf{b})\|^2] = \mathbf{b}^\top\mathbb{E}[\mathbf{S}^\top(\mathbf{I} - \widehat{\mathbf{H}})\mathbf{S}]\mathbf{b}$.

Experiments   In Appendix B we empirically show that our estimate for the expected approximation error of sketched OLS with a LESS embedding is even more accurate than suggested by the theory, on a range of benchmark regression tasks. Our experiments also show that this problem-independent expected approximation error is not always shared by other fast sketching methods that are considered to have strong guarantees for least squares, such as SRHTs.

3.2. Lasso and constrained least squares

In this section, we show how our results can be used to extend existing sketching guarantees in constrained least squares optimization from dense sub-gaussian sketches to LESS embeddings. This general problem setting includes such standard tasks as Lasso regression, sparse recovery and training a support vector machine.

We consider the following constrained optimization task for $\mathbf{A} \in \mathbb{R}^{N \times d}$ and $\mathbf{b} \in \mathbb{R}^N$:

$$\text{find} \qquad \mathbf{w}^* = \underset{\mathbf{w} \in \mathcal{C}}{\operatorname{argmin}} \, L(\mathbf{w}), \qquad \text{for} \qquad L(\mathbf{w}) = \|\mathbf{A}\mathbf{w} - \mathbf{b}\|^2,$$

where $\mathcal{C}$ is some convex subset of $\mathbb{R}^d$. Our goal is to approximately solve this task by solving a sketched version of the problem: $\widehat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathcal{C}} \|\mathbf{S}(\mathbf{A}\mathbf{w} - \mathbf{b})\|^2$, for an $n \times N$ sketching

matrix $\mathbf{S}$. When the sketching matrix is a sub-gaussian embedding, then Pilanci and Wainwright (2015) showed that the sketch size n needed for $\widehat{\mathbf{w}}$ to achieve a $(1+\epsilon)$-approximation, i.e., $L(\widehat{\mathbf{w}}) \leq (1+\epsilon) \cdot L(\mathbf{w}^*)$, is controlled by the so-called Gaussian width, which can be viewed as a measure of the degrees of freedom for the constrained optimization task (in particular, it can be much smaller than the actual dimension d), and is defined as follows:

$$\mathbb{W} := \mathbb{W}(\mathbf{A}\mathcal{K}) = \mathbb{E}_{\mathbf{g}}\Big[\sup_{\mathbf{u} \in \mathbf{A}\mathcal{K} \cap \mathcal{S}^{N-1}} |\mathbf{g}^\top \mathbf{u}| \Big],$$

where $\mathbf{g}$ has i.i.d. standard Gaussian entries, $\mathcal{S}^{N-1}$ is the unit sphere in $\mathbb{R}^N$, and $\mathcal{K}$ is the tangent cone of the constraint set $\mathcal{C}$ at the optimum $\mathbf{w}^*$, i.e., the closed convex hull of $\{\Delta \in \mathbb{R}^d \mid \Delta(\mathbf{w} - \mathbf{w}^*)$ for $t \geq 0, \mathbf{w} \in \mathcal{C}\}$. Specifically, they showed that if $\mathbf{S}$ consists of i.i.d. mean zero isotropic sub-gaussian rows and $n \geq O(\mathbb{W}^2/\epsilon^2)$, then with high probability $\widehat{\mathbf{w}}$ is a $(1+\epsilon)$-approximation of $\mathbf{w}^*$. We use Theorem 3 to extend this result to LESS embeddings (see Appendix F).

Theorem 9 *If $\mathbf{S}$ is a LESS embedding matrix for $\mathbf{A}$, with size $n \geq \tilde{O}(\mathbb{W}^2/\epsilon^2)$, then with high probability $\widehat{\mathbf{w}}$ is a $(1+\epsilon)$-approximation of $\mathbf{w}^*$.*

We illustrate this claim with a corollary for Lasso regression, where the constraint set is an ℓ_1 ball, i.e., $\mathcal{C} = \{\mathbf{w} : \|\mathbf{w}\|_1 \leq R\}$. Crucially, here the sketch size needed to obtain a good approximation scales with the sparsity of the solution $\mathbf{w}^*$, rather than with the dimension of the data. First, we formulate an ℓ_1-restricted condition number of $\mathbf{A}$, which is always smaller than the actual condition number, and which naturally arises in bounding the Gaussian width:

$$\kappa = \max_i \|\mathbf{A}_{:,i}\|/\gamma_s^-(\mathbf{A}), \qquad \text{where} \qquad \gamma_s^-(\mathbf{A}) = \min_{\|\mathbf{v}\|=1, \, \|\mathbf{v}\|_1 \leq 2\sqrt{s}} \|\mathbf{A}\mathbf{v}\|.$$

Corollary 10 *Suppose that the Lasso solution $\mathbf{w}^*$ is s-sparse. Then, using a LESS embedding of size $n \geq \tilde{O}(\kappa^2 s/\epsilon^2)$, we can obtain a $(1+\epsilon)$-approximation of $\mathbf{w}^*$ in time $O(\mathrm{nnz}(\mathbf{A})\log N + nd^2)$.*

Finally, we note that similar results for approximately solving Lasso and other constrained least squares problems can be obtained for SRHT sketches. However, there are two key differences: (1) the bounds obtained for SRHT are weaker than the ones achieved for LESS or sub-gaussian sketches, e.g., in addition to the Gaussian width, they also scale with other problem-dependent quantities such as the Rademacher width (see Theorem 2 of Pilanci and Wainwright, 2015); and (2) the cost of SRHT is $O(Nd\log N)$, which is higher than the cost of LESS when $\mathrm{nnz}(\mathbf{A}) \ll Nd$.

4. Key technical result: A Hanson-Wright limit theorem

In this section, we present the key technical result that we use to prove Theorem 3. This result shows that, by relying on the central limit phenomenon, we can establish that a sum of bounded random vectors gets very close, in terms of the total variation distance d_{tv}, to a random vector that satisfies the Hanson-Wright inequality (2) with a small constant K, even though the sum may not satisfy this property directly. There has been extensive literature dedicated to extending and adapting the Hanson-Wright inequality (e.g., Hsu et al., 2012; Adamczak, 2015; Vershynin, 2020; Bamberger et al., 2021), and our result is of independent interest in the context of this line of works.

Theorem 11 *Let $\mathbf{x}_1, \mathbf{x}_2, ...$ be i.i.d. random vectors, where $\mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\top] = \mathbf{\Sigma}$ is full-rank and $\|\mathbf{\Sigma}^{-1/2}\mathbf{x}_i\|_2$ is M-sub-gaussian. Define a k-gaussianized sample $\tilde{\mathbf{x}} = \frac{1}{\sqrt{k}} \sum_{i=1}^k r_i \mathbf{x}_i$, where r_i are independent mean zero unit variance R-sub-gaussian random variables. For any $\delta > 0$, there is $L = O\big(1 + M \log(M/\delta)/\sqrt{k}\big)$, a $d \times d$ psd matrix $\widetilde{\mathbf{\Sigma}}$, and a mean zero isotropic vector $\mathbf{z}$ satisfying the Hanson-Wright inequality* (2) *with $K = LR$, such that:*

$$d_{\mathrm{tv}}\big(\tilde{\mathbf{x}}, \widetilde{\mathbf{\Sigma}}^{1/2}\mathbf{z}\big) \le \delta \qquad and \qquad \widetilde{\mathbf{\Sigma}} \approx_{1+\delta} \mathbf{\Sigma}.$$

Remark 12 *Note that $M^2 \ge \mathbb{E}[\|\mathbf{\Sigma}^{-1/2}\mathbf{x}_i\|_2^2] = d$, where d is the dimension of $\mathbf{x}_i$. If we make a stronger assumption that $\|\mathbf{\Sigma}^{-1/2}\mathbf{x}_i\|_2 \le M$ almost surely, then we can obtain a slightly sharper result: $L = O(1 + M\sqrt{\log(M/\delta)/k})$, which is what we use to establish Theorem 3. Either of the boundedness assumptions includes all distributions over finite populations in general position.*

The proof, which can be found in Appendix D, is briefly summarized here. We start by defining a high-probability event $\mathcal{E}$ ensuring that the sample covariance matrix $\frac{1}{k} \sum_i \mathbf{x}_i \mathbf{x}_i^\top$ is sufficiently well bounded. In the process, we establish a bound on the matrix moments of the random covariates $\mathbf{x}_i \mathbf{x}_i^\top$ (Lemma 20). Then, we construct the vector $\mathbf{z}$ so that it is coupled with $\tilde{\mathbf{x}}$ when the event $\mathcal{E}$ holds. This requires introducing a correction term designed to minimize the distortion of the covariance matrix of $\mathbf{z}$, so as to ensure that $\widetilde{\mathbf{\Sigma}} \approx_{1+\delta} \mathbf{\Sigma}$. We then establish the Hanson-Wright inequality for $\mathbf{z}$ in three steps: First, analyzing the error coming from the r_i's; then accounting for the randomness in $\mathbf{x}_i$'s; and finally, bounding the noise coming from the correction term.

At a high level, Theorem 11 states that as k goes to infinity, the total variation distance of $\tilde{\mathbf{x}}$ from a certain family of sub-gaussian distributions goes to zero. Using the total variation distance to measure central limit behavior (instead of, say, the Wasserstein distance) is unusual because of how strong of a guarantee it requires (see Section 5.2 for discussion). This is possible here because we are comparing to a sub-gaussian distribution, instead of directly to a Gaussian.

Note that the lower bound we provide in Theorem 5 for p-sparsified sketches also constitutes a lower bound for Theorem 11. In particular, Theorem 5 implies that there are distributions of $\mathbf{x}_i$ such that $M = \Theta(\sqrt{d})$, for which $K = \tilde{\Omega}(1 + M/\sqrt{k})$. This, up to logarithmic factors, matches our upper bound of $K = \tilde{O}(1 + M/\sqrt{k})$ samples, and can be easily extended to $M \gg \sqrt{d}$.

Proof of Theorem 3 We now briefly explain how Theorem 3 follows from Theorem 11. Consider an $N \times d$ matrix $\mathbf{A}$ with rank d, and let $\mathbf{\Sigma} = \mathbf{A}^\top \mathbf{A}$. Suppose that the random vectors $\mathbf{x}_i = \frac{1}{\sqrt{p_{I_i}}}\mathbf{a}_i^\top$ are defined as scaled row samples from $\mathbf{A}$, drawn using the index distribution $I_i \sim p$ and scaled so that $\mathbf{\Sigma} := \mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\top] = \mathbf{A}^\top \mathbf{A}$. Now, let us construct an $n \times d$ random matrix $\widetilde{\mathbf{X}}$ consisting of rows $\tilde{\mathbf{x}}_1, ..., \tilde{\mathbf{x}}_n$ which are independent k-gaussianized samples defined as in Theorem 11. This sketch can be obtained by applying a p-sparsified sketching matrix $\sqrt{n}\mathbf{S}$ to $\mathbf{A}$, where each row of $\sqrt{n}\mathbf{S}$ is given by $\sum_{i=1}^k \frac{r_i}{\sqrt{kp_{I_i}}}\mathbf{e}_{I_i}^\top$. Notice that $\|\mathbf{\Sigma}^{-1/2}\mathbf{x}_i\|_2 \le \sqrt{\ell_{I_i}(\mathbf{A})/p_{I_i}} \le \sqrt{\mu}$ (as defined in Theorem 3), so we can let the constant M in Theorem 11 be $\sqrt{\mu}$, where recall that $\ell_i(\mathbf{A})$ is the ith leverage score of matrix $\mathbf{A}$, i.e., $\ell_i(\mathbf{A}) = \mathbf{a}_i^\top(\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{a}_i = \|\mathbf{\Sigma}^{-1/2}\mathbf{a}_i\|^2$. Note that we can rely here on the stronger boundedness condition as in Remark 12, so that we obtain $d_{\mathrm{tv}}(\tilde{\mathbf{x}}_i, \widetilde{\mathbf{\Sigma}}^{1/2}\mathbf{z}) \le \delta$ for each i, with $K = O(1 + \sqrt{\mu \log(\mu/\delta)/k})$. Recall that small total variation distance implies that we can couple $\tilde{\mathbf{x}}_i$ with a corresponding $\widetilde{\mathbf{\Sigma}}^{1/2}\mathbf{z}_i$, where $\mathbf{z}_i$ is an independent sample of $\mathbf{z}$, so that they are identical with probability $1 - \delta$. It remains to apply a union bound over the n random vectors $\tilde{\mathbf{x}}_1, ..., \tilde{\mathbf{x}}_n$ and their corresponding coupled samples $\mathbf{z}_1, ..., \mathbf{z}_n$, to show the total variation distance of $n\delta$ between

Examples
$$\mathbf{x} \in \mathbb{R}^d$$

Concentration
$$\mathcal{F} \subseteq \{\mathbb{R}^d \to \mathbb{R}\}$$

Gaussian vectors

uniform on the sphere

Lipschitz functions

i.i.d. bounded entries

Convex functions

i.i.d. sub-gaussian entries

Euclidean functions
$f(\mathbf{x}) = \sqrt{\mathbf{x}^\top \mathbf{B} \mathbf{x}}$

Hanson-Wright vectors

sub-gaussian vectors

Linear functions
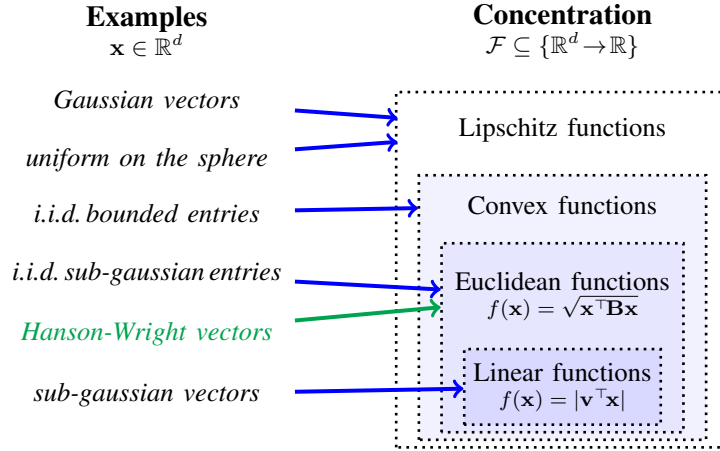$f(\mathbf{x}) = |\mathbf{v}^\top \mathbf{x}|$

Figure 3: Hierarchy of sub-gaussian concentration for mean zero isotropic random vectors. Right column represents different function classes $\mathcal{F}$ used in Definition 13, and left column provides examples of random vectors that satisfy the property with a particular function class (the larger the function class, the stronger the concentration property). "Hanson-Wright vectors" refers to vectors satisfying the Hanson-Wright inequality (2), whereas "sub-gaussian vectors" have bounded multivariate sub-gaussian norm $\|\cdot\|_{\psi_2}$.

the entire sketch $\sqrt{n}\mathbf{S}\mathbf{A}$ and the corresponding sub-gaussian design $\mathbf{Z}\widetilde{\boldsymbol{\Sigma}}^{1/2}$. Replacing δ with δ/n, we obtain the claim of Theorem 3.

5. Background and related work

In this section, we put our results in context by discussing different types of sub-gaussianity that have been considered in the literature. Here, we rely on the hierarchy of sub-gaussian concentration, shown in Figure 3, which is based on a range of classical results in high-dimensional probability (e.g., see Vershynin, 2018). We then discuss how these concepts arise in related work, particularly in the context of randomized sketching.

5.1. Sub-gaussian concentration hierarchy for random vectors

While the notion of sub-gaussian concentration for scalar random variables can be naturally represented by the Orlicz norm $\|\cdot\|_{\psi_2}$, the landscape of sub-gaussian concentration becomes more complex when we consider multivariate distributions. As discussed in Section 2, the natural multivariate extension of the Orlicz norm, i.e., the sub-gaussian norm $\|\mathbf{x}\|_{\psi_2} = \sup_{\mathbf{v}:\|\mathbf{v}\|=1} \|\mathbf{v}^\top \mathbf{x}\|_{\psi_2}$, while useful for some applications (such as Theorem 9), does not suffice for other Gaussian-like guarantees (such as Theorem 7). A more general approach to quantifying sub-gaussianity of random vectors is to analyze the concentration of real-valued functions of these vectors. In this context, let $\mathcal{F}$ be some subset of functions $f : \mathbb{R}^d \to \mathbb{R}$ with a bounded Lipschitz constant defined as: $\|f\|_{\mathrm{Lip}} := \sup_{\mathbf{u},\mathbf{v}} |f(\mathbf{u}) - f(\mathbf{v})|/\|\mathbf{u} - \mathbf{v}\|$.

Definition 13 *We say that a d-dimensional random vector $\mathbf{x}$ has the concentration property over $\mathcal{F}$ with constant K if $\|f(\mathbf{x}) - \mathbb{E}\, f(\mathbf{x})\|_{\psi_2} \leq K \cdot \|f\|_{\mathrm{Lip}}$ for all $f \in \mathcal{F}$.*

For the sake of simplicity, we focus on random vectors that are mean zero, $\mathbb{E}[\mathbf{x}] = \mathbf{0}$, and isotropic, i.e., $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbf{I}$. Gaussian vectors (as well as, e.g., random vectors on the sphere) satisfy the concentration property over all Lipschitz functions, with an absolute constant K. However, as this condition is very restrictive, smaller function families have been considered. If we restrict $\mathcal{F}$ to convex functions, then vectors with independent bounded entries satisfy the concentration property (but vectors with i.i.d. sub-gaussian entries may not). On the other end of the concentration hierarchy, illustrated in Figure 3, is the family of linear functions, i.e., $\mathcal{F} = \{f : f(\mathbf{x}) = |\mathbf{v}^\top\mathbf{x}|, \text{ for some } \mathbf{v} \in \mathbb{R}^d\}$, for which the concentration constant of $\mathbf{x}$ becomes simply its sub-gaussian norm. Of primary interest to this work is the family of Euclidean functions (Vershynin, 2020), defined as $\mathcal{F} = \{f : f(\mathbf{x}) = \sqrt{\mathbf{x}^\top\mathbf{B}\mathbf{x}} \text{ for psd } \mathbf{B}\}$, which falls between convex and linear in the hierarchy, and can be used to characterize vectors with i.i.d. sub-gaussian entries.

Proposition 14 *(Vershynin, 2018, Section 6.3) If an isotropic mean zero vector satisfies the Hanson-Wright inequality (2), then it has $O(K)$ sub-gaussian norm and $O(K^2)$-Euclidean concentration.*

5.2. Related work

Randomized sketching has emerged out of theoretical computer science, as part of the broader area of Randomized Numerical Linear Algebra (RandNLA; Woodruff, 2014; Drineas and Mahoney, 2016; Dereziński and Mahoney, 2021), including computationally efficient techniques such as the Subsampled Randomized Hadamard Transform (SRHT; Ailon and Chazelle, 2009), Leverage Score Sampling (Drineas et al., 2006), the CountSketch (Clarkson and Woodruff, 2017), Sparse Johnson-Lindenstrauss Transforms (SJLT; Nelson and Nguyên, 2013), Leverage Score Sparsified embeddings (LESS; Dereziński et al., 2021b) and Determinantal Point Processes (DPPs; Dereziński, 2019). More recently, there has been an increased interest in the statistical analysis of sketching and importance sampling techniques, for example in the context of linear regression (Ma et al., 2014), kernel ridge regression (Alaoui and Mahoney, 2015), model averaging (Wang et al., 2017), and the bootstrap (Lopes et al., 2019). Some works have shown deep connections between the performance of certain sketching methods and Gaussian/sub-gaussian embeddings: for DPPs, e.g., in the context of low-rank approximation and stochastic optimization (Dereziński et al., 2020b); for LESS embeddings in the context of inverse covariance estimation and stochastic optimization (Dereziński et al., 2021a,b); and for SRHT, a similar phenomenon has been studied asymptotically in linear regression (Dobriban and Liu, 2019) and stochastic optimization (Lacotte et al., 2020). Despite this extensive literature, no general non-asymptotic equivalence result was known between any of these fast sketching methods and sub-gaussian embeddings. Our results show such equivalence for LESS embeddings, while also providing a negative result for leverage score sampling.

Statistical analysis of random design models has a long history, with important connections to both asymptotic and non-asymptotic random matrix theory (e.g., see Bai and Silverstein, 2010; Vershynin, 2018). A variety of random designs have been considered, of which the most relevant to our work are Gaussian and sub-gaussian designs. These models have proven extremely useful in understanding the performance of a variety of linear regression estimators (Dobriban and Wager, 2018; Hastie et al., 2019; Bayati and Montanari, 2011; Miolane and Montanari, 2021), sample covariance estimators (Koltchinskii and Lounici, 2017; Ledoit and Péché, 2011) and others. Here,

a sub-gaussian random design typically refers to a matrix $\mathbf{Z}\mathbf{\Sigma}^{1/2}$, where $\mathbf{Z}$ consists of independent rows which have either bounded sub-gaussian norm or i.i.d. sub-gaussian entries, or more generally, satisfy a concentration property for a family of Lipschitz functions (e.g., Louart and Couillet, 2018). Our gaussianization framework opens the possibility of extending many of these results to sketching.

Our results are related to the study of the rates of convergence in the multivariate central limit theorem (CLT), with the key difference that we relax our notion of gaussianity, which allows us to use a stronger notion of distance (i.e., the total variation distance, together with the Hanson-Wright constant). Nevertheless, it is helpful to compare our approach with the convergence rates for the multivariate CLT in terms of the Wasserstein distance (Chen and Shao, 2005; Bonis, 2020; Fang and Koike, 2022). Using the setup from our key technical result, Theorem 11, for a random vector $\mathbf{x}$ with covariance matrix $\mathbf{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ that satisfies $\|\mathbf{\Sigma}^{-1/2}\mathbf{x}\| \leq M$ almost surely, the currently best known Wasserstein CLT of order 2 yields $W_2(\tilde{\mathbf{x}}, \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})) = O(M\sqrt{d\log(k)/k})$ (Eldan et al., 2020), whereas our bound on the Hanson-Wright constant is $O(1 + M\sqrt{\log(M)/k})$. The guarantees are not directly comparable, but one could argue that both bounds are useful primarily when k is large enough to absorb the dependence on M and d, i.e., $k = O(M^2 d\log(M))$ for the Wasserstein bound and $k = O(M^2 \log(M))$ for our result. Also, while the Wasserstein distance can be used to bound the difference between expectations of Lipschitz functions, it is non-trivial to effectively bound the Lipschitz constant for most of the quantities considered in this work, such as the least squares approximation error, and using this approach would require a more complex case-by-case analysis.

6. Conclusions and open questions

We provided the first general characterization of Algorithmic Gaussianization, which refers to the phenomenon that many algorithmic techniques used to construct small random representations (sketches) of large datasets produce data samples that are more Gaussian-like than the original data. In our main result, we showed that a sparse sketching matrix can be used to produce a sample that is nearly indistinguishable from a sub-gaussian random design. We used this to show a reduction between a fast sketching technique called LESS embeddings and a sub-gaussian random matrix whose rows satisfy the classical Hanson-Wright inequality. Our lower bound showed that the result is nearly tight up to logarithmic factors, and that the sub-gaussian reduction is not possible for leverage score sampling. We demonstrated how our techniques can be used to provide improved guarantees for sketched estimators in least squares, Lasso regression, and low-rank approximation. More broadly, our results point to new open questions related to the complexity of generating repeated samples from a product between a matrix and a Gaussian-like random vector:

Open question. Given an $N \times d$ matrix $\mathbf{A}$ and a function family $\mathcal{F}$ in $\mathbb{R}^d \to \mathbb{R}$ (as in Figure 3), what is the complexity of producing n independent samples of a d-dimensional random vector that is, up to total variation distance δ, distributed according to $\mathbf{A}^\top\mathbf{z}$, where $\mathbf{z}$ is isotropic and has the concentration property over $\mathcal{F}$ with constant $K = O(1)$, as given by Definition 13.

Acknowledgments

References

Radoslaw Adamczak. A note on the hanson-wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20:1–13, 2015.

Nir Ailon and Bernard Chazelle. The fast johnson–lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on computing*, 39(1):302–322, 2009.

Ahmed El Alaoui and Michael W. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 775–783, Montreal, Canada, December 2015.

Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Stefan Bamberger, Felix Krahmer, and Rachel Ward. The hanson-wright inequality for random tensors. *arXiv preprint arXiv:2106.13345*, 2021.

Mohsen Bayati and Andrea Montanari. The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2011.

Thomas Bonis. Stein's method for normal approximation in wasserstein distances with application to the multivariate central limit theorem. *Probability Theory and Related Fields*, 178(3):827–860, 2020.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

Louis HY Chen and Qi-Man Shao. Stein's method for normal approximation. *An introduction to Stein's method*, 4:1–59, 2005.

Xue Chen and Eric Price. Active regression via linear-sample sparsification. In *Proceedings of the 32nd Conference on Learning Theory*, 2019.

Kenneth L. Clarkson and David P. Woodruff. Low-rank approximation and regression in input sparsity time. *J. ACM*, 63(6):54:1–54:45, January 2017.

Michał Dereziński. Fast determinantal point processes via distortion-free intermediate sampling. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99, pages 1029–1049, 2019.

Michał Dereziński and Michael W Mahoney. Determinantal point processes in randomized numerical linear algebra. *Notices of the American Mathematical Society*, 68(1):34–45, 2021.

Michał Dereziński, Rajiv Khanna, and Michael W Mahoney. Improved guarantees and a multiple-descent curve for the column subset selection problem and the nyström method. In *Advances in Neural Information Processing Systems*, volume 33, pages 4953–4964, 2020a.

Michał Dereziński, Feynman Liang, Zhenyu Liao, and Michael W Mahoney. Precise expressions for random projections: Low-rank approximation and randomized Newton. In *Advances in Neural Information Processing Systems*, volume 33, pages 18272–18283, 2020b.

Michał Dereziński, Jonathan Lacotte, Mert Pilanci, and Michael W Mahoney. Newton-LESS: Sparsification without trade-offs for the sketched newton update. In *Advances in Neural Information Processing Systems*, volume 34, pages 2835–2847, 2021a.

Michał Dereziński, Zhenyu Liao, Edgar Dobriban, and Michael W Mahoney. Sparse sketches with small inversion bias. In *Proceedings of the 34th Conference on Learning Theory*, 2021b.

Edgar Dobriban and Sifan Liu. Asymptotics for sketching in least squares regression. In *Advances in Neural Information Processing Systems*, pages 3675–3685, 2019.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

Petros Drineas and Michael W. Mahoney. RandNLA: Randomized numerical linear algebra. *Communications of the ACM*, 59:80–90, 2016.

Petros Drineas, Michael W Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the Symposium on Discrete algorithms*, pages 1127–1136, 2006.

Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *J. Mach. Learn. Res.*, 13(1):3475–3506, December 2012.

Ronen Eldan, Dan Mikulincer, and Alex Zhai. The clt in high dimensions: quantitative bounds via martingale embedding. *The Annals of Probability*, 48(5):2494–2524, 2020.

Xiao Fang and Yuta Koike. From p-wasserstein bounds to moderate deviations. *arXiv preprint arXiv:2205.13307*, 2022.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53 (2):217–288, 2011.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. Technical Report Preprint: arXiv:1903.08560, 2019.

Pawel Hitczenko. Best constants in martingale version of rosenthal's inequality. *The Annals of Probability*, 18(4):1656–1668, 1990.

Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017.

Jonathan Lacotte, Sifan Liu, Edgar Dobriban, and Mert Pilanci. Optimal iterative sketching methods with the subsampled randomized hadamard transform. *Advances in Neural Information Processing Systems*, 33:9725–9735, 2020.

Olivier Ledoit and Sandrine Péché. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1-2):233–264, 2011.

Miles E Lopes, N Benjamin Erichson, and Michael W Mahoney. Bootstrapping the operator norm in high dimensions: Error estimation for covariance matrices and sketching. *arXiv preprint arXiv:1909.06120*, 2019.

Cosme Louart and Romain Couillet. Concentration of measure and large random matrices with an application to sample covariance matrices. *arXiv preprint arXiv:1805.08295*, 2018.

Ping Ma, Michael Mahoney, and Bin Yu. A statistical perspective on algorithmic leveraging. In *Proceedings of the 31st International Conference on Machine Learning*, pages 91–99, 2014.

Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis*, 17(4):1248–1282, 2007.

Léo Miolane and Andrea Montanari. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics*, 49(4):2313–2335, 2021.

Jelani Nelson and Huy L. Nguyên. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, FOCS '13, pages 117–126. IEEE Computer Society, 2013.

Mert Pilanci and Martin J Wainwright. Randomized sketches of convex programs with sharp guarantees. *IEEE Transactions on Information Theory*, 61(9):5096–5115, 2015.

Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Symposium on Foundations of Computer Science*, pages 143–152, 2006.

Joel A Tropp. Improved analysis of the subsampled randomized hadamard transform. *Advances in Adaptive Data Analysis*, 3(01n02):115–126, 2011.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, August 2012.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Roman Vershynin. Concentration inequalities for random tensors. *Bernoulli*, 26(4):3139–3162, 2020.

Shusen Wang, Alex Gittens, and Michael W. Mahoney. Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3608–3616, 06–11 Aug 2017.

David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Appendix A. More applications: Randomized SVD and low-distortion embeddings

In this section, we provide further examples of how Theorem 3 can be applied in combination with existing results for sub-gaussian random designs. In each of these examples, we can extend existing results from sub-gaussian embeddings (which are dense, and therefore not very efficient) to analogous results for LESS embeddings (which are sparse, and thus can be implemented much more efficiently).

A.1. Randomized SVD

An important application of randomized sketching is low-rank approximation, where our goal is to estimate a small number of top principal directions of an $N \times d$ data matrix $\mathbf{A}$. One of the most popular techniques in this area is the Randomized SVD algorithm (Halko et al., 2011), where we construct the approximation by projecting the dataset onto a subspace spanned by the rows of the sketch $\mathbf{SA}$, with $\mathbf{S}$ denoting the $n \times N$ sketching matrix, and $n \ll d$. The error of such an approximation is often measured by the sum of squared lengths of the residuals from the projection: $\|\mathbf{A} - \mathbf{A} \cdot \mathrm{Proj}_{\mathbf{SA}}\|_F^2$, where $\mathrm{Proj}_{\mathbf{SA}}$ denotes the projection onto the row-span of $\mathbf{SA}$ and $\|\cdot\|_F$ is the Frobenius norm. Prior work has shown strong approximation guarantees for Gaussian embeddings in this context, such as the following, given by Halko et al. (2011), comparing the expected error to the best rank k approximation, for some $k < n - 2$:

$$\mathbb{E}\Big[\big\|\mathbf{A} - \mathbf{A} \cdot \mathrm{Proj}_{\mathbf{SA}}\big\|_F^2\Big] \le \Big(1 + \frac{k}{n-k-1}\Big) \cdot \min_{\mathbf{B}\,:\,\mathrm{rank}(\mathbf{B})=k} \|\mathbf{A} - \mathbf{B}\|_F^2.$$

Those guarantees were later extended (and, in some regimes, improved) for sub-gaussian embeddings (Dereziński et al., 2020b) and Determinantal Point Processes (Dereziński et al., 2020a). The results for Gaussian and sub-gaussian embeddings rely heavily on the Hanson-Wright property of the rows of the sketching matrices. Thus, our techniques can be used to extend them to LESS embeddings by using Theorem 3. Consequently, we can obtain bounds for the expected approximation error of Randomized SVD with LESS embeddings that, in a certain regime of small sketch sizes, nearly match those achieved by Gaussian embeddings.

Corollary 15 *Consider an $N \times d$ matrix $\mathbf{A}$ with stable rank $r = \|\mathbf{A}\|_F^2/\|\mathbf{A}\|^2$. Given a LESS embedding $\mathbf{S}$ of size $n \le r/2$, and a Gaussian embedding $\widetilde{\mathbf{S}}$ of the same size, we have:*

$$\mathbb{E}\Big[\big\|\mathbf{A} - \mathbf{A} \cdot \mathrm{Proj}_{\mathbf{SA}}\big\|_F^2\Big] \approx_{1+\epsilon} \mathbb{E}\Big[\big\|\mathbf{A} - \mathbf{A} \cdot \mathrm{Proj}_{\widetilde{\mathbf{S}}\mathbf{A}}\big\|_F^2\Big] \quad \textit{with} \quad \epsilon = \tilde{O}(1/\sqrt{r}).$$

Proof We rely on Theorem 2 of Dereziński et al. (2020b), characterizing the expected approximation error for a sub-gaussian embedding of size n via an implicit analytic formula:

$$\mathbb{E}\Big[\big\|\mathbf{A} - \mathbf{A} \cdot \mathrm{Proj}_{\widetilde{\mathbf{S}}\mathbf{A}}\big\|_F^2\Big] \approx_{1+\epsilon} n\lambda_n, \quad \text{where} \quad \lambda_n = f_{\mathbf{A}}^{-1}(n) \quad \text{for} \quad f_{\mathbf{A}}(\lambda) = \mathrm{tr}\,\mathbf{A}^\top\mathbf{A}(\mathbf{A}^\top\mathbf{A} + \lambda\mathbf{I})^{-1}.$$
$$(3)$$

Here, $f_{\mathbf{A}}^{-1}(n)$ denotes the function inverse of $f_{\mathbf{A}}$ at n. The function $f_{\mathbf{A}}(\lambda)$ is known in the literature as λ-statistical dimension, a notion of degrees of freedom that arises in the analysis of ridge regression. Naturally, their result applies to the Gaussian embedding $\widetilde{\mathbf{S}}\mathbf{A}$, but also to any random design $\mathbf{Z}\mathbf{\Sigma}^{1/2}$ (where $\mathbf{\Sigma} = \mathbf{A}^\top\mathbf{A}$) such that the rows satisfy the Hanson-Wright inequality with

a constant K (with ϵ having a polynomial dependence on K). So, we can apply our reduction to LESS embeddings by coupling $\sqrt{n}\mathbf{S}\mathbf{A}$ with $\mathbf{Z}\widetilde{\mathbf{\Sigma}}^{1/2}$ for some $\widetilde{\mathbf{\Sigma}} \approx_{1+\delta} \mathbf{\Sigma}$, as in Theorem 3. Letting $\mathrm{Er}(\mathbf{S}\mathbf{A})$ denote $\|\mathbf{A} - \mathbf{A}\cdot\mathrm{Proj}_{\mathbf{S}\mathbf{A}}\|_F^2$ and $\mathcal{E}$ be the $(1-\delta)$-probability event where $\sqrt{n}\mathbf{S}\mathbf{A} = \mathbf{Z}\widetilde{\mathbf{\Sigma}}^{1/2}$, we have:

$$\big|\mathbb{E}\big[\mathrm{Er}(\mathbf{S}\mathbf{A})\big] - \mathbb{E}\big[\mathrm{Er}(\mathbf{Z}\widetilde{\mathbf{\Sigma}}^{1/2})\big]\big| = \delta \cdot \big|\mathbb{E}\big[\mathrm{Er}(\mathbf{S}\mathbf{A}) - \mathrm{Er}(\mathbf{Z}\widetilde{\mathbf{\Sigma}}^{1/2}) \mid \neg\mathcal{E}\big]\big| \le \delta \cdot \|\mathbf{A}\|_F^2.$$

Since $\|\mathbf{A}\|_F^2 \le r \cdot \|\mathbf{A}\|^2 \le \frac{r}{r-n} \cdot \min_{\mathbf{Z}} \mathrm{Er}(\mathbf{Z}\widetilde{\mathbf{\Sigma}}^{1/2})$, setting $\delta \le \epsilon/2$, we get $\mathbb{E}\big[\mathrm{Er}(\mathbf{S}\mathbf{A})\big] \approx_{1+\epsilon} \mathbb{E}\big[\mathrm{Er}(\mathbf{Z}\widetilde{\mathbf{\Sigma}}^{1/2})\big]$. Next, since $\widetilde{\mathbf{\Sigma}} \approx_{1+\delta} \mathbf{\Sigma}$, we can use the implicit error formula for sketching $\widetilde{\mathbf{\Sigma}}^{1/2}$ with $\mathbf{Z}$, and relate that back to the original problem as follows:

$$\mathbb{E}\big[\mathrm{Er}(\mathbf{Z}\widetilde{\mathbf{\Sigma}}^{1/2})\big] \approx_{1+\delta} \mathbb{E}\big[\|\widetilde{\mathbf{\Sigma}}^{1/2}(\mathbf{I} - \mathrm{Proj}_{\mathbf{Z}\widetilde{\mathbf{\Sigma}}})\|_F^2\big] \approx_{1+\epsilon} n\tilde{\lambda}_n,$$

where $\tilde{\lambda}_n$ is defined as in (3), but with $\mathbf{A}^\top\mathbf{A}$ replaced by $\widetilde{\mathbf{\Sigma}}$. To close the loop, we must relate the implicit analytic expression based on $\widetilde{\mathbf{\Sigma}}$ back to λ_n. This can be done easily by bounding the derivative of $f_{\mathbf{A}}^{-1}$, showing that $\tilde{\lambda}_n \approx_{1+O(\delta)} \lambda_n$. Now, it remains to observe that since both the LESS embedding error $\mathbb{E}\big[\mathrm{Er}(\mathbf{S}\mathbf{A})\big]$ and the Gaussian embedding error $\mathbb{E}\big[\mathrm{Er}(\widetilde{\mathbf{S}}\mathbf{A})\big]$ are approximated by the same quantity $n\lambda_n$ up to a $1 + O(\epsilon)$ factor, they are also approximated by each other, concluding the proof. ∎

A.2. Low-distortion embeddings

Another important property of sub-gaussian random matrices is that they can be used to construct low-dimensional embeddings that preserve the geometry of a high-dimensional space, such as the subspace embedding property and the Johnson-Lindenstrauss property, which are central to many applications of sketching (e.g., see Woodruff, 2014). We briefly mention some classical examples and discuss how they can be extended to gaussianized sketches, including LESS embeddings.

We say that an $n \times N$ matrix $\mathbf{S}$ is an ϵ-low-distortion embedding for some set of points $\mathcal{X} \subseteq \mathbb{R}^N$ if there is a fixed scalar $\alpha > 0$ such that $\|\alpha\mathbf{S}\mathbf{v}\| \approx_{1+\epsilon} \|\mathbf{v}\|$ for all $\mathbf{v} \in \mathcal{X}$, i.e., the embedding approximately preserves the Euclidean norm of $\mathbf{v}$. When $\mathbf{S}$ is a sketching matrix applied to an $N \times d$ data matrix $\mathbf{A}$, we are typically interested in vectors $\mathbf{v}$ from the column span of $\mathbf{A}$ (i.e., such that $\mathbf{v} = \mathbf{A}\mathbf{x}$ for some $\mathbf{x} \in \mathbb{R}^d$), denoted by $\mathrm{span}(\mathbf{A})$. Relying on Theorem 3 and standard properties of sub-gaussian random matrices Vershynin (2018), we can establish the following low-distortion embedding properties for LESS embeddings.

Corollary 16 *Let $\mathbf{A}$ be $N \times d$, and $\mathbf{S}$ be a LESS embedding for $\mathbf{A}$ of size n with $d\log(nd/(\epsilon\delta))$ non-zeros per row. Then, there is an absolute constant C such that the following claims are true:*

1. *(Johnson-Lindenstrauss property)   For any finite set $\mathcal{X} \subseteq \mathrm{span}(\mathbf{A})$, if $n \ge C\log(|\mathcal{X}|/\delta)/\epsilon^2$, then with probability $1 - \delta$, matrix $\mathbf{S}$ is an ϵ-low-distortion embedding for $\mathcal{X}$.*

2. *(Subspace embedding property)   If $n \ge C(d + \log(1/\delta))/\epsilon^2$, then with probability $1 - \delta$, matrix $\mathbf{S}$ is an ϵ-low-distortion embedding for $\mathcal{X} = \mathrm{span}(\mathbf{A})$.*

Proof Whenever, $\mathcal{X} \subseteq \mathrm{span}(\mathbf{A})$, the low-distortion property can be formulated as a property of the sketch $\mathbf{S}\mathbf{A}$, treated as a linear transformation: for any $\mathbf{v} \in \mathrm{span}(\mathbf{A})$, we have $\mathbf{v} = \mathbf{A}\mathbf{x}$ for

some $\mathbf{x} \in \mathbb{R}^d$, and we need to ensure that $\|\alpha\mathbf{S}\mathbf{A}\mathbf{x}\| \approx_{1+\epsilon} \|\mathbf{A}\mathbf{x}\|$. Here, we can once again rely on Theorem 3, coupling $\sqrt{n}\mathbf{S}\mathbf{A}$ with a sub-gaussian design $\mathbf{Z}\widetilde{\mathbf{\Sigma}}^{1/2}$ for some $\widetilde{\mathbf{\Sigma}} \approx_{1+\delta} \mathbf{A}^\top\mathbf{A}$. If $\mathbf{Z}$ is an ϵ-low-distortion embedding for $\widetilde{\mathcal{X}} = \{\widetilde{\mathbf{\Sigma}}^{1/2}\mathbf{x} : \mathbf{A}\mathbf{x} \in \mathcal{X}\}$ with constant α, then with probability $1 - \delta$ (specifically, when $\sqrt{n}\mathbf{S}\mathbf{A} = \mathbf{Z}\widetilde{\mathbf{\Sigma}}^{1/2}$), for any $\mathbf{A}\mathbf{x} \in \mathcal{X}$, we have:

$$\|\alpha\sqrt{n}\mathbf{S}\mathbf{A}\mathbf{x}\| = \|\alpha\mathbf{Z}\widetilde{\mathbf{\Sigma}}^{1/2}\mathbf{x}\| \approx_{1+\epsilon} \|\widetilde{\mathbf{\Sigma}}^{1/2}\mathbf{x}\| \approx_{1+\delta} \|\mathbf{A}\mathbf{x}\|,$$

where in the last step we used that $\widetilde{\mathbf{\Sigma}} \approx_{1+\delta} \mathbf{A}^\top\mathbf{A}$. Thus, as long as $\delta \leq \epsilon$, we reduced the problem to showing the corresponding property for $\mathbf{Z}$. Both the Johnson-Lindenstrauss and subspace embedding properties can be shown for sub-gaussian embeddings (with sizes as given in the statement) using standard arguments (e.g., see the proof of Theorem 4.6.1 in Vershynin (2018)), thus concluding the proof. ∎

We note that low-distortion embedding properties have been shown for other sketching operators, including SRHTs (Tropp, 2011) and other sparse sketching matrices like CountSketch (Clarkson and Woodruff, 2017) and SJLTs (Nelson and Nguyên, 2013). These results often circumvent sub-gaussian analysis by relying on matrix concentration inequalities such as the matrix Bernstein inequality (Lemma 18) or other more involved arguments, however this inevitably incurs additional overhead factors in the required sketch size n. For example, an SRHT requires sketch size of at least $n \geq O(d\log(d)/\epsilon^2)$ to establish the subspace embedding property, instead of $O(d/\epsilon^2)$ shown above (the additional log factor is a well-known limitation of the matrix concentration-style analysis). The $O(d\log(d)/\epsilon^2)$ guarantee was also previously shown for LESS embeddings with d non-zeros per row, also using matrix concentration inequalities instead of sub-gaussian concentration (Dereziński et al., 2021b, Lemma 12). Thus, the above corollary suggests that, when requiring a subspace embedding, we can trade an additional log-factor in the sketch size for a log-factor in the density of a LESS embedding matrix. This could be a useful trade-off when we wish to minimize the storage space of the sketched data.

Appendix B. Experiments

In this section, we aim to evaluate the degree of Algorithmic Gaussianization for various sketching methods, using the sketched least squares task from Section 3.1 as an example. In this task, we are given an $N \times d$ matrix $\mathbf{A}$ and an N-dimensional vector $\mathbf{b}$ which define the regression loss $L(\mathbf{w}) = \|\mathbf{A}\mathbf{w} - \mathbf{b}\|^2$. We then use an $n \times N$ sketching matrix $\mathbf{S}$ to generate a sketched least squares estimate $\widehat{\mathbf{w}} = \operatorname{argmin}_\mathbf{w} \|\mathbf{S}(\mathbf{A}\mathbf{w} - \mathbf{b})\|^2$ of the exact solution $\mathbf{w}^*$. Our goal is to compare the normalized expected approximation error $\mathbb{E}[L(\widehat{\mathbf{w}}) - L(\mathbf{w}^*)]/L(\mathbf{w}^*)$ of the sketched estimate to the problem-independent expression $\frac{d}{n-d-1}$ achieved by a Gaussian embedding.

In Figure 4, we plot the empirically estimated expected approximation error with varying sketch size for several sketching techniques on four benchmark datasets. We compare them against the expected error for Gaussian embeddings, which is given by (1). Two of the sketching techniques, Subsampled Randomized Hadamard Transform (SRHT, Ailon and Chazelle, 2009) and Leverage Score Sparsified embeddings (LESS, Dereziński et al., 2021b), have strong theoretical guarantees for the expected error, and have comparable nearly-linear time complexity.[2] However, there are some key

---

2. Recall that the complexity of sketched least squares with LESS is $O(\text{nnz}(\mathbf{A})\log N + nd^2)$, compared to $O(Nd\log N + nd^2)$ for SRHT. So, the complexity of LESS is better than SRHT when $\mathbf{A}$ is a sparse matrix, i.e., when $\text{nnz}(\mathbf{A}) \ll Nd$, and otherwise, the complexities are comparable.
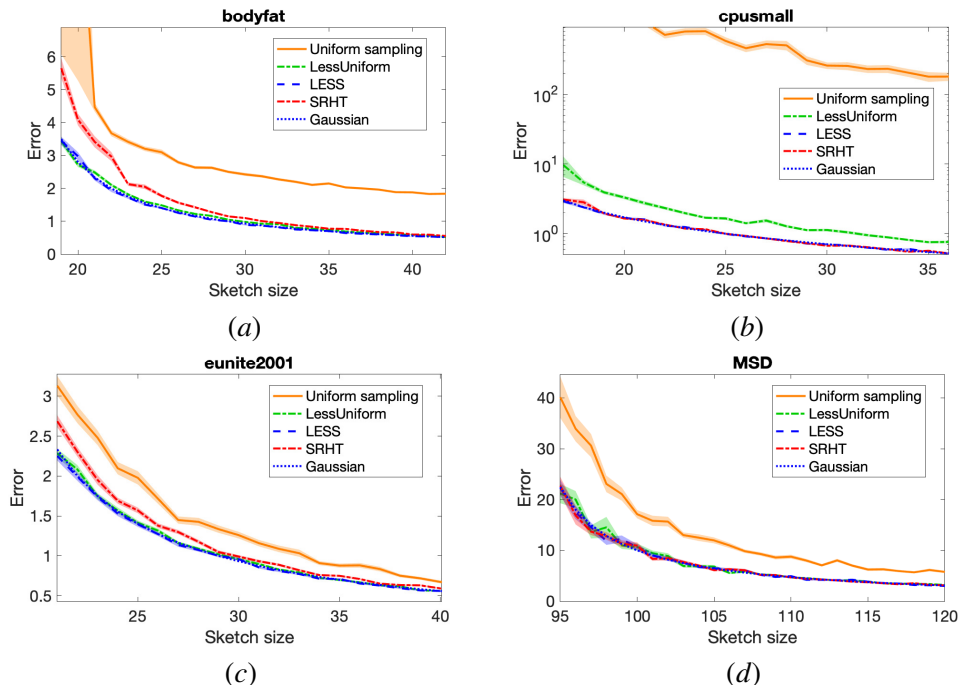
Figure 4: Comparison of expected approximation error for sketched least squares on Libsvm datasets (Chang and Lin, 2011), with shading indicating standard error of the mean, and "Gaussian" showing the theoretical $\frac{d}{n-d-1}$ expression from Theorem 7.

differences: LESS has a precise non-asymptotic guarantee given in this work (Theorem 7), whereas SRHT only has a non-asymptotic upper bound (Sarlos, 2006) and a precise asymptotic guarantee under some additional assumptions (Dobriban and Liu, 2019). We compare these two sketches with two much cheaper baselines: (1) uniform sampling, and (2) a simplified variant of LESS, called LessUniform (Dereziński et al., 2021a), which eliminates the leverage score approximation pre-processing step (but still uses d non-zeros per row). The overall cost for both LessUniform and for uniform sampling is the same as the cost of solving the sketched sub-problem, i.e., $O(nd^2)$, where n is the sketch size. Our Theorems 3 and 7 can also be applied to LessUniform, but the Hanson-Wright constant becomes dependent on the maximum leverage score of matrix $\mathbf{A}$ (i.e., the coherence of $\mathbf{A}$; see Remark 4).

From Figure 4, we first confirm that LESS enjoys a Gaussian-like problem-independent expected approximation error that matches our theory. Remarkably, we can verify this for all datasets and all sketch sizes, downto the precision of our empirical mean estimates. Thus, these results suggest that the Gaussian error estimate for LESS embeddings is even more accurate and broadly applicable than promised by our theory. Next, we observe that for small sketch sizes, SRHT does not always exhibit Gaussian-like error (see left two plots), but as the sketch size increases, the gap decreases. This appears roughly in line with the existing asymptotic theory for SRHTs. Finally, LessUniform performs much better than uniform sampling, despite having the same time complexity. The error for LessUniform appears Gaussian-like in three out of four cases. In the case of dataset *cpusmall*, the problem exhibits very high coherence, and as a result both uniform sampling

and LessUniform perform worse. This aligns with Theorem 3 applied to LessUniform, since the Hanson-Wright constant scales with the largest leverage score of $\mathbf{A}$.

Appendix C. Notation and preliminaries for the proofs

Notation. We say that $a \approx_\alpha b$ for $\alpha \geq 1$, if $b/\alpha \leq a \leq \alpha b$. We analogously define $\mathbf{A} \approx_\alpha \mathbf{B}$ for positive semidefinite matrices using the Loewner ordering. We use $\|\mathbf{A}\|_F = \sqrt{\operatorname{tr}(\mathbf{A}^\top \mathbf{A})}$ to denote the Frobenius norm and $\|\mathbf{A}\|$ to denote the spectral norm. Also, we let $a \lesssim b$ denote that there exists an absolute constant $C > 0$ such that $a \leq Cb$, and $a = \operatorname{poly}(b, c)$ means that a is bounded by a polynomial function of b and c. Moreover, we use $\tilde{O}(\cdot)$ to refer to big-O notation where polylogarithmic terms are ignored. For random variables/vectors X and Y defined over the same domain with measures μ and ν, respectively, we define the total variation distance between them as $d_{\mathrm{tv}}(X, Y) = \sup_{E \in \mathcal{B}} |\mu(E) - \nu(E)|$, where $\mathcal{B}$ denotes all measurable events. Note that the total variation distance can also be defined as the infimum over δ such that there exists a coupling between X and Y for which $\Pr(X \neq Y) = \delta$. Finally, we define the sub-gaussian Orlicz norm as: $\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E} \exp(X^2/t^2) \leq 2\}$.

Matrix concentration inequalities. In the proof of our main results, we use the following versions of the matrix Bernstein concentration inequality for the sums of independent symmetric random matrices.

Lemma 17 (Sub-exponential matrix Bernstein) *(Tropp, 2012, Theorem 6.2) For $i = 1, 2, ...,$ consider a finite sequence $\mathbf{M}_i$ of $d \times d$ independent and symmetric random matrices such that*

$$\mathbb{E}[\mathbf{M}_i] = \mathbf{0}, \quad \mathbb{E}[\mathbf{M}_i^p] \preceq \frac{p!}{2} \cdot R^{p-2} \mathbf{A}_i^2 \quad for \quad p = 2, 3, ...$$

Then, defining the variance parameter $\sigma^2 = \|\sum_i \mathbf{A}_i^2\|$, for any $t > 0$ we have:

$$\Pr\left\{\lambda_{\max}\left(\sum_i \mathbf{M}_i\right) \geq t\right\} \leq d \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt}\right).$$

Lemma 18 (Bounded matrix Bernstein) *(Tropp, 2012, Theorem 6.1) For $i = 1, 2, ...,$ consider a finite sequence $\mathbf{M}_i$ of $d \times d$ independent and symmetric random matrices such that*

$$\mathbb{E}[\mathbf{M}_i] = \mathbf{0}, \quad \lambda_{\max}(\mathbf{M}_i) \leq R \qquad almost\ surely.$$

Then, defining the variance parameter $\sigma^2 = \|\sum_i \mathbb{E}[\mathbf{M}_i^2]\|$, for any $t > 0$ we have:

$$\Pr\left\{\lambda_{\max}\left(\sum_i \mathbf{M}_i\right) \geq t\right\} \leq d \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right).$$

Linear algebraic identities. Our proof of the least squares approximation guarantee for LESS embeddings relies on the following standard rank-one update formula for the matrix inverse.

Lemma 19 (Sherman-Morrison formula) *For a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ such that both $\mathbf{A}$ and $\mathbf{A} + \mathbf{u}\mathbf{v}^\top$ are invertible, we have:*

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^\top)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^\top \mathbf{A}^{-1}}{1 + \mathbf{v}^\top \mathbf{A}^{-1}\mathbf{u}}.$$

In particular, it follows that:

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^\top)^{-1}\mathbf{u} = \frac{\mathbf{A}^{-1}\mathbf{u}}{1 + \mathbf{v}^\top \mathbf{A}^{-1}\mathbf{u}}.$$

Appendix D. Hanson-Wright limit theorem: Proof of Theorem 11

In this section, we prove our main result, Theorem 11, showing that a k-gaussianized sample $\tilde{\mathbf{x}} = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} r_i \mathbf{x}_i$ is close in total variation distance to a random vector that satisfies the Hanson-Wright inequality with a small constant.

We start by using the matrix Bernstein inequality to verify how well the sample covariance $\frac{1}{k} \sum_{i=1}^{k} \mathbf{x}_i \mathbf{x}_i^\top$ can be bounded in terms of the true covariance $\boldsymbol{\Sigma}$. For this we establish the following lemma, proven in Appendix H. The lemma is used to show the matrix moment condition required by the sub-exponential matrix Bernstein (Lemma 17).

Lemma 20 *There is an absolute constant C such that any d-dimensional random vector $\mathbf{x}$ with covariance $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \boldsymbol{\Sigma}$, such that $\|\boldsymbol{\Sigma}^{-1/2}\mathbf{x}\|$ is M-sub-gaussian, satisfies the following sub-exponential matrix moment bound:*

$$\left\| \mathbb{E}\left[\left(\boldsymbol{\Sigma}^{-1/2}\mathbf{x}\mathbf{x}^\top\boldsymbol{\Sigma}^{-1/2} - \mathbf{I} \right)^p \right] \right\| \leq \left(CM^2(p + \log M) \right)^{p-1}.$$

We can now use Lemma 20 in conjunction with the sub-exponential matrix Bernstein inequality (Lemma 17) by setting $\mathbf{M}_i = \boldsymbol{\Sigma}^{-1/2}\mathbf{x}_i\mathbf{x}_i^\top\boldsymbol{\Sigma}^{-1/2} - \mathbf{I}$. From Lemma 20, we can set $R = CM^2 \log M$ and $\mathbf{A}_i^2 = R\mathbf{I}$, with $\sigma^2 = kR$ (appropriately adjusting constant C), concluding that:

$$\Pr\left\{ \lambda_{\max}\left(\sum_{i=1}^{k} \mathbf{M}_i \right) \geq kt \right\} \leq d \cdot \exp\left(\frac{-t^2 k}{2(1+t)CM^2 \log M} \right).$$

Now, define the $k \times d$ matrix $\mathbf{U}$ with ith row $\frac{1}{\sqrt{k}}\mathbf{x}_i^\top\boldsymbol{\Sigma}^{-1/2}$. The above concentration inequality implies that if $k \geq 4CM^2 \log^2(M/\delta)/t$ for $t \geq 1$, then $\|\mathbf{U}\|^2 = \|\mathbf{U}^\top\mathbf{U}\| \leq 1 + t$ with probability $1 - \delta$. We mention that, if we assumed $\|\boldsymbol{\Sigma}^{-1/2}\mathbf{x}\| \leq M$ almost surely, as in Remark 12, then we can use the bounded matrix Bernstein (Lemma 18) to obtain a slightly sharper guarantee of $k \geq O(M^2 \log(d/\delta)/t)$, without relying on Lemma 20.

We next define the random variable $\mathbf{z}$ and covariance matrix $\widetilde{\boldsymbol{\Sigma}}$ discussed in the theorem:

$$\mathbf{z} = \mathbf{1}_{\mathcal{E}} \cdot \widetilde{\boldsymbol{\Sigma}}^{-1/2}\tilde{\mathbf{x}} + \mathbf{1}_{\neg\mathcal{E}} \cdot \alpha\mathbf{V}^\top\mathbf{r}, \qquad \widetilde{\boldsymbol{\Sigma}} = \frac{1}{\Pr\{\mathcal{E}\}} \mathbb{E}[\mathbf{1}_{\mathcal{E}}\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top],$$

$$\text{for} \qquad \mathcal{E} = \left[\|\mathbf{U}\| \leq L \right], \qquad \alpha = \max\left\{ 1, \sqrt{\frac{d}{k}} \right\}, \qquad \mathbf{V} = \mathbf{G}((\mathbf{G}^\top\mathbf{G})^\dagger)^{1/2},$$

where $\mathbf{r}$ is the vector of R-sub-gaussian random variables r_i from the definition of $\tilde{\mathbf{x}}$; $\mathbf{G}$ denotes a $k \times d$ matrix with i.i.d. Gaussian entries and $(\cdot)^\dagger$ is the Moore-Penrose pseudoinverse (so that either $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$ or $\mathbf{V}\mathbf{V}^\top = \mathbf{I}$); and $\mathcal{E}$ denotes an event in the probability space of $\tilde{\mathbf{x}}$, with $\mathbf{1}_{\mathcal{E}}$ being the characteristic function of $\mathcal{E}$. Here, $L = O(1 + M\log(M/\delta)/\sqrt{k})$ is chosen so that $\Pr\{\neg\mathcal{E}\} \leq \delta/L^2$.[3] We can do this by simply adjusting the constants, because $M\log(ML/\delta) = O(M\log(M/\delta))$. Also, we can easily make sure that $L \geq \max\{2, \alpha\}$ (recall that $\alpha \leq 1 + \sqrt{d/k} \leq 1 + M/\sqrt{k}$). Note that $\mathbf{z}$ is an isotropic random vector since:

$$\mathbb{E}[\mathbf{z}\mathbf{z}^\top] = \widetilde{\boldsymbol{\Sigma}}^{-1/2}\mathbb{E}[\mathbf{1}_{\mathcal{E}}\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top]\widetilde{\boldsymbol{\Sigma}}^{-1/2} + \Pr\{\neg\mathcal{E}\} \cdot \mathbb{E}[\alpha^2\mathbf{V}^\top\mathbf{V}] = \Pr\{\mathcal{E}\}\mathbf{I} + \Pr\{\neg\mathcal{E}\}\,\mathbf{I} = \mathbf{I}.$$

---

3. If $\|\boldsymbol{\Sigma}^{-1/2}\mathbf{x}_i\| \leq M$ a.s., as in Remark 12, then we can let $L = O(1 + M\sqrt{\log(M/\delta)/k})$ by using bounded matrix Bernstein (Lemma 18).

Also, note that we can write $\mathbf{z} = \widetilde{\mathbf{U}}^\top \mathbf{r}$, where

$$\widetilde{\mathbf{U}} = \mathbf{1}_{\mathcal{E}} \cdot \mathbf{U}\boldsymbol{\Sigma}^{1/2}\widetilde{\boldsymbol{\Sigma}}^{-1/2} + \mathbf{1}_{\neg\mathcal{E}} \cdot \alpha\mathbf{V},$$

and since $\mathbf{r}$ is mean zero and independent of $\widetilde{\mathbf{U}}$, then $\mathbf{z}$ is also mean zero.

Next we show that $\widetilde{\boldsymbol{\Sigma}}$ approximates $\boldsymbol{\Sigma}$. First, by definition we immediately have $\widetilde{\boldsymbol{\Sigma}} \preceq \frac{1}{1-\delta/2}\boldsymbol{\Sigma} \preceq (1+\delta)\boldsymbol{\Sigma}$. Next, observe that we have:

$$\begin{aligned}
\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}})\boldsymbol{\Sigma}^{-\frac{1}{2}}\| &= \|\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbb{E}\big[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top \cdot \mathbf{1}_{\neg\mathcal{E}}\big]\boldsymbol{\Sigma}^{-\frac{1}{2}}\| = \big\|\mathbb{E}\big[\mathbf{U}^\top\mathbf{r}\mathbf{r}^\top\mathbf{U} \cdot \mathbf{1}_{\neg\mathcal{E}}\big]\big\| \\
&= \|\mathbb{E}[\mathbf{U}^\top\mathbf{U} \cdot \mathbf{1}_{\neg\mathcal{E}}]\| \leq \mathbb{E}\big[\|\mathbf{U}\|^2 \cdot \mathbf{1}_{\neg\mathcal{E}}\big] \\
&\overset{(*)}{=} \int_0^\infty \Pr\{\|\mathbf{U}\|^2 \cdot \mathbf{1}_{\neg\mathcal{E}} > x\}dx \\
&\leq \frac{\delta}{L^2} \cdot L^2 + \int_{L^2}^\infty \Pr\{\|\mathbf{U}\|^2 > x\}dx,
\end{aligned}$$

where $(*)$ is the integral formula for the expectation of a non-negative random variable via its cdf, and the last step uses the observation that we have $\Pr\{\|\mathbf{U}\|^2 \cdot \mathbf{1}_{\neg\mathcal{E}} > x\} \leq \Pr\{\neg\mathcal{E}\} \leq \delta/L^2$. To bound the integral, note that for $x \geq L^2$:

$$\Pr\{\|\mathbf{U}\|^2 > x\} \leq d \cdot \exp\left(\frac{-(x-1)k}{4CM^2\log d}\right) \leq (\delta/L^2)^{(x-1)/L^2}.$$

Thus, using the formula $\int (\delta/L^2)^{x/L^2}dx = -L^2(\delta/L^2)^{x/L^2}/\ln(L^2/\delta)$, we obtain:

$$\int_{L^2}^\infty \Pr\{\|\mathbf{U}\|^2 > x\}dx \leq (\delta/L^2)^{-1/L^2}\int_{L^2}^\infty (\delta/L^2)^{x/L^2} = (\delta/L^2)^{-1/L^2} \cdot L^2(\delta/L^2)/\ln(L^2/\delta) \leq \delta.$$

We conclude that $\|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}})\boldsymbol{\Sigma}^{-\frac{1}{2}}\| \leq \delta + \delta \leq 2\delta$, obtaining:

$$(1 - 2\delta) \cdot \boldsymbol{\Sigma} \preceq \widetilde{\boldsymbol{\Sigma}} \preceq (1+\delta)\boldsymbol{\Sigma},$$

so $\widetilde{\boldsymbol{\Sigma}} \approx_{1+O(\delta)} \boldsymbol{\Sigma}$. Note that the above also implies that $\|\widetilde{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\Sigma}^{1/2}\|^2 = \|\widetilde{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\Sigma}\widetilde{\boldsymbol{\Sigma}}^{-1/2}\| \leq \frac{1}{1-2\delta}$. As a consequence we have:

$$\begin{aligned}
\|\widetilde{\mathbf{U}}\|^2 &= \mathbf{1}_{\mathcal{E}} \cdot \|\widetilde{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\Sigma}^{1/2}\mathbf{U}^\top\mathbf{U}\boldsymbol{\Sigma}^{1/2}\widetilde{\boldsymbol{\Sigma}}^{-1/2}\| + \mathbf{1}_{\neg\mathcal{E}} \cdot \|\alpha^2\mathbf{V}^\top\mathbf{V}\| \\
&\leq \max\big\{\|\widetilde{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\Sigma}^{1/2}\|^2\|\mathbf{1}_{\mathcal{E}}\mathbf{U}\|^2, \alpha^2\big\} \leq \frac{L^2}{1-2\delta} \leq 2L^2.
\end{aligned}$$

We are now ready to establish the Hanson-Wright inequality for $\mathbf{z}$. Specifically, consider a $d \times d$ psd matrix $\mathbf{B}$. We study the concentration of the quadratic form $\mathbf{z}^\top\mathbf{B}\mathbf{z}$ around its mean $\mathrm{tr}\mathbf{B}$. We start with the following decomposition:

$$|\mathbf{z}^\top\mathbf{B}\mathbf{z} - \mathrm{tr}\mathbf{B}| = |\mathbf{r}^\top\widetilde{\mathbf{U}}\mathbf{B}\widetilde{\mathbf{U}}^\top\mathbf{r} - \mathrm{tr}\mathbf{B}| \leq |\mathbf{r}^\top\widetilde{\mathbf{U}}\mathbf{B}\widetilde{\mathbf{U}}^\top\mathbf{r} - \mathrm{tr}(\widetilde{\mathbf{U}}\mathbf{B}\widetilde{\mathbf{U}}^\top)| + |\mathrm{tr}(\widetilde{\mathbf{U}}\mathbf{B}\widetilde{\mathbf{U}}^\top) - \mathrm{tr}\mathbf{B}|. \quad (4)$$

Since $\widetilde{\mathbf{U}}$ and $\mathbf{r}$ are independent, and we have $\mathbb{E}[\mathbf{r}^\top\widetilde{\mathbf{U}}\mathbf{B}\widetilde{\mathbf{U}}^\top\mathbf{r}] = \mathrm{tr}(\widetilde{\mathbf{U}}\mathbf{B}\widetilde{\mathbf{U}}^\top)$, we can show concentration for the first term in (4) by using the classical Hanson-Wright inequality (Lemma 1) applied

to the random vector $\mathbf{r}$ and the matrix $\widetilde{\mathbf{U}}\mathbf{B}\widetilde{\mathbf{U}}^\top$:

$$\Pr\Big\{|\mathbf{r}^\top\widetilde{\mathbf{U}}\mathbf{B}\widetilde{\mathbf{U}}^\top\mathbf{r} - \mathrm{tr}(\widetilde{\mathbf{U}}\mathbf{B}\widetilde{\mathbf{U}}^\top)| \geq t \mid \widetilde{\mathbf{U}}\Big\} \leq 2\exp\Big(-c\min\Big\{\frac{t^2}{R^4\|\widetilde{\mathbf{U}}\mathbf{B}\widetilde{\mathbf{U}}\|_F^2}, \frac{t}{R^2\|\widetilde{\mathbf{U}}\mathbf{B}\widetilde{\mathbf{U}}^\top\|}\Big\}\Big)$$

$$\leq 2\exp\Big(-c\min\Big\{\frac{t^2}{4L^4R^4\|\mathbf{B}\|_F^2}, \frac{t}{2L^2R^2\|\mathbf{B}\|}\Big\}\Big),$$

where we used that $\|\widetilde{\mathbf{U}}\|^2 \leq 2L^2$. Next, to show concentration for the second term in (4), we decompose it further as follows, letting $\widetilde{\mathbf{B}} = \boldsymbol{\Sigma}^{1/2}\widetilde{\boldsymbol{\Sigma}}^{-1/2}\mathbf{B}\widetilde{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\Sigma}^{1/2}$:

$$|\mathrm{tr}(\widetilde{\mathbf{U}}\mathbf{B}\widetilde{\mathbf{U}}^\top) - \mathrm{tr}\mathbf{B}| \leq \mathbf{1}_{\mathcal{E}} \cdot |\mathrm{tr}(\mathbf{U}\widetilde{\mathbf{B}}\mathbf{U}^\top) - \mathrm{tr}\mathbf{B}| + \mathbf{1}_{\neg\mathcal{E}} \cdot |\alpha^2\mathrm{tr}\mathbf{V}\mathbf{B}\mathbf{V}^\top - \mathrm{tr}\mathbf{B}|$$

$$\leq |\mathrm{tr}(\mathbf{U}\widetilde{\mathbf{B}}\mathbf{U}^\top) - \mathrm{tr}\widetilde{\mathbf{B}}| + |\mathrm{tr}\widetilde{\mathbf{B}} - \mathrm{tr}\mathbf{B}| + |\alpha^2\mathrm{tr}\mathbf{V}\mathbf{B}\mathbf{V}^\top - \mathrm{tr}\mathbf{B}|. \quad (5)$$

Note that $\mathrm{tr}(\mathbf{U}\widetilde{\mathbf{B}}\mathbf{U}) - \mathrm{tr}\widetilde{\mathbf{B}} = \frac{1}{k}\sum_{i=1}^k X_i$ for $X_i = \|\widetilde{\mathbf{B}}^{1/2}\boldsymbol{\Sigma}^{-1/2}\mathbf{x}_i\|^2 - \mathrm{tr}\widetilde{\mathbf{B}}$. We are going to again use Bernstein's inequality, this time the scalar version, i.e., Lemma 17 with $d = 1$.[4] Note that we have $\mathbb{E}[X_i] = 0$, and also in Lemma 28 (along the same lines as Lemma 20, see Appendix H) we show that:

$$\mathbb{E}[|X_i|^p] \leq (CM^2(p + \log M))^{p-1}\|\widetilde{\mathbf{B}}\|^{p-2}\|\widetilde{\mathbf{B}}\|_F^2.$$

Using Lemma 17 with $R = O(\|\widetilde{\mathbf{B}}\|M^2\log M)$ and $\sigma^2 = k \cdot O(\|\widetilde{\mathbf{B}}\|_F^2 M^2\log M)$, we obtain (for some absolute constants $c, c' > 0$):

$$\Pr\Big\{|\mathrm{tr}(\mathbf{U}\widetilde{\mathbf{B}}\mathbf{U}^\top) - \mathrm{tr}\widetilde{\mathbf{B}}| \geq t\Big\} \leq 2\exp\Big(-c\min\Big\{\frac{t^2 k}{\|\widetilde{\mathbf{B}}\|_F^2 M^2\log M}, \frac{tk}{\|\widetilde{\mathbf{B}}\|M^2\log M}\Big\}\Big)$$

$$\leq 2\exp\Big(-c'\min\Big\{\frac{t^2}{L^2\|\mathbf{B}\|_F^2}, \frac{t}{L^2\|\mathbf{B}\|}\Big\}\Big),$$

where we used the facts that $\|\widetilde{\mathbf{B}}\| \leq 2\|\mathbf{B}\|$ and $L^2 \geq M^2\log(M)/k$. We now turn to the second term in (5). Note that this is in fact a deterministic quantity which can be bounded as follows: $|\mathrm{tr}\widetilde{\mathbf{B}} - \mathrm{tr}\mathbf{B}| \leq 4\delta \cdot \mathrm{tr}\mathbf{B}$. Now, assuming that $\delta \leq 1/(4d)$ (which is w.l.o.g. by adjusting the constants), we have that for any $0 \leq t \leq 4\delta\mathrm{tr}\mathbf{B}$:

$$\min\Big\{\frac{t^2}{\|\mathbf{B}\|_F^2}, \frac{t}{\|\mathbf{B}\|}\Big\} \leq \min\Big\{\frac{(\mathrm{tr}\mathbf{B})^2}{d^2\|\mathbf{B}\|_F^2}, \frac{\mathrm{tr}\mathbf{B}}{d\|\mathbf{B}\|}\Big\} \leq \min\Big\{\frac{\|\mathbf{B}\|^2}{\|\mathbf{B}\|_F^2}, \frac{\|\mathbf{B}\|}{\|\mathbf{B}\|}\Big\} \leq 1,$$

where we used that $\mathrm{tr}\mathbf{B}/d \leq \|\mathbf{B}\|$. This means that $\Pr\{\|\mathrm{tr}\widetilde{\mathbf{B}} - \mathrm{tr}\mathbf{B}| \geq t\} = \mathbf{1}_{[t \leq 4\delta\cdot\mathrm{tr}\mathbf{B}]} \leq 2\exp(-c\min\{\frac{t^2}{\|\mathbf{B}\|_F^2}, \frac{t}{\|\mathbf{B}\|}\})$, so the desired concentration inequality trivially holds. Finally, it remains to establish the concentration for $|\alpha^2\mathrm{tr}(\mathbf{V}\mathbf{B}\mathbf{V}^\top) - \mathrm{tr}\mathbf{B}|$, the final term in (5). If $k \geq d$, then $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$ and $\alpha^2 = 1$, so the term is 0. Now, suppose that $k < d$. In this case, $\mathbf{V}^\top\mathbf{V}$ is a projection onto a uniformly random k-dimensional subspace of $\mathbb{R}^d$. Here, we will use a simple form of the Johnson-Lindenstrauss lemma for uniformly random projections (Lemma 5.3.2, Vershynin, 2018), which states that for any fixed unit vector $\mathbf{b} \in \mathbb{R}^d$:

$$\Pr\Big\{\big|\|\alpha\mathbf{V}\mathbf{b}\| - 1\big| \geq \epsilon\Big\} \leq 2\exp(-c\epsilon^2 k).$$

---

4. Here, again, if we have the boundedness assumption from Remark 12, then we can use bounded Bernstein (scalar version of Lemma 18).

Let $\mathbf{B} = \sum_{i=1}^{d} \lambda_i \mathbf{b}_i \mathbf{b}_i^\top$ be the eigendecomposition of $\mathbf{B}$. The Johnson-Lindenstrauss concentration lemma implies that with probability $1 - 2d \exp(-c\epsilon^2 k)$ we have $\max_i |\|\alpha \mathbf{V} \mathbf{b}_i\| - 1| \leq \epsilon$, which in turn implies that:

$$|\alpha^2 \mathrm{tr}(\mathbf{V}\mathbf{B}\mathbf{V}^\top) - \mathrm{tr}\mathbf{B}| \leq \sum_{i=1}^{d} \lambda_i |\|\alpha \mathbf{V}\mathbf{b}_i\|^2 - 1| \leq \mathrm{tr}\mathbf{B} \cdot \max_i |\|\alpha \mathbf{V}\mathbf{b}_i\|^2 - 1| \leq \epsilon(\epsilon+2) \cdot \mathrm{tr}\mathbf{B}.$$

Setting $t = \epsilon(\epsilon+2)\mathrm{tr}\mathbf{B}$ and solving for ϵ, we convert this to a concentration inequality:

$$\Pr\Big\{|\alpha^2 \mathrm{tr}(\mathbf{V}\mathbf{B}\mathbf{V}^\top) - \mathrm{tr}\mathbf{B}| \geq t\Big\} \leq 2d \exp\Big(-c \min\Big\{\frac{t^2 k}{\mathrm{tr}\mathbf{B}}, \frac{tk}{\mathrm{tr}\mathbf{B}}\Big\}\Big)$$

$$\leq 2\exp\Big(-c\min\Big\{\frac{t^2 k}{\|\mathbf{B}\| d \log d}, \frac{tk}{\|\mathbf{B}\| d \log d}\Big\}\Big)$$

$$\leq 2\exp\Big(-c\min\Big\{\frac{t^2}{L^2 \|\mathbf{B}\|_F^2}, \frac{t}{L^2 \|\mathbf{B}\|}\Big\}\Big).$$

Putting everything together, we combine the four inequalities to obtain that for any psd matrix $\mathbf{B}$:

$$\Pr\big\{\mathbf{z}^\top \mathbf{B}\mathbf{z} - \mathrm{tr}\mathbf{B}| \geq t\big\} \leq 4 \cdot 2\exp\Big(-c\min\Big\{\frac{t^2}{(LR)^4 \|\mathbf{B}\|_F^2}, \frac{t}{(LR)^2 \|\mathbf{B}\|}\Big\}\Big).$$

We can easily extend this to arbitrary matrices $\mathbf{B}$. First, observe that if $\mathbf{B}$ is symmetric, then we can write it as $\mathbf{B} = \mathbf{B}_+ - \mathbf{B}_-$ where $\mathbf{B}_+$ and $\mathbf{B}_-$ are psd, and it suffices to apply the concentration inequality to both $\mathbf{B}_+$ and $\mathbf{B}_-$. Next, consider an arbitrary $\mathbf{B}$. Then, it suffices to apply the result to $(\mathbf{B} + \mathbf{B}^\top)/2$, which is symmetric.

Appendix E. Sketched least squares: Proof of Theorem 7

In this section, we prove Theorem 7, giving a precise estimate for the expected approximation error for sketched least squares with LESS embeddings. Recall that we are given an $N \times d$ data matrix $\mathbf{A}$ and an N-dimensional vector $\mathbf{b}$. We let $L(\mathbf{w}) = \|\mathbf{A}\mathbf{w} - \mathbf{b}\|^2$ and $\mathbf{w}^* = \mathrm{argmin}_{\mathbf{w}} L(\mathbf{w})$. Given an $n \times N$ LESS embedding matrix $\mathbf{S}$, we define the sketched least squares estimator as:

$$\widehat{\mathbf{w}} = \underset{\mathbf{w} \in \mathcal{R}^d}{\mathrm{argmin}} \|\mathbf{S}(\mathbf{A}\mathbf{w} - \mathbf{b})\|^2,$$

where $\mathcal{R} = [-D, D]$ is only used to protect against the corner cases where $\mathbf{S}\mathbf{A}$ is very ill-conditioned, which arise with a small but non-zero probability when computing the exact expectation of the loss. In the proof, we will use our reduction between a LESS embedding and a sub-gaussian design to define an event that holds with a high $1 - \delta$ probability (Lemma 22). This event ensures certain regularity conditions on the sketch, such as near-uniformity of the leverage scores, and avoids the aforementioned corner cases. We then proceed to analyze the conditional expectation, conditioned on the high-probability event. To obtain the final unconditional expectation, we must choose δ sufficiently small to absorb the worst-case loss that might occur if the high-probability event fails. This is where we incur a dependence on the range diameter D, which is in general unavoidable when computing the expected loss, but it only affects the logarithmic factors in the result.

Let us briefly comment on those logarithmic dependencies. If we let the LESS embedding have d non-zeros per row, as recommended by Dereziński et al. (2021b), then our reduction to a

sub-gaussian sketch (which is used in Lemma 22) incurs a logarithmic dependence in the Hanson-Wright constant, i.e., $K = O(\sqrt{\log(nd/\delta)})$, and also, the lower bound on the sketch size required for the claim becomes $n \geq O(d\log(d/\delta))$. This also results in a polylogarithmic factor in the error ϵ. We note that these logarithmic terms can be avoided if we slightly increase the density of the sketch, from d to $d\log(nd/\delta)$ non-zeros per row. Then, the result can be established for $n \geq O(d)$.

Instead of computing the expectation of $L(\widehat{\mathbf{w}})$ directly, we will define an unbiased estimate of the expected loss via leave-one-out cross-validation (CV):

$$L_{\mathrm{CV}} = \sum_{i=1}^{n} \left(\mathbf{s}_i^\top \mathbf{A}\widehat{\mathbf{w}}^{-i} - \mathbf{s}_i^\top \mathbf{b}\right)^2, \qquad \text{for} \quad \widehat{\mathbf{w}}^{-i} = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{R}^d} \|\mathbf{S}_{-i}(\mathbf{Aw} - \mathbf{b})\|^2,$$

where $\mathbf{s}_i^\top$ is the ith row of $\mathbf{S}$ and $\mathbf{S}_{-i}$ denotes $\mathbf{S}$ without the ith row. We use the following standard "shortcut formula" for the CV estimate, which holds as long as the scalar range $\mathcal{R} = [-D, D]$ is large enough so that each $\widehat{\mathbf{w}}^{-i}$ is the same as the unconstrained least squares solution (we include the proof in Appendix I).

Lemma 21 *Suppose that $(\mathbf{A}^\top \mathbf{S}_{-i}\mathbf{S}_{-i}\mathbf{A})^{-1} \preceq O(1) \cdot (\mathbf{A}^\top \mathbf{A})^{-1}$ for all i. Then, for sufficiently large D, the sketched estimator is given by $\widehat{\mathbf{w}} = (\mathbf{A}^\top \mathbf{S}^\top \mathbf{S}\mathbf{A})^{-1}\mathbf{A}^\top \mathbf{S}^\top \mathbf{S}\mathbf{b}$, and the leave-one-out cross-validation loss can be computed as:*

$$L_{\mathrm{CV}} = \sum_{i=1}^{n} \left(\frac{\mathbf{s}_i^\top \mathbf{A}\widehat{\mathbf{w}} - \mathbf{s}_i^\top \mathbf{b}}{1 - \ell_i(\mathbf{SA})}\right)^2,$$

where $\ell_i(\mathbf{SA}) = \mathbf{s}_i^\top \mathbf{A}(\mathbf{A}^\top \mathbf{S}^\top \mathbf{S}\mathbf{A})^{-1}\mathbf{A}^\top \mathbf{s}_i$ is the ith leverage score of the sketch.

Crucially, the CV loss is an unbiased estimate of the expected loss based on the sketch of size $n-1$:

$$\mathbb{E}[L_{\mathrm{CV}}] = \mathbb{E}[L(\widehat{\mathbf{w}}^{-n})],$$

Thus, for the rest of the proof, we will actually estimate $\mathbb{E}[L(\widehat{\mathbf{w}}^{-n})]$ rather than $\mathbb{E}[L(\widehat{\mathbf{w}})]$, but the difference between the estimates is sufficiently small to be absorbed by ϵ since $\frac{d}{n-d} \approx_{1+\epsilon} \frac{d}{n-d-1}$.

A key part of the analysis is showing that the leverage scores of the sketched matrix are nearly uniform with high probability. This is established in the following lemma.

Lemma 22 *With probability $1 - \delta$, for every $i \in [n]$ we have:*

$$(\mathbf{A}^\top \mathbf{S}_{-i}^\top \mathbf{S}_{-i}\mathbf{A})^{-1} \preceq O(1) \cdot (\mathbf{A}^\top \mathbf{A})^{-1} \qquad \text{and} \qquad \mathbf{s}_i^\top \mathbf{A}(\mathbf{A}^\top \mathbf{S}_{-i}^\top \mathbf{S}_{-i}\mathbf{A})^{-1}\mathbf{A}^\top \mathbf{s}_i \approx_{1+\epsilon} \frac{d}{n-d},$$

for $\epsilon = \tilde{O}(1/\sqrt{d})$. In particular, this implies that $\ell_i(\mathbf{SA}) \approx_{1+2\epsilon} \frac{d}{n}$ for all i.

We now define a high probability event $\mathcal{E}$ that ensures uniformity of the leverage scores:

$$\mathcal{E} = \left[(\mathbf{A}^\top \mathbf{S}_{-i}^\top \mathbf{S}_{-i}\mathbf{A})^{-1} \preceq O(1) \cdot (\mathbf{A}^\top \mathbf{A})^{-1}, \quad \mathbf{s}_i^\top \mathbf{A}(\mathbf{A}^\top \mathbf{S}_{-i}^\top \mathbf{S}_{-i}\mathbf{A})^{-1}\mathbf{A}^\top \mathbf{s}_i \approx_{1+\epsilon} \frac{d}{n-d} \quad \forall i\right].$$

In particular, the event implies that $1 - \ell_i(\mathbf{SA}) \approx_{1+\epsilon'} 1 - \frac{d}{m}$ for $\epsilon' = \tilde{O}(\sqrt{d}/n)$, which allows us to simplify the CV loss by approximating all the leverage scores with a uniform estimate. In what follows, we will use $\mathbb{E}_{\mathcal{E}}$ to denote expectation conditioned on the event $\mathcal{E}$:

$$\mathbb{E}_{\mathcal{E}}[L_{\mathrm{CV}}] \approx_{1+\epsilon'} \mathbb{E}_{\mathcal{E}}\left[\sum_{i=1}^{n} \left(\frac{\mathbf{s}_i^\top \mathbf{A}\widehat{\mathbf{w}} - \mathbf{s}_i^\top \mathbf{b}}{1 - d/n}\right)^2\right] = \left(\frac{n}{n-d}\right)^2 \cdot \mathbb{E}_{\mathcal{E}}\left[\|\mathbf{S}(\mathbf{A}\widehat{\mathbf{w}} - \mathbf{b})\|^2\right].$$

Thus, it remains to approximate $\mathbb{E}_{\mathcal{E}}\left[\min_{\mathbf{w}}\|\mathbf{S}(\mathbf{A}\mathbf{w}-\mathbf{b})\|^2\right]$ in terms of $L(\mathbf{w}^*)=\min_{\mathbf{w}}\|\mathbf{A}\mathbf{w}-\mathbf{b}\|^2$. Here, we are going to take advantage of the similar form of the two quantities, in that they can be both written as squared distance between a vector and a subspace. In the case of $L(\mathbf{w}^*)$, it is the squared distance between $\mathbf{b}$ and the column-span of $\mathbf{A}$, and it can be written as:

$$L(\mathbf{w}^*) = \mathbf{b}^\top(\mathbf{I}-\mathbf{H})\mathbf{b}, \qquad \text{for} \quad \mathbf{H} = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A},$$

where $\mathbf{H}$ is the so-called hat matrix, an orthogonal projection onto the column-span of $\mathbf{A}$. In the sketched version, we get a squared distance between $\mathbf{S}\mathbf{b}$ and the column-span of $\mathbf{S}\mathbf{A}$, which can be written similarly, using a sketched version of the hat matrix:

$$\min_{\mathbf{w}}\|\mathbf{S}(\mathbf{A}\mathbf{w}-\mathbf{b})\|^2 = \mathbf{b}^\top\mathbf{S}^\top(\mathbf{I}-\widehat{\mathbf{H}})\mathbf{S}\mathbf{b}, \qquad \text{for} \quad \widehat{\mathbf{H}} = \mathbf{S}\mathbf{A}(\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{S}^\top.$$

Thus, both quantities are quadratic forms applied to the vector $\mathbf{b}$. We relate these two quadratic forms in expectation through the following result.

Lemma 23 *The sketched hat matrix $\widehat{\mathbf{H}}$ satisfies the following approximate expectation formula:*

$$\mathbb{E}_{\mathcal{E}}\left[\mathbf{S}^\top(\mathbf{I}-\widehat{\mathbf{H}})\mathbf{S}\right] \approx_{1+\epsilon'} \left(1-\frac{d}{n}\right)\cdot(\mathbf{I}-\mathbf{H}), \qquad \text{for} \quad \epsilon' = \tilde{O}(\sqrt{d}/n).$$

Remark 24 *Here, $\approx_{1+\epsilon}$ refers to upper/lower bounds in terms of the positive semidefinite ordering. This result extends an exact expectation formula that holds when $\mathbf{S}$ is a Gaussian embedding.*

Putting everything together, we obtain the following approximation of the expected CV loss conditioned on $\mathcal{E}$:

$$\mathbb{E}_{\mathcal{E}}[L_{\mathrm{CV}}] \approx_{1+\epsilon'} \left(\frac{n}{n-d}\right)^2 \mathbf{b}^\top\mathbb{E}_{\mathcal{E}}\left[\mathbf{S}^\top(\mathbf{I}-\widehat{\mathbf{H}})\mathbf{S}\right]\mathbf{b}$$

$$\approx_{1+\epsilon'} \left(\frac{n}{n-d}\right)^2\left(1-\frac{d}{n}\right)\cdot L(\mathbf{w}^*) = \frac{n}{n-d}\cdot L(\mathbf{w}^*).$$

From this it follows that

$$\mathbb{E}_{\mathcal{E}}[L_{\mathrm{CV}}] - L(\mathbf{w}^*) \approx_{1+\epsilon} \frac{d}{n-d}\cdot L(\mathbf{w}^*) \approx_{1+\epsilon} \frac{d}{n-d-1}\cdot L(\mathbf{w}^*),$$

for $\epsilon = \tilde{O}(1/\sqrt{d})$. Note that for this we have to switch from $\epsilon' = \tilde{O}(\sqrt{d}/n)$ to $\epsilon = \tilde{O}(1/\sqrt{d})$, because the right-hand side gets scaled by d/n. Finally, we convert from conditional expectation as follows:

$$(1-\delta)\mathbb{E}_{\mathcal{E}}[L_{\mathrm{CV}}] \leq \mathbb{E}[L_{\mathrm{CV}}] \leq \mathbb{E}_{\mathcal{E}}[L_{\mathrm{CV}}] + \delta\sup_{\mathbf{S}} L_{\mathrm{CV}}$$

$$\leq \mathbb{E}_{\mathcal{E}}[L_{\mathrm{CV}}] + \delta\cdot n\cdot\sup_{\mathbf{S}}\|\mathbf{s}_n\|^2\|\mathbf{A}\widehat{\mathbf{w}}^{-n}-\mathbf{b}\|^2$$

$$\leq \mathbb{E}_{\mathcal{E}}[L_{\mathrm{CV}}] + \delta\cdot\mathrm{poly}(N,D),$$

where in the last step, to avoid the corner cases where sketch $\mathbf{S}\mathbf{A}$ is extremely ill-conditioned, we used that $\widehat{\mathbf{w}}$, as well as $\mathbf{A}$ and $\mathbf{b}$, have entries in $[-D, D]$, and that, by our assumption on the sparsifying distribution p, a LESS embedding matrix $\mathbf{S}$ has entries uniformly bounded by $O(\sqrt{Nk})$.

Choosing sufficiently small $\delta \leq \epsilon L(\mathbf{w}^*)/\text{poly}(N, D)$, we can obsorb that bound into a logarithmic factor that appears in the upper bound on ϵ (if we use d non-zeros per row in LESS). Alternatively, we can increase the number of non-zeros per row in LESS to $d \log(nd/\delta)$, in which case, we avoid the dependence of ϵ on D. We note that another simple way to impose a strict bound on the worst-case loss is to use a small ℓ_2 regularizer when computing the sketched least squares, ensuring that the sketched problem is never ill-conditioned. Extending our analysis to the regularized least squares setting is an interesting direction for future work.

E.1. Proof of Lemma 23

Let $\mathbf{U} = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1/2}$ and note that $\mathbf{U}$ is an $N \times d$ matrix that satisfies $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$. We can rewrite both $\mathbf{H}$ and $\widehat{\mathbf{H}}$ by replacing $\mathbf{A}$ with $\mathbf{U}$, namely: $\mathbf{H} = \mathbf{U}\mathbf{U}^\top$ and $\widehat{\mathbf{H}} = \mathbf{S}\mathbf{U}(\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{S}^\top$. To show the claim, it suffices to show that for every vector $\mathbf{v} \in \mathbb{R}^N$, the following approximation holds:

$$\mathbb{E}_{\mathcal{E}}[\mathbf{v}^\top\mathbf{S}^\top(\mathbf{I} - \widehat{\mathbf{H}})\mathbf{S}\mathbf{v}] \approx_{1+\epsilon'} \left(1 - \frac{d}{n}\right)\mathbf{v}^\top(\mathbf{I} - \mathbf{H})\mathbf{v}. \tag{6}$$

First, if $\mathbf{v}$ lies in the column-span of $\mathbf{U}$, i.e., $\mathbf{v} = \mathbf{U}\mathbf{x}$ for some $\mathbf{x}$, then clearly the right-hand side is zero, and also:

$$\mathbf{v}^\top\mathbf{S}^\top(\mathbf{I} - \widehat{\mathbf{H}})\mathbf{S}\mathbf{v} = \mathbf{x}^\top\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U}\mathbf{x} - \mathbf{x}^\top\mathbf{U}^\top\mathbf{S}^\top\widehat{\mathbf{H}}\mathbf{S}\mathbf{U}\mathbf{x} = 0 = \left(1 - \frac{d}{n}\right)\mathbf{v}^\top(\mathbf{I} - \mathbf{H})\mathbf{v}.$$

Thus, it suffices to show (6) for $\mathbf{v}$ orthogonal to the column-span of $\mathbf{U}$, i.e., one that satisfies $\mathbf{U}^\top\mathbf{v} = \mathbf{0}$. In that case,

$$\left|\mathbf{v}^\top\mathbb{E}_{\mathcal{E}}[\mathbf{S}^\top(\mathbf{I} - \widehat{\mathbf{H}})\mathbf{S}]\mathbf{v} - \left(1 - \frac{d}{n}\right)\mathbf{v}^\top(\mathbf{I} - \mathbf{H})\mathbf{v}\right| = \left|\mathbf{v}^\top\mathbb{E}_{\mathcal{E}}[\mathbf{S}^\top\mathbf{S}]\mathbf{v} - \mathbf{v}^\top\mathbb{E}_{\mathcal{E}}[\mathbf{S}^\top\widehat{\mathbf{H}}\mathbf{S}]\mathbf{v} - \left(1 - \frac{d}{n}\right)\right|.$$

We now focus on analyzing $\mathbb{E}_{\mathcal{E}}[\mathbf{S}^\top\widehat{\mathbf{H}}\mathbf{S}]$. To that end, let us use the shorthands $\mathbf{Q} = (\gamma\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U})^{-1}$ and $\mathbf{Q}_{-i} = (\gamma\mathbf{U}^\top\mathbf{S}_{-i}^\top\mathbf{S}_{-i}\mathbf{U})^{-1}$, where $\gamma = \frac{n}{n-d}$, so that $\widehat{\mathbf{H}} = \mathbf{S}\mathbf{U}\gamma\mathbf{Q}\mathbf{U}^\top\mathbf{S}^\top$. We also use the formula $\mathbf{s}_i^\top\mathbf{U}\mathbf{Q} = \mathbf{s}_i^\top\mathbf{U}\mathbf{Q}_{-i}/\gamma_i$, where $\gamma_i = 1 + \gamma\mathbf{s}_i^\top\mathbf{U}\mathbf{Q}_{-i}\mathbf{U}^\top\mathbf{s}_i$, which is a consequence of the Sherman-Morrison formula (Lemma 19). Recall that, from Lemma 22, the event $\mathcal{E}$ implies that $\gamma_i \approx \gamma$. Also, let us use a simplifying shorthand of $\tilde{\mathbf{s}}_i = \sqrt{n}\,\mathbf{s}_i$, so that $\mathbb{E}[\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i^\top] = \mathbf{I}$. We can now rewrite the expectation as follows (where i is any fixed index):

$$\mathbb{E}_{\mathcal{E}}[\mathbf{S}^\top\widehat{\mathbf{H}}\mathbf{S}] = \mathbb{E}_{\mathcal{E}}[\mathbf{S}^\top\mathbf{S}\mathbf{U}\gamma\mathbf{Q}\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}] = \mathbb{E}_{\mathcal{E}}[\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i^\top\mathbf{U}\gamma\mathbf{Q}\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}] = \mathbb{E}_{\mathcal{E}}\left[\frac{\gamma}{\gamma_i}\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i^\top\mathbf{U}\mathbf{Q}_{-i}\mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\right]$$

$$= \frac{1}{n}\mathbb{E}_{\mathcal{E}}\left[\frac{\gamma}{\gamma_i}\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i^\top\mathbf{U}\mathbf{Q}_{-i}\mathbf{U}^\top\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i^\top\right] + \mathbb{E}_{\mathcal{E}}\left[\frac{\gamma}{\gamma_i}\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i^\top\mathbf{U}\mathbf{Q}_{-i}\mathbf{U}^\top\mathbf{S}_{-i}^\top\mathbf{S}_{-i}\right]$$

$$= \mathbb{E}_{\mathcal{E}}\left[\frac{\gamma_i - 1}{\gamma_i}\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i^\top\right] + \mathbb{E}_{\mathcal{E}}\left[\frac{\gamma}{\gamma_i}\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i^\top\mathbf{U}\mathbf{Q}_{-i}\mathbf{U}^\top\mathbf{S}_{-i}^\top\mathbf{S}_{-i}\right]$$

$$= \mathbb{E}_{\mathcal{E}}\left[\frac{\gamma - 1}{\gamma} \cdot \tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i^\top\right] + \mathbb{E}_{\mathcal{E}}\left[\mathbf{U}\mathbf{Q}_{-i}\mathbf{U}^\top\mathbf{S}_{-i}^\top\mathbf{S}_{-i}\right]$$

$$+ \mathbb{E}_{\mathcal{E}}\left[\left(\frac{1}{\gamma} - \frac{1}{\gamma_i}\right)\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i^\top\right] + \mathbb{E}_{\mathcal{E}}\left[\left(\frac{\gamma}{\gamma_i} - 1\right)\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i^\top\mathbf{U}\mathbf{Q}_{-i}\mathbf{U}^\top\mathbf{S}_{-i}^\top\mathbf{S}_{-i}\right] =: \mathbf{T}_1 + \mathbf{T}_2 + \mathbf{T}_3 + \mathbf{T}_4.$$

Thus, it follows that:

$$\mathbb{E}_{\mathcal{E}}\big[\mathbf{S}^\top(\mathbf{I} - \widehat{\mathbf{H}})\mathbf{S}\big] = \mathbb{E}_{\mathcal{E}}\big[\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i^\top\big] - \mathbb{E}_{\mathcal{E}}\Big[\frac{\gamma - 1}{\gamma}\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i^\top\Big] - (\mathbf{T}_2 + \mathbf{T}_3 + \mathbf{T}_4)$$

$$= \Big(1 - \frac{d}{n}\Big)\mathbb{E}_{\mathcal{E}}[\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i^\top] - (\mathbf{T}_2 + \mathbf{T}_3 + \mathbf{T}_4).$$

In the following steps, we will show that $\mathbb{E}_{\mathcal{E}}[\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i^\top] \approx \mathbf{I}$ and that $\mathbf{v}^\top\mathbf{T}_k\mathbf{v} \approx 0$ for $k \in \{2, 3, 4\}$. First, denoting $\delta = \Pr(\neg\mathcal{E})$, we have:

$$\|\mathbb{E}_{\mathcal{E}}[\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i^\top] - \mathbf{I}\| = \frac{1}{1 - \delta}\|\mathbb{E}[\mathbf{1}_{\neg\mathcal{E}}\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i]\| \leq \frac{\delta}{1 - \delta} \cdot O(Nk),$$

where we used that $\|\tilde{\mathbf{s}}_i\|^2 \leq O(Nk)$. Next, observe that for $\mathbf{v}$ orthogonal to the column-span of $\mathbf{U}$, we have $\mathbf{v}^\top\mathbf{T}_2\mathbf{v} = \mathbf{v}^\top\mathbf{U}\mathbf{Q}_{-i}\mathbf{U}^\top\mathbf{S}_{-i}^\top\mathbf{S}_{-i}\mathbf{v} = 0$ because $\mathbf{v}^\top\mathbf{U} = \mathbf{0}$. Furthermore, using Lemma 22, conditioned on $\mathcal{E}$ we know that $|\gamma_i - \gamma| = \tilde{O}(\sqrt{d}/n)$, which allows us to control the remaining terms. In the following, we also use the fact that $\mathbb{E}_{\mathcal{E}}[X] \leq \frac{1}{1-\delta}\mathbb{E}[X]$ for any non-negative X:

$$\big|\mathbf{v}^\top\mathbf{T}_3\mathbf{v}\big| \leq \mathbb{E}_{\mathcal{E}}\Big[\frac{|\gamma_i - \gamma|}{\gamma\gamma_i}(\mathbf{v}^\top\tilde{\mathbf{s}}_i)^2\Big] \leq \tilde{O}(\sqrt{d}/n) \cdot \mathbb{E}_{\mathcal{E}}[(\mathbf{v}^\top\tilde{\mathbf{s}}_i)^2]$$

$$\leq \tilde{O}(\sqrt{d}/n) \cdot \mathbf{v}^\top\mathbb{E}[\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i]\mathbf{v} = \tilde{O}(\sqrt{d}/n) \cdot \|\mathbf{v}\|^2.$$

Similarly, we bound the last term, using the Cauchy-Schwartz inequality:

$$\big|\mathbf{v}^\top\mathbf{T}_4\mathbf{v}\big| \leq \tilde{O}(\sqrt{d}/n) \cdot \mathbb{E}_{\mathcal{E}}\big[|\mathbf{v}^\top\tilde{\mathbf{s}}_i| \cdot |\tilde{\mathbf{s}}_i^\top\mathbf{U}\mathbf{Q}_{-i}\mathbf{U}^\top\mathbf{S}_{-i}^\top\mathbf{S}_{-i}\mathbf{v}|\big]$$

$$\leq \tilde{O}(\sqrt{d}/n) \cdot \sqrt{\mathbb{E}_{\mathcal{E}}\big[(\mathbf{v}^\top\tilde{\mathbf{s}}_i)^2\big]} \cdot \sqrt{\mathbb{E}_{\mathcal{E}}\big[(\tilde{\mathbf{s}}_i^\top\mathbf{U}\mathbf{Q}_{-i}\mathbf{U}^\top\mathbf{S}_{-i}^\top\mathbf{S}_{-i}\mathbf{v})^2\big]}.$$

As before, we have $\mathbb{E}_{\mathcal{E}}[(\mathbf{v}^\top\tilde{\mathbf{s}}_i)^2] \leq O(1) \cdot \|\mathbf{v}\|^2$. For the last term in the product, we use a LESS embedding property of the random vector vector $\mathbf{U}^\top\tilde{\mathbf{s}}_i$ that was shown in (Dereziński et al., 2021b, Lemma 28), which they called the Bai-Silverstein property. In particular, this property implies that for any vector $\mathbf{v}$, we have $\mathbb{E}[(\tilde{\mathbf{s}}_i^\top\mathbf{U}\mathbf{v})^2] \leq O(1) \cdot \|\mathbf{v}\|^2$ (this property also holds for any sub-gaussian vector, so it could be easily inferred from our Theorem 3, after conditioning on a high-probability event). We also use that the event $\mathcal{E}$ implies that $\|\mathbf{Q}_{-i}\| = O(1)$. It follows that:

$$\mathbb{E}_{\mathcal{E}}\big[(\tilde{\mathbf{s}}_i^\top\mathbf{U}\mathbf{Q}_{-i}\mathbf{U}^\top\mathbf{S}_{-i}^\top\mathbf{S}_{-i}\mathbf{v})^2\big] \leq \frac{1}{1 - \delta}\mathbb{E}\big[(\tilde{\mathbf{s}}_i^\top\mathbf{U}\mathbf{Q}_{-i}\mathbf{U}^\top\mathbf{S}_{-i}^\top\mathbf{S}_{-i}\mathbf{v})^2 \mid \|\mathbf{Q}_{-i}\| = O(1)\big]$$

$$\leq O(1) \cdot \mathbb{E}\big[\|\mathbf{Q}_{-i}\mathbf{U}^\top\mathbf{S}_{-i}^\top\mathbf{S}_{-i}\mathbf{v}\|^2 \mid \|\mathbf{Q}_{-i}\| = O(1)\big]$$

$$\leq O(1) \cdot \mathbb{E}\big[\|\mathbf{U}^\top\mathbf{S}_{-i}^\top\mathbf{S}_{-i}\mathbf{v}\|^2\big]$$

$$= O(1) \cdot \mathbb{E}\bigg[\Big\|\frac{1}{n}\sum_{j\neq i}\mathbf{U}^\top\tilde{\mathbf{s}}_j\tilde{\mathbf{s}}_j^\top\mathbf{v}\Big\|^2\bigg] \leq O(1) \cdot \frac{1}{n}\mathbb{E}\big[\|\mathbf{U}^\top\tilde{\mathbf{s}}_j\tilde{\mathbf{s}}_j^\top\mathbf{v}\|^2\big],$$

where in the last step we used that $\tilde{\mathbf{s}}_j$ are independent, and $\mathbf{U}^\top\mathbf{v} = \mathbf{0}$. To bound $\mathbb{E}[\|\mathbf{U}^\top\tilde{\mathbf{s}}_j\tilde{\mathbf{s}}_j^\top\mathbf{v}\|^2]$, we will use the definition of a p-sparsified embedding (Definition 2), where we let I_i be the k random indices that indicate the non-zero entries of $\tilde{\mathbf{s}}_j = \sum_{i=1}^k \frac{r_i\mathbf{e}_{I_i}}{\sqrt{kp_{I_i}}}$ and r_i's are the random signs.

$$\mathbb{E}\big[\|\mathbf{U}^\top\tilde{\mathbf{s}}_j\tilde{\mathbf{s}}_j^\top\mathbf{v}\|^2\big] \leq k\,\mathbb{E}\bigg[\frac{\|\mathbf{U}^\top\mathbf{e}_{I_1}\|^2(\mathbf{e}_{I_1}^\top\mathbf{v})^2}{k^2p_{I_1}^2}\bigg] + k^2\,\mathbb{E}\bigg[\frac{\|\mathbf{U}^\top\mathbf{e}_{I_1}\|^2}{kp_{I_1}}\bigg]\mathbb{E}\bigg[\frac{(\mathbf{e}_{I_2}^\top\mathbf{v})^2}{kp_{I_2}}\bigg] \leq \big(Cd/k + Cd\big) \cdot \|\mathbf{v}\|^2,$$

where, to eliminate the cross-terms when expanding the square, we used independence of I_i's and that $\mathbf{U}^\top \mathbf{v} = \mathbf{0}$. In the second step, we used that $p_{I_i} \geq \|\mathbf{U}^\top \mathbf{e}_{I_i}\|^2/(Cd)$. Thus, $|\mathbf{v}^\top \mathbf{T}_4 \mathbf{v}| \leq \tilde{O}(d/n^{1.5})\|\mathbf{v}\|^2 = \tilde{O}(\sqrt{d}/n)\|\mathbf{v}\|^2$. We conclude that for sufficiently small δ, claim (6) is satisfied with $\epsilon = \tilde{O}(\sqrt{d}/n)$, which completes the proof.

E.2. Proof of Lemma 22

We note that the first claim, i.e, that $(\mathbf{A}^\top \mathbf{S}_{-i} \mathbf{S}_{-i} \mathbf{A})^{-1} \preceq O(1) \cdot (\mathbf{A}^\top \mathbf{A})^{-1}$ follows easily for LESS embeddings with d non-zeros per row of size $n \geq O(d \log(d/\delta))$, by relying on the subspace embedding property, shown in (Dereziński et al., 2021b, Lemma 12), which essentially follows from the matrix Bernstein inequality (Lemma 18). An alternative way of showing this claim is to rely on the subspace embedding guarantee for sub-gaussian embeddings, and then convert that to a result for LESS embeddings via our Theorem 3. This approach is given in Corollary 16. Interestingly, this leads to a slightly different trade-off in terms of the logarithmic factors. Namely, if we consider a LESS embedding with $d \log(nd/\delta)$ non-zeros per row, slightly more than recommended by Dereziński et al. (2021b), then the subspace embedding guarantee holds for $n \geq O(d)$ samples, rather than $n \geq O(d \log d)$. Such a guarantee is new for LESS embeddings, since it cannot be obtained via the matrix Bernstein inequality.

We now focus on establishing the second part of the claim, i.e., the uniformity of the leverage scores, which is where the Hanson-Wright inequality is essential. Since we want to show a statement that holds with high probability, we can use Theorem 3 to show this claim for an $n \times d$ sub-gaussian design $\mathbf{X} = \mathbf{Z}\boldsymbol{\Sigma}^{1/2}$ with $\mathbf{Z}$ consisting of mean zero isotropic rows that satisfy the Hanson-Wright inequality (2) with constant $\tilde{O}(1)$, and then convert the claim to LESS embeddings. In this case, our goal is to show that for $n \geq \tilde{O}(d)$, with probability $1 - \delta$:

$$\mathbf{x}_i^\top \left(\mathbf{X}_{-i}^\top \mathbf{X}_{-i}\right)^{-1} \mathbf{x}_i \approx_{1+\epsilon} \frac{d}{n-d}.$$

Note that, by Hanson-Wright, with high probability we immediately have that $\mathbf{x}_i^\top \left(\mathbf{X}_{-i}^\top \mathbf{X}_{-i}\right)^{-1} \mathbf{x}_i \approx \operatorname{tr} \boldsymbol{\Sigma}(\mathbf{X}_{-i}^\top \mathbf{X}_{-i})^{-1}$ (this will be formalized later on), so the main challenge is in showing that the trace functional is also concentrated. Below, we provide a slightly more general result, regarding the concentration properties of a certain class of trace functionals for a sub-gaussian design, which may be of independent interest. (Below, we use $a \lesssim b$ to denote $a \leq cb$ for an absolute constant c).

Lemma 25 *Suppose that $\mathbf{X}$ consists of rows $\mathbf{x}_1^\top, \mathbf{x}_2^\top, ..., \mathbf{x}_n^\top$ for $n \geq \tilde{O}(d)$, that are i.i.d. sampled from a d-variate distribution with covariance $\mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\top] = \boldsymbol{\Sigma}$ such that $\boldsymbol{\Sigma}^{-1/2}\mathbf{x}_i$ satisfies the Hanson-Wright inequality (2) with constant $K = \tilde{O}(1)$. Consider $F(\mathbf{X}) = \operatorname{tr} \mathbf{B}(\frac{1}{n}\mathbf{X}^\top \mathbf{X} + \mathbf{C})^{-1}$ for some psd matrices $\mathbf{B}$ and $\mathbf{C}$. Then, there is a scalar $\tilde{F}$ such that with probability $1 - \delta$:*

$$|F(\mathbf{X}) - \tilde{F}| \leq C\|\mathbf{M}\| \cdot \frac{\operatorname{r}(\mathbf{M}) + K^4}{\sqrt{n}} \log 1/\delta,$$

where $\mathbf{M} = \mathbf{B}^{1/2}(\boldsymbol{\Sigma} + \mathbf{C})^{-1}\mathbf{B}^{1/2}$, $\operatorname{r}(\mathbf{M}) = \operatorname{tr}(\mathbf{M})/\|\mathbf{M}\|$, and C is an absolute constant.

Proof Note that $\mathbf{C}$ may not be positive definite (in fact, of our primary interest is when $\mathbf{C} = \mathbf{0}$), so function $F(\mathbf{X})$ may not be well defined everywhere because of the inverse. In particular, it may not have a bounded expectation. So, to define our scalar $\tilde{F}$, we first construct a high-probability event

$\mathcal{E}$ which will ensure that $F(\mathbf{X})$ is not only well-defined, but also sufficiently well behaved. Without loss of generality, assume that n is even. Now, define:

$$\mathcal{E}_1 = \Big[\frac{2}{n}\sum_{i=1}^{n/2}\mathbf{x}_i\mathbf{x}_i^\top \succeq \frac{1}{C}\boldsymbol{\Sigma}\Big], \qquad \mathcal{E}_2 = \Big[\frac{2}{n}\sum_{i=n/2+1}^{n}\mathbf{x}_i\mathbf{x}_i^\top \succeq \frac{1}{C}\boldsymbol{\Sigma}\Big],$$

with $\mathcal{E} = \mathcal{E}_1 \wedge \mathcal{E}_2$. Since $n \geq \tilde{O}(d)$, by relying on the subspace embedding property (as discussed at the beginning of the section), each of the events $\mathcal{E}_1$ and $\mathcal{E}_2$ holds with probability at least $1 - \delta$. We will use $\tilde{F} = \mathbb{E}_{\mathcal{E}}[F(\mathbf{X})]$ as our scalar. Note that, conditioned on $\mathcal{E}$, the function F satisfies $0 \leq F(\mathbf{X}) \leq C\operatorname{tr}\mathbf{M}$, so in particular the conditional expectation is well-defined.

Next, we proceed with the following decomposition of the trace (following standard literature in random matrix theory). Let $\mathbf{Q} = (\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \mathbf{C})^{-1}$ and $\mathbf{Q}_{-i} = (\frac{1}{n}\mathbf{X}_{-i}^\top\mathbf{X}_{-i} + \mathbf{C})^{-1}$ where $\mathbf{X}_{-i}$ denotes $\mathbf{X}$ without the i row. Also, let $\mathbb{E}_i$ denote expectation conditioned on $\mathbf{x}_1, ..., \mathbf{x}_i$. We have, conditioned on $\mathcal{E}$ that:

$$F(\mathbf{X}) - \tilde{F} = \mathbb{E}_n[\operatorname{tr}\mathbf{B}\mathbf{Q} \mid \mathcal{E}] - \mathbb{E}_0[\operatorname{tr}\mathbf{B}\mathbf{Q} \mid \mathcal{E}] = \sum_{i=1}^{n}(\mathbb{E}_i - \mathbb{E}_{i-1})[\operatorname{tr}\mathbf{B}\mathbf{Q} \mid \mathcal{E}]$$

$$= \sum_{i=1}^{n}(\mathbb{E}_i - \mathbb{E}_{i-1})[\operatorname{tr}\mathbf{B}(\mathbf{Q} - \mathbf{Q}_{-i}) \mid \mathcal{E}] + (\mathbb{E}_i - \mathbb{E}_{i-1})[\operatorname{tr}\mathbf{B}\mathbf{Q}_{-i} \mid \mathcal{E}].$$

We proceed to bound each of the two terms. The latter can be bounded straightforwardly. Note that one of $\mathcal{E}_1$ or $\mathcal{E}_2$ is independent of $\mathbf{x}_i$. Without loss of generality suppose that it is $\mathcal{E}_1$. Then, we have:

$$\big|(\mathbb{E}_i - \mathbb{E}_{i-1})[\operatorname{tr}\mathbf{B}\mathbf{Q}_{-i} \mid \mathcal{E}]\big| \leq \frac{\delta}{1-\delta} \cdot \big|(\mathbb{E}_i - \mathbb{E}_{i-1})[\operatorname{tr}\mathbf{B}\mathbf{Q}_{-i} \mid \mathcal{E}_1 \wedge \neg\mathcal{E}_2]\big| \lesssim \delta \cdot \operatorname{tr}\mathbf{M}.$$

To bound the first term, we use the Sherman-Morrison rank-one update formula (Lemma 19), observing that:

$$|\operatorname{tr}\mathbf{B}(\mathbf{Q} - \mathbf{Q}_{-i})| = \frac{\frac{1}{n}\operatorname{tr}\mathbf{B}\mathbf{Q}_{-i}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{Q}_{-i}}{1 + \frac{1}{n}\mathbf{x}_i^\top\mathbf{Q}_{-i}\mathbf{x}_i} \leq \frac{1}{n}\mathbf{x}_i^\top\mathbf{Q}_{-i}\mathbf{B}\mathbf{Q}_{-i}\mathbf{x}_i = \frac{1}{n}\mathbf{z}_i^\top\tilde{\mathbf{M}}_i\mathbf{z}_i,$$

where we let $\mathbf{z}_i = \boldsymbol{\Sigma}^{-1/2}\mathbf{x}_i$ and $\tilde{\mathbf{M}}_i = \boldsymbol{\Sigma}^{1/2}\mathbf{Q}_{-i}\mathbf{B}\mathbf{Q}_{-i}\boldsymbol{\Sigma}^{1/2}$. Now, since $\mathbf{z}_i$ satisfies Hanson-Wright inequality with constant K, it also satisfies Euclidean concentration with constant $O(K^2)$ (see Proposition 14), and in particular, letting $Z_i = (\mathbf{z}_i^\top\tilde{\mathbf{M}}_i\mathbf{z}_i)^{1/2}$ and conditioning on $\tilde{\mathbf{M}}_i$ we have: $\|Z_i - \mathbb{E}[Z_i \mid \tilde{\mathbf{M}}_i]\|_{\psi_2} \lesssim K^2\|\tilde{\mathbf{M}}_i\|^{1/2}$. Thus, defining $M_i = \mathbb{E}[Z_i \mid \tilde{\mathbf{M}}_i] \leq (\operatorname{tr}\tilde{\mathbf{M}}_i)^{1/2}$, we have:

$$\mathbb{E}\big[|\mathbf{z}_i^\top\tilde{\mathbf{M}}_i\mathbf{z}_i|^p \mid \mathcal{E}\big]^{1/p} \leq \frac{1}{1-\delta}\mathbb{E}\big[|\mathbf{z}_i^\top\tilde{\mathbf{M}}_i\mathbf{z}_i|^p \mid \mathcal{E}_1\big]^{1/p} = \frac{1}{1-\delta}\mathbb{E}\big[(M_i - M_i + Z_i)^{2p} \mid \mathcal{E}_1\big]^{1/p}$$

$$\lesssim \mathbb{E}\big[\mathbb{E}[M_i^{2p} \mid \tilde{\mathbf{M}}_i] \mid \mathcal{E}_1\big]^{1/p} + \mathbb{E}\big[\mathbb{E}[|Z_i - M_i|^{2p} \mid \tilde{\mathbf{M}}_i] \mid \mathcal{E}_1\big]^{1/p}$$

$$\lesssim \mathbb{E}\big[(\operatorname{tr}\tilde{\mathbf{M}}_i)^p \mid \mathcal{E}_1\big]^{1/p} + \mathbb{E}\big[(\sqrt{p}K^2\|\tilde{\mathbf{M}}_i\|^{1/2})^{2p} \mid \mathcal{E}_1\big]^{1/p}$$

$$\lesssim \operatorname{tr}\mathbf{M} + pK^4\|\mathbf{M}\|.$$

where the last step follows because conditioned on $\mathcal{E}_1$ we have $\operatorname{tr}\tilde{\mathbf{M}}_i \lesssim \operatorname{tr}\mathbf{M}$ and $\|\tilde{\mathbf{M}}_i\| \lesssim \|\mathbf{M}\|$. Using Burkholder's inequality (e.g., see Hitczenko, 1990) for the martingale difference sequence

$X_i = (\mathbb{E}_i - \mathbb{E}_{i-1})[\text{tr}\mathbf{B}\mathbf{Q} \mid \mathcal{E}]$, we get:

$$\mathbb{E}\big[|F(X) - \tilde{F}|^p\big]^{1/p} = \mathbb{E}\Big[\Big(\sum_{i=1}^{n} X_i\Big)^p\Big]^{1/p} \leq Cp \cdot \Big(\mathbb{E}\Big[\Big(\sum_{i=1}^{n} \mathbb{E}_{i-1}[X_i^2]\Big)^{p/2}\Big] + \sum_{i=1}^{n} \mathbb{E}\big[X_i^p\big]\Big)^{1/p}$$

$$\lesssim \frac{p}{n} \cdot \Big(\Big(n((\text{tr}\,\mathbf{M})^2 + K^4\|\mathbf{M}\|)\Big)^{p/2} + n(\text{tr}\,\mathbf{M} + pK^4\|\mathbf{M}\|)^p\Big)^{1/p}$$

$$\lesssim p\big(\text{tr}\,\mathbf{M} + K^4\|\mathbf{M}\|\big) \cdot \Big(\frac{1}{\sqrt{n}} + \frac{p}{n^{1-1/p}}\Big).$$

Applying Markov's inequality with $p = \log(1/\delta)$, we obtain the desired result. ∎

Note that in Lemma 25 we did not explicitly state what is the scalar $\tilde{F}$ around which the trace functional $F(\mathbf{X})$ concentrates. Normally, one would expect this quantity to be the expectation of $F(\mathbf{X})$, however since this expectation may not exist, we instead realy on the notion of a nearly-unbiased estimator. For the trace functional we are interested in, i.e., $F(\mathbf{X}) = \text{tr}\,\mathbf{\Sigma}(\mathbf{X}^\top\mathbf{X})^{-1}$, a nearly-unbiased estimator has been derived by Dereziński et al. (2021b), and we use that result in the following corollary.

Corollary 26 *Suppose that $\mathbf{X}$ consists of rows $\mathbf{x}_1^\top, \mathbf{x}_2^\top, ..., \mathbf{x}_n^\top$ that are i.i.d. sampled from a d-variate distribution with covariance $\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\top] = \mathbf{\Sigma}$ such that $\mathbf{\Sigma}^{-1/2}\mathbf{x}_i$ satisfies the Hanson-Wright inequality (2) with $K = \tilde{O}(1)$ and $n \geq \tilde{O}(d)$. Then, with probability $1 - \delta$:*

$$\text{tr}\,\mathbf{\Sigma}(\mathbf{X}^\top\mathbf{X})^{-1} \approx_{1+\epsilon} \frac{d}{n-d} \quad \text{where} \quad \epsilon \leq \tilde{O}\Big(\frac{\log 1/\delta}{\sqrt{n}}\Big).$$

Proof Here, we use the result from Dereziński et al. (2021b) which shows that $\mathbf{\Sigma}^{1/2}(\frac{1}{n}\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{\Sigma}^{1/2}$ is an (ϵ', δ)-unbiased estimator of $\frac{n}{n-d}\mathbf{I}$, where $\epsilon' = O(K^4\sqrt{d}/n)$. Namely, this means that there is an event $\mathcal{E}$ with probability $1 - \delta$ such that $\mathbb{E}[\mathbf{\Sigma}^{1/2}(\frac{1}{n}\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{\Sigma}^{1/2} \mid \mathcal{E}] \approx_{1+\epsilon} \frac{n}{n-d}\mathbf{I}$ and conditioned on $\mathcal{E}$ we have $\mathbf{\Sigma}^{1/2}(\frac{1}{n}\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{\Sigma}^{1/2} \preceq O(1) \cdot \frac{n}{n-d}\mathbf{I}$. As a consequence, following the argument analogous to Lemma 34 in Dereziński et al. (2021b), we can easily show that if $F(\mathbf{X}) = \text{tr}\,\mathbf{\Sigma}(\frac{1}{n}\mathbf{X}^\top\mathbf{X})^{-1}$, then our scalar $\tilde{F}$ used in Lemma 25 also satisfies $\tilde{F} \approx_{1+\epsilon} \frac{n}{n-d} \cdot d$. So, combining this with Lemma 25, and observing that in this case we have $\mathbf{M} = \mathbf{I}$ completes the proof. ∎

To conclude the proof of Lemma 22, we combine the corollary with the Hanson-Wright inequality, which states that, conditioned on $\mathbf{X}_{-i}^\top\mathbf{X}_{-i}$:

$$\Pr\Big\{|\mathbf{x}_i^\top(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\mathbf{x}_i - \text{tr}\,\mathbf{\Sigma}(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}| \geq \epsilon \cdot \text{tr}\,\mathbf{\Sigma}(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\Big\}$$

$$\leq 2\exp\Big(-c\min\Big\{\frac{\epsilon^2 \cdot (\text{tr}\,\mathbf{\Sigma}(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1})^2}{K^4\|\mathbf{\Sigma}^{1/2}(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\mathbf{\Sigma}^{1/2}\|_F^2}, \frac{\epsilon \cdot \text{tr}\,\mathbf{\Sigma}(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}}{K^2\|\mathbf{\Sigma}^{1/2}(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\mathbf{\Sigma}^{1/2}\|}\Big\}\Big)$$

$$\leq \exp\Big(-c\epsilon^2\frac{\text{tr}\,\mathbf{\Sigma}(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}}{K^4\|\mathbf{\Sigma}^{1/2}(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\mathbf{\Sigma}^{1/2}\|}\Big).$$

Using that $K = \tilde{O}(1)$, as well as that with high probability we have $\operatorname{tr}\boldsymbol{\Sigma}(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1} \approx \frac{d}{n-d}$ and $\|\boldsymbol{\Sigma}^{1/2}(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\boldsymbol{\Sigma}^{1/2}\| = O(1/n)$, we conclude that with high probability, we also have:

$$\mathbf{x}_i^\top(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})^{-1}\mathbf{x}_i \approx_{1+\epsilon} \frac{d}{n-d} \qquad \text{for} \qquad \epsilon = \tilde{O}(1/\sqrt{d}).$$

By coupling a LESS embedding with the sub-gaussian design, we obtain the desired claim.

Appendix F. Constrained least squares: Proof of Theorem 9

The claim can be shown by using our reduction from a LESS embedding to a sub-gaussian embedding. Namely, let $\mathbf{Z}\widetilde{\boldsymbol{\Sigma}}^{1/2}$ be the sub-gaussian design that is coupled with the LESS embedding $\sqrt{n}\mathbf{S}\mathbf{A}$ in Theorem 3. Recall that we have $\widetilde{\boldsymbol{\Sigma}} \approx_{1+\delta} \boldsymbol{\Sigma}$. To simplify the reduction, let us also define the following sketching matrix:

$$\widetilde{\mathbf{S}} = \frac{1}{\sqrt{n}}\big(\mathbf{Z}\widetilde{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{\Sigma}^{-1/2}\mathbf{U}^\top + \mathbf{G}(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\big),$$

where $\mathbf{U} = \mathbf{A}\boldsymbol{\Sigma}^{-1/2}$ and $\mathbf{G}$ is an $n \times N$ Gaussian matrix. It is easy to verify that $\sqrt{n}\widetilde{\mathbf{S}}\mathbf{A} = \mathbf{Z}\boldsymbol{\Sigma}^{1/2} = \sqrt{n}\mathbf{S}\mathbf{A}$ with probability $1 - \delta$, since $\mathbf{I} - \mathbf{U}\mathbf{U}^\top$ is the projection onto the complement of the column-span of $\mathbf{A}$. Moreover, we have $\mathbb{E}[\widetilde{\mathbf{S}}^\top\widetilde{\mathbf{S}}] \approx_{1+\delta} \mathbf{I}$, with each row of $\sqrt{n}\widetilde{\mathbf{S}}$ being mean zero and $\tilde{O}(1)$-sub-gaussian. Thus, it suffices to verify that the analysis of Pilanci and Wainwright (2015) still works when the sketching matrix is only approximately isotropic. Note that the difference here is very small since we can easily let $\delta \ll \epsilon$. It suffices to show that their Proposition 1 can be adapted. This proposition is itself a corollary of Theorem D from Mendelson et al. (2007). We can easily obtain a variant that suits our setup.

Proposition 27 *Let $\tilde{\mathbf{s}}_1, ..., \tilde{\mathbf{s}}_n$ be i.i.d. N-dimensional samples from a zero mean K-sub-gaussian distribution with covariance $\mathbb{E}[\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i^\top] \approx_{1+\delta} \mathbf{I}$. For any subset $\mathcal{Y} \subseteq \mathcal{S}^{N-1}$, if $n \geq CK^2\mathbb{W}^2(\mathcal{Y})/\epsilon^2$ and $\delta \leq c\epsilon$, then:*

$$\sup_{y\in\mathcal{Y}}\Big|\mathbf{y}^\top\Big(\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i^\top - \mathbf{I}\Big)\mathbf{y}\Big| \leq \epsilon,$$

with probability at least $1 - 2\exp(-c\epsilon^2 n/K^4)$.

The proposition follows from Theorem D of Mendelson et al. (2007), with functions $f_y(\mathbf{s}) = \frac{\mathbf{s}^\top\mathbf{y}}{\|\mathbf{y}\|_\mathbf{M}}$, where $\mathbf{M} = \mathbb{E}[\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i^\top]$. Remainder of the analysis in Pilanci and Wainwright (2015) stays the same.

Appendix G. Lower bound: Proof of Theorem 5

Without loss of generality, assume that N is a multiple of d (otherwise we can pad the matrix with zeros), and consider matrix $\mathbf{A}$ consisting of stacked $d \times d$ identity matrices, scaled by $\sqrt{d/N}$ so that $\mathbf{A}^\top\mathbf{A} = \mathbf{I}$. It suffices to prove the result for $n = 1$, in which case the sketching matrix $\mathbf{S}$ and the design matrix $\mathbf{Z}$ both consist of single random vectors $\mathbf{s}$ and $\mathbf{z}$, respectively.

We will in fact show a strictly stronger claim that $\|\mathbf{z}\|_{\psi_2} \geq \min\big\{1, c\sqrt{\frac{d/k}{\log d}}\big\}$, for some constant $c > 0$. This claim is stronger because, by Proposition 14, the Hanson-Wright constant K upper

bounds the sub-gaussian norm as follows: $\|\mathbf{z}\|_{\psi_2} = O(K)$. Observe that since $\mathbf{z}$ is isotropic, it has to satisfy $\|\mathbf{z}\|_{\psi_2} \geq 1$, so it suffices to show $\|\mathbf{z}\|_{\psi_2} \geq c\sqrt{\frac{d/k}{\log d}}$ for $k \leq d/C$, where C is some absolute constant.

Let $\tilde{\mathbf{x}} = \mathbf{A}^\top \mathbf{s} = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \frac{r_i}{\sqrt{p_{I_i}}} \mathbf{a}_{I_i}$, and let $\tilde{\mathbf{z}} = \boldsymbol{\Sigma}^{1/2} \mathbf{z}$ be the random vector coupled with $\tilde{\mathbf{x}}$ so that there is an event $\mathcal{E}$ with probability at least $1/2$ such that, conditioned on $\mathcal{E}$, we have $\tilde{\mathbf{z}} = \tilde{\mathbf{x}}$.

Next, in the joint probability space of $(\tilde{\mathbf{x}}, \tilde{\mathbf{z}})$, for each $j = 1, ..., d$, define an event $\mathcal{A}_j$, which holds when exactly one of the vectors $\mathbf{a}_{I_1}, ..., \mathbf{a}_{I_k}$ has a non-zero jth coordinate. Also, let $\mathcal{A} = \mathcal{A}_1 \vee ... \vee \mathcal{A}_d$. Let us start by showing the $\Pr\{\mathcal{A}\} \geq 2/3$. Each row of $\mathbf{A}$ has only one non-zero coordinate, and each coordinate is equally represented among the rows, which means all of the leverage scores of $\mathbf{A}$ are the same. So, since the probability distribution p is approximately uniform, i.e., $p_i \approx_{O(1)} 1/N$ for all i, after sampling $k - 1$ rows, we have at least $1 - O(k/d)$ chance that the last sample will produce a row $\mathbf{a}_{I_k}$ with a non-zero in a coordinate we have not seen before. Choosing constant C appriioriately, we can ensure that this probability is at least $2/3$. Thus, we showed that $\Pr\{\mathcal{A}\} \geq 2/3$.

Next, our goal is to show that one of the events $\mathcal{A}_j$ has a positive intersection with the event $\mathcal{E}$ (i.e., when $\tilde{\mathbf{x}}$ coincides with $\tilde{\mathbf{z}}$). We proceed as follows:

$$\sum_{j=1}^{d} \Pr\{\mathcal{A}_j \wedge \mathcal{E}\} \geq \Pr\{\mathcal{A} \wedge \mathcal{E}\} \geq \Pr\{\mathcal{A}\} - \Pr\{\neg\mathcal{E}\} \geq 2/3 - 1/2 \geq 1/6,$$

where we used the union bound. We conclude that there exists j such that $\Pr\{\mathcal{A}_j \wedge \mathcal{E}\} \geq 1/(6d)$. Note that as long as $\mathcal{A}_j$ holds, then $|\tilde{x}_j| \geq c\sqrt{d/k}$ (here, $\tilde{x}_j$ denotes the jth coordinate of $\tilde{\mathbf{x}}$), so:

$$\|\tilde{\mathbf{x}} \cdot \mathbf{1}_\mathcal{E}\|_{\psi_2} \geq \left\| |\tilde{x}_j| \cdot \mathbf{1}_\mathcal{E} \right\|_{\psi_2} \geq \left\| c\sqrt{d/k} \cdot \mathbf{1}_{\mathcal{A}_i \wedge \mathcal{E}} \right\|_{\psi_2} = \Omega\left(\sqrt{\frac{d/k}{\log d}}\right),$$

where we used a simple lower bound for the sub-gaussian norm of a Bernoulli random variable. This immediately implies the same bound for $\tilde{\mathbf{z}}$ since $\|\tilde{\mathbf{z}}\|_{\psi_2} \geq \|\tilde{\mathbf{x}} \cdot \mathbf{1}_\mathcal{E}\|_{\psi_2}$. We can finally return to $\mathbf{z} = \widetilde{\boldsymbol{\Sigma}}^{-1/2} \tilde{\mathbf{z}}$. Using the fact that $\widetilde{\boldsymbol{\Sigma}} \preceq O(1) \cdot \mathbf{I}$, we get:

$$\|\mathbf{z}\|_{\psi_2} = \sup_{\mathbf{v}:\|\mathbf{v}\|=1} \|\mathbf{v}^\top \widetilde{\boldsymbol{\Sigma}}^{-1/2} \tilde{\mathbf{z}}\|_{\psi_2} = \sup_{\mathbf{v}:\|\mathbf{v}\|=1} \|\mathbf{v}^\top \widetilde{\boldsymbol{\Sigma}}^{-1/2}\| \left\| \frac{\mathbf{v}^\top \widetilde{\boldsymbol{\Sigma}}^{-1/2}}{\|\mathbf{v}^\top \widetilde{\boldsymbol{\Sigma}}^{-1/2}\|} \tilde{\mathbf{z}} \right\|_{\psi_2}$$

$$= \Omega(1) \sup_{\mathbf{v}:\|\mathbf{v}\|=1} \left\| \frac{\mathbf{v}^\top \widetilde{\boldsymbol{\Sigma}}^{-1/2}}{\|\mathbf{v}^\top \widetilde{\boldsymbol{\Sigma}}^{-1/2}\|} \tilde{\mathbf{z}} \right\|_{\psi_2} = \Omega\big(\|\tilde{\mathbf{z}}\|_{\psi_2}\big) = \Omega\left(\sqrt{\frac{d/k}{\log d}}\right).$$

Appendix H. Proof of Lemma 20

Let $\mathbf{u} = \boldsymbol{\Sigma}^{-1/2} \mathbf{x}$. We will use the following simple decomposition:

$$(\mathbf{u}\mathbf{u}^\top - \mathbf{I})^p = (\|\mathbf{u}\|^2 - 1)^{p-1} \mathbf{u}\mathbf{u}^\top - (\mathbf{u}\mathbf{u}^\top - \mathbf{I})^{p-1}. \tag{7}$$

Using the fact that $\|\mathbf{u}\|$ is M-sub-gaussian, for any $t > 0$ we have that:

$$\Pr\{\|\mathbf{u}\|^2 \geq t\} \leq \exp(-ct/M^2),$$

for some absolute constant c. Define the following event: $\mathcal{E} = \left[\|\mathbf{u}\|^2 \leq L\right]$ where $L = CM^2(p + 3\log M)$. We conclude that $\Pr\{\neg\mathcal{E}\} \leq \exp(-c'(p + 2\log M)) \leq M^{-2}\exp(-c'(p + \log M))$. Next, we use this to bound the first term from (7):

$$
\begin{aligned}
\left\|\mathbb{E}\left[(\|\mathbf{u}\|^2 - 1)^{p-1}\mathbf{u}\mathbf{u}^\top\right]\right\| &\leq \left\|\mathbb{E}\left[\big|\|\mathbf{u}\|^2 - 1\big|^{p-1}\mathbf{u}\mathbf{u}^\top\mathbf{1}_{\mathcal{E}}\right]\right\| + \left\|\mathbb{E}\left[\big|\|\mathbf{u}\|^2 - 1\big|^{p-1}\mathbf{u}\mathbf{u}^\top\mathbf{1}_{\neg\mathcal{E}}\right]\right\| \\
&\leq L^{p-1}\|\mathbb{E}[\mathbf{u}\mathbf{u}^\top]\| + \mathbb{E}\left[\|\mathbf{u}\|^{2p}\mathbf{1}_{\neg\mathcal{E}}\right] \\
&\leq L^{p-1} + \int_0^\infty px^{p-1}\Pr\{\|\mathbf{u}\|^2\mathbf{1}_{\neg\mathcal{E}} > x\}dx \\
&\leq L^{p-1} + L^p \cdot M^{-2}\exp(-c(p + \log M)) + \int_L^\infty px^{p-1}\exp(-cx/M^2)\,dx \\
&\leq C'L^{p-1} + (C'M^2p)^p\exp(-c'L/M^2) \\
&\leq (C''L)^{p-1},
\end{aligned}
$$

where C, C' and C'' all denote absolute constants. Thus, we obtain that $\|\mathbb{E}[(\mathbf{u}\mathbf{u}^\top - \mathbf{I})^p]\| \leq (C''L)^{p-1} + \|\mathbb{E}[(\mathbf{u}\mathbf{u}^\top - \mathbf{I})^{p-1}]\| \leq (C'''L)^{p-1}$ (by expanding the recursion). This completes the proof of Lemma 20.

Here, we also establish a closely related result used in the proof of Theorem 11.

Lemma 28 *There is an absolute constant $C > 0$ such that any d-dimensional random vector $\mathbf{x}$ with covariance $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \Sigma$, where $\|\Sigma^{-1/2}\mathbf{x}\|$ is M-sub-gaussian, satisfies the following sub-exponential moment bound for any psd matrix $\mathbf{B}$:*

$$
\mathbb{E}\left[\big|\|\mathbf{B}^{1/2}\Sigma^{-1/2}\mathbf{x}\|^2 - \mathrm{tr}\mathbf{B}\big|^p\right] \leq (CM^2(p + \log M))^{p-1}\|\mathbf{B}\|^{p-2}\|\mathbf{B}\|_F^2.
$$

Proof We proceed similarly to the proof of Lemma 20, letting $L = CM^2(p + \log M)$ and $\mathcal{E} = \left[\|\Sigma^{-1/2}\mathbf{x}\|^2 \leq L\right]$:

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{B}^{1/2}\Sigma^{-1/2}\mathbf{x}\|^{2p}\right] &\leq \mathbb{E}\left[\|\Sigma^{-1/2}\mathbf{x}\|^{2(p-1)}\mathbf{x}^\top\Sigma^{-1/2}\mathbf{B}^p\Sigma^{-1/2}\mathbf{x}\right] \\
&\leq L^{p-1}\mathrm{tr}\mathbf{B}^p + \|\mathbf{B}\|^p \cdot \mathbb{E}\left[\mathbf{1}_{\neg\mathcal{E}}\|\Sigma^{-1/2}\mathbf{x}\|^{2p}\right] \\
&\leq (CL\|\mathbf{B}\|)^{p-2}L\|\mathbf{B}\|_F^2,
\end{aligned}
$$

where we used the integral bound from the proof of Lemma 20 and that $\mathrm{tr}\mathbf{B}^p \leq \|\mathbf{B}\|^{p-2}\|\mathbf{B}\|_F^2$. Finally, note that

$$
\begin{aligned}
\mathbb{E}\left[\big|\|\mathbf{B}^{1/2}\Sigma^{-1/2}\mathbf{x}\|^2 - \mathrm{tr}\mathbf{B}\big|^p\right] &\leq 2^p\left(\mathbb{E}\left[\big|\|\mathbf{B}^{1/2}\Sigma^{-1/2}\mathbf{x}\|^2\big|\right] + (\mathrm{tr}\mathbf{B})^p\right) \\
&\leq 2^p(L\|\mathbf{B}\|)^{p-2}L\|\mathbf{B}\|_F^2 + 2^p(d\|\mathbf{B}\|)^{p-2}\|\mathbf{B}\|_F^2 \\
&\leq (C'L\|\mathbf{B}\|)^{p-2}L\|\mathbf{B}\|_F^2,
\end{aligned}
$$

where we used the fact that $L \geq M^2 \geq d$. $\blacksquare$

Appendix I. Leave-one-out cross-validation formula: Proof of Lemma 21

First, note that, assuming $(\mathbf{A}^\top \mathbf{S}_{-i} \mathbf{S}_{-i} \mathbf{A})^{-1} \preceq O(1) \cdot (\mathbf{A}^\top \mathbf{A})^{-1}$, we have:

$$
\left\| \operatorname*{argmin}_{\mathbf{w}} \|\mathbf{S}_{-i}(\mathbf{A}\mathbf{w} - \mathbf{b})\| \right\| = \|(\mathbf{A}^\top \mathbf{S}_{-i}^\top \mathbf{S}_{-i} \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{S}_{-i}^\top \mathbf{S}_{-i} \mathbf{b}\|
$$

$$
\leq \|(\mathbf{A}^\top \mathbf{S}_{-i}^\top \mathbf{S}_{-i} \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{S}_{-i}^\top\| \cdot \|\mathbf{S}_{-i} \mathbf{b}\|
$$

$$
\leq O(1)\sqrt{\|(\mathbf{A}^\top \mathbf{A})^{-1}\|} \cdot O(\sqrt{Nk})\|\mathbf{b}\| \leq \operatorname{poly}(N, \kappa(\mathbf{A}), \|\mathbf{A}\|, \|\mathbf{b}\|),
$$

where we used that $\|\mathbf{S}_{-i}\|^2 = O(Nk)$ for LESS embeddings satisfying the assumptions from Theorem 7. The same bound holds for the unconstrained version of $\widehat{\mathbf{w}}$. Thus, it follows that, for sufficiently large D, each $\widehat{\mathbf{w}}^{-i}$, as well as $\widehat{\mathbf{w}}$, can be computed using the standard formulas for the unconstrained least squares, without restricting to $\mathcal{R}^d = [-D, D]^d$.

Proof of the shortcut formula. This is a standard derivation, which we include for completeness. As shorthands, we let $\mathbf{X} = \mathbf{S}\mathbf{A}$, $\mathbf{y} = \mathbf{S}\mathbf{b}$, and $\ell_i = \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i$. From Sherman-Morrison (Lemma 19), we have:

$$
(\mathbf{X}_{-i}^\top \mathbf{X}_{-i})^{-1} \mathbf{x}_i = \frac{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i}{1 - \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i} = \frac{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i}{1 - \ell_i},
$$

where $\mathbf{X}_{-i}$ is $\mathbf{X}$ without the ith row. Next, we use that to obtain:

$$
\widehat{\mathbf{w}}^{-i} = (\mathbf{X}_{-i}^\top \mathbf{X}_{-i})^{-1} \mathbf{X}^\top \mathbf{y} - (\mathbf{X}_{-i}^\top \mathbf{X}_{-i})^{-1} \mathbf{x}_i y_i
$$

$$
= \left((\mathbf{X}^\top \mathbf{X})^{-1} + \frac{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1}}{1 - \ell_i} \right) \mathbf{X}^\top \mathbf{y} - \frac{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i y_i}{1 - \ell_i}
$$

$$
= \widehat{\mathbf{w}} + \frac{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i}{1 - \ell_i} \left(\mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} - y_i \right) = \widehat{\mathbf{w}} + \frac{\mathbf{x}_i^\top \widehat{\mathbf{w}} - y_i}{1 - \ell_i} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i.
$$

Finally, we plug this into the ith component of the leave-one-out cross-validation estimate:

$$
\mathbf{x}_i^\top \widehat{\mathbf{w}}^{-i} - y_i = \mathbf{x}_i^\top \widehat{\mathbf{w}} + \frac{\mathbf{x}_i^\top \widehat{\mathbf{w}} - y_i}{1 - \ell_i} \cdot \ell_i - y_i = \frac{\mathbf{x}_i^\top \widehat{\mathbf{w}} - y_i}{1 - \ell_i}.
$$

Squaring and summing the components concludes the proof.