SIAM J. MATH. DATA SCI.
Vol. 6, No. 1, pp. 199–225

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Asymptotics of the Sketched Pseudoinverse*

Daniel LeJeune†, Pratik Patil‡, Hamid Javadi§, Richard G. Baraniuk¶, and Ryan J. Tibshirani‡

Abstract. We take a random matrix theory approach to random sketching and show an asymptotic first-order equivalence of the regularized sketched pseudoinverse of a positive semidefinite matrix to a certain evaluation of the resolvent of the same matrix. We focus on real-valued regularization and extend previous results on an asymptotic equivalence of random matrices to the real setting, providing a precise characterization of the equivalence even under negative regularization, including a precise characterization of the smallest nonzero eigenvalue of the sketched matrix, which may be of independent interest. We then further characterize the second-order equivalence of the sketched pseudoinverse. We also apply our results to the analysis of the sketch-and-project method and to sketched ridge regression. Last, we prove that these results generalize to asymptotically free sketching matrices, obtaining the resulting equivalence for orthogonal sketching matrices and comparing our results to several common sketches used in practice.

Key words. sketching, random projections, pseudoinverse, proportional asymptotics, random matrix theory

MSC codes. 15B52, 46L54, 62J07

DOI. 10.1137/22M1530264

1. Introduction.

In large-scale data processing systems, *sketching* or *random projections* play an essential role in making computation efficient and tractable. The basic idea is to replace high-dimensional data by relatively low-dimensional random linear projections of the data such that distances are preserved. It is well-known that sketching can significantly reduce the size of the data without harming statistical performance, while providing a dramatic computational advantage [1, 24, 30, 54]. For a summary of results on the applications of sketching in optimization and numerical linear algebra, we refer the reader to [38, 55].

In this work, we present a different kind of result than the usual sketching guarantee. Typically, sketching is guaranteed to preserve the output or statistical performance of computational methods with an error term that vanishes for sufficiently large sketch sizes [5, 6, 10, 27, 44, 56]. In contrast, we characterize the precise way in which the solution to a

*Received by the editors December 1, 2022; accepted for publication (in revised form) November 20, 2023; published electronically March 29, 2024. A preliminary version of this work was part of the author's Ph.D. thesis *Ridge Regularization by Randomization in Linear Ensembles*, prepared at Rice University, Houston, Texas, in 2022.

https://doi.org/10.1137/22M1530264

Funding: This work was sponsored by Office of Naval Research MURI grant N00014-20-1-2787. The first, third, and fourth authors were also supported by NSF grants CCF-1911094, IIS-1838177, and IIS-1730574; ONR grants N00014-18-12571 and N00014-20-1-2534; AFOSR grant FA9550-22-1-0060; and a Vannevar Bush Faculty Fellowship, ONR grant N00014-18-1-2047. The first author was partially supported by ARO grant 2003514594.

†Department of Statistics, Stanford University, Stanford, CA 94305 USA (daniel@dlej.net).

‡Department of Statistics, University of California, Berkeley, CA 94720 USA (pratikpatil@berkeley.edu, ryantibs@berkeley.edu).

§Google, Mountain View, CA 94043 USA (hamid71reza@gmail.com).

¶Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005 USA (richb@rice.edu).

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

computational problem changes when operating on a sketched version of data instead of the original data, showing that sketching induces a specific type of regularization.

Our primary contribution is a statement about the effect of sketching on the (regularized) pseudoinverse of a matrix. An informal statement of our result is as follows. Here the notation $\mathbf{A} \simeq \mathbf{B}$ for two matrices $\mathbf{A}$ and $\mathbf{B}$ indicates an asymptotic first-order equivalence, which we define in section 2, and $\lambda_{\min}^{+}(\mathbf{A})$ is the smallest nonzero eigenvalue of a matrix $\mathbf{A}$. We refer to $\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}}$ as the sketched (regularized) pseudoinverse of $\mathbf{A}$, because when $\mathbf{S}$ has orthonormal columns, the pseudoinverse of $\mathbf{S}\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S}\mathbf{S}^{\mathsf{H}}$ is equal to $\mathbf{S}(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S})^{-1}\mathbf{S}^{\mathsf{H}}$. This expression is also related to the Nyström approximation of $\mathbf{A}$.

Theorem 1.1 (Theorems 4.1 and 7.2, informal). *Given a positive semidefinite matrix $\mathbf{A} \in \mathbb{C}^{p\times p}$ and sketching matrix $\mathbf{S} \in \mathbb{C}^{p\times q}$, for any $\lambda > -\lambda_{\min}^{+}(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S})$, there exists $\mu \in \mathbb{R}$ such that*

$$\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}} \simeq (\mathbf{A} + \mu \mathbf{I}_p)^{-1}.$$

The general implication of this result is that when we do computation using the sketched version of a matrix, there is a sense in which it is as if we were using additional ridge regularization. More precisely, when we solve (regularized) linear systems on a sketched version of the data and apply this solution to the sketched data, it is equivalent in a first-order sense to solving a regularized linear system in the original space. To see this, consider, for example, a least squares problem $\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$. The first-order optimality condition is $\mathbf{X}^{\mathsf{H}}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^{\mathsf{H}}\mathbf{y}$, and if we replace $\mathbf{X}$ by a sketch $\mathbf{X}\mathbf{S}$, we have the solution in the sketched domain $\widehat{\boldsymbol{\beta}}_{\mathbf{S}} = (\mathbf{S}^{\mathsf{H}}\mathbf{X}^{\mathsf{H}}\mathbf{X}\mathbf{S})^{-1}\mathbf{S}^{\mathsf{H}}\mathbf{X}^{\mathsf{H}}\mathbf{y}$. If we then measure this solution in some sketched direction $\mathbf{S}^{\mathsf{H}}\mathbf{u}$ for some independent unit vector $\mathbf{u}$, we obtain $\hat{\beta}_{\mathbf{u}} = \mathbf{u}^{\mathsf{H}}\mathbf{S}\widehat{\boldsymbol{\beta}}_{\mathbf{S}} = \mathbf{u}^{\mathsf{H}}\mathbf{S}(\mathbf{S}^{\mathsf{H}}\mathbf{X}^{\mathsf{H}}\mathbf{X}\mathbf{S})^{-1}\mathbf{S}^{\mathsf{H}}\mathbf{X}^{\mathsf{H}}\mathbf{y}$. By our result, this is asymptotically equivalent to measuring $\hat{\beta}_{\mathbf{u}} \simeq \mathbf{u}^{\mathsf{H}}(\mathbf{X}^{\mathsf{H}}\mathbf{X} + \mu\mathbf{I})^{-1}\mathbf{X}^{\mathsf{H}}\mathbf{y}$—that is, as if we had solved the original least squares problem using some regularization μ.

Summary of contributions. Below we summarize the main contributions of the paper.

1. **Real-valued equivalence.** We extend previous results from random matrix theory [45] for independent and identically distributed (i.i.d.) random matrices to real-valued regularization, explicitly characterizing the behavior of the associated fixed-point equation extended from the complex half-plane to the reals, allowing for consideration of negative regularization. This result includes what is to the best of our knowledge the first characterization of the limiting smallest nonzero eigenvalue of arbitrary Wishart type sample covariance matrices, which may be of independent interest.
2. **First-order equivalence.** Applying the real-valued equivalence, we obtain a first-order equivalence for the ridge-regularized i.i.d. sketched pseudoinverse.
3. **Second-order equivalence.** Using the calculus of asymptotic equivalents, we also obtain a second-order equivalence for the ridge-regularized i.i.d. sketched pseudoinverse that captures a variance-like inflation due to the randomness of sketching.
4. **Equivalence properties.** We provide a thorough investigation of the theoretical properties of the equivalence relationship, such as how the induced regularization depends on the original applied regularization, sketch size, and matrix rank.
5. **Applications.** We demonstrate how to apply our results by performing novel analysis of sketch-and-project [24] and sketched ridge regression.

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

6. **Free sketching.** Finally, we extend the scope of our results for first-order equivalence of the sketched pseudoinverse beyond i.i.d. sketching to general asymptotically free sketching and specialize to orthogonal sketching matrices.

Related work. The existence of an implicit regularization effect of sketching or random projections has been known for some time [17, 31, 46, 50]. While prior works have demonstrated clear theoretical and empirical statistical advantages of sketching, our understanding of the precise nature of this implicit regularization has been largely limited to quantities such as error bounds. We provide, in contrast, a precise asymptotic characterization of the solution obtained by a sketching-based solver, not only enabling the understanding of the statistical performance of sketching-based methods, but also opening the door for exploiting the specific regularization induced by sketching in future algorithms.

Our results in this work provide a general extension of a few results appearing in recent works that have revealed explicit characterizations of the implicit regularization effects induced by random subsampling. To the best of our knowledge, the first such result was presented by [34], who showed that ensembles of (unregularized) ordinary least squares predictors on randomly subsampled observations and features converge in an ℓ_2 metric to an (optimal) ridge regression solution in the proportional asymptotics regime. This result was limited in several aspects: (a) it required a strong isotropic Gaussian data assumption; (b) it required the subsampled data to have more observations than features; (c) it considered only unregularized base learners in the ensemble; (d) it required an ensemble of infinite size to show the ridge regression equivalence; (e) it provided only a marginal guarantee of convergence over the data distribution rather than a single-instance convergence guarantee; and (f) it did not provide the relationship between the subsampling ratio and the amount of induced ridge regularization. In addition, the proof relied on rote computation of expectations of matrix quantities, providing limited insight into the underlying mathematical principles at work. The result we present in this work in Theorem 4.1 addresses all of these issues.

Around the same time, [42] showed the remarkably simple result that the expected value of the pseudoinverse of any positive definite matrix sampled by a determinantal point process is equal to a resolvent of the matrix. Similarly to the result by [34], this result demonstrated that when random subsampling is applied in techniques without any regularization, the resulting solution is as if a regularized technique was used on the original data. This result provided a simple form of the argument of the induced resolvent as a solution to a matrix trace equation, which is analogous to the results we present in this work for sketching. The same authors later empirically demonstrated that the same effects occur when using i.i.d. Gaussian and Rademacher sketches [12] and obtained a first-order equivalent for certain sub-Gaussian sketched projection operators [14] and first- and second-order moments for certain debiased sketches [13]. Our work generalizes these later developments and also differs from these works in that we provide a single-instance equivalent ridge regularization in the asymptotic regime, rather than an expectation over the random projections.

Our results also echo the finite-sample results of [16], who showed that the unregularized inverse of a particular sketched matrix form has a merely multiplicative bias for sketch size minimally larger than the rank of the original matrix. This is captured by Theorem 3.1 in our work when $z \to 0$, combined with Remark 5.7, in which we observe that there is asymptotically no spectral distortion in the range of the original matrix for sketches larger than the rank.

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Our work leverages techniques from random matrix theory [45], and the techniques employed bear some resemblance to other recent work in high dimensional statistical analysis [15, 22, 26]. In particular, we leverage the calculus of deterministic equivalences as presented by [21]. However, instead of characterizing only very specific quantities such as in-distribution generalization error, requiring tedious updates to the proof to adapt to other quantities of interest, we have isolated the expressions that will be needed to analyze *any* quadratic functional of the sketched pseudoinverse. In addition, instead of characterizing $\mathbf{A}^{1/2}\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S}+\lambda\mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}}\mathbf{A}^{1/2}$ (as considered, e.g., by [14] for $\lambda=0$), which is a simple reparameterization of $(\mathbf{A}^{1/2}\mathbf{S}^{\mathsf{H}}\mathbf{S}\mathbf{A}^{1/2}+\lambda\mathbf{I}_p)^{-1}$ and therefore straightforwardly understood through equivalences for sample covariance matrices [32, 45], we characterize the quantity $\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S}+\lambda\mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}}$, which is essential for asymmetric applications such as ridge regression without data assumptions (see the example in subsection 6.2).

Our application of our results to sketch-and-project [24] improves upon recent work by [14] in that we are also able to calculate asymptotic computational complexity as a function of sketch size thanks to the uniformity of convergence over bounded sketching ratios and the ability to consider sparse sketches that can be applied in $O(q^2)$ time (see Remark 4.5).

Other works have considered other types of sketches that do not have the same random matrix properties as the matrices we consider in our main results. In particular, fast sketching techniques such as CountSketch [9] and the subsampled randomized Hadamard transform (SRHT) [51] are among the most popular random projections in practice, since they can be applied in only $O(p\log p)$ time rather than $O(pq)$ or $O(q^2)$ for i.i.d. sketches. Very little is known about the properties of these sketches under proportional asymptotics; we know only of [29], who analyzed specific first and second moments in the isotropic case for the SRHT. Other prior work has shown universality of certain sketching inversion bias behavior under any rotationally invariant sketch [16]. We show that our results generalize to the broader class of "free" sketches in Theorem 7.2 using free probability [53, 40] and specialize to an exact formula for orthogonal sketching in Corollary 7.3. Then we empirically show that fast sketches commonly used in practice behave according to our generalization.

A few works have shown that under certain data geometry and noise, the optimal ridge regression parameter can be negative [28, 57]. For this reason, we take special care to determine the limit of allowable negative regularization in sketched settings. Then in a ridge regression example in subsection 6.2, we demonstrate how negative regularization can be optimal for standard noisy learning problems in undersampled distributed optimization settings.

Organization. The rest of the paper is structured as follows. In section 2, we start with some preliminaries on the language of asymptotic equivalence of random matrices that we will use to state our results. In section 3, we extend a previous result on asymptotic equivalence for a ridge regularized resolvent to include real-valued negative regularization and provide a precise limiting lower limit of the permitted negative regularization. In section 4, we provide our main results about the first- and second-order equivalence of the sketched pseudoinverse. Then, in section 5, we explore properties of the equivalence and present illustrative examples. In section 6, we perform novel analysis of two sketching-based optimization methods. Finally, in section 7, we conclude by giving various extensions and providing a generalization of the asymptotic behavior of sketched pseudoinverse for a broad family of sketching matrices using

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

the insights obtained from the proof of our main result and experimentally compare sketches commonly used in practice to our theory. Our code for generating all figures can be found at https://github.com/dlej/sketched-pseudoinverse.

Notation. We denote the real line by $\mathbb{R}$ and the complex plane by $\mathbb{C}$. For a complex number $z = x + iy$, $\mathrm{Re}(z)$ denotes its real part x, $\mathrm{Im}(z)$ denotes its imaginary part y, and $\overline{z} = x - iy$ denotes its conjugate. We use $\mathbb{R}_{\geq 0}$ and $\mathbb{R}_{>0}$ to denote the set of nonnegative and positive real numbers, respectively; similarly, $\mathbb{R}_{\leq 0}$ and $\mathbb{R}_{<0}$ respectively denote the set of nonpositive and negative real numbers. We use $\mathbb{C}^+ = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$ to denote the upper half of the complex plane and $\mathbb{C}^- = \{z \in \mathbb{C} : \mathrm{Im}(z) < 0\}$ to denote the lower half of the complex plane.

We denote vectors in lowercase bold letters (e.g., $\mathbf{y}$) and matrices in uppercase bold letters (e.g., $\mathbf{X}$). For a vector $\mathbf{y}$, $\|\mathbf{y}\|_2$ denotes its ℓ_2 norm. For a rectangular matrix $\mathbf{S} \in \mathbb{C}^{p \times q}$, $\mathbf{S}^{\mathsf{H}} \in \mathbb{C}^{q \times p}$ denotes its conjugate or Hermitian transpose (such that $[\mathbf{S}^{\mathsf{H}}]_{ij} = \overline{[\mathbf{S}]_{ji}}$), $\|\mathbf{S}\|_{\mathrm{tr}}$ denotes its trace norm (or nuclear norm), that is, $\|\mathbf{S}\|_{\mathrm{tr}} = \mathrm{tr}[(\mathbf{S}^{\mathsf{H}}\mathbf{S})^{1/2}]$, and $\|\mathbf{S}\|_{\mathrm{op}}$ denotes the operator norm with respect to the ℓ_2 vector norm (which is also its spectral norm). For a square matrix $\mathbf{A} \in \mathbb{C}^{p \times p}$, $\mathrm{tr}[\mathbf{A}]$ denotes its trace, $\mathrm{rank}(\mathbf{A})$ denotes its rank, $r(\mathbf{A}) = \frac{1}{p}\mathrm{rank}(\mathbf{A})$ denotes its relative rank, and $\mathbf{A}^{-1} \in \mathbb{C}^{p \times p}$ denotes its inverse, if it is invertible. For any matrix $\mathbf{A} \in \mathbb{C}^{p \times q}$, $\mathbf{A}^{\dagger}$ denotes the Moore–Penrose inverse. For a positive semidefinite matrix $\mathbf{A} \in \mathbb{C}^{p \times p}$, $\mathbf{A}^{1/2} \in \mathbb{C}^{p \times p}$ denotes its positive semidefinite principal square root, $\lambda_{\min}(\mathbf{A})$ its smallest eigenvalue, and $\lambda^{+}_{\min}(\mathbf{A})$ its smallest positive eigenvalue.

A sequence x_n converging to x_∞ from the left or right is denoted by $x \nearrow x_\infty$ or $x \searrow x_\infty$, respectively. We denote almost sure convergence by $\xrightarrow{\text{a.s.}}$.

2. Preliminaries.

We will use the language of asymptotic equivalence of sequences of random matrices to state our main results. In this section, we define the notion of asymptotic equivalence, review some of the basic properties that such equivalence satisfies, and present an asymptotic equivalence for the ridge resolvent. We then extend that result to handle real-valued resolvents, which will form the building block for our subsequent results.

To begin, consider two sequences $\mathbf{A}_n$ and $\mathbf{B}_n$ of $p(n) \times q(n)$ matrices, where p and q are increasing in n. We will say that $\mathbf{A}_n$ and $\mathbf{B}_n$ are asymptotically equivalent if for any sequence of deterministic matrices $\mathbf{\Theta}_n$ with trace norm uniformly bounded in n, we have $\mathrm{tr}[\mathbf{\Theta}_n(\mathbf{A}_n - \mathbf{B}_n)] \xrightarrow{\text{a.s.}} 0$ as $n \nearrow \infty$. We write $\mathbf{A}_n \simeq \mathbf{B}_n$ to denote this asymptotic equivalence.[1] The notion of *deterministic* equivalence, where the right-hand sequence is a sequence of deterministic matrices, has been typically used in random matrix theory to obtain limiting behavior of functionals of random matrices; for example, see [11, 25, 47], among others. More recently, the notion of deterministic equivalence has been popularized and developed further in [20, 21].[2] We will use a slightly more general notion of asymptotic equivalence in this paper, where both sequences of matrices may be random.

The notion of asymptotic equivalence enjoys some properties that we list next. The majority of these are stated in the context of deterministic equivalence in [20, 21], but they

[1] When we use the same notation for a vector or scalar equivalence, it can be understood as applying this definition to a $p(n) \times 1$ or 1×1 matrix, respectively.

[2] Note that [20, 21] use the notation $\mathbf{A}_n \asymp \mathbf{B}_n$ to denote deterministic equivalence of sequence $\mathbf{A}_n$ to $\mathbf{B}_n$. We instead use the notation $\mathbf{A}_n \simeq \mathbf{B}_n$ to emphasize that this equivalence is asymptotically exact, rather than up to constants.

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

also hold more generally for asymptotic equivalence. For the statements to follow, let $\mathbf{A}_n$, $\mathbf{B}_n$, $\mathbf{C}_n$, and $\mathbf{D}_n$ be sequences of random or deterministic matrices (of appropriate dimensions). Then the following properties hold:

1. **Equivalence.** The relation $\simeq$ is an equivalence relation.
2. **Sum.** If $\mathbf{A}_n \simeq \mathbf{B}_n$ and $\mathbf{C}_n \simeq \mathbf{D}_n$, then $\mathbf{A}_n + \mathbf{C}_n \simeq \mathbf{B}_n + \mathbf{D}_n$.
3. **Product.** If $\mathbf{A}_n \simeq \mathbf{B}_n$, and $\mathbf{C}_n$ is independent of $\mathbf{A}_n$ and $\mathbf{B}_n$ with operator norm bounded in n almost surely, then $\mathbf{A}_n\mathbf{C}_n \simeq \mathbf{B}_n\mathbf{C}_n$.
4. **Trace.** If $\mathbf{A}_n \simeq \mathbf{B}_n$ for square matrices $\mathbf{A}_n$ and $\mathbf{B}_n$ of dimension $p(n) \times p(n)$, then $\frac{1}{p(n)}\mathrm{tr}[\mathbf{A}_n] \simeq \frac{1}{p(n)}\mathrm{tr}[\mathbf{B}_n]$.
5. **Elements.** If $\mathbf{A}_n \simeq \mathbf{B}_n$ for $\mathbf{A}_n, \mathbf{B}_n$ of dimension $p(n) \times q(n)$ and $i(n) \in \{1, \ldots, p(n)\}$ and $j(n) \in \{1, \ldots, q(n)\}$, then $[\mathbf{A}_n]_{i(n),j(n)} \simeq [\mathbf{B}_n]_{i(n),j(n)}$.
6. **Differentiation.** Suppose $f(z, \mathbf{A}_n) \simeq g(z, \mathbf{B}_n)$ where the entries of f and g are analytic functions in $z \in D$ and D is an open connected subset of $\mathbb{C}$. Furthermore, suppose for any sequence $\mathbf{\Theta}_n$ of deterministic matrices with trace norm uniformly bounded in n, we have that $|\mathrm{tr}[\mathbf{\Theta}_n(f(z, \mathbf{A}_n) - g(z, \mathbf{B}_n))]| \leq M$ for every n and $z \in D$ for some constant $M < \infty$. Then we have that $f'(z, \mathbf{A}_n) \simeq g'(z, \mathbf{B}_n)$ for every $z \in D$, where the derivatives are taken entrywise with respect to z.

The almost sure convergence in the statements above is with respect to the entire randomness in the random variables involved. One can also consider the notion of conditional asymptotic equivalence wherein we condition on a sequence of random matrices. More precisely, suppose $\mathbf{A}_n$, $\mathbf{B}_n$ are a sequence of random matrices that may depend of another sequence of random matrices $\mathbf{Z}_n$. We call $\mathbf{A}_n$ and $\mathbf{B}_n$ to be asymptotically equivalent conditioned on $\mathbf{Z}_n$ if for any sequence of deterministic matrices $\mathbf{\Theta}_n$ with trace norm uniformly bounded in n, we have $\lim_{n \nearrow \infty} \mathrm{tr}[\mathbf{\Theta}_n(\mathbf{A}_n - \mathbf{B}_n)] = 0$ almost surely conditioned on $\mathbf{Z}_n$. Properties similar to those listed above for unconditional asymptotic equivalence also hold for conditional equivalence by considering all the statements conditioned on the sequence $\mathbf{Z}_n$. In particular, for the product rule, we require that the sequence $\mathbf{C}_n$ be *conditionally* independent of $\mathbf{A}_n$ and $\mathbf{B}_n$ given $\mathbf{Z}_n$. Finally, for our asymptotic statements, we will work with sequences of matrices, indexed by either n or p. However, for notational brevity, we will drop the index from now on whenever it is clear from the context.

Equipped with the notion of asymptotic equivalence, below we state a result on the asymptotic deterministic equivalence for ridge resolvents of Wishart type matrices, adapted from Theorem 1 of [45] and Theorem 3.1 of [21], that will form a base for our results.

Lemma 2.1 (basic asymptotic equivalent for ridge resolvents, complex-valued regularization). *Let $\mathbf{Z} \in \mathbb{C}^{n \times p}$ be a random matrix consisting of i.i.d. random variables that have mean 0, variance 1, and finite absolute moment of order $8 + \delta$ for some $\delta > 0$. Let $\mathbf{\Sigma} \in \mathbb{C}^{p \times p}$ be a positive semidefinite matrix with operator norm uniformly bounded in p, and let $\mathbf{X} = \mathbf{Z}\mathbf{\Sigma}^{1/2}$. Then, for $z \in \mathbb{C}^+$, as $n, p \nearrow \infty$ such that $0 < \liminf \frac{p}{n} \leq \limsup \frac{p}{n} < \infty$, we have*

$$(\tfrac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} - z\mathbf{I}_p)^{-1} \simeq (c(z)\mathbf{\Sigma} - z\mathbf{I}_p)^{-1}, \tag{2.1}$$

where $c(z)$ is the unique solution in $\mathbb{C}^-$ to the fixed-point equation

$$\frac{1}{c(z)} - 1 = \tfrac{1}{n}\mathrm{tr}\left[\mathbf{\Sigma}(c(z)\mathbf{\Sigma} - z\mathbf{I}_p)^{-1}\right]. \tag{2.2}$$

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Furthermore, $\frac{1}{p}\mathrm{tr}\left[\boldsymbol{\Sigma}(c(z)\boldsymbol{\Sigma} - z\mathbf{I}_p)^{-1}\right]$ *is a Stieltjes transform of a certain positive measure on* $\mathbb{R}_{\geq 0}$ *with total mass* $\frac{1}{p}\mathrm{tr}[\boldsymbol{\Sigma}]$.

Strictly speaking, the results in [45] and [21] require that the sequence $\boldsymbol{\Sigma}$ be deterministic. However, one can take $\boldsymbol{\Sigma}$ to be a random sequence of matrices that are independent of $\mathbf{Z}$; see, for example, [32]. In this case, the asymptotic equivalence is treated conditionally on $\boldsymbol{\Sigma}$.

3. Real-valued equivalence. For real-valued negative z, corresponding to positive ridge regularization, we remark that one can use Lemma 2.1 to derive limits of linear and certain nonlinear functionals (through the calculus rules of asymptotic equivalence) of the ridge resolvent $(\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} - z\mathbf{I}_p)^{-1}$ by considering $z \in \mathbb{C}^+$ with $\mathrm{Re}(z) < 0$ and letting $\mathrm{Im}(z) \searrow 0$. This follows because a short calculation (see the proof of Theorem 3.1) shows that $\mathrm{Im}(c(z)) \nearrow 0$ as $\mathrm{Im}(z) \searrow 0$ for $z \in \mathbb{C}^+$ with $\mathrm{Re}(z) < 0$. Thus one can recover a real limit from the right-hand side of (2.1) through a limiting argument. Moreover, it is easy to see that the fixed-point equation (2.2) has a unique (real) solution $c(z) > 0$ for $z \in \mathbb{R}_{<0}$.

However, it has recently been pointed out that under certain special data geometry, negative regularization is often beneficial, in real data experiments [28] as well as in theoretical formulations where it can achieve optimal squared prediction risk [57]. One can still recover such a case by considering $z \in \mathbb{C}^+$ with $\mathrm{Re}(z) > 0$ over a valid range, and taking the limit as $\mathrm{Im}(z) \searrow 0$. However, solving the fixed-point equation (2.2) over reals directly in this case, which is the most efficient way to compute the solution numerically, poses certain subtleties as we no longer can guarantee a unique real solution for $c(z)$.

Our next theorem shows how to handle this case. We will make use of this for our results on sketching in section 4, but we believe the result to be of independent interest and worth stating on its own. In addition to enabling the computation of the asymptotic equivalence for nonnegative real-valued z, it also provides the asymptotic value of $\lambda_{\min}^{+}(\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X})$ (given by z_0 in the theorem statement) for arbitrary $\boldsymbol{\Sigma}$, which to our knowledge is the first explicit general characterization of the smallest nonzero eigenvalue of Wishart-type matrices, although the underlying principles are known in random matrix theory [49] and have been applied algorithmically [19]. We note that our characterization enables an extremely efficient and simple approach for computing z_0 via direct root finding in ζ_0. Furthermore, z_0 improves significantly on the naïve lower bounds commonly used in theoretical works [43, 57], as seen in Figure 1.

Theorem 3.1 (basic asymptotic equivalent for ridge resolvents, real-valued regularization). *Assume the setting of Lemma 2.1. Let* $\zeta_0, z_0 \in \mathbb{R}$ *be the unique solutions, satisfying* $\zeta_0 < \lambda_{\min}^{+}(\boldsymbol{\Sigma})$*, to the system of equations*

$$1 = \tfrac{1}{n}\mathrm{tr}\left[\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} - \zeta_0\mathbf{I}_p)^{-2}\right], \quad z_0 = \zeta_0\left(1 - \tfrac{1}{n}\mathrm{tr}\left[\boldsymbol{\Sigma}(\boldsymbol{\Sigma} - \zeta_0\mathbf{I}_p)^{-1}\right]\right). \tag{3.1}$$

Then, for each $z \in \mathbb{R}$ *satisfying* $z < \liminf z_0$*, as* $n, p \nearrow \infty$ *such that* $0 < \liminf \frac{p}{n} \leq \limsup \frac{p}{n} < \infty$*, we have*

$$z(\tfrac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} - z\mathbf{I}_p)^{-1} \simeq \zeta(\boldsymbol{\Sigma} - \zeta\mathbf{I}_p)^{-1}, \tag{3.2}$$

where $\zeta \in \mathbb{R}$ *is the unique solution in* $(-\infty, \zeta_0)$ *to the fixed-point equation*

$$z = \zeta\left(1 - \tfrac{1}{n}\mathrm{tr}\left[\boldsymbol{\Sigma}(\boldsymbol{\Sigma} - \zeta\mathbf{I}_p)^{-1}\right]\right). \tag{3.3}$$

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

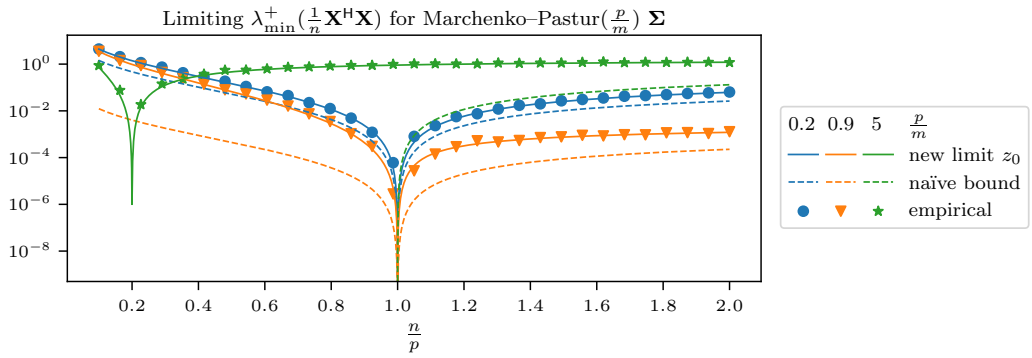

Figure 1. *Plots showing how z_0 (solid) from (3.1) matches the empirical minimum nonzero eigenvalue (markers) of $\frac{1}{n}\mathbf{X}^{\top}\mathbf{X}$ when $\mathbf{\Sigma} = \frac{1}{m}\mathbf{Y}^{\top}\mathbf{Y}$ for $\mathbf{Y} \in \mathbb{R}^{m\times p}$ with i.i.d. $\mathcal{N}(0,1)$ elements, such that the limiting spectrum of $\mathbf{\Sigma}$ follows the* Marchenko–Pastur($\frac{p}{m}$) *distribution for $\frac{p}{m} \in \{0.2, 0.9, 5\}$. In contrast, the commonly used naïve bound (dashed) $\liminf \lambda_{\min}^{+}(\frac{1}{n}\mathbf{X}^{\top}\mathbf{X}) \geq (1-\sqrt{\frac{p}{m}})^2(1-\sqrt{\frac{p}{n}})^2\mathbb{1}\{p < \max\{m,n\}\}$, obtained by multiplying the minimum nonzero eigenvalues of $\frac{1}{n}\mathbf{Z}^{\top}\mathbf{Z}$ and $\mathbf{\Sigma}$ when at most one of them is singular, is quite loose outside of the $m \gg p$ and $n \gg p$ cases and fails to capture the correct behavior at all when both are singular ($p > \max\{m,n\}$). Empirical values are computed for $p = 500$ for a single trial.*

Furthermore, as $n, p \nearrow \infty$, $\zeta \simeq -\frac{1}{v(z)}$, where $v(z)$ is the companion Stieltjes transform of the spectrum of $\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X}$ given by

$$v(z) = \tfrac{1}{n}\mathrm{tr}\Big[(\tfrac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{H}} - z\mathbf{I}_n)^{-1}\Big],$$

and $z_0 \simeq \lambda_{\min}^{+}(\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X})$.

Proof sketch. To prove this corollary, we define $\zeta \triangleq \frac{z}{c(z)}$ to obtain (3.2) from (2.1) for $z \in \mathbb{C}^{+}$ and also observe that $-\frac{1}{\zeta}$ is the limiting companion Stieltjes transform $v(z)$ of $\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{H}}$ at z. This implies that $\zeta \in \mathbb{C}^{+}$ and that the mapping $z \mapsto \zeta$ is a holomorphic function on its domain, which includes all real $z < \liminf \lambda_{\min}^{+}(\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{H}})$. We then identify the analytic continuation of the mapping $z \mapsto \zeta$ to the reals, which consists of careful bookkeeping to determine z_0, the least positive value of z for which ζ does not exist, which must be asymptotically equal to $\lambda_{\min}^{+}(\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{H}})$. The proof details can be found in section SM2 of the supplementary material. ∎

Remark 3.2 (the case of $z = 0$). The form of the equivalence (2.1) is slightly different as compared with (3.2) in that the resolvent $(\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} - z\mathbf{I}_p)^{-1}$ has a normalizing multiplier of z in the latter case. This enables continuity of the left-hand side at $z = 0$, in contrast to specializing the equivalence (2.1) to real z, where both the left- and right-hand sides may diverge as $z \nearrow 0$.

Our main result in the next section for sketching follows directly from this theorem and shares a very similar form. For this reason, we defer discussion about the interpretation of the solutions to the above equations for our reformulation under the sketching setting; however, analogous interpretations will apply to the above theorem.

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

4. Main results. One way to think about Theorem 3.1 is that the data matrix $\mathbf{X} = \mathbf{Z}\mathbf{\Sigma}^{1/2}$ is a sketched version of the (square root) covariance matrix $\mathbf{\Sigma}^{1/2}$, where $\mathbf{Z}$ acts as a sketching matrix. The sketching is done by "nature" in the form of the n observations, rather than by the statistician, but is otherwise mathematically identical to sketching. Using this insight, along with the Woodbury identity, we can adapt the random matrix resolvent equivalence in Theorem 3.1 to a sketched (regularized) pseudoinverse equivalence. To emphasize the shift in perspective, we denote the dimensionality of the sketched data as q (replacing n), replace $\mathbf{\Sigma}$ with $\mathbf{A}$, and absorb the normalization by $\frac{1}{q}$ (replacing $\frac{1}{n}$) into the sketching matrix $\mathbf{S}$ (replacing $\mathbf{Z}$), so that the sketching transformation is norm-preserving (see Remark 4.2 for more details).

4.1. First-order equivalence. Our first result provides a first-order equivalence for the sketched regularized pseudoinverse. By first-order equivalence, we refer to equivalence for matrices that involve the *first* power of the ridge resolvent. We also present a second-order equivalence for matrices that involve the *second* power of the ridge resolvent in subsection 4.2.

In preparation for the statements to follow, recall that $r(\mathbf{A}) = \frac{1}{p}\sum_{i=1}^{p} \mathbb{1}\{\lambda_i(\mathbf{A}) > 0\}$, or in other words, the normalized number of nonzero eigenvalues of $\mathbf{A}$. Note that $0 \le r(\mathbf{A}) \le 1$.

Theorem 4.1 (isotropic sketching equivalence). *Let $\mathbf{A} \in \mathbb{C}^{p\times p}$ be a positive semidefinite matrix such that $\|\mathbf{A}\|_{\mathrm{op}}$ is uniformly bounded in p and $\liminf \lambda_{\min}^{+}(\mathbf{A}) > 0$. Let $\sqrt{q}\mathbf{S} \in \mathbb{C}^{p\times q}$ be a random matrix consisting of i.i.d. random variables that have mean 0, variance 1, and finite $8+\delta$ moment for some $\delta > 0$. Let $\lambda_0, \mu_0 \in \mathbb{R}$ be the unique solutions, satisfying $\mu_0 > -\lambda_{\min}^{+}(\mathbf{A})$, to the system of equations*

$$1 = \tfrac{1}{q}\mathrm{tr}\left[\mathbf{A}^2(\mathbf{A}+\mu_0\mathbf{I}_p)^{-2}\right], \quad \lambda_0 = \mu_0\left(1 - \tfrac{1}{q}\mathrm{tr}\left[\mathbf{A}(\mathbf{A}+\mu_0\mathbf{I}_p)^{-1}\right]\right). \tag{4.1}$$

Then, as $q, p \nearrow \infty$ such that $0 < \liminf \frac{q}{p} \le \limsup \frac{q}{p} < \infty$, the following asymptotic equivalences hold:

(i) *for any $\lambda > \limsup \lambda_0$, we have*

$$\mathbf{A}^{1/2}\mathbf{S}(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda\mathbf{I}_q)^{-1}\mathbf{S}^{\mathsf{H}} \simeq \mathbf{A}^{1/2}(\mathbf{A}+\mu\mathbf{I}_p)^{-1}; \tag{4.2}$$

(ii) *if furthermore either $\lambda \neq 0$ or $\limsup \frac{q}{p} < \liminf r(\mathbf{A})$, we have*

$$\mathbf{S}(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda\mathbf{I}_q)^{-1}\mathbf{S}^{\mathsf{H}} \simeq (\mathbf{A}+\mu\mathbf{I}_p)^{-1}, \tag{4.3}$$

where μ is the unique solution in (μ_0, ∞) to the fixed-point equation

$$\lambda = \mu\left(1 - \tfrac{1}{q}\mathrm{tr}\left[\mathbf{A}(\mathbf{A}+\mu\mathbf{I}_p)^{-1}\right]\right). \tag{4.4}$$

Furthermore, as $p, q \to \infty$, $\mu \simeq \frac{1}{\widetilde{v}(\lambda)}$, where

$$\widetilde{v}(\lambda) = \tfrac{1}{q}\mathrm{tr}\left[(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda\mathbf{I}_q)^{-1}\right],$$

and $\lambda_0 \simeq -\lambda_{\min}^{+}(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S})$.

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

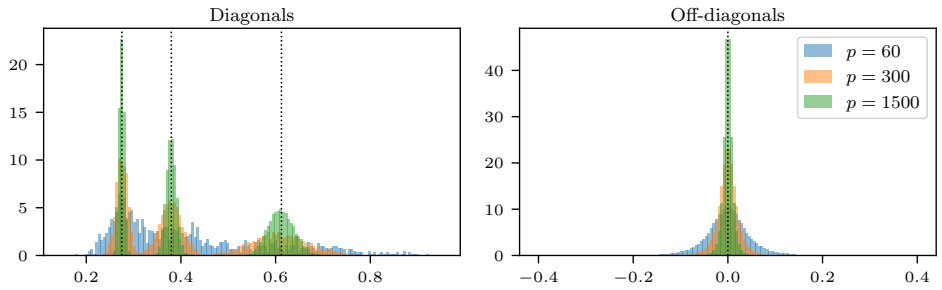

Figure 2. *Empirical density histograms over* 20 *trials demonstrating the concentration of the elements of* $\mathbf{S}(\mathbf{S}^\top\mathbf{A}\mathbf{S}+\lambda\mathbf{I})^{-1}\mathbf{S}^\top$ *for real Gaussian* $\mathbf{S}$ *and diagonal* $\mathbf{A}$ *taking values* $\{0,1,2\}$ *with equal frequency along the diagonal. We choose* $\lambda=1$ *and* $q=\lfloor\alpha p\rfloor$ *for* $\alpha=0.8$ *over* $p\in\{60,300,1500\}$. *As expected by Theorem* 4.1, *the individual elements of the sketched pseudoinverse converge to those of* $(\mathbf{A}+\mu\mathbf{I})^{-1}$, *where for this problem* $\mu\approx 1.63$. *Therefore, the diagonals concentrate with equal mass around* $\{1/(a+\mu):a\in\{0,1,2\}\}$ *(black, dotted), and the off-diagonals concentrate around* 0.

Proof sketch. We begin by considering the case that $\mathbf{A}$ satisfies $\limsup\|\mathbf{A}^{-1}\|_{\mathrm{op}}<\infty$. Then we can rewrite the left-hand side of (4.2) or (4.3) such that we can apply Theorem 3.1 with $\mathbf{X}=\sqrt{q}\mathbf{S}^{\mathsf{H}}\mathbf{A}^{1/2}$, $\lambda=-z$, and $\mu=-\zeta$. For any $\lambda>-\liminf z_0$,

$$\begin{aligned}\mathbf{A}^{1/2}\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S}+\lambda\mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}}\mathbf{A}^{1/2}&=\mathbf{A}^{1/2}\mathbf{S}\mathbf{S}^{\mathsf{H}}\mathbf{A}^{1/2}\left(\mathbf{A}^{1/2}\mathbf{S}\mathbf{S}^{\mathsf{H}}\mathbf{A}^{1/2}+\lambda\mathbf{I}_p\right)^{-1}\\&=\mathbf{I}_p-\lambda\left(\mathbf{A}^{1/2}\mathbf{S}\mathbf{S}^{\mathsf{H}}\mathbf{A}^{1/2}+\lambda\mathbf{I}_p\right)^{-1}\\&\simeq\mathbf{I}_p-\mu\left(\mathbf{A}+\mu\mathbf{I}_p\right)^{-1}\\&=\mathbf{A}^{1/2}\left(\mathbf{A}+\mu\mathbf{I}_p\right)^{-1}\mathbf{A}^{1/2}.\end{aligned}$$

We can then multiply on the right, or both left and right, by $\mathbf{A}^{-1/2}$ to obtain the results in (4.2) and (4.3), respectively, by the product rule of asymptotic equivalences. If $\mathbf{A}$ does not have a norm-bounded inverse, we can apply the above result for $\mathbf{A}_\delta\triangleq\mathbf{A}+\delta\mathbf{I}_p$ for $\delta>0$ and make a uniform convergence argument for interchanging limits of p and δ to prove the equivalence in (4.3). We then multiply by $\mathbf{A}^{1/2}$ and make another uniform convergence argument to extend this equivalence to the case $\lambda=0$ to obtain the equivalence in (4.2). The details can be found in section SM3 of the supplementary material. ∎

In words, the sketched pseudoinverse of $\mathbf{A}$ with regularization λ is asymptotically equivalent to the regularized inverse of $\mathbf{A}$ with regularization μ, and the relationship between λ and μ asymptotically depends only on $\mathbf{A}$, p, and q. As mentioned in section 2, this implies, for example, that the elements of the sketched pseudoinverse converge to the elements of the ridge-regularized inverse. We illustrate this in Figure 2, where for a diagonal $\mathbf{A}$, the off-diagonals of the sketched pseudoinverse quickly converge to zero as p increases, while the diagonals converge to the diagonals of the regularized inverse of $\mathbf{A}$.

Below we provide several remarks on the assumptions and implications of Theorem 4.1. It will be useful to interpret the equations in terms of the sketching aspect ratio $\alpha\triangleq\frac{q}{p}$.

Remark 4.2 (normalization choice for the sketching matrix). We remark that the normalization factor $\sqrt{q}$ in $\sqrt{q}\mathbf{S}$ of the sketching matrix is such that the norm of the rows of $\mathbf{S}$ is 1

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

in expectation. This is done so that $\mathbb{E}[\|\mathbf{S}^{\mathsf{H}}\mathbf{x}\|_2^2] = \|\mathbf{x}\|_2^2$ as $\mathbb{E}[\mathbf{S}\mathbf{S}^{\mathsf{H}}] = \mathbf{I}_p$. One can alternately consider sketching matrices with normalization $\sqrt{p}\mathbf{S}$ such that the columns have norm 1 in expectation. It is easy to write an equivalent version of Theorem 4.1 with such a normalization. We choose to focus on the former scaling because it is more common in practice.

Remark 4.3 (on assumptions). The assumptions imposed in Theorem 4.1 are quite mild. In particular, the sequences of matrices $\mathbf{A}$ being sketched can be random, so long as they are independent of $\mathbf{S}$. Furthermore, the spectrum of the sequences of matrices $\mathbf{A}$ need not converge to a fixed spectrum. The aspect ratio α of the sketching matrices $\mathbf{S}$ also need not converge to a fixed number. The reason this is possible is because we are not expressing the sketched resolvent in terms of the limiting spectrum of $\mathbf{S}$ and $\mathbf{A}$, but rather relating it through $\mathbf{A}$ and a parameter μ that depends on α and $\mathbf{A}$ (and the original regularization level λ), which allows us to keep our assumptions weak.

Remark 4.4 (rotationally invariant unregularized sketching). When $\lambda = 0$, the first-order equivalence in fact holds for any sketching matrix $\mathbf{S}$ that is rotationally invariant on the left and is not limited to i.i.d. sketching matrices. That is, if we look at the singular value decomposition of $\mathbf{S} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathsf{H}}$, the left singular vectors $\mathbf{U}$ are drawn from the Haar distribution over matrices with orthonormal columns. For $q \le \operatorname{rank}(\mathbf{A})$, $\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S}\right)^{-1}\mathbf{S}^{\mathsf{H}} = \mathbf{U}\left(\mathbf{U}^{\mathsf{H}}\mathbf{A}\mathbf{U}\right)^{-1}\mathbf{U}^{\mathsf{H}}$, and so the sketched pseudoinverse does not depend on the spectrum of $\mathbf{S}\mathbf{S}^{\mathsf{H}}$ at all and we can without any loss of generality apply Theorem 4.1. Given the universality of this result, it is no surprise that essentially all prior results for unregularized random projections [14, 34, 42] agree even for sketches of varying spectra or determinantal point processes. However, this universality does not extend to $\lambda \neq 0$ or to higher order equivalences; see Theorem 7.2.

Remark 4.5 (proportionally sparse sketching). Although i.i.d. sketching is commonly referred to as "dense sketching," Theorem 4.1 easily accommodates relatively sparse sketches that are faster to apply. We can draw $[\mathbf{S}]_{ij}$ from a distribution taking value 0 with probability $1 - \frac{q}{p}$ and still satisfy the bounded $8+\delta$ moment condition, leading to an $\mathbf{S}$ with $O(q^2)$ nonzero elements with high probability. This means that a vector multiply $\mathbf{S}^{\mathsf{H}}\mathbf{u}$ has cost $O(q^2)$ rather than $O(pq)$, which can be sufficient in many cases to make the cost of sketching negligible (see an example in subsection 6.1). This approach is essentially identical to the LESS-uniform embedding proposed by [13] as a special case, although LESS-uniform sketches can be "truly sparse" (less than $O(q^2)$) with additional incoherence assumptions on $\mathbf{A}$. It is worth recalling that since the ratio $\frac{q}{p}$ is bounded, strictly speaking all of these costs are $O(p^2)$; however, the relative advantages are often still computationally meaningful (see Figure 6). Faster $O(p \log p)$ sketches are not covered by this theorem, but we expect most such sketches to be covered by our extension in Theorem 7.2.

Remark 4.6 (the case of $\lambda = 0$). While the form in (4.3) is the most general, it does not hold for $\lambda = 0$ if the sketch size is larger than the rank of $\mathbf{A}$, since the inverse is unbounded. However, in machine learning settings such as ridge(less) regression, we only need to evaluate the regularized pseudoinverse $\mathbf{S}(\mathbf{S}^{\mathsf{H}}\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X}\mathbf{S} + \lambda\mathbf{I}_p)^{-1}\mathbf{S}^{\mathsf{H}}\frac{1}{\sqrt{n}}\mathbf{X}$. Thus, we can apply the form in (4.2) with $\mathbf{A}^{1/2} = (\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X})^{1/2}$, which is sufficient for any downstream analysis.

Remark 4.7 (alternate form of equivalence representation). Expressed in terms of $\widetilde{v}(\lambda)$, the equivalence (4.3) becomes

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

$$\mathbf{S}(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S}+\lambda\mathbf{I}_q)^{-1}\mathbf{S}^{\mathsf{H}} \simeq \widetilde{v}(\lambda)(\widetilde{v}(\lambda)\mathbf{A}+\mathbf{I}_p)^{-1},$$

and the fixed-point equation (4.4) becomes

$$\lambda = \frac{1}{\widetilde{v}(\lambda)} - \tfrac{1}{q}\mathrm{tr}\left[\mathbf{A}(\widetilde{v}(\lambda)\mathbf{A}+\mathbf{I}_p)^{-1}\right].$$

4.2. Second-order equivalence. Although the equivalence in Theorem 4.1 holds for first-order trace functionals, this equivalence does not hold for higher order functionals. To intuitively understand why, it is helpful to reason about the asymptotic equivalence similarly to an equivalence of expectation in classical random variables. That is, we may have two random variables X, Y with $\mathbb{E}[X] = \mathbb{E}[Y]$, but this does not allow us to make any conclusions about the relationship between $\mathbb{E}\left[X^k\right]$ and $\mathbb{E}\left[Y^k\right]$ for $k > 1$. In the same way, our first-order asymptotic equivalence does not directly tell us higher order equivalences.

Fortunately, however, because of the resolvent structure of the regularized pseudoinverse, we can cleverly apply the derivative rule of the calculus of asymptotic equivalences to obtain a second-order equivalence from the first- order equivalence. Such a derivative trick has been employed in several prior works [22, 26, 23, 32, 37] for computing some specific second-order functionals, but we extend to generic second-order functionals. This approach could in principle be repeated for higher order functionals.

Theorem 4.8 (second-order isotropic sketching equivalence). *Consider the setting of Theorem 4.1. If $\mathbf{\Psi} \in \mathbb{C}^{p\times p}$ is a deterministic or random positive semidefinite matrix independent of $\mathbf{S}$ with $\|\mathbf{\Psi}\|_{\mathrm{op}}$ uniformly bounded in p, then if either $\lambda \neq 0$ or $\limsup \frac{q}{p} < \liminf r(\mathbf{A})$,*

$$\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S}+\lambda\mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}}\mathbf{\Psi}\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S}+\lambda\mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}} \simeq (\mathbf{A}+\mu\mathbf{I}_p)^{-1}(\mathbf{\Psi}+\mu'\mathbf{I}_p)(\mathbf{A}+\mu\mathbf{I}_p)^{-1},$$

where μ is as in Theorem 4.1, and

$$\mu' = \frac{\frac{1}{q}\mathrm{tr}\left[\mu^3(\mathbf{A}+\mu\mathbf{I}_p)^{-1}\mathbf{\Psi}(\mathbf{A}+\mu\mathbf{I}_p)^{-1}\right]}{\lambda + \frac{1}{q}\mathrm{tr}\left[\mu^2\mathbf{A}\,(\mathbf{A}+\mu\mathbf{I}_p)^{-2}\right]} \geq 0. \tag{4.5}$$

Proof. By assumption, there exists $M < \infty$ such that $M > \limsup \|(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S}+\lambda\mathbf{I}_q)^{-1}\|_{\mathrm{op}}$ and $M > \limsup \|(\mathbf{A}+\mu\mathbf{I}_p)^{-1}\|_{\mathrm{op}}$ almost surely (see proof details for Theorem 4.1 in the supplementary material). Define $\mathbf{B}_z \triangleq \mathbf{A} + z\mathbf{\Psi}$. Then for all $z \in D$, where

$$D = \{z \in \mathbb{C} : \limsup\left(|z|M\|\mathbf{\Psi}\|_{\mathrm{op}}\max\{\|\mathbf{S}\|_{\mathrm{op}}^2, 1\}\right) < \tfrac{1}{2}\},$$

we have that $\max\{\limsup \|(\mathbf{S}^{\mathsf{H}}\mathbf{B}_z\mathbf{S}+\lambda\mathbf{I}_q)^{-1}\|_{\mathrm{op}}, \limsup \|(\mathbf{B}_z+\mu\mathbf{I}_p)^{-1}\|_{\mathrm{op}}\} \leq 2M$. Therefore, we can apply the differentiation rule of asymptotic equivalences for all $z \in D$:

$$\begin{aligned}
&-\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{B}_z\mathbf{S}+\lambda\mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}}\mathbf{\Psi}\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{B}_z\mathbf{S}+\lambda\mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}} = \tfrac{\partial}{\partial z}\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{B}_z\mathbf{S}+\lambda\mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}} \\
&\quad\simeq \tfrac{\partial}{\partial z}\left(\mathbf{B}_z+\mu(z)\mathbf{I}_p\right)^{-1} \\
&\quad= -\left(\mathbf{B}_z+\mu(z)\mathbf{I}_p\right)^{-1}\left(\mathbf{\Psi}+\tfrac{\partial}{\partial z}\mu(z)\mathbf{I}_p\right)\left(\mathbf{B}_z+\mu(z)\mathbf{I}_p\right)^{-1}.
\end{aligned}$$

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

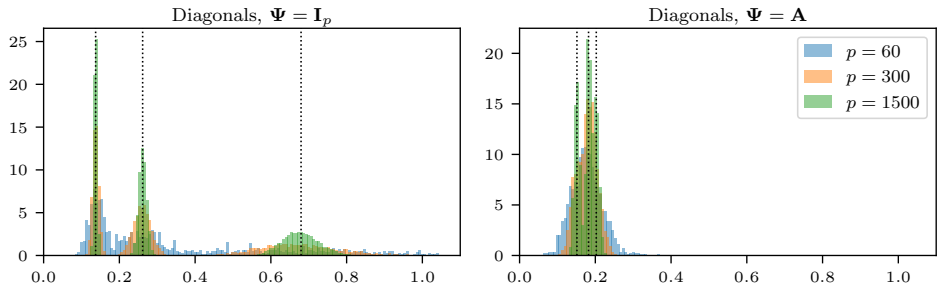

Figure 3. *Empirical density histograms over* 20 *trials demonstrating the concentration of diagonal elements of* $\mathbf{S}(\mathbf{S}^\top\mathbf{A}\mathbf{S}+\lambda\mathbf{I})^{-1}\mathbf{S}^\top\mathbf{\Psi}\mathbf{S}(\mathbf{S}^\top\mathbf{A}\mathbf{S}+\lambda\mathbf{I})^{-1}\mathbf{S}^\top$ *for* $(\mathbf{S},\mathbf{A},\lambda)$ *as in Figure 2 and* $\mathbf{\Psi}\in\{\mathbf{I}_p,\mathbf{A}\}$. *As expected by Theorem 4.8, the individual elements of the sketched pseudoinverse converge to those of* $(\mathbf{A}+\mu\mathbf{I})^{-1}(\mathbf{\Psi}+\mu'\mathbf{I})(\mathbf{A}+\mu\mathbf{I})^{-1}$ *(black, dotted), where* $\mu'\approx 0.813$ *and* 0.403 *for* $\mathbf{\Psi}=\mathbf{I}_p$ *and* $\mathbf{A}$, *respectively.*

We let $\mu'(z)=\frac{\partial}{\partial z}\mu(z)$, and then we can divide (4.4) by $\mu(z)$ and differentiate to obtain

$$\frac{\lambda\mu'(z)}{\mu(z)^2}=\tfrac{1}{q}\mathrm{tr}\left[\mathbf{\Psi}(\mathbf{B}_z+\mu(z)\mathbf{I}_p)^{-1}-\mathbf{B}_z(\mathbf{B}_z+\mu(z)\mathbf{I}_p)^{-1}\left(\mathbf{\Psi}+\mu'(z)\mathbf{I}_p\right)(\mathbf{B}_z+\mu(z)\mathbf{I}_p)^{-1}\right].$$

Solving for $\mu'(0)$ gives the expression in (4.5). For the nonnegativity of μ', see Remark 5.6 and its proof. ∎

That is, the second-order equivalence is the same as plugging in the first-order equivalence and then adding a nonnegative inflation $\mu'(\mathbf{A}+\mu\mathbf{I})^{-2}$. The inflation factor μ' depends linearly on the matrix $\mathbf{\Psi}$, but the inflation is always isotropic, rather than in the direction of $\mathbf{\Psi}$. It is nonnegative in the same way that the variance of an estimator is also nonnegative. Examples of quadratic forms where this second-order equivalence can be used include estimation error ($\mathbf{\Psi}=\mathbf{I}$) and prediction error ($\mathbf{\Psi}=\mathbf{\Sigma}$, the population covariance) in ridge regression problems. We give a demonstration of the concentration in Figure 3. While typically $\mu'>0$, it can go to 0 in the special case of $\mu=0$ and $\mathbf{\Psi}$ sharing a subspace with $\mathbf{A}$, as we discuss in Remark 5.7.

Remark 4.9 (the case of $\lambda=0$). Similar to the variant form in (4.2) of Theorem 4.1, if we consider the slightly different form

$$\mathbf{A}^{1/2}\mathbf{S}\left(\mathbf{S}^\mathsf{H}\mathbf{A}\mathbf{S}+\lambda\mathbf{I}_q\right)^{-1}\mathbf{S}^\mathsf{H}\mathbf{\Psi}\mathbf{S}\left(\mathbf{S}^\mathsf{H}\mathbf{A}\mathbf{S}+\lambda\mathbf{I}_q\right)^{-1}\mathbf{S}^\mathsf{H}\mathbf{A}^{1/2}$$
$$\simeq\mathbf{A}^{1/2}(\mathbf{A}+\mu\mathbf{I}_p)^{-1}(\mathbf{\Psi}+\mu'\mathbf{I}_p)(\mathbf{A}+\mu\mathbf{I}_p)^{-1}\mathbf{A}^{1/2}$$

for the second-order resolvent, we do not need the $\lambda\neq 0$ or $\limsup\frac{q}{p}<\liminf r(\mathbf{A})$ restriction as stated in the theorem. Because the proof of this case is entirely analogous to the results in Theorems 4.1 and 4.8, we omit the proof.

5. Properties and examples.

Below we provide various analytical properties of the quantities that appear in Theorems 4.1 and 4.8. See section SM4 in the supplementary material for their proofs.

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Table 1
Sign patterns of λ_0 and μ_0.

α vs. $r(\mathbf{A})$	μ_0	α vs. $\frac{1}{p}\mathrm{tr}[\mathbf{A}(\mathbf{A}+\mu_0\mathbf{I})^{-1}]$	λ_0
$\alpha > r(\mathbf{A})$	<0	$\alpha = \frac{1}{p}\mathrm{tr}[\mathbf{A}^2(\mathbf{A}+\mu_0\mathbf{I})^{-2}] > \frac{1}{p}\mathrm{tr}[\mathbf{A}(\mathbf{A}+\mu_0\mathbf{I})^{-1}]$	<0
$\alpha = r(\mathbf{A})$	0	$\alpha = \lim_{x\searrow 0}\frac{1}{p}\mathrm{tr}[\mathbf{A}^2(\mathbf{A}+x\mathbf{I})^{-2}] = \lim_{x\searrow 0}\frac{1}{p}\mathrm{tr}[\mathbf{A}(\mathbf{A}+x\mathbf{I})^{-1}]$	0
$\alpha < r(\mathbf{A})$	>0	$\alpha = \frac{1}{p}\mathrm{tr}[\mathbf{A}^2(\mathbf{A}+\mu_0\mathbf{I})^{-2}] < \frac{1}{p}\mathrm{tr}[\mathbf{A}(\mathbf{A}+\mu_0\mathbf{I})^{-1}]$	<0

5.1. Lower limits. The quantities λ_0 and μ_0 provide the lower limits of regularization in Theorem 4.1. The following two remarks describe their behavior in terms of α.

Remark 5.1 (dependence of μ_0 and λ_0 on α). Writing the first equation in (4.1) as

$$\alpha = \tfrac{1}{p}\mathrm{tr}\left[\mathbf{A}^2\left(\mathbf{A}+\mu_0\mathbf{I}_p\right)^{-2}\right], \tag{5.1}$$

note that for fixed $\mathbf{A}$, μ_0 only depends on α. Furthermore, the equation indeed admits a unique solution for μ_0 for a given α. This can be seen by noting that the function $f:\mu_0 \mapsto \frac{1}{p}\mathrm{tr}[\mathbf{A}^2(\mathbf{A}+\mu_0\mathbf{I}_p)^{-2}]$ is monotonically decreasing in μ_0, and

$$\tfrac{1}{p}\lim_{\mu_0\searrow -\lambda^+_{\min}(\mathbf{A})}\mathrm{tr}[\mathbf{A}^2(\mathbf{A}+\mu_0\mathbf{I}_p)^{-2}] = \infty, \quad \text{and} \quad \tfrac{1}{p}\lim_{\mu_0\nearrow\infty}\mathrm{tr}[\mathbf{A}^2(\mathbf{A}+\mu_0\mathbf{I}_p)^{-2}] = 0.$$

In addition, because $\mu_0(\alpha) = f^{-1}(\alpha)$, μ_0 is monotonically decreasing in α, and $\lim_{\alpha\searrow 0}\mu_0(\alpha) = \infty$ and $\lim_{\alpha\nearrow\infty}\mu_0(\alpha) = -\lambda^+_{\min}(\mathbf{A})$.

Given μ_0, the second equation in (4.1) then provides λ_0 as

$$\lambda_0 = \mu_0\left(1 - \tfrac{1}{\alpha}\tfrac{1}{p}\mathrm{tr}\left[\mathbf{A}(\mathbf{A}+\mu_0\mathbf{I})^{-1}\right]\right). \tag{5.2}$$

For $\alpha \in (0, r(\mathbf{A}))$, $\lambda_0 : \alpha \mapsto \lambda_0(\alpha)$ is monotonically increasing, and $\lim_{\alpha\searrow 0}\lambda_0(\alpha) = -\infty$ and $\lim_{\alpha\to r(\mathbf{A})}\lambda_0(\alpha) = 0$. When $\alpha = r(\mathbf{A})$, $\mu_0 = 0$ and consequently $\lambda_0 = 0$. Finally, for $\alpha \in (r(\mathbf{A}), \infty)$, $\lambda_0 : \alpha \mapsto \lambda_0(\alpha)$ is monotonically decreasing in α, and $\lim_{\alpha\nearrow\infty}\lambda_0(\alpha) = -\lambda^+_{\min}(\mathbf{A})$. This follows from a short limiting calculation.

Remark 5.2 (joint sign patterns of μ_0 and λ_0). Observe from (4.1) the sign pattern summarized in Table 1.

5.2. First-order equivalence. In general, the exact μ depends on λ, α, and $\mathbf{A}$ via the fixed-point equation (4.4). However, we can infer several properties of the behavior of μ as a function of λ and α as summarized below.

Proposition 5.3 (monotonicities of μ in λ and α). *For a fixed $\alpha \geq 0$, the map $\lambda \mapsto \mu(\lambda)$, where $\mu(\lambda)$ is as defined in (4.4), is monotonically increasing in λ over (λ_0, ∞), and $\lim_{\lambda\searrow\lambda_0}\mu(\lambda) = \mu_0$, while $\lim_{\lambda\nearrow\infty}\mu(\lambda) = \infty$. For a fixed $\lambda \geq 0$, the map $\alpha \mapsto \mu(\alpha)$ where $\mu(\alpha)$ is as defined in (4.4) is monotonically decreasing in α over $(0,\infty)$; when $\lambda < 0$, the map $\alpha \to \mu(\alpha)$ is monotonically decreasing over $(0, r(\mathbf{A}))$ and monotonically increasing over $(r(\mathbf{A}), \infty)$. Furthermore, for any $\lambda \in (\lambda_0, \infty)$, $\lim_{\alpha\searrow 0}\mu(\alpha) = \infty$, and $\lim_{\alpha\nearrow\infty}\mu(\alpha) = \lambda$.*

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Remark 5.4 (joint signs of λ and μ). When $\lambda \geq 0$, for any $\alpha > 0$, we have $\mu \geq 0$, where μ is the unique solution to (4.4) in (μ_0, ∞). When $\lambda < 0$, for $\alpha \leq r(\mathbf{A})$, we have $\mu \geq 0$, while for $\alpha > r(\mathbf{A})$, we have $\operatorname{sign}(\mu) = \operatorname{sign}(\lambda)$.

Proposition 5.5 (concavity, bounds, and asymptotic behavior of μ in λ). *The function $\lambda \mapsto \mu(\lambda)$, where $\mu(\lambda)$ is the solution to (4.4), is a concave function over (λ_0, ∞). Furthermore, for any $\alpha \in (0, \infty)$, $\mu(\lambda) \leq \lambda + \frac{1}{q}\operatorname{tr}[\mathbf{A}]$ for all $\lambda \in (\lambda_0, \infty)$, and when $\alpha \leq r(\mathbf{A})$, $\mu(\lambda) \geq \lambda$ for all $\lambda \in (\lambda_0, \infty)$; otherwise $\mu(\lambda) \geq \lambda$ for $\lambda \geq 0$. Additionally, $\lim_{\lambda \nearrow \infty} |\mu(\lambda) - (\lambda + \frac{1}{q}\operatorname{tr}[\mathbf{A}])| = 0$.*

5.3. Second-order equivalence. Below we provide a few additional properties related to the inflation factor μ' in (4.5), which appears in the statement of Theorem 4.8.

Remark 5.6. We have the following alternative form for μ':

$$\mu' = \tfrac{1}{q}\operatorname{tr}\left[\mu^2(\mathbf{A} + \mu\mathbf{I}_p)^{-1}\mathbf{\Psi}(\mathbf{A} + \mu\mathbf{I}_p)^{-1}\right]\frac{\partial \mu}{\partial \lambda}.$$

Note that the term $\frac{\partial \mu}{\partial \lambda}$ does not depend in any way on $\mathbf{\Psi}$ and that the remaining term is well-controlled for any $\mu > \mu_0$. Therefore, μ' will only diverge when $\frac{\partial \mu}{\partial \lambda}$ diverges, which occurs as $\lambda \to \lambda_0$. This is clearly visible in Figure 4 (top) as λ approaches λ_0, where the slope of the curve tends to infinity. Additionally, because μ is increasing in λ, this decomposition shows that $\mu' \geq 0$.

Remark 5.7 (vanishing μ'). If $\operatorname{Ker}(\mathbf{A}) \subseteq \operatorname{Ker}(\mathbf{\Psi})$, then as $\mu \to 0$, $\mu' \searrow 0$. The best intuition for this is in the case $\mathbf{\Psi} = \mathbf{A}$. Because we can only have $\mu = 0$ for $\alpha > r(\mathbf{A})$ and $\lambda = 0$, we have $\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda\mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda\mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}}|_{\lambda=0} = \mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda\mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}}|_{\lambda=0}$, and the second-order equivalence reduces to the first-order equivalence with no inflation factor. This remarkable property means that sketching leads to extremely accurate estimates with no spectral distortion, but only in low-rank settings with little regularization.

5.4. Illustrative examples. In order to better understand Theorems 4.1 and 4.8, we consider a few examples with special choices of the matrix $\mathbf{A}$. When the spectrum of $\mathbf{A}$ converges to a particular distribution of eigenvalues, μ will converge to a value that is deterministic given $\mathbf{A}$.

5.4.1. Isotropic rank-deficient matrix. For the first example, let $0 < r \leq 1$ be a real number. We then consider $\mathbf{A} = \left[\begin{smallmatrix} \mathbf{I}_{\lfloor rp \rfloor} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right]$ such that $r(\mathbf{A}) \to r$ as $p \nearrow \infty$. We have chosen the standard basis representation of this matrix, but the following results also hold for any $\mathbf{A}$ that is isotropic on a subspace, regardless of basis. Such an $\mathbf{A}$ includes settings such as $\mathbf{A} = \mathbf{X}^\top\mathbf{X}$ where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is an orthogonal design matrix with orthonormal rows. In this case,

$$\mu = \frac{\lambda + \frac{r}{\alpha} - 1 + \sqrt{(\lambda + \frac{r}{\alpha} - 1)^2 + 4\lambda}}{2}.$$

Furthermore, we have simple forms for μ_0 and λ_0:

$$\mu_0 = \sqrt{\tfrac{r}{\alpha}} - 1, \quad \lambda_0 = -\left(\sqrt{\tfrac{r}{\alpha}} - 1\right)^2.$$

The expression for λ_0 can also be obtained directly from the minimum nonzero eigenvalue of the Marchenko–Pastur distribution with aspect ratio $\frac{\alpha}{r}$ and variance scaling $\frac{r}{\alpha}$, which describes $\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S}$. In the case $\lambda = 0$, we have a very simple expression for μ:

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

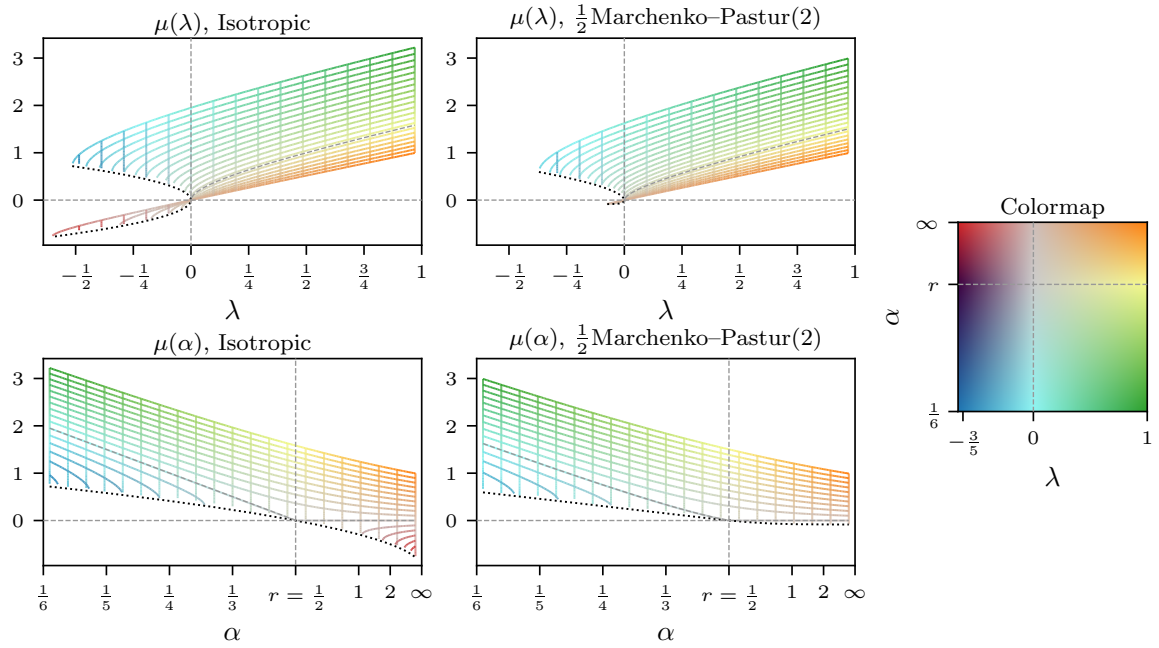

Figure 4. *Plots of μ as a function of λ and α for rank-deficient isotropic (left) and Marchenko–Pastur (middle) spectra, normalized so that $\frac{1}{p}\mathrm{tr}[\mathbf{A}] = r = 1/2$. The values of λ and α in each location of the plot are indicated by the colormap (right), shared between the two views of each plot. As we sweep α, we also plot $(\alpha, \lambda_0, \mu_0)$ (black, dotted). We also plot the lines $\mu = 0$, $\lambda = 0$, and $\alpha = r$ (gray, dashed). The scaling of the μ and λ axes are linear, and the scaling of the α axis is proportional to $1/\alpha$. In this way we can clearly capture the general $\mu \approx \lambda + \frac{1}{p}\mathrm{tr}[A]/\alpha$ relationship for $\lambda > 0$, as well as the limiting behavior of $\mu = \lambda$ for large α. The most significant difference between the two distributions is that for the isotropic distribution, $\lambda^+_{\min}(\mathbf{A}) = 1$, while for the Marchenko–Pastur case, $\lambda^+_{\min}(\mathbf{A}) = (\sqrt{2}-1)^2/2 \approx 0.0859$, limiting the achievable negative values of μ when $\lambda < 0$ and $\alpha > r$.*

$$\mu = \begin{cases} \frac{r}{\alpha} - 1 & \text{if } \alpha < r, \\ 0 & \text{otherwise.} \end{cases}$$

We can also obtain the limiting behavior of μ for large λ or small α:

$$\lim_{\lambda + \frac{r}{\alpha} \nearrow \infty} \frac{\mu}{\lambda + \frac{r}{\alpha}} = 1.$$

In Figure 4 (left), we plot μ as a function of both λ and α. We see that even for modest values of $\lambda > 0$ or $\alpha < r$, the relationship $\mu \sim \lambda + \frac{r}{\alpha}$ holds quite accurately. We see a clear transition point at $\alpha = r$ where $\lambda_0 = 0$, and on either side of which λ_0 decreases. Other properties from the previous sections, such as monotonicity, concavity in λ, and sign patterns, are clearly visible in this plot as well. We also plot μ' as a function of μ and α in Figure 5, where we see that the inflation vanishes for $\mathbf{\Psi} = \mathbf{A}$ only if $\alpha > r$ and $\mu = 0$. It is nonnegligible otherwise and tends to infinity as μ tends to μ_0 for each α.

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

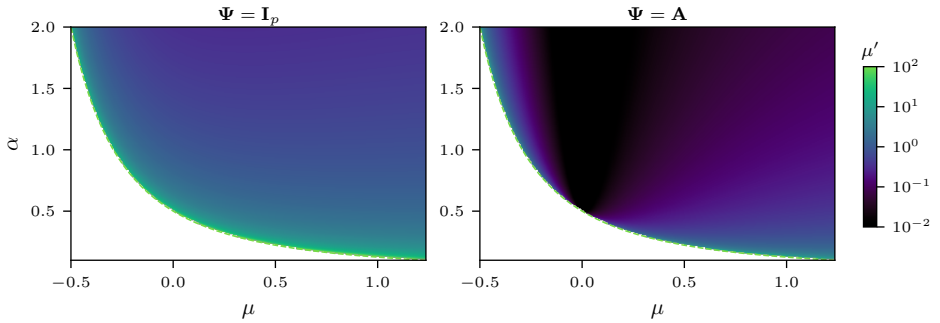

Figure 5. *Plot of μ' as a function of μ and α for the rank-deficient isotropic spectrum with $r = 1/2$ for $\mathbf{\Psi} \in \{\mathbf{I}_p, \mathbf{A}\}$. In both cases, as $\mu \searrow \mu_0$ (dashed), $\mu' \nearrow \infty$. Otherwise, μ' is not too large. For $\mathbf{\Psi} = \mathbf{I}_p$, μ' decays slowly in α and μ. However, for $\mathbf{\Psi} = \mathbf{A}$, there is a regime for $\alpha > r$ around $\mu = 0$ for which μ' tends to zero. Thus, the unregularized pseudoinverse preserves $\mathbf{A}$ remarkably well on its range when the sketch size is greater than the rank of the matrix, but outside of the range of $\mathbf{A}$, it has nonnegligible error.*

5.4.2. Marchenko–Pastur spectrum. We also consider the case when $\mathbf{A}$ is a random matrix of the form $\mathbf{A} = \frac{1}{n}\mathbf{Z}^\top\mathbf{Z}$, where $\mathbf{Z} \in \mathbb{R}^{n\times p}$ contains i.i.d. entries of mean 0, variance 1, and bounded moments of order $4+\delta$ for some $\delta > 0$. This case is of interest for real data settings where $\mathbf{A}$ will be a sample covariance matrix. In this case, the spectrum of $\mathbf{A}$ can be computed explicitly and is given by the Marchenko–Pastur law. Computing μ explicitly in this case is possible, but cumbersome. We instead provide numerical illustrations on the behavior of μ as a function of α and λ.

From Figure 4 (middle), we can see that the behavior of μ for the Marchenko–Pastur spectrum is not substantially different from the rank-deficient isotropic spectrum. The only regime that differs significantly is when $\alpha > r(\mathbf{A})$ and $\lambda < 0$, where λ_0 is much closer to 0 than in the isotropic case, and so there is no equivalence for more negative values of λ.

It is also worth noting that when $\alpha < r(\mathbf{A}) < 1$, the naïve bound on the smallest regularization λ permissible is 0 (as explained in the caption of Figure 1). However, from Figure 4 we observe that the equivalence in Theorem 4.1 holds even for quite negative λ (blue region), contrary to this naïve bound. In fact, the true bound is almost the same as the rank-deficient isotropic case, $\lambda_0 = -(\sqrt{\frac{r}{\alpha}} - 1)^2$.

6. Applications.

To demonstrate how to apply our theory to sketching-based algorithms, we give two concrete examples, demonstrating when the first-order equivalence can be sufficient to characterize performance and when the second-order equivalence is necessary. We leave proof details to section SM5 in the supplementary material.

6.1. Sketch-and-project.

The sketch-and-project algorithm, also known as the generalized Kaczmarz method, solves the satisfiable linear system $\mathbf{L}\mathbf{x} = \mathbf{b}$ for some $\mathbf{L} \in \mathbb{C}^{n\times p}$ via the following iterations:

$$\mathbf{x}_t = \mathbf{x}_{t-1} - \mathbf{L}^{\mathsf{H}}\mathbf{S}_t(\mathbf{S}_t^{\mathsf{H}}\mathbf{L}\mathbf{L}^{\mathsf{H}}\mathbf{S}_t)^{\dagger}\mathbf{S}_t^{\mathsf{H}}(\mathbf{L}\mathbf{x}_{t-1} - \mathbf{b}).$$

Here $\mathbf{S}_t \in \mathbb{C}^{n\times m}$ are independently drawn random sketching matrices. This algorithm classically enjoys linear convergence of $\mathbb{E}[\|\mathbf{x}_t - \mathbf{x}_*\|_2^2]$ where $\mathbf{x}_* = \mathbf{L}^{\dagger}\mathbf{b}$ that depends only on the

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

smallest eigenvalue of $\mathbb{E}\left[\mathbf{L}^{\mathsf{H}}\mathbf{S}_t(\mathbf{S}_t^{\mathsf{H}}\mathbf{L}\mathbf{L}^{\mathsf{H}}\mathbf{S}_t)^{\dagger}\mathbf{S}_t^{\mathsf{H}}\mathbf{L}\right]$ [24, 18]. Since this is the same quantity of interest as in our sketching equivalence, we obtain a similar convergence guarantee in the asymptotic limit almost surely by applying Theorem 4.1 with $\mathbf{A} = \mathbf{L}\mathbf{L}^{\mathsf{H}}$ (see subsection SM5.1):

$$\|\mathbf{x}_t - \mathbf{x}_*\|_2^2 \lesssim \rho^t \|\mathbf{x}_0 - \mathbf{x}_*\|_2^2, \quad \text{where} \quad \rho \triangleq \frac{\mu}{\lambda_{\min}^{+}(\mathbf{L}\mathbf{L}^{\mathsf{H}}) + \mu}. \tag{6.1}$$

Here by $a_{n,t} \lesssim b_{n,t}$, we mean that for any fixed t, $\liminf_{n\to\infty} b_{n,t} - a_{n,t} \geq 0$, and the result holds for an implicit sequence of $\mathbf{x}_0$, $\mathbf{x}_*$ with increasing dimensions and uniformly bounded norms such that Theorem 4.1 can be applied. Since there are no second-order effects, and we use $\lambda = 0$, this convergence result holds in fact for any rotationally invariant sketch by Remark 4.4. Asymptotically, assuming we can compute the product $\mathbf{L}^{\mathsf{H}}\mathbf{S}_t$ efficiently, the computational bottleneck comes from evaluating the pseudoinverse $\mathbf{L}^{\mathsf{H}}\mathbf{S}_t(\mathbf{S}_t^{\mathsf{H}}\mathbf{L}\mathbf{L}^{\mathsf{H}}\mathbf{S}_t)^{\dagger}$, which typically has complexity $O(mp\min\{m,p\})$.[3] To reach a desired residual $\|\mathbf{L}\mathbf{x}_t - \mathbf{b}\|_2^2 \leq \varepsilon$, we must run the algorithm for at most $t_\varepsilon = \lceil \log(\varepsilon/\lambda_{\max}(\mathbf{L}\mathbf{L}^{\mathsf{H}})\|\mathbf{x}_0 - \mathbf{x}_*\|_2^2)/\log(\rho)\rceil$ iterations. The total complexity of the algorithm is therefore $O(m^2 p t_\varepsilon)$ for $m < p$, compared to $O(np\min\{n,p\})$ to solve the system directly. Since both of these quantities diverge in the asymptotic limit, it is of more interest to study their quotient. To that end, we define the *relative computation factor* $\alpha^2 t_\varepsilon$ for $\alpha = \frac{m}{n}$, which is equal to the quotient up to a factor of $\frac{\min\{n,p\}}{n}$, which does not depend on α.

Remark 6.1 (optimal sketch size for minimizing computation). The asymptotic relative computation factor $\alpha^2 t_\varepsilon$ is characterized as follows. For $\alpha \geq r(\mathbf{L})$, $t_\varepsilon = 1$ for all ε, and so $\alpha^2 t_\varepsilon = \alpha^2$. For all sufficiently small ε, $\lim_{\alpha \searrow 0} \alpha^2 t_\varepsilon = 0$. For $0 < \alpha < r(\mathbf{L})$, $\lim_{\varepsilon \searrow 0} \alpha^2 t_\varepsilon = \infty$. Thus, for small ε, the computational complexity of sketch-and-project is minimized globally by letting $\alpha \searrow 0$ and locally by choosing $\alpha = r(\mathbf{L})$.

We demonstrate this observation empirically in Figure 6. In order to keep the cost of evaluating $\mathbf{L}^{\mathsf{H}}\mathbf{S}_t$ to $O(m^2 p)$, we sample sparse Gaussian matrices $\mathbf{S}_t$ according to Remark 4.5

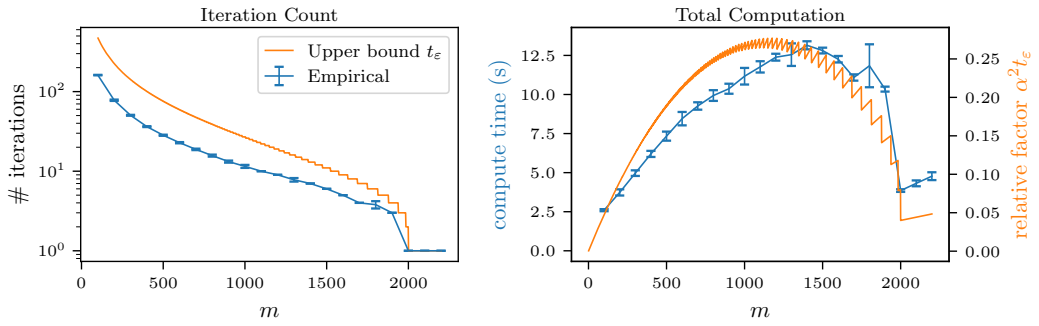

Figure 6. *Empirical computation time of sketch-and-project as a function of sketch size m. We sample a fixed $[\mathbf{L}]_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $\mathbf{x}_* \sim \mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{I}_p)$ for $n = 10^4$, $p = 2000$. We run the algorithm until $\frac{1}{n}\|\mathbf{L}\mathbf{x}_t - \mathbf{b}\|_2^2 \leq 10^{-3}$. We find that the number of iterations (blue) matches our upper bound t_ε (orange) up to a constant factor (left). Additionally (right), we find that the trend of the wall-clock time of the algorithm (blue) matches the relative computation factor $\alpha^2 t_\varepsilon$ (orange), and that the computation time is minimized by taking α as small as possible. Error bars denote standard deviation over 10 random trials.*

[3] Our remarks here also hold directly for any possible "galactic" matrix inversion algorithm of complexity $O(mp\min\{m,p\}^{\delta})$ for some $\delta > 0$ [3], provided $\mathbf{L}^{\mathsf{H}}\mathbf{S}_t$ can be computed in similar time.

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

having elements drawn from $\mathcal{N}(0, \frac{n}{m^2})$ with probability $\frac{m}{n}$ and 0 otherwise, such that there are $O(m^2)$ nonzero elements of $\mathbf{S}_t$ with high probability.

6.2. Sketched ridge regression. In sketch-and-project, we introduced new randomness in each iteration, and as a result the first-order equivalence was sufficient to characterize the algorithm's performance. However, with less randomness, the second-order effects are much more pronounced. We illustrate this in the setting of sketched ridge regression, also known as sketch-and-solve, which is an important problem in randomized numerical linear algebra [41].

Concretely, we can define the sketched ridge regression problem for design matrix $\mathbf{L} \in \mathbb{C}^{n\times p}$, targets $\mathbf{b} \in \mathbb{C}^n$, and sketching matrix $\mathbf{S} \in \mathbb{C}^{n\times m}$ as

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \tfrac{1}{n}\|\mathbf{S}^{\mathsf{H}}(\mathbf{L}\mathbf{x} - \mathbf{b})\|_2^2 + \lambda\|\mathbf{x}\|^2.$$

To connect back to sketch-and-project from the previous section, a single iteration of sketch-and-project solves this exact problem if we set $\lambda = 0$ and replace $\mathbf{b}$ by $\mathbf{b} - \mathbf{L}\mathbf{x}_t$. For brevity and parallelism with sketch-and-project, we only consider this formulation of sketched ridge regression. However, similar analyses can be performed for "dual" sketching where we consider residuals $\mathbf{L}\mathbf{S}'\mathbf{x} - \mathbf{b}$, as well as joint sketching with residuals $\mathbf{S}^{\mathsf{H}}(\mathbf{L}\mathbf{S}'\mathbf{x} - \mathbf{b})$; see [33].

The solution $\widehat{\mathbf{x}}$ is given in terms of the sketched (regularized) pseudoinverse, which means we can obtain its first-order asymptotic equivalent from Theorem 4.1 with $\mathbf{A} = \frac{1}{n}\mathbf{L}\mathbf{L}^{\mathsf{H}}$:

$$\widehat{\mathbf{x}} = \tfrac{1}{n}\mathbf{L}^{\mathsf{H}}\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\tfrac{1}{n}\mathbf{L}\mathbf{L}^{\mathsf{H}}\mathbf{S} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{S}^{\mathsf{H}}\mathbf{b} \simeq \tfrac{1}{n}\mathbf{L}^{\mathsf{H}}\left(\tfrac{1}{n}\mathbf{L}\mathbf{L}^{\mathsf{H}} + \mu\mathbf{I}_p\right)^{-1}\mathbf{b} \triangleq \widehat{\mathbf{x}}_{\text{equiv}}.$$

Furthermore, we can characterize second-order errors; if we define the quadratic error

$$\mathcal{E}_{\mathbf{\Phi}}(\mathbf{x}, \mathbf{x}') \triangleq (\mathbf{x} - \mathbf{x}')^{\mathsf{H}}\mathbf{\Phi}(\mathbf{x} - \mathbf{x}'),$$

we can apply Theorem 4.8 with $\mathbf{\Psi} = \frac{1}{n}\mathbf{L}\mathbf{\Phi}\mathbf{L}^{\mathsf{H}}$ to obtain

$$\mathcal{E}_{\mathbf{\Phi}}\left(\widehat{\mathbf{x}}, \mathbf{x}'\right) \simeq \mathcal{E}_{\mathbf{\Phi}}\left(\widehat{\mathbf{x}}_{\text{equiv}}, \mathbf{x}'\right) + \frac{\mu'}{n}\mathbf{b}^{\mathsf{H}}\left(\tfrac{1}{n}\mathbf{L}\mathbf{L}^{\mathsf{H}} + \mu\mathbf{I}_n\right)^{-2}\mathbf{b}, \tag{6.2}$$

where

$$\mu' = \frac{\frac{1}{m}\operatorname{tr}\left[\mu^3\left(\frac{1}{n}\mathbf{L}\mathbf{L}^{\mathsf{H}} + \mu\mathbf{I}_n\right)^{-1}\frac{1}{n}\mathbf{L}\mathbf{\Phi}\mathbf{L}^{\mathsf{H}}\left(\frac{1}{n}\mathbf{L}\mathbf{L}^{\mathsf{H}} + \mu\mathbf{I}_n\right)^{-1}\right]}{\lambda + \frac{1}{m}\operatorname{tr}\left[\mu^2\frac{1}{n}\mathbf{L}\mathbf{L}^{\mathsf{H}}\left(\frac{1}{n}\mathbf{L}\mathbf{L}^{\mathsf{H}} + \mu\mathbf{I}_n\right)^{-2}\right]} \geq 0.$$

In other words, the error of the sketched solution can be decomposed into the error of the first-order equivalent solution plus an inflation quantity. Note that this inflation is only the additional effect due to sketching. This should not be conflated with estimate variance, which is generally defined to include the effect of noise in $\mathbf{b}$, which will appear in both the $\mathcal{E}_{\mathbf{\Phi}}(\widehat{\mathbf{x}}_{\text{equiv}}, \mathbf{x}')$ and inflation terms.

The inflation term can be quite large when λ is near λ_0, meaning the sketched solution is quite poor; however, by averaging K independently sketched solutions we can replace μ' by $\frac{\mu'}{K}$, allowing us to control the inflation via randomized parallelization, such as in distributed settings. We demonstrate this theoretically and empirically in Figure 7. Note how in the

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

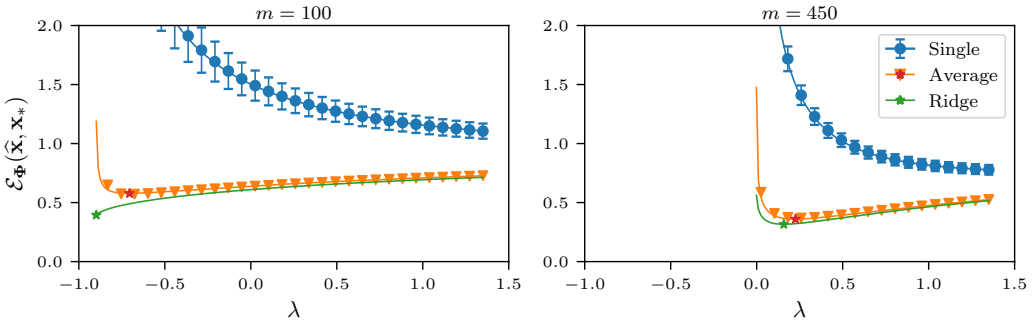

Figure 7. *Estimation error* $\mathcal{E}_\Phi(\widehat{\mathbf{x}}, \mathbf{x}_*) = \|\widehat{\mathbf{x}} - \mathbf{x}_*\|_2^2$ *for a sketched ridge regression problem as a function of* λ. *We sample a fixed* $[\mathbf{L}]_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ *and* $\mathbf{x}_* \sim \mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{I}_p)$ *and generate a fixed* $\mathbf{b} = \mathbf{L}\mathbf{x}_* + \mathbf{h}$ *with* $\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$ *for* $n = 2000$, $p = 400$, *and* $\sigma = 1.5$. *We plot the theoretical asymptotic error from* (6.2) *(lines) as well as empirical values (circles and triangles), averaging over* $K = 30$ *random sketches* $\mathbf{S}$. *We plot the single estimate error (blue), average of* K *estimates (orange), and equivalent ridge predictor (green) for an undersampled setting* ($m = 100$, *left) and an oversampled setting* ($m = 450$, *right). In the undersampled setting, the optimal error (stars) for the averaged estimate is obtained by using negative* λ. *We emphasize that the data model here is underparameterized with a moderate signal-to-noise ratio and is not contrived to make negative regularization optimal as seen in some overparameterized settings* [28, 57].

undersampled regime with $m = 100$, which is the regime of interest for distributed optimization as it reduces the computational cost per worker, the optimal regularization penalty λ can in fact be *negative,* even if the optimal ridge penalty μ for the equivalent problem is positive. Our theoretical characterization enables us to handle this case elegantly. The intuition behind this is that the smaller the sketch size is, the more regularization is added, and so to achieve a target regularization (the optimal ridge penalty), negative regularization may be required.

7. Discussion and extensions. In this paper, we have provided a detailed look at the asymptotic effects of i.i.d. sketching on matrix inverses. We have provided an extension of existing asymptotic equivalence results to real-valued regularization (including negative) and used this result to obtain both first- and second-order asymptotic equivalences for the sketched regularized pseudoinverse. We have also described how to apply these equivalences to analyze algorithms based on random sketching, providing novel insights into sketch-and-project and ridge regression as concrete examples.

Our work is far from a complete characterization of sketching. We now list some natural extensions to our results.

Relaxing assumptions, strengthening conclusions. As mentioned in section 4, we make minimal assumptions on the base matrix $\mathbf{A}$. In particular, we do not assume that the empirical spectral distribution of $\mathbf{A}$ converges to any fixed limit. The assumption that the maximum and minimum eigenvalues of $\mathbf{A}$ are bounded away from 0 and ∞ can be weakened. In particular, one can let some eigenvalues escape to ∞ and have some eigenvalues decay to 0, provided certain functionals of the eigenvalues remain bounded. Our assumptions on the sketching matrix $\mathbf{S}$ are also weak. We do not assume any distributional structure on its entries and only require bounded moments of order $8 + \delta$ for some $\delta > 0$. Using a truncation strategy, one can push this to only requiring moments of order $4 + \delta$ for some $\delta > 0$ for almost sure equivalences up to order 2 that we show in this paper. Finally, while our asymptotic results give practically

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

relevant insights for finite systems, we lack a precise characterization for nonasymptotic settings. In particular, the rate of convergence depends on a number of factors including the choice of λ and the higher order moments of the elements of $\mathbf{S}$.

Generalized sketching. Our assumption that the elements of the matrix $\mathbf{S}$ are i.i.d. draws from some distribution limits its application in practical settings on two key fronts: the effect of a rotationally invariant sketch is isotropic regularization, i.i.d. sketches can be slow to apply, and there is unnecessary distortion of the spectrum of $\mathbf{A}$ for $q \nearrow p$. We now discuss how to extend our framework to extend to more general classes of sketches that more closely align with those used in practice.

We may desire to use generalized nonisotropic ridge regularization to perform Bayes-optimal regression (see, e.g., Chapter 3 of [52]) or to avoid multiple descent [39, 58], or we may find ourselves using nonisotropic sketching matrices, such as in adaptive sketching [30], where the sketching matrix depends on the data. We can cover these cases with the following extension of Theorem 4.1.

Corollary 7.1 (nonisotropic sketching equivalence). *Assume the setting of Theorem 4.1. Let $\mathbf{R}$ be an invertible $p \times p$ positive semidefinite matrix, either deterministic or random but independent of $\mathbf{S}$ with $\limsup \|\mathbf{R}\|_{\mathrm{op}} < \infty$, and let $\widetilde{\mathbf{S}} = \mathbf{R}^{1/2}\mathbf{S}$. Then, for each $\lambda > -\liminf \lambda_{\min}^{+}(\widetilde{\mathbf{S}}^\top \mathbf{A}\widetilde{\mathbf{S}})$, as $p, q \nearrow \infty$ such that $0 < \liminf \frac{q}{p} \le \limsup \frac{q}{p} < \infty$,*

$$\widetilde{\mathbf{S}}\left(\widetilde{\mathbf{S}}^\top \mathbf{A}\widetilde{\mathbf{S}} + \lambda \mathbf{I}_q\right)^{-1}\widetilde{\mathbf{S}}^\top \simeq \left(\mathbf{A} + \mu \mathbf{R}^{-1}\right)^{-1},$$

where μ is the most positive solution to

$$\lambda = \mu\left(1 - \tfrac{1}{q}\mathrm{tr}\left[\mathbf{A}\left(\mathbf{A} + \mu\mathbf{R}^{-1}\right)^{-1}\right]\right).$$

Proof. The proof uses simple algebraic manipulations. Observe that, since the operator norm is submultiplicative, and $\|\mathbf{R}\|_{\mathrm{op}}$, $\|\mathbf{A}\|_{\mathrm{op}}$ are uniformly bounded in p, $\|\mathbf{R}^{1/2}\mathbf{A}\mathbf{R}^{1/2}\|_{\mathrm{op}}$ is also uniformly bounded in p. Using Theorem 4.1, we then have that

$$\mathbf{S}\left(\mathbf{S}^\top \mathbf{R}^{1/2}\mathbf{A}\mathbf{R}^{1/2}\mathbf{S} + \lambda \mathbf{I}_q\right)^{-1}\mathbf{S}^\top \simeq \left(\mathbf{R}^{1/2}\mathbf{A}\mathbf{R}^{1/2} + \mu \mathbf{I}_q\right)^{-1}.$$

Right and left multiplying both sides by $\mathbf{R}^{1/2}$, and writing $\widetilde{\mathbf{S}} = \mathbf{R}^{1/2}\mathbf{S}$, we get

$$\widetilde{\mathbf{S}}\left(\widetilde{\mathbf{S}}^\top \mathbf{A}\widetilde{\mathbf{S}} + \lambda \mathbf{I}_q\right)^{-1}\widetilde{\mathbf{S}}^\top \simeq \mathbf{R}^{1/2}\left(\mathbf{R}^{1/2}\mathbf{A}\mathbf{R}^{1/2} + \mu \mathbf{I}_p\right)^{-1}\mathbf{R}^{1/2} = \left(\mathbf{A} + \mu\mathbf{R}^{-1}\right)^{-1}$$

as desired, completing the proof. ∎

Because nonisotropic sketching can be used to induce generalized ridge regularization, this can be exploited adaptively to induce a wide range of structure-promoting regularization via iteratively reweighted least squares, in a manner similar to adaptive dropout methods (see [35] and references therein). Additionally, this result shows that methods applying ridge regularization to adaptive sketching methods, using, for example, $\mathbf{R} = \mathbf{A}$ as in [30], are not equivalent to ridge regression but instead to generalized ridge regression.

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Free sketching. Even among isotropic sketches, there can be a wide range of behavior beyond i.i.d. sketches. It turns out that a more general result holds for *free* sketching matrices (a notion from free probability that generalizes independence of random variables; see [40] for an introductory text). We state a complex version of the result in the following theorem and defer the general extension to real arguments and investigation of properties to future work.[4]

Theorem 7.2 (general free sketching). *Let $\mathbf{A} \in \mathbb{C}^{p\times p}$ be a positive semidefinite matrix and $\mathbf{S} \in \mathbb{C}^{p\times q}$ be a sketch such that the spectral distributions of $\mathbf{A}$ and $\mathbf{S}\mathbf{S}^{\mathsf{H}}$ converge almost surely to bounded distributions, and $\mathbf{S}\mathbf{S}^{\mathsf{H}}$ is asymptotically free from any other matrices*[5] *with respect to the average trace $\frac{1}{p}\mathrm{tr}[\cdot]$ and has limiting S-transform $\mathcal{S}_{\mathbf{S}\mathbf{S}^{\mathsf{H}}}$ analytic on $\mathbb{C}^-$. Then, for all $z \in \mathbb{C}^+$, there exists $\zeta \in \mathbb{C}^+$ such that*

$$\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} - z\mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}} \simeq (\mathbf{A} - \zeta\mathbf{I}_p)^{-1}.$$

Furthermore,

$$\zeta \simeq z\mathcal{S}_{\mathbf{S}\mathbf{S}^{\mathsf{H}}}\left(-\tfrac{1}{p}\mathrm{tr}[\mathbf{A}\,(\mathbf{A} - \zeta\mathbf{I}_p)^{-1}]\right) \quad \text{and} \quad \zeta \simeq z\mathcal{S}_{\mathbf{S}\mathbf{S}^{\mathsf{H}}}\left(-\tfrac{1}{p}\mathrm{tr}\left[\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} - z\mathbf{I}_q\right)^{-1}\right]\right).$$

Proof sketch. The key idea of the proof is to use Jacobi's formula for a parameterized matrix: $\frac{\partial}{\partial t}\log\det(\mathbf{B}_t) = \mathrm{tr}[\mathbf{B}_t^{-1}\frac{\partial \mathbf{B}_t}{\partial t}]$. First we simplify by considering self-adjoint $\boldsymbol{\Theta}$ and $\widetilde{\mathbf{S}} = (\mathbf{S}\mathbf{S}^{\mathsf{H}})^{1/2}$ so that we can work entirely in dimension p. We can then define $\mathbf{B}_{t,\zeta} = \mathbf{A} + t\boldsymbol{\Theta} - \zeta\mathbf{I}_p$ and $\mathbf{B}_{t,z}^{\widetilde{\mathbf{S}}} = \widetilde{\mathbf{S}}(\mathbf{A} + t\boldsymbol{\Theta})\widetilde{\mathbf{S}} - z\mathbf{I}_p$. What we need to prove is that $\frac{\partial}{\partial t}\frac{1}{p}\log\det(\mathbf{B}_{t,z}^{\widetilde{\mathbf{S}}}) \simeq \frac{\partial}{\partial t}\frac{1}{p}\log\det(\mathbf{B}_{t,\zeta})$ for some appropriate ζ at $t = 0$. We can eliminate the complexity introduced by $\boldsymbol{\Theta}$ by instead first differentiating with respect to z and controlling the derivative with respect to t using the second derivative. In the process, the choice of ζ presented in the statement naturally arises and can be shown to be correct using differential calculus. The details can be found in section SM6 of the supplementary material. ∎

That is, a more general version of Theorem 4.1 holds for any $\mathbf{S}$ that has the rotational invariance properties associated with freeness. By the same reasoning as in Remark 4.4, we expect that in the special case of $z \to 0$, free sketches will generally have the exact same first-order properties as the i.i.d. sketching case, since all spectral properties of $\mathbf{S}\mathbf{S}^{\mathsf{H}}$ except the rank (sketch size) become irrelevant. In general, however, the mapping $z \mapsto \zeta$ depends on the spectrum of $\mathbf{S}\mathbf{S}^{\mathsf{H}}$ and is not the same as in the i.i.d. sketching case.

A particularly important sketching matrix that fits this broader definition is the orthogonal sketch. For example, randomized Fourier transforms are orthogonal and asymptotically free [4, 29]. Unlike the i.i.d. sketch, an orthogonal sketch does not distort the spectrum near $q = p$ and so has less implicit regularization. We give proof details in section SM6.

[4] After a preprint of this article was made available, a reader pointed out connections of Theorem 7.2 to the multiplicative subordination result in Theorem 3.6 of [7]. Exploring these connections and possible generalizations of Theorem 7.2 further is left for future work.

[5] Standard zero-order freeness suffices when $p\boldsymbol{\Theta}$ has uniformly bounded operator norm. For general trace norm bounded $\boldsymbol{\Theta}$, first-order (infinitesimal) freeness [48] is also required; see proof details. Unitarily invariant ensembles such as the orthogonal sketches in Corollary 7.3 are known to satisfy all the necessary properties [8].

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

Corollary 7.3 (orthogonal sketching). *For $q \le p$ with $\lim \frac{q}{p} = \alpha$, let $\sqrt{\frac{q}{p}}\mathbf{Q} \in \mathbb{C}^{p\times q}$ be a Haar-distributed matrix with orthonormal columns, and let $\mathbf{A} \in \mathbb{C}^{p\times p}$ be positive semidefinite with eigenvalues converging to a bounded limiting spectral measure. Then, for any $\lambda > 0$,*

$$\mathbf{Q}(\mathbf{Q}^{\mathsf{H}}\mathbf{A}\mathbf{Q} + \lambda\mathbf{I}_q)^{-1}\mathbf{Q}^{\mathsf{H}} \simeq (\mathbf{A} + \gamma\mathbf{I}_p)^{-1},$$

where γ is the most positive solution to

$$\frac{1}{p}\mathrm{tr}\left[(\mathbf{A} + \gamma\mathbf{I}_p)^{-1}\right](\gamma - \alpha\lambda) = 1 - \alpha. \tag{7.1}$$

Furthermore, for μ from Theorem 4.1 applied to the same $(\mathbf{A}, \alpha, \lambda)$, we have $\gamma < \mu$.

Proof. First, $\mathbf{Q}\mathbf{Q}^{\mathsf{H}}$ and $\mathbf{A}$ are almost surely asymptotically free [40, Theorem 4.9]. We can therefore apply Theorem 7.2. It is straightforward to obtain the analytic limiting S-transform $\mathcal{S}_{\mathbf{Q}\mathbf{Q}^{\mathsf{H}}}(w) = \frac{\alpha(1+w)}{\alpha+w}$, from which we can obtain (7.1) from the equation $\gamma = \lambda\mathcal{S}_{\mathbf{Q}\mathbf{Q}^{\mathsf{H}}}(-\frac{1}{p}\mathrm{tr}[\mathbf{A}(\mathbf{A} + \gamma\mathbf{I}_p)^{-1}])$. That is, if we take $z \to -\lambda$, which is a well-defined limit for $\mathrm{Im}(z) \searrow 0$ for any $\lambda > 0$, we have $\zeta \simeq -\gamma$. To see that $\gamma < \mu$, observe that we can write (4.4) and (7.1) as

$$\begin{aligned}\frac{\mu}{p}\mathrm{tr}\left[(\mathbf{A} + \mu\mathbf{I}_p)^{-1}\right] &= 1 - \alpha + \frac{\alpha\lambda}{\mu},\\ \frac{\gamma}{p}\mathrm{tr}\left[(\mathbf{A} + \gamma\mathbf{I}_p)^{-1}\right] &= 1 - \alpha + \alpha\lambda\frac{1}{p}\mathrm{tr}\left[(\mathbf{A} + \gamma\mathbf{I}_p)^{-1}\right].\end{aligned}$$

The left-hand sides of these two equations are the same increasing function of μ and γ, respectively, while the right-hand sides are decreasing functions, with the function of μ being strictly greater than the function of γ, since $\frac{1}{p}\mathrm{tr}[(\mathbf{A} + \mu\mathbf{I}_p)^{-1}] < \frac{1}{\mu}$ for $\mu > 0$. This means that the intersection with the decreasing function for γ must occur for a smaller value than the intersection for μ, proving the claim. ∎

In the statement, $\gamma < \mu$ means that the orthogonal sketch has less effective regularization than the i.i.d. sketch. For settings in which we desire to solve a linear system with as little distortion as possible, we therefore would much prefer an orthogonal sketch to an i.i.d. sketch, especially for $q \approx p$. With additional work, one could extend this result to negative regularization as we have done in the i.i.d. sketching case. We leave it for future work.

In Figure 8, we repeat the experiment from Figure 2 for a variety of normalized non-i.i.d. sketches used frequently in practice. Both CountSketch [9] and the fast Johnson–Lindenstrauss transform (FJLT) [2] behave similarly to i.i.d. sketching, with the FJLT slightly overregularizing. As predicted by Corollary 7.1, adaptive sketching with $\mathbf{R} = \mathbf{A}$ [30] behaves very differently from the other sketches, showing only two point masses instead of three since $\mathbf{A}^{-1}$ is not well-defined for its eigenvalues of 0. Last, the SRHT [51] is an orthogonal version of the FJLT, and our experiment elucidates the effect of zero padding on the Hadamard transform of the SRHT. The fast Hadamard transform is defined only for powers of 2, so for other dimensions, the common approach is to simply zero-pad the data to the nearest power of 2. However, from this experiment we can see that this zero-padding can have a significant impact on the effective regularization; for p slightly smaller than a power of 2, the SRHT performs almost identically to an orthogonal sketch as expected. However, for p slightly larger than a power of 2, there is significant effective regularization induced, even though the sketch is

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

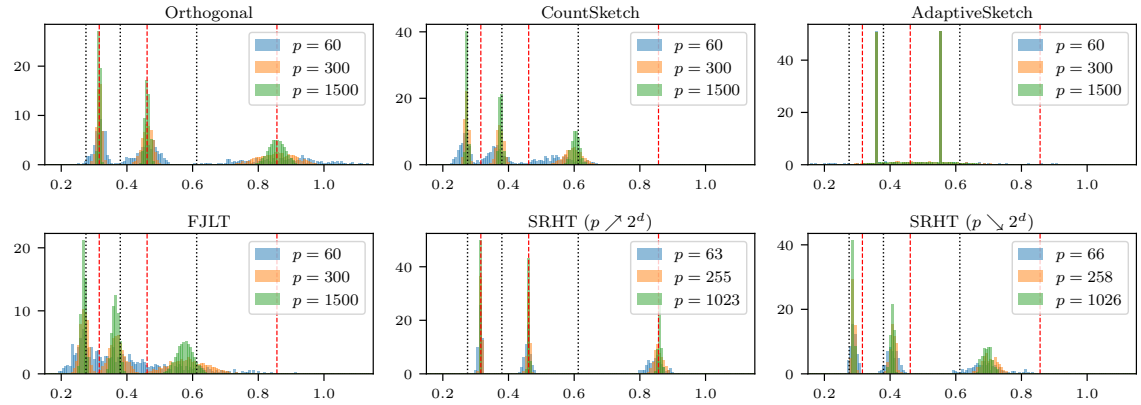

Figure 8. *Empirical density histograms over* 20 *trials demonstrating the concentration of diagonal elements of* $\mathbf{S}(\mathbf{S}^\top \mathbf{A}\mathbf{S} + \lambda \mathbf{I})^{-1}\mathbf{S}^\top$ *for* $\mathbf{A}$ *as in Figure* 2 *with* $q \approx 0.8p$, $\lambda = 1$, *and several normalized sketches* $\mathbf{S}$ *commonly used in practice. We also plot the diagonals of the i.i.d. sketching equivalence* $(\mathbf{A} + \mu \mathbf{I})^{-1}$ *(black, dotted) and the orthogonal sketching equivalence* $(\mathbf{A} + \gamma \mathbf{I})^{-1}$ *from Corollary* 7.3 *(red, dashed), where* $\mu \approx 1.63$ *and* $\gamma \approx 1.17$.

still norm-preserving. This is because zero padding changes the spectrum, so the S-transform deviates from the orthogonal case.

Our proposed framework of first- and second-order equivalence promises to provide a principled means of comparison of different sketching techniques. Once ζ from Theorem 7.2 can be determined for a given sketch (which depends on its spectral properties), an analogous result to Theorem 4.8 will directly follow to yield inflation with a factor of ζ'. Armed with both ζ and ζ' for a collection of sketches, we can compare them using these bias and variance-style decompositions and make principled choices analogously to classical estimation techniques. Our best guidance to practitioners from the insights presented in this work would be to apply a fast sketch with an isotropic spectrum to minimize computation time and distortion, such as the SRHT, but to be aware of issues arising from zero-padding; for this reason we suggest that other Fourier transforms be used instead of the standard fast Hadamard transform.

Future work. As alluded to in the introduction, the first- and second-order equivalences developed in this work can be used directly to analyze the asymptotics of the predicted values and quadratic errors of sketched ridge regression. We leave a complete detailed analysis of sketched ridge regression for a companion paper, in which we use the results in this work to study both primal (observation-side) and dual (feature-side) sketching of the data matrix, as well as joint primal and dual sketching. We believe that our results can also be combined with the techniques in [36], who obtain deterministic equivalents for the Hessian of generalized linear models, enabling precise asymptotics for the implicit regularization due to sketching in nonlinear prediction models such as classification with logistic regression.

Acknowledgments. We are grateful to Arun Kumar Kuchibhotla, Alessandro Rinaldo, Yuting Wei, Jin-Hong Du, and other members of the Operational Overparameterized Statistics (OOPS) Working Group at Carnegie Mellon University for helpful conversations. We are also grateful to Edgar Dobriban, Mert Pilanci, Benson Au, Elad Romanov, and Dimitri Shlyakhtenko, as well as participants of the Deep Learning ONR MURI seminar series for

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

useful discussions and feedback on this work. We thank the anonymous reviewers for their thoughtful suggestions, which have strengthened this work, and the associate editor for the swift review process.

REFERENCES

[1] A. Aghazadeh, R. Spring, D. LeJeune, G. Dasarathy, A. Shrivastava, and R. G. Baraniuk, *MISSION: Ultra large-scale feature selection using count-sketches*, in Proceedings of the 35th International Conference on Machine Learning, Proc. Mach. Learn. Res. 80, 2018, pp. 80–88.

[2] N. Ailon and B. Chazelle, *The fast Johnson–Lindenstrauss transform and approximate nearest neighbors*, SIAM J. Comput., 39 (2009), pp. 302–322, https://doi.org/10.1137/060673096.

[3] J. Alman and V. V. Williams, *A refined laser method and faster matrix multiplication*, in Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, 2021, pp. 522–539, https://doi.org/10.1137/1.9781611976465.32.

[4] G. W. Anderson and B. Farrell, *Asymptotically liberating sequences of random unitary matrices*, Adv. Math., 255 (2014), pp. 381–413, https://doi.org/10.1016/j.aim.2013.12.026.

[5] H. Avron, K. L. Clarkson, and D. P. Woodruff, *Faster kernel ridge regression using sketching and preconditioning*, SIAM J. Matrix Anal. Appl., 38 (2017), pp. 1116–1138, https://doi.org/10.1137/16M1105396.

[6] A. Bakshi, N. Chepurko, and D. P. Woodruff, *Robust and sample optimal algorithms for PSD low rank approximation*, in Proceedings of the 61st Annual Symposium on Foundations of Computer Science, IEEE, 2020, pp. 506–516, https://doi.org/10.1109/FOCS46700.2020.00054.

[7] P. Biane, *Processes with free increments*, Math. Z., 227 (1998), pp. 143–174, https://doi.org/10.1007/PL00004363.

[8] G. Cébron, A. Dahlqvist, and F. Gabriel, *Freeness of Type B and Conditional Freeness for Random Matrices*, preprint, arXiv:2205.01926, 2022.

[9] M. Charikar, K. Chen, and M. Farach-Colton, *Finding frequent items in data streams*, Theoret. Comput. Sci., 312 (2004), pp. 3–15, https://doi.org/10.1016/S0304-3975(03)00400-6.

[10] K. L. Clarkson and D. P. Woodruff, *Sketching for M-estimators: A unified approach to robust regression*, in Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms, 2015, pp. 921–939, https://doi.org/10.1137/1.9781611973730.63.

[11] R. Couillet, M. Debbah, and J. W. Silverstein, *A deterministic equivalent for the analysis of correlated MIMO multiple access channels*, IEEE Trans. Inform. Theory, 57 (2011), pp. 3493–3514, https://doi.org/10.1109/TIT.2011.2133151.

[12] M. Dereziński, B. Bartan, M. Pilanci, and M. W. Mahoney, *Debiasing distributed second order optimization with surrogate sketching and scaled regularization*, in Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 6684–6695.

[13] M. Dereziński, J. Lacotte, M. Pilanci, and M. W. Mahoney, *Newton-LESS: Sparsification without trade-offs for the sketched Newton update*, in Advances in Neural Information Processing Systems, Vol. 34, 2021, pp. 2835–2847.

[14] M. Dereziński, F. T. Liang, Z. Liao, and M. W. Mahoney, *Precise expressions for random projections: Low-rank approximation and randomized Newton*, in Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 18272–18283.

[15] M. Dereziński, F. T. Liang, and M. W. Mahoney, *Exact expressions for double descent and implicit regularization via surrogate random design*, in Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 5152–5164.

[16] M. Dereziński, Z. Liao, E. Dobriban, and M. Mahoney, *Sparse sketches with small inversion bias*, in Proceedings of 34th Conference on Learning Theory, Proc. Mach. Learn. Res. 134, 2021, pp. 1467–1510.

[17] M. Dereziński and M. W. Mahoney, *Determinantal point processes in randomized numerical linear algebra*, Notices Amer. Math. Soc., 68 (2021), pp. 34–45, https://doi.org/10.1090/noti2202.

[18] M. Dereziński and E. Rebrova, *Sharp analysis of sketch-and-project methods via a connection to randomized singular value decomposition*, SIAM J. Math. Data Sci., 6 (2024), pp. 127–153, https://doi.org/10.1137/23M1545537.

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

[19] E. Dobriban, *Efficient computation of limit spectra of sample covariance matrices*, Random Matrices Theory Appl., 4 (2015), 1550019, https://doi.org/10.1142/S2010326315500197.
[20] E. Dobriban and Y. Sheng, *WONDER: Weighted one-shot distributed ridge regression in high dimensions*, J. Mach. Learn. Res., 21 (2020), pp. 1–52.
[21] E. Dobriban and Y. Sheng, *Distributed linear regression by averaging*, Ann. Statist., 49 (2021), pp. 918–943, https://doi.org/10.1214/20-AOS1984.
[22] E. Dobriban and S. Wager, *High-dimensional asymptotics of prediction: Ridge regression and classification*, Ann. Statist., 46 (2018), pp. 247–279, https://doi.org/10.1214/17-AOS1549.
[23] N. El Karoui and H. Kösters, *Geometric Sensitivity of Random Matrix Results: Consequences for Shrinkage Estimators of Covariance and Related Statistical Methods*, preprint, https://arxiv.org/abs/1105.1404, 2011.
[24] R. M. Gower and P. Richtárik, *Randomized iterative methods for linear systems*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 1660–1690, https://doi.org/10.1137/15M1025487.
[25] W. Hachem, P. Loubaton, and J. Najim, *The empirical distribution of the eigenvalues of a Gram matrix with a given variance profile*, Ann. Inst. Henri Poincaré Probab. Stat., 42 (2006), pp. 649–670, https://doi.org/10.1016/j.anihpb.2005.10.001.
[26] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, *Surprises in high-dimensional ridgeless least squares interpolation*, Ann. Statist., 50 (2022), pp. 949–986.
[27] N. Ivkin, D. Rothchild, E. Ullah, V. Braverman, I. Stoica, and R. Arora, *Communication-efficient distributed SGD with sketching*, in Advances in Neural Information Processing Systems, Vol. 32, 2019.
[28] D. Kobak, J. Lomond, and B. Sanchez, *The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization*, J. Mach. Learn. Res., 21 (2020), pp. 1–16.
[29] J. Lacotte, S. Liu, E. Dobriban, and M. Pilanci, *Optimal iterative sketching methods with the subsampled randomized Hadamard transform*, in Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 9725–9735.
[30] J. Lacotte, M. Pilanci, and M. Pavone, *High-dimensional optimization in adaptive random subspaces*, in Advances in Neural Information Processing Systems, Vol. 32, 2019.
[31] E. E. Leamer and G. Chamberlain, *A Bayesian interpretation of pretesting*, J. R. Stat. Soc. Ser. B. Stat. Methodol., 38 (1976), pp. 85–94, https://doi.org/10.1111/j.2517-6161.1976.tb01570.x.
[32] O. Ledoit and S. Péché, *Eigenvectors of some large sample covariance matrix ensembles*, Probab. Theory Related Fields, 151 (2011), pp. 233–264, https://doi.org/10.1007/s00440-010-0298-3.
[33] D. LeJeune, *Ridge Regularization by Randomization in Linear Ensembles*, Ph.D. thesis, Rice University, 2022.
[34] D. LeJeune, H. Javadi, and R. G. Baraniuk, *The implicit regularization of ordinary least squares ensembles*, in Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, Proc. Mach. Learn. Res. 108, 2020, pp. 3525–3535.
[35] D. LeJeune, H. Javadi, and R. G. Baraniuk, *The flip side of the reweighted coin: Duality of adaptive dropout and regularization*, in Advances in Neural Information Processing Systems, Vol. 34, 2021, pp. 23401–23412.
[36] Z. Liao and M. W. Mahoney, *Hessian eigenspectra of more realistic nonlinear models*, in Advances in Neural Information Processing Systems, Vol. 34, 2021, pp. 20104–20117.
[37] S. Liu and E. Dobriban, *Ridge regression: Structure, cross-validation, and sketching*, in Proceedings of the 8th International Conference on Learning Representations, 2020.
[38] M. W. Mahoney, *Randomized algorithms for matrices and data*, Found. Trends Mach. Learn., 3 (2011), pp. 123–224, https://doi.org/10.1561/2200000035.
[39] G. Mel and S. Ganguli, *A theory of high dimensional regression with arbitrary correlations between input features and target functions: Sample complexity, multiple descent curves and a hierarchy of phase transitions*, in Proceedings of the 38th International Conference on Machine Learning, Proc. Mach. Learn. Res. 139, 2021, pp. 7578–7587.
[40] J. Mingo and R. Speicher, *Free Probability and Random Matrices*, Fields Inst. Monogr., Springer New York, 2017, https://doi.org/10.1007/978-1-4939-6942-5.

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license

[41] R. Murray, J. Demmel, M. W. Mahoney, N. B. Erichson, M. Melnichenko, O. A. Malik, L. Grigori, P. Luszczek, M. Dereziński, M. E. Lopes, T. Liang, H. Luo, and J. Dongarra, *Randomized Numerical Linear Algebra: A Perspective on the Field with an Eye to Software*, preprint, arXiv:2302.11474, 2023.

[42] M. Mutny, M. Dereziński, and A. Krause, *Convergence analysis of block coordinate algorithms with determinantal sampling*, in Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, Proc. Mach. Learn. Res. 108, 2020, pp. 3110–3120.

[43] P. Patil, A. Rinaldo, and R. Tibshirani, *Estimating functionals of the out-of-sample error distribution in high-dimensional ridge regression*, in Proceedings of the 25th International Conference on Artificial Intelligence and Statistics, Vol. 151, 2022, pp. 6087–6120.

[44] M. Pilanci and M. J. Wainwright, *Iterative Hessian sketch: Fast and accurate solution approximation for constrained least-squares*, J. Mach. Learn. Res., 17 (2016), pp. 1–38.

[45] F. Rubio and X. Mestre, *Spectral convergence for a general class of random matrices*, Statist. Probab. Lett., 81 (2011), pp. 592–602, https://doi.org/10.1016/j.spl.2011.01.004.

[46] A. Rudi, R. Camoriano, and L. Rosasco, *Less is more: Nyström computational regularization*, in Advances in Neural Information Processing Systems, Vol. 28, 2015.

[47] V. Serdobolskii, *Multivariate Statistical Analysis: A High-dimensional Approach*, Theory Decis. Lib. Ser. B 41, Springer, Dordrecht, 2000, https://doi.org/10.1007/978-94-015-9468-4.

[48] D. Shlyakhtenko, *Free probability of type-B and asymptotics of finite-rank perturbations of random matrices*, Indiana Univ. Math. J., 67 (2018), pp. 971–991, https://doi.org/10.1512/iumj.2018.67.7294.

[49] J. W. Silverstein and S. I. Choi, *Analysis of the limiting spectral distribution of large dimensional random matrices*, J. Multivariate Anal., 54 (1995), pp. 295–309, https://doi.org/10.1006/jmva.1995.1058.

[50] G.-A. Thanei, C. Heinze, and N. Meinshausen, *Random projections for large-scale regression*, in Big and Complex Data Analysis, Contrib. Stat., Springer, Cham, 2017, pp. 51–68, https://doi.org/10.1007/978-3-319-41573-4_3.

[51] J. A. Tropp, *Improved analysis of the subsampled randomized Hadamard transform*, Adv. Adapt. Data Anal., 3 (2011), pp. 115–126, https://doi.org/10.1142/S1793536911000787.

[52] W. N. van Wieringen, *Lecture Notes on Ridge Regression*, preprint, arXiv:1509.09169, 2015.

[53] D. V. Voiculescu, K. J. Dykema, and A. Nica, *Introduction to the Theory of Linear Nonselfadjoint Operators in Hilbert Space*, CRM Monogr. Ser. 1, AMS, Providence, RI, 1992, https://doi.org/10.1090/crmm/001.

[54] J. Wang, J. Lee, M. Mahdavi, M. Kolar, and N. Srebro, *Sketching meets random projection in the dual: A provable recovery algorithm for big and high-dimensional data*, in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Proc. Mach. Learn. Res. 54, 2017, pp. 1150–1158.

[55] D. P. Woodruff, *Sketching as a tool for numerical linear algebra*, Found. Trends Theor. Comput. Sci., 10 (2014), pp. 1–157, https://doi.org/10.1561/0400000060.

[56] D. P. Woodruff, *A very sketchy talk*, in 48th International Colloquium on Automata, Languages, and Programming, LIPIcs Leibniz Int. Proc. Inform. 198, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021, pp. 6:1–6:8, https://doi.org/10.4230/LIPIcs.ICALP.2021.6.

[57] D. Wu and J. Xu, *On the optimal weighted ℓ_2 regularization in overparameterized linear regression*, in Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 10112–10123.

[58] F. F. Yilmaz and R. Heckel, *Regularization-wise double descent: Why it occurs and how to eliminate it*, in Proceedings of the International Symposium on Information Theory, IEEE, 2022, pp. 426–431, https://doi.org/10.1109/ISIT50566.2022.9834569.

Downloaded 10/31/24 to 99.34.180.178 . Redistribution subject to CCBY license

© 2024 SIAM. Published by SIAM under the terms of the Creative Commons 4.0 license