Chapter 1

Exact expectation expressions for sub-Gaussian random projections

It is often desirable to reduce the dimensionality of a large dataset by projecting it onto a low-dimensional subspace. Matrix sketching has emerged as a powerful technique for performing such dimensionality reduction very efficiently. Even though there is an extensive literature on the worst-case performance of sketching, existing guarantees are typically very different from what is observed in practice. Building on the Stieltjes transform methods employed in **??** while establishing asymptotic consistency of surrogate design's MSE, this chapter develops novel techniques that provide provably accurate expressions for the expected value of random projection matrices obtained via sketching. These expressions can be used to characterize the performance of dimensionality reduction in a variety of common machine learning tasks, ranging from low-rank approximation to iterative stochastic optimization. Our results apply to several popular sketching methods, including Gaussian and Rademacher sketches, and they enable precise analysis of these methods in terms of spectral properties of the data. Empirical results show that the expressions we derive reflect the practical performance of these sketching methods, down to lower-order effects and even constant factors. Some of the results here were originally published in **precise-expressions**.

1.1 Introduction

Many settings in modern machine learning, optimization and scientific computing require us to work with data matrices that are so large that some form of dimensionality reduction is a necessary component of the process. One of the most popular families of methods for dimensionality reduction, coming from the literature on Randomized Numerical Linear Algebra (RandNLA), consists of data-oblivious sketches [**Mah-mat-rev_JRNL**; **tropp2011structure**; **woodruff2014sketching**]. Consider a large $m \times n$ matrix $\mathbf{A}$. A *data-oblivious sketch* of size k is the matrix $\mathbf{SA}$, where $\mathbf{S}$ is a $k \times m$ random matrix such that $\mathbb{E}[\frac{1}{k}\mathbf{S}^\top\mathbf{S}] = \mathbf{I}$, whose distribution does not depend on $\mathbf{A}$. This sketch reduces the first

dimension of $\mathbf{A}$ from m to a much smaller k (we assume without loss of generality that $k \ll n \leq m$), and an analogous procedure can be defined for reducing the second dimension as well. This approximate representation of $\mathbf{A}$ is central to many algorithms in areas such as linear regression, low-rank approximation, kernel methods, and iterative second-order optimization. While there is a long line of research aimed at bounding the worst-case approximation error of such representations, these bounds are often too loose to reflect accurately the practical performance of these methods. In this paper, we develop new theory which enables more precise analysis of the accuracy of sketched data representations.

A common way to measure the accuracy of the sketch $\mathbf{SA}$ is by considering the k-dimensional subspace spanned by its rows. The goal of the sketch is to choose a subspace that best aligns with the distribution of all of the m rows of $\mathbf{A}$ in $\mathbb{R}^n$. Intuitively, our goal is to minimize the (norm of the) residual when projecting a vector $\mathbf{a} \in \mathbb{R}^n$ onto that subspace, i.e., $\mathbf{a} - \mathbf{Pa} = (\mathbf{I} - \mathbf{P})\mathbf{a}$, where $\mathbf{P} = (\mathbf{SA})^\dagger\mathbf{SA}$ is the orthogonal projection matrix onto the subspace spanned by the rows of $\mathbf{SA}$ (and $(\cdot)^\dagger$ denotes the Moore-Penrose pseudoinverse). For this reason, the quantity that has appeared ubiquitously in the error analysis of RandNLA sketching is what we call the residual projection matrix:

$$\textbf{(residual projection matrix)} \quad \mathbf{P}_\perp \;:=\; \mathbf{I} - \mathbf{P} \;=\; \mathbf{I} - (\mathbf{SA})^\dagger\mathbf{SA}.$$

Since $\mathbf{P}_\perp$ is random, the average performance of the sketch can often be characterized by its expectation, $\mathbb{E}[\mathbf{P}_\perp]$. For example, the low-rank approximation error of the sketch can be expressed as $\mathbb{E}[\|\mathbf{A} - \mathbf{AP}\|_F^2] = \operatorname{tr}\mathbf{A}^\top\mathbf{A}\,\mathbb{E}[\mathbf{P}_\perp]$, where $\|\cdot\|_F$ denotes the Frobenius norm. A similar formula follows for the trace norm error of a sketched Nyström approximation [**Williams01Nystrom**; **revisiting-nystrom**]. Among others, this approximation error appears in the analysis of sketched kernel ridge regression [**fanuel2020diversity**] and Gaussian process regression [**sparse-variational-gp**]. Furthermore, a variety of iterative algorithms, such as randomized second-order methods for convex optimization [**Qu2015Feb**; **Qu2016**; **Gower2019**; **jacsketch**] and linear system solvers based on the generalized Kaczmarz method [**generalized-kaczmarz**], have convergence guarantees which depend on the extreme eigenvalues of $\mathbb{E}[\mathbf{P}_\perp]$. Finally, a generalized form of the expected residual projection has been recently used to model the implicit regularization of the interpolating solutions in over-parameterized linear models [**surrogate-design**; **BLLT19_TR**].

Main result

Despite its prevalence in the literature, the expected residual projection is not well understood, even in such simple cases as when $\mathbf{S}$ is a Gaussian sketch (i.e., with i.i.d. standard normal entries). We address this by providing a surrogate expression, i.e., a simple analytically tractable approximation, for this matrix quantity:

$$\mathbb{E}[\mathbf{P}_\perp] \;\overset{\epsilon}{\simeq}\; \bar{\mathbf{P}}_\perp := (\gamma\mathbf{A}^\top\mathbf{A} + \mathbf{I})^{-1}, \quad \text{with} \;\; \gamma > 0 \;\; \text{s.t.} \;\; \operatorname{tr}\bar{\mathbf{P}}_\perp = n - k. \tag{1.1}$$

Here, $\overset{\epsilon}{\simeq}$ means that while the surrogate expression is not exact, it approximates the true quantity up to some ϵ accuracy. Our main result provides a rigorous approximation guarantee

for this surrogate expression with respect to a range of sketching matrices $\mathbf{S}$, including the standard Gaussian and Rademacher sketches. We state the result using the positive semi-definite ordering denoted by $\preceq$.

Theorem 1.1 *Let* $\mathbf{S}$ *be a sketch of size* k *with i.i.d. mean-zero sub-gaussian entries and let* $r = \|\mathbf{A}\|_F^2/\|\mathbf{A}\|^2$ *be the stable rank of* $\mathbf{A}$*. If we let* $\rho = r/k$ *be a fixed constant larger than* 1*, then*

$$(1-\epsilon)\,\bar{\mathbf{P}}_\perp \preceq \mathbb{E}[\mathbf{P}_\perp] \preceq (1+\epsilon)\,\bar{\mathbf{P}}_\perp \quad \textit{for} \quad \epsilon = O(\tfrac{1}{\sqrt{r}}).$$

In other words, when the sketch size k is smaller than the stable rank r of $\mathbf{A}$, then the discrepancy between our surrogate expression $\bar{\mathbf{P}}_\perp$ and $\mathbb{E}[\mathbf{P}_\perp]$ is of the order $1/\sqrt{r}$, where the big-O notation hides only the dependence on ρ and on the sub-gaussian constant (see Theorem 1.2 for more details). Our proof of Theorem 1.1 is inspired by the techniques from random matrix theory which have been used to analyze the asymptotic spectral distribution of large random matrices by focusing on the associated matrix resolvents and Stieltjes transforms [**hachem2007deterministic**; **bai2010spectral**]. However, our analysis is novel in several respects:

1. The residual projection matrix can be obtained from the appropriately scaled resolvent matrix $z(\mathbf{A}^\top\mathbf{S}^\top\mathbf{S}\mathbf{A} + z\mathbf{I})^{-1}$ by taking $z \to 0$. Prior work (e.g., [**HMRT19_TR**]) combined this with an exchange-of-limits argument to analyze the asymptotic behavior of the residual projection. This approach, however, does not allow for a precise control in finite-dimensional problems. We are able to provide a more fine-grained, non-asymptotic analysis by working directly with the residual projection itself, instead of the resolvent.

2. We require no assumptions on the largest and smallest singular value of $\mathbf{A}$. Instead, we derive our bounds in terms of the stable rank of $\mathbf{A}$ (as opposed to its actual rank), which implicitly compensates for ill-conditioned data matrices.

3. We obtain upper/lower bounds for $\mathbb{E}[\mathbf{P}_\perp]$ in terms of the positive semi-definite ordering $\preceq$, which can be directly converted to guarantees for the precise expressions of expected low-rank approximation error derived in the following section.

It is worth mentioning that the proposed analysis is significantly different from the sketching literature based on subspace embeddings (e.g., [**sarlos-sketching**; **cw-sparse**; **nn-sparse**; **projection-cost-preserving**; **optimal-matrix-product**]), in the sense that here our object of interest is not to obtain a worst-case approximation with high probability, but rather, our analysis provides *precise* characterization on the *expected* residual projection matrix that goes *beyond worst-case bounds*. From an application perspective, the subspace embedding property is neither sufficient nor necessary for many numerical implementations of sketching [**blendenpik**; **lsrn**], or statistical results [**GarveshMahoney_JMLR**; **dobriban2019asymptotics**; **yang2020reduce**], as well as in the context of iterative optimization and implicit regularization (see Sections 1.1 and 1.1 below), which are discussed in detail as concrete applications of the proposed analysis.

Low-rank approximation

We next provide some immediate corollaries of Theorem 1.1, where we use $x \overset{\epsilon}{\simeq} y$ to denote a multiplicative approximation $|x - y| \leq \epsilon y$. Note that our analysis is new even for the classical Gaussian sketch where the entries of $\mathbf{S}$ are i.i.d. standard normal. However the results apply more broadly, including a standard class of data-base friendly Rademacher sketches where each entry s_{ij} is a ± 1 Rademacher random variable [**achlioptas2003database**]. We start by analyzing the Frobenius norm error $\|\mathbf{A} - \mathbf{AP}\|_F^2 = \operatorname{tr} \mathbf{A}^\top \mathbf{A}\, \mathbf{P}_\perp$ of sketched low-rank approximations. Note that by the definition of γ in (1.1), we have $k = \operatorname{tr}(\mathbf{I} - \bar{\mathbf{P}}_\perp) = \operatorname{tr} \gamma \mathbf{A}^\top \mathbf{A}(\gamma \mathbf{A}^\top \mathbf{A} + \mathbf{I})^{-1}$, so the surrogate expression we obtain for the expected error is remarkably simple.

Corollary 1.1 *Let σ_i be the singular values of $\mathbf{A}$. Under the assumptions of Theorem 1.1, we have:*

$$\mathbb{E}\big[\|\mathbf{A} - \mathbf{AP}\|_F^2\big] \;\overset{\epsilon}{\simeq}\; k/\gamma \quad \textit{for} \;\; \gamma > 0 \;\; \textit{s.t.} \;\; \sum_i \frac{\gamma \sigma_i^2}{\gamma \sigma_i^2 + 1} = k.$$

Remark 1.1 *The parameter $\gamma = \gamma(k)$ increases at least linearly as a function of k, which is why the expected error will always decrease with increasing k. For example, when the singular values of $\mathbf{A}$ exhibit exponential decay, i.e., $\sigma_i^2 = C \cdot \alpha^{i-1}$ for $\alpha \in (0, 1)$, then the error also decreases exponentially, at the rate of $k/(\alpha^{-k} - 1)$. We discuss this further in Section 1.5, giving explicit formulas for the error as a function of k under both exponential and polynomial spectral decay profiles.*

The above result is important for many RandNLA methods, and it is also relevant in the context of kernel methods, where the data is represented via a positive semi-definite $m \times m$ kernel matrix $\mathbf{K}$ which corresponds to the matrix of dot-products of the data vectors in some reproducible kernel Hilbert space. In this context, sketching can be applied directly to the matrix $\mathbf{K}$ via an extended variant of the Nyström method [**revisiting-nystrom**]. A Nyström approximation constructed from a sketching matrix $\mathbf{S}$ is defined as $\tilde{\mathbf{K}} = \mathbf{C}^\top \mathbf{W}^\dagger \mathbf{C}$, where $\mathbf{C} = \mathbf{SK}$ and $\mathbf{W} = \mathbf{SKS}^\top$, and it is applicable to a variety of settings, including Gaussian Process regression, kernel machines and Independent Component Analysis [**sparse-variational-gp**; **Williams01Nystrom**; **Bach2003**]. By setting $\mathbf{A} = \mathbf{K}^{\frac{1}{2}}$, it is easy to see [**nystrom-multiple-descent**] that the trace norm error $\|\mathbf{K} - \tilde{\mathbf{K}}\|_*$ is identical to the squared Frobenius norm error of the low-rank sketch $\mathbf{SA}$, so Corollary 1.1 implies that

$$\mathbb{E}\big[\|\mathbf{K} - \tilde{\mathbf{K}}\|_*\big] \;\overset{\epsilon}{\simeq}\; k/\gamma \quad \text{for} \;\; \gamma > 0 \;\; \text{s.t.} \;\; \sum_i \frac{\gamma \lambda_i}{\gamma \lambda_i + 1} = k, \tag{1.2}$$

with any sub-gaussian sketch, where λ_i denote the eigenvalues of $\mathbf{K}$. Our error analysis given in Section 1.5 is particularly relevant here, since commonly used kernels such as the Radial Basis Function (RBF) or the Matérn kernel induce a well-understood eigenvalue decay [**Santa97gaussianregression**; **RasmussenWilliams06**].

Metrics other than the aforementioned Frobenius norm error, such as the spectral norm error [**tropp2011structure**], are also of significant interest in the low-rank approximation literature. We leave these directions for future investigation.

Randomized iterative optimization

We next turn to a class of iterative methods which take advantage of sketching to reduce the per iteration cost of optimization. These methods have been developed in a variety of settings, from solving linear systems to convex optimization and empirical risk minimization, and in many cases the residual projection matrix appears as a black box quantity whose spectral properties determine the convergence behavior of the algorithms [**generalized-kaczmarz**]. With our new results, we can precisely characterize not only the rate of convergence, but also, in some cases, the complete evolution of the parameter vector, for the following algorithms:

1. *Generalized Kaczmarz method* [**generalized-kaczmarz**] for approximately solving a linear system $\mathbf{Ax} = \mathbf{b}$;
2. *Randomized Subspace Newton* [**Gower2019**], a second order method, where we sketch the Hessian matrix.
3. *Jacobian Sketching* [**jacsketch**], a class of first order methods which use additional information via a weight matrix $\mathbf{W}$ that is sketched at every iteration.

We believe that extensions of our techniques will apply to other algorithms, such as that of [**lacotte2019high**].

We next give a result in the context of linear systems for the generalized Kaczmarz method [**generalized-kaczmarz**], but a similar convergence analysis is given for the methods of [**Gower2019**; **jacsketch**] in Section 1.2.

Corollary 1.2 *Let* $\mathbf{x}^*$ *be the unique solution of* $\mathbf{Ax}^* = \mathbf{b}$ *and consider the iterative algorithm:*

$$\mathbf{x}^{t+1} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{x}^t\|^2 \quad \text{subject to} \quad \mathbf{SAx} = \mathbf{Sb}.$$

Under the assumptions of Theorem 1.1, with γ *defined in* (1.1) *and* $r = \|\mathbf{A}\|_F^2/\|\mathbf{A}\|^2$*, we have:*

$$\mathbb{E}\big[\mathbf{x}^{t+1} - \mathbf{x}^*\big] \overset{\epsilon}{\simeq} (\gamma \mathbf{A}^\top \mathbf{A} + \mathbf{I})^{-1}\, \mathbb{E}\big[\mathbf{x}^t - \mathbf{x}^*\big] \quad \textit{for} \quad \epsilon = O(\tfrac{1}{\sqrt{r}}).$$

The corollary follows from Theorem 1.1 combined with Theorem 4.1 in [**generalized-kaczmarz**]. Note that when $\mathbf{A}^\top\mathbf{A}$ is positive definite then $(\gamma\mathbf{A}^\top\mathbf{A}+\mathbf{I})^{-1} \prec \mathbf{I}$, so the algorithm will converge from any starting point, and the worst-case convergence rate of the above method can be obtained by evaluating the largest eigenvalue of $(\gamma\mathbf{A}^\top\mathbf{A} + \mathbf{I})^{-1}$. However the result itself is much stronger, in that it can be used to describe the (expected) trajectory of the iterates for any starting point $\mathbf{x}^0$. Moreover, when the spectral decay profile of $\mathbf{A}$ is known, then the explicit expressions for γ as a function of k derived in Section 1.5 can be used to characterize the convergence properties of generalized Kaczmarz as well as other methods discussed above.

Implicit regularization

Setting $\mathbf{x}^t = \mathbf{0}$, we can view one step of the iterative method in Corollary 1.2 as finding a minimum norm interpolating solution of an under-determined linear system $(\mathbf{SA}, \mathbf{Sb})$. Recent interest in the generalization capacity of over-parameterized machine learning models has motivated extensive research on the statistical properties of such interpolating solutions [**BLLT19_TR**; **HMRT19_TR**; **surrogate-design**]. In this context, Theorem 1.1 provides new evidence for the implicit regularization conjecture posed by [**surrogate-design**] (see their Theorem 2 and associated discussion), with the amount of regularization equal $\frac{1}{\gamma}$, where γ is implicitly defined in (1.1):

$$\underbrace{\mathbb{E}\Big[\operatorname*{argmin}_{\mathbf{x}} \|\mathbf{x}\|^2 \;\text{ s.t. }\; \mathbf{SAx} = \mathbf{Sb}\Big] - \mathbf{x}^*}_{\text{Bias of sketched minimum norm solution}} \;\overset{\epsilon}{\simeq}\; \underbrace{\operatorname*{argmin}_{\mathbf{x}} \Big\{\|\mathbf{Ax}-\mathbf{b}\|^2 + \tfrac{1}{\gamma}\|\mathbf{x}\|^2\Big\} - \mathbf{x}^*}_{\text{Bias of } l_2\text{-regularized solution}}.$$

While implicit regularization has received attention recently in the context of SGD algorithms for overparameterized machine learning models, it was originally discussed in the context of approximation algorithms more generally [**Mah12**]. Recent work has made precise this notion in the context of RandNLA [**surrogate-design**], and our results here can be viewed in terms of implicit regularization of scalable RandNLA methods.

Related work

A significant body of research has been dedicated to understanding the guarantees for low-rank approximation via sketching, particularly in the context of RandNLA [**DM16_CACM**; **RandNLA_PCMIchapter_chapter**]. This line of work includes i.i.d. row sampling methods [**BoutsidisMD08**; **ridge-leverage-scores**] which preserve the structure of the data, and data-oblivious methods such as Gaussian and Rademacher sketches [**Mah-mat-rev_JRNL**; **tropp2011structure**; **woodruff2014sketching**]. However, all of these results focus on worst-case upper bounds on the approximation error. One exception is a recent line of works on non-i.i.d. row sampling with Determinantal Point Processes (DPP, [**dpps-in-randnla**]). In this case, exact analysis of the low-rank approximation error [**nystrom-multiple-descent**], as well as precise convergence analysis of stochastic second order methods [**randomized-newton**], have been obtained. Remarkably, the expressions they obtain are analogous to (1.1), despite using completely different techniques. However, their analysis is limited only to DPP-based sketches, which are considerably more expensive to construct and thus much less widely used. The connection between DPPs and Gaussian sketches was recently explored by [**surrogate-design**] in the context of analyzing the implicit regularization effect of choosing a minimum norm solution in under-determined linear regression. They conjectured that the expectation formulas obtained for DPPs are a good proxy for the corresponding quantities obtained under a Gaussian distribution. Similar observations were made by [**debiasing-second-order**] in the context of sketching for regularized least squares and second order optimization. While both of these works only provide empirical evidence for

this particular claim, our Theorem 1.1 can be viewed as the first theoretical non-asymptotic justification of that conjecture.

The effectiveness of sketching has also been extensively studied in the context of second order optimization. These methods differ depending on how the sketch is applied to the Hessian matrix, and whether or not it is applied to the gradient as well. The class of methods discussed in Section 1.1, including Randomized Subspace Newton and the Generalized Kaczmarz method, relies on projecting the Hessian downto a low-dimensional subspace, which makes our results directly applicable. A related family of methods uses the so-called Iterative Hessian Sketch (IHS) approach [**pilanci2016iterative**; **lacotte2019faster**]. The similarities between IHS and the Subspace Newton-type methods (see [**Qu2015Feb**] for a comparison) suggest that our techniques could be extended to provide precise convergence guarantees also to the IHS. Finally, yet another family of Hessian sketching methods has been studied by [**roosta2019sub**; **sketched-ridge-regression**; **XRM17_theory_TR**; **YXRM18_TR**; **fred_newtonMR_TR**; **distributed-newton**; **determinantal-averaging**]. These methods preserve the rank of the Hessian, and so their convergence guarantees do not rely on the residual projection.

1.2 Convergence analysis of randomized iterative methods

Here, we discuss how our surrogate expressions for the expected residual projection can be used to perform convergence analysis for several randomized iterative optimization methods discussed in Section 1.1.

Generalized Kaczmarz method

Generalized Kaczmarz [**generalized-kaczmarz**] is an iterative method for solving an $m \times n$ linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$, which uses a $k \times m$ sketching matrix $\mathbf{S}_t$ to reduce the linear system and update an iterate $\mathbf{x}^t$ as follows:

$$\mathbf{x}^{t+1} = \operatorname*{argmin}_{\mathbf{x}} \|\mathbf{x} - \mathbf{x}^t\|^2 \quad \text{subject to} \quad \mathbf{S}_t\mathbf{A}\mathbf{x} = \mathbf{S}_t\mathbf{b}.$$

Assume that $\mathbf{x}^*$ is the unique solution to the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$. In Theorems 4.1 and 4.6, [**generalized-kaczmarz**] show that the expected trajectory of the generalized Kaczmarz iterates, as they converge to $\mathbf{x}^*$, is controlled by the projection matrix $\mathbf{P} = (\mathbf{S}_t\mathbf{A})^\dagger\mathbf{S}_t\mathbf{A}$ as follows:

([**generalized-kaczmarz**], Theorem 4.1) $\qquad \mathbb{E}[\mathbf{x}^{t+1} - \mathbf{x}^*] = \big(\mathbf{I} - \mathbb{E}[\mathbf{P}]\big)\,\mathbb{E}[\mathbf{x}^t - \mathbf{x}^*],$

([**generalized-kaczmarz**], Theorem 4.6) $\qquad \mathbb{E}\big[\|\mathbf{x}^{t+1} - \mathbf{x}^*\|^2\big] \leq (1-\kappa)\,\mathbb{E}\big[\|\mathbf{x}^t - \mathbf{x}^*\|^2\big],$ where $\kappa = \lambda_{\min}\big(\mathbb{E}$

Both of these results depend on the expected projection $\mathbb{E}[\mathbf{P}]$. The first one describes the expected trajectory of the iterate, whereas the second one gives the worst-case convergence

rate in terms of the so-called *stochastic condition number* κ. We next demonstrate how Theorem 1.1 can be used in combination with the above results to obtain convergence analysis for generalized Kaczmarz which is formulated in terms of the spectral properties of $\mathbf{A}$. This includes precise expressions for both the expected trajectory and κ. The following result is a more detailed version of Corollary 1.2 from Section 1.1.

Corollary 1.3 *Let σ_i denote the singular values of $\mathbf{A}$, and let k denote the size of sketch $\mathbf{S}_t$. Define:*

$$\Delta_t = \mathbf{x}^t - \mathbf{x}^* \quad \textit{and} \quad \bar{\Delta}_{t+1} = (\gamma \mathbf{A}^\top \mathbf{A} + \mathbf{I})^{-1} \mathbb{E}[\Delta_t] \quad \textit{s.t.} \quad \sum_i \frac{\gamma \sigma_i^2}{\gamma \sigma_i^2 + 1} = k.$$

Suppose that $\mathbf{S}_t$ has i.i.d. mean-zero sub-gaussian entries and let $r = \|\mathbf{A}\|_F^2 / \|\mathbf{A}\|^2$ be the stable rank of $\mathbf{A}$. Assume that $\rho = r/k$ is a constant larger than 1. Then, the expected trajectory satisfies:

$$\big\|\mathbb{E}[\Delta_{t+1}] - \bar{\Delta}_{t+1}\big\| \leq \epsilon \cdot \|\bar{\Delta}_{t+1}\|, \quad \textit{for} \quad \epsilon = O\big(\tfrac{1}{\sqrt{r}}\big). \tag{1.3}$$

Moreover, we obtain the following worst-case convergence guarantee:

$$\mathbb{E}\big[\|\Delta_{t+1}\|^2\big] \leq \big(1 - (\bar{\kappa} - \epsilon)\big)\, \mathbb{E}\big[\|\Delta_t\|^2\big], \quad \textit{where} \quad \bar{\kappa} = \frac{\sigma_{\min}^2}{\sigma_{\min}^2 + 1/\gamma}. \tag{1.4}$$

Remark 1.2 *Our worst-case convergence guarantee* (1.4) *requires the matrix $\mathbf{A}$ to be sufficiently well-conditioned so that $\bar{\kappa} - \epsilon > 0$. However, we believe that our surrogate expression $\bar{\kappa}$ for the stochastic condition number is far more accurate than suggested by the current analysis.*

Proof of Corollary 1.3 Using Theorem 1.1, for $\bar{\mathbf{P}}_\perp$ as defined in (1.1), we have

$$(1-\epsilon)\bar{\mathbf{P}}_\perp \preceq \mathbf{I} - \mathbb{E}[\mathbf{P}] = \mathbb{E}[\mathbf{P}_\perp] \preceq (1+\epsilon)\bar{\mathbf{P}}_\perp, \quad \text{where} \quad \epsilon = O\big(\tfrac{1}{\sqrt{r}}\big).$$

In particular, this implies that $\|\bar{\mathbf{P}}_\perp^{-\frac{1}{2}}(\mathbb{E}[\mathbf{P}_\perp] - \bar{\mathbf{P}}_\perp)\bar{\mathbf{P}}_\perp^{-\frac{1}{2}}\| \leq \epsilon$. Moreover, in the proof of Theorem 1.2 we showed that $\frac{\rho-1}{\rho}\mathbf{I} \preceq \bar{\mathbf{P}}_\perp \preceq \mathbf{I}$, see (1.6), so it follows that:

$$\bar{\mathbf{P}}_\perp^{-1}(\mathbb{E}[\mathbf{P}_\perp] - \bar{\mathbf{P}}_\perp)^2\bar{\mathbf{P}}_\perp^{-1} \preceq \frac{\rho}{\rho-1}\big(\bar{\mathbf{P}}_\perp^{-\frac{1}{2}}(\mathbb{E}[\mathbf{P}_\perp] - \bar{\mathbf{P}}_\perp)\bar{\mathbf{P}}_\perp^{-\frac{1}{2}}\big)^2 \preceq \frac{\rho}{\rho-1}\,\epsilon^2 \cdot \mathbf{I},$$

where note that $\frac{\rho}{\rho-1}\,\epsilon^2 = O(1/r)$, since ρ is treated as a constant. Thus we conclude that:

$$\begin{aligned}\|\mathbb{E}[\Delta_{t+1}] - \bar{\Delta}_{t+1}\|^2 &= \mathbb{E}[\Delta_t]^\top(\mathbb{E}[\mathbf{P}_\perp] - \bar{\mathbf{P}}_\perp)^2\mathbb{E}[\Delta_t] \\ &\leq O(1/r) \cdot \mathbb{E}[\Delta_t]^\top \bar{\mathbf{P}}_\perp^2 \mathbb{E}[\Delta_t] = O(1/r) \cdot \|\bar{\Delta}_{t+1}\|^2,\end{aligned}$$

which completes the proof of (1.3). To show (1.4), it suffices to observe that

$$\lambda_{\min}(\mathbb{E}[\mathbf{P}]) = 1 - \lambda_{\max}(\mathbb{E}[\mathbf{P}_\perp]) \geq 1 - (1+\epsilon)\lambda_{\max}(\bar{\mathbf{P}}_\perp) \geq \lambda_{\min}(\mathbf{I} - \bar{\mathbf{P}}_\perp) - \epsilon,$$

which completes the proof since $\mathbf{I} - \bar{\mathbf{P}}_\perp = \gamma \mathbf{A}^\top \mathbf{A}(\gamma \mathbf{A}^\top \mathbf{A} + \mathbf{I})^{-1}$. ■ Corollaries 1.4 and 1.5 follow analogously from Theorem 1.1.

Randomized Subspace Newton

Randomized Subspace Newton (RSN, [**Gower2019**]) is a randomized Newton-type method for minimizing a smooth, convex and twice differentiable function $f : \mathbb{R}^d \times \mathbb{R}$. The iterative update for this algorithm is defined as follows:

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \frac{1}{L}\mathbf{S}_t^\top(\mathbf{S}_t\mathbf{H}(\mathbf{x}^t)\mathbf{S}_t^\top)^\dagger\mathbf{S}_t\mathbf{g}(\mathbf{x}^t),$$

where $\mathbf{H}(\mathbf{x}^t)$ and $\mathbf{g}(\mathbf{x}^t)$ are the Hessian and gradient of f at $\mathbf{x}^t$, respectively, whereas $\mathbf{S}_t$ is a $k \times d$ sketching matrix (with $k \ll d$) which is refreshed at every iteration. Here, L denotes the *relative smoothness* constant defined by [**Gower2019**] in Assumption 1, which also defines relative strong convexity, denoted by μ. In Theorem 2, they prove the following convergence guarantee for RSN:

$$\mathbb{E}[f(\mathbf{x}^t)] - f(\mathbf{x}^*) \leq \left(1 - \kappa\frac{\mu}{L}\right)^t (f(\mathbf{x}^0) - f(\mathbf{x}^*)),$$

where $\kappa = \min_{\mathbf{x}} \kappa(\mathbf{x})$ and $\kappa(\mathbf{x}) = \lambda_{\min}^+(\mathbb{E}[\mathbf{P}(\mathbf{x})])$ is the smallest positive eigenvalue of the expectation of the projection matrix $\mathbf{P}(\mathbf{x}) = \mathbf{H}^{\frac{1}{2}}(\mathbf{x})\mathbf{S}_t^\top(\mathbf{S}_t\mathbf{H}(\mathbf{x})\mathbf{S}_t^\top)^\dagger\mathbf{S}_t\mathbf{H}^{\frac{1}{2}}(\mathbf{x})$. Our results lead to the following surrogate expression for this expected projection when the sketch is sub-gaussian:

$$\mathbb{E}[\mathbf{P}(\mathbf{x})] \simeq \mathbf{H}(\mathbf{x})\big(\mathbf{H}(\mathbf{x}) + \tfrac{1}{\gamma(\mathbf{x})}\mathbf{I}\big)^{-1} \quad \text{for} \quad \gamma(\mathbf{x}) > 0 \quad \text{s.t.} \quad \operatorname{tr}\mathbf{H}(\mathbf{x})\big(\mathbf{H}(\mathbf{x}) + \tfrac{1}{\gamma(\mathbf{x})}\mathbf{I}\big)^{-1} = k.$$

Thus, the condition number κ of RSN can be estimated using the following surrogate expression:

$$\kappa \simeq \bar{\kappa} := \min_{\mathbf{x}} \frac{\lambda_{\min}^+(\mathbf{H}(\mathbf{x}))}{\lambda_{\min}^+(\mathbf{H}(\mathbf{x})) + 1/\gamma(\mathbf{x})}.$$

Just as in Corollary 1.3, an approximation of the form $|\bar{\kappa} - \kappa| \leq \epsilon$ can be shown from Theorem 1.1.

Corollary 1.4 *Suppose that sketch $\mathbf{S}_t$ has size k and i.i.d. mean-zero sub-gaussian entries. Let $r = \min_{\mathbf{x}} \operatorname{tr}\mathbf{H}(\mathbf{x})/\|\mathbf{H}(\mathbf{x})\|$ be the (minimum) stable rank of the (square root) Hessian and assume that $\rho = r/k$ is a constant larger than 1. Then,*

$$|\kappa - \bar{\kappa}| \leq O\big(\tfrac{1}{\sqrt{r}}\big).$$

Jacobian Sketching

Jacobian Sketching (JacSketch, [**jacsketch**]) defines an $n \times n$ positive semi-definite weight matrix $\mathbf{W}$, and combines it with an $k \times n$ sketching matrix $\mathbf{S}$ (which is refreshed at every iteration of the algorithm), to implicitly construct the following projection matrix:

$$\Pi_{\mathbf{S}} = \mathbf{S}^\top(\mathbf{S}\mathbf{W}\mathbf{S}^\top)^\dagger\mathbf{S}\mathbf{W},$$

which is used to sketch the Jacobian at the current iterate (for the complete method, we refer to their Algorithm 1). The convergence rate guarantee given in their Theorem 3.6 for JacSketch is given in terms of the Lyapunov function:

$$\Psi^t = \|\mathbf{x}^t - \mathbf{x}^*\|^2 + \frac{\alpha}{2\mathcal{L}_2}\|\mathbf{J}^t - \nabla F(\mathbf{x}^*)\|^2_{\mathbf{W}^{-1}},$$

where α is the step size used by the algorithm. Under appropriate choice of the step-size, Theorem 3.6 states that:

$$\mathbb{E}[\Psi^t] \leq \left(1 - \mu \min\left\{\frac{1}{4\mathcal{L}_1}, \frac{\kappa}{4\mathcal{L}_2\rho/n^2 + \mu}\right\}\right)^t \cdot \Psi^0,$$

where $\kappa = \lambda_{\min}(\mathbb{E}[\Pi_{\mathbf{S}}])$ is the *stochastic condition number* analogous to the one defined for the Generalized Kaczmarz method, n is the data size and parameters ρ, $\mathcal{L}_1$, $\mathcal{L}_2$ and μ are problem dependent constants defined in Theorem 3.6. Similarly as before, we can use our surrogate expressions for the expected residual projection to obtain a precise estimate for the stochastic condition number κ under sub-gaussian sketching:

$$\kappa \simeq \bar{\kappa} := \frac{\lambda_{\min}(\mathbf{W})}{\lambda_{\min}(\mathbf{W}) + 1/\gamma} \quad \text{for} \quad \gamma > 0 \quad \text{s.t.} \quad \operatorname{tr}\mathbf{W}(\mathbf{W} + \tfrac{1}{\gamma}\mathbf{I})^{-1} = k.$$

Corollary 1.5 *Suppose $\mathbf{S}_t$ has size k and i.i.d. mean-zero sub-gaussian entries. Let $r = \operatorname{tr}\mathbf{W}/\|\mathbf{W}\|$ be the stable rank of $\mathbf{W}^{\frac{1}{2}}$ and assume that $\rho = r/k$ is a constant larger than 1. Then,*

$$|\kappa - \bar{\kappa}| \leq O\big(\tfrac{1}{\sqrt{r}}\big).$$

1.3 Precise analysis of the residual projection

In this section, we give a detailed statement of our main technical result, along with a sketch of the proof. First, recall the definition of sub-gaussian random variables and vectors.

Definition 1.1 *We say that x is a K-sub-gaussian random variable if its sub-gaussian Orlicz norm $\|x\|_{\psi_2} \leq K$, where $\|x\|_{\psi_2} := \inf\{t > 0 : \ \mathbb{E}[\exp(x^2/t^2)] \leq 2\}$. Similarly, we say that a random vector $\mathbf{x}$ is K-sub-gaussian if for all $\|\mathbf{a}\| \leq 1$ we have $\|\mathbf{x}^\top\mathbf{a}\|_{\psi_2} \leq K$.*

For convenience, we state the main result in a slightly different form than Theorem 1.1. Namely, we replace the $m \times n$ matrix $\mathbf{A}$ with a positive semi-definite $n \times n$ matrix $\mathbf{\Sigma}^{\frac{1}{2}}$. Furthermore, instead of a sketch $\mathbf{S}$ with i.i.d. sub-gaussian entries, we use a random matrix $\mathbf{Z}$ with i.i.d. sub-gaussian *rows*, which is a strictly weaker condition because it allows for the entries of each row to be correlated. Since the rows of $\mathbf{Z}$ are also assumed to have mean zero and identity covariance, each row of $\mathbf{Z}\mathbf{\Sigma}^{\frac{1}{2}}$ has covariance $\mathbf{\Sigma}$. In Section 1.3 we show how to convert this statement back to the form of Theorem 1.1.

Theorem 1.2 *Let* $\mathbf{P}_\perp = \mathbf{I} - \mathbf{X}^\dagger\mathbf{X}$ *for* $\mathbf{X} = \mathbf{Z}\boldsymbol{\Sigma}^{\frac{1}{2}}$*, where* $\mathbf{Z} \in \mathbb{R}^{k\times n}$ *has i.i.d. K-sub-gaussian rows with zero mean and identity covariance, and* $\boldsymbol{\Sigma}$ *is an* $n \times n$ *positive semi-definite matrix. Define:*

$$\bar{\mathbf{P}}_\perp = (\gamma\boldsymbol{\Sigma} + \mathbf{I})^{-1}, \quad \textit{such that} \quad \operatorname{tr}\bar{\mathbf{P}}_\perp = n - k.$$

Let $r = \operatorname{tr}(\boldsymbol{\Sigma})/\|\boldsymbol{\Sigma}\|$ *be the stable rank of* $\boldsymbol{\Sigma}^{\frac{1}{2}}$ *and fix* $\rho = r/k > 1$*. There exists a constant* $C_\rho > 0$*, depending only on* ρ *and* K*, such that if* $r \geq C_\rho$*, then*

$$\Big(1 - \frac{C_\rho}{\sqrt{r}}\Big) \cdot \bar{\mathbf{P}}_\perp \preceq \mathbb{E}[\mathbf{P}_\perp] \preceq \Big(1 + \frac{C_\rho}{\sqrt{r}}\Big) \cdot \bar{\mathbf{P}}_\perp. \tag{1.5}$$

We first provide the following informal derivation of the expression for $\bar{\mathbf{P}}_\perp$ given in Theorem 1.2. Let us use $\mathbf{P}$ to denote the matrix $\mathbf{X}^\dagger\mathbf{X} = \mathbf{I} - \mathbf{P}_\perp$. Using a rank-one update formula for the Moore-Penrose pseudoinverse (see Lemma 1.1 in the appendix) we have

$$\mathbf{I} - \mathbb{E}[\mathbf{P}_\perp] = \mathbb{E}[\mathbf{P}] = \mathbb{E}\big[(\mathbf{X}^\top\mathbf{X})^\dagger\mathbf{X}^\top\mathbf{X}\big] = \sum_{i=1}^{k} \mathbb{E}[(\mathbf{X}^\top\mathbf{X})^\dagger\mathbf{x}_i\mathbf{x}_i^\top] = k\,\mathbb{E}\left[\frac{(\mathbf{I} - \mathbf{P}_{-k})\mathbf{x}_k\mathbf{x}_k^\top}{\mathbf{x}_k^\top(\mathbf{I} - \mathbf{P}_{-k})\mathbf{x}_k}\right],$$

where we use $\mathbf{x}_i^\top$ to denote the i-th row of $\mathbf{X}$, and $\mathbf{P}_{-k} = \mathbf{X}_{-k}^\dagger\mathbf{X}_{-k}$, where $\mathbf{X}_{-i}$ is the matrix $\mathbf{X}$ without its i-th row. Due to the sub-gaussianity of $\mathbf{x}_k$, the quadratic form $\mathbf{x}_k^\top(\mathbf{I} - \mathbf{P}_{-k})\mathbf{x}_k$ in the denominator concentrates around its expectation (with respect to $\mathbf{x}_k$), i.e., $\operatorname{tr}\boldsymbol{\Sigma}(\mathbf{I} - \mathbf{P}_{-k})$, where we use $\mathbb{E}[\mathbf{x}_k\mathbf{x}_k^\top] = \boldsymbol{\Sigma}$. Further note that, with $\mathbf{P}_{-k} \simeq \mathbf{P}$ for large k and $\frac{1}{k}\operatorname{tr}\boldsymbol{\Sigma}(\mathbf{I} - \mathbf{P}_{-k}) \simeq \frac{1}{k}\operatorname{tr}\boldsymbol{\Sigma}\mathbb{E}[\mathbf{P}_\perp]$ from a concentration argument, we conclude that

$$\mathbf{I} - \mathbb{E}[\mathbf{P}_\perp] \simeq \frac{k\mathbb{E}[\mathbf{P}_\perp]\boldsymbol{\Sigma}}{\operatorname{tr}\boldsymbol{\Sigma}\mathbb{E}[\mathbf{P}_\perp]} \qquad \Longrightarrow \qquad \mathbb{E}[\mathbf{P}_\perp] \simeq \Big(\frac{k\boldsymbol{\Sigma}}{\operatorname{tr}\boldsymbol{\Sigma}\mathbb{E}[\mathbf{P}_\perp]} + \mathbf{I}\Big)^{-1},$$

and thus $\mathbb{E}[\mathbf{P}_\perp] \simeq \bar{\mathbf{P}}_\perp$ for $\bar{\mathbf{P}}_\perp = (\gamma\boldsymbol{\Sigma} + \mathbf{I})^{-1}$ and $\gamma^{-1} = \frac{1}{k}\operatorname{tr}\boldsymbol{\Sigma}\bar{\mathbf{P}}_\perp$. This leads to the (implicit) expression for $\bar{\mathbf{P}}_\perp$ and γ given in Theorem 1.2.

Proof sketch of Theorem 1.2

To make the above intuition rigorous, we next present a proof sketch for Theorem 1.2, with the detailed proof deferred to Appendix 1.4. The proof can be divided into the following three steps.

Step 1. First note that, to obtain the lower and upper bound for $\mathbb{E}[\mathbf{P}_\perp]$ in the sense of symmetric matrix as in Theorem 1.2, it suffices to bound the spectral norm $\|\mathbf{I} - \mathbb{E}[\mathbf{P}_\perp]\bar{\mathbf{P}}_\perp^{-1}\| \leq \frac{C_\rho}{\sqrt{r}}$, so that, with $\frac{\rho-1}{\rho}\mathbf{I} \preceq \bar{\mathbf{P}}_\perp \preceq \mathbf{I}$ for $\rho = r/k > 1$ from the definition of $\bar{\mathbf{P}}_\perp$, we have

$$\|\mathbf{I} - \bar{\mathbf{P}}_\perp^{-\frac{1}{2}}\mathbb{E}[\mathbf{P}_\perp]\bar{\mathbf{P}}_\perp^{-\frac{1}{2}}\| = \|\bar{\mathbf{P}}_\perp^{-\frac{1}{2}}(\mathbf{I} - \mathbb{E}[\mathbf{P}_\perp]\bar{\mathbf{P}}^{-1})\bar{\mathbf{P}}_\perp^{\frac{1}{2}}\| \leq \frac{C_\rho}{\sqrt{r}}\sqrt{\frac{\rho}{\rho-1}} =: \epsilon.$$

This means that all eigenvalues of the p.s.d. matrix $\bar{\mathbf{P}}_\perp^{-\frac{1}{2}}\mathbb{E}[\mathbf{P}_\perp]\bar{\mathbf{P}}_\perp^{-\frac{1}{2}}$ lie in the interval $[1-\epsilon, 1+\epsilon]$, so $(1-\epsilon)\mathbf{I} \preceq \bar{\mathbf{P}}_\perp^{-\frac{1}{2}}\mathbb{E}[\mathbf{P}_\perp]\bar{\mathbf{P}}_\perp^{-\frac{1}{2}} \preceq (1+\epsilon)\mathbf{I}$. Multiplying by $\bar{\mathbf{P}}_\perp^{\frac{1}{2}}$ from both sides, we obtain the desired bound.

Step 2. Then, we carefully design an event E that (i) is provable to occur with high probability and (ii) ensures that the denominators in the following decomposition are bounded away from zero:

$$\begin{aligned}\mathbf{I} - \mathbb{E}[\mathbf{P}_\perp]\bar{\mathbf{P}}_\perp^{-1} &= \mathbb{E}[\mathbf{P}] - \gamma\mathbb{E}[\mathbf{P}_\perp]\boldsymbol{\Sigma} = \mathbb{E}[\mathbf{P}\cdot\mathbf{1}_E] + \mathbb{E}[\mathbf{P}\cdot\mathbf{1}_{\neg E}] - \gamma\mathbb{E}[\mathbf{P}_\perp]\boldsymbol{\Sigma} \\ &= \gamma\,\underbrace{\mathbb{E}\left[(\bar{s}-\hat{s})\cdot\frac{(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k\mathbf{x}_k^\top}{\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k}\cdot\mathbf{1}_E\right]}_{\mathbf{T}_1} - \gamma\underbrace{\mathbb{E}[(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k\mathbf{x}_k^\top\cdot\mathbf{1}_{\neg E}]}_{\mathbf{T}_2} \\ &\quad + \gamma\underbrace{\mathbb{E}[\mathbf{P}-\mathbf{P}_{-k}]\boldsymbol{\Sigma}}_{\mathbf{T}_3} + \underbrace{\mathbb{E}[\mathbf{P}\cdot\mathbf{1}_{\neg E}]}_{\mathbf{T}_4},\end{aligned}$$

where we let $\hat{s} = \mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k$ and $\bar{s} = k/\gamma$.

Step 3. It then remains to bound the spectral norms of $\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3, \mathbf{T}_4$ respectively to reach the conclusion. More precisely, the terms $\|\mathbf{T}_2\|$ and $\|\mathbf{T}_4\|$ are proportional to $\Pr(\neg E)$, while the term $\|\mathbf{T}_3\|$ can be bounded using the rank-one update formula for the pseudoinverse (Lemma 1.1 in the appendix). The remaining term $\|\mathbf{T}_1\|$ is more subtle and can be bounded with a careful application of the Hanson-Wright type [**rudelson2013hanson**] sub-gaussian concentration inequalities (Lemmas 1.2 and 1.3 in the appendix). This allows for a bound on the operator norm $\|\mathbf{I} - \mathbb{E}[\mathbf{P}_\perp]\bar{\mathbf{P}}_\perp^{-1}\|$ and hence the conclusion.

Proof of Theorem 1.1

We now discuss how Theorem 1.1 can be obtained from Theorem 1.2. The crucial difference between the statements is that in Theorem 1.1 we let $\mathbf{A}$ be an arbitrary rectangular matrix, whereas in Theorem 1.2 we instead use a square, symmetric and positive semi-definite matrix $\boldsymbol{\Sigma}$. To convert between the two notations, consider the SVD decomposition $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ of $\mathbf{A} \in \mathbb{R}^{m\times n}$ (recall that we assume $m \geq n$), where $\mathbf{U} \in \mathbb{R}^{m\times n}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$ have orthonormal columns and $\mathbf{D}$ is a diagonal matrix. Now, let $\mathbf{Z} = \mathbf{S}\mathbf{U}$, $\boldsymbol{\Sigma} = \mathbf{D}^2$ and $\mathbf{X} = \mathbf{Z}\boldsymbol{\Sigma}^{\frac{1}{2}} = \mathbf{S}\mathbf{U}\mathbf{D}$. Using the fact that $\mathbf{V}^\top\mathbf{V} = \mathbf{V}\mathbf{V}^\top = \mathbf{I}$, it follows that:

$$\mathbf{I} - (\mathbf{S}\mathbf{A})^\dagger\mathbf{S}\mathbf{A} = \mathbf{V}(\mathbf{I}-\mathbf{X}^\dagger\mathbf{X})\mathbf{V}^\top \quad \text{and} \quad (\gamma\mathbf{A}^\top\mathbf{A}+\mathbf{I})^{-1} = \mathbf{V}(\gamma\boldsymbol{\Sigma}+\mathbf{I})^{-1}\mathbf{V}^\top.$$

Note that since $\|\mathbf{U}\mathbf{v}\| = \|\mathbf{v}\|$, the rows of $\mathbf{Z}$ are sub-gaussian with the same constant as the rows of $\mathbf{S}$. Moreover, using the fact that $\mathbf{B} \preceq \mathbf{C}$ implies $\mathbf{V}\mathbf{B}\mathbf{V}^\top \preceq \mathbf{V}\mathbf{C}\mathbf{V}^\top$ for any p.s.d. matrices $\mathbf{B}$ and $\mathbf{C}$, Theorem 1.1 follows as a corollary of Theorem 1.2.

1.4 Proof of Theorem 1.2

We first introduce the following technical lemmas.

Lemma 1.1 *For $\mathbf{X} \in \mathbb{R}^{k\times n}$ with $k < n$, denote $\mathbf{P} = \mathbf{X}^\dagger\mathbf{X}$ and $\mathbf{P}_{-k} = \mathbf{X}_{-k}^\dagger\mathbf{X}_{-k}$, with $\mathbf{X}_{-i} \in \mathbb{R}^{(k-1)\times n}$ the matrix $\mathbf{X}$ without its i-th row $\mathbf{x}_i \in \mathbb{R}^n$. Then, conditioned on the event $E_k : \left\{\left|\frac{\mathrm{tr}\boldsymbol{\Sigma}(\mathbf{I}-\mathbf{P}_{-k})}{\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k} - 1\right| \le \frac{1}{2}\right\}$:*

$$(\mathbf{X}^\top\mathbf{X})^\dagger\mathbf{x}_k = \frac{(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k}{\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k}, \quad \mathbf{P}-\mathbf{P}_{-k} = \frac{(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})}{\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k}.$$

Proof Since conditioned on E_k we have $\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k \neq 0$, from [**10.2307/2099767**] we deduce

$$\begin{aligned}(\mathbf{X}^\top\mathbf{X})^\dagger = (\mathbf{A}+\mathbf{x}_k\mathbf{x}_k^\top)^\dagger = \mathbf{A}^\dagger - \frac{\mathbf{A}^\dagger\mathbf{x}_k\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})}{\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k} - \frac{(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k\mathbf{x}_k^\top\mathbf{A}^\dagger}{\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k} \\ + (1+\mathbf{x}_k^\top\mathbf{A}^\dagger\mathbf{x}_k)\frac{(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})}{(\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k)^2}\end{aligned}$$

for $\mathbf{A} = \mathbf{X}_{-k}^\top\mathbf{X}_{-k}$ so that $\mathbf{I}-\mathbf{P}_{-k} = \mathbf{I}-\mathbf{A}^\dagger\mathbf{A}$, where we used the fact that $\mathbf{I}-\mathbf{P}_{-k}$ is a projection matrix so that $(\mathbf{I}-\mathbf{P}_{-k})^2 = \mathbf{I}-\mathbf{P}_{-k}$. As a consequence, multiplying by $\mathbf{x}_k$ and simplifying we get

$$(\mathbf{X}^\top\mathbf{X})^\dagger\mathbf{x}_k = \frac{(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k}{\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k}.$$

By definition of the pseudoinverse, $\mathbf{P} = \mathbf{X}^\dagger\mathbf{X} = (\mathbf{X}^\top\mathbf{X})^\dagger\mathbf{X}^\top\mathbf{X}$ so that

$$\mathbf{P}-\mathbf{P}_{-k} = \mathbf{X}^\dagger\mathbf{X} - \mathbf{X}_{-k}^\dagger\mathbf{X}_{-k} = \frac{(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})}{\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k}$$

where we used $\mathbf{A}(\mathbf{I}-\mathbf{P}_{-k}) = \mathbf{A}-\mathbf{A}\mathbf{A}^\dagger\mathbf{A} = 0$ and thus the conclusion. ■

Lemma 1.2 *For a K-sub-gaussian random vector $\mathbf{x} \in \mathbb{R}^n$ with $\mathbb{E}[\mathbf{x}] = 0$, $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbf{I}_n$ and positive semi-definite matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, we have*

$$\Pr\left[|\mathbf{x}^\top\mathbf{A}\mathbf{x} - \mathrm{tr}\mathbf{A}| \ge \frac{1}{3}\mathrm{tr}\mathbf{A}\right] \le 2\exp\left(-\min\left\{\frac{r_\mathbf{A}}{9C^2K^4}, \frac{\sqrt{r_\mathbf{A}}}{3CK^2}\right\}\right)$$

with $r_\mathbf{A} = \mathrm{tr}\mathbf{A}/\|\mathbf{A}\|$ the stable rank of $\mathbf{A}$, and

$$\mathbb{E}\left[(\mathbf{x}^\top\mathbf{A}\mathbf{x} - \mathrm{tr}\mathbf{A})^2\right] \le c\ K^4\ \mathrm{tr}\mathbf{A}^2$$

for some $C, c > 0$ independent of K.

Proof This follows from a Hanson-Wright type [**rudelson2013hanson**] sub-gaussian concentration inequality. More precisely, from [**zajkowski2018bounds**] we have, for K-sub-gaussian $\mathbf{x} \in \mathbb{R}^n$ with $\mathbb{E}[\mathbf{x}] = 0$, $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbf{I}_n$ and symmetric positive semi-definite $\mathbf{A} \in \mathbb{R}^{n\times n}$ that

$$\Pr\left\{|\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathrm{tr}\mathbf{A}| \geq t\right\} \leq 2\exp\left(-\min\left\{\frac{t^2}{C^2K^4\mathrm{tr}\mathbf{A}^2}, \frac{t}{CK^2\sqrt{\mathrm{tr}\mathbf{A}^2}}\right\}\right)$$

for some universal constant $C > 0$. Taking $t = \frac{1}{3}\mathrm{tr}\mathbf{A}$ we have

$$\frac{t^2}{C^2K^4\mathrm{tr}\mathbf{A}^2} = \frac{(\mathrm{tr}\mathbf{A})^2}{9C^2K^4\mathrm{tr}\mathbf{A}^2} \geq \frac{\mathrm{tr}\mathbf{A}}{9C^2K^4\|\mathbf{A}\|} = \frac{r_{\mathbf{A}}}{9C^2K^4}, \quad \frac{t}{CK^2\sqrt{\mathrm{tr}\mathbf{A}^2}} \geq \frac{\sqrt{r_{\mathbf{A}}}}{3CK^2}$$

where we use the fact that $\mathrm{tr}\mathbf{A}^2 \leq \|\mathbf{A}\|\mathrm{tr}\mathbf{A}$.

Integrating this bound yields:

$$\mathbb{E}\left[(\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathrm{tr}\mathbf{A})^2\right] \leq c\ K^4\ \mathrm{tr}\mathbf{A}^2$$

and thus the conclusion. ■

Lemma 1.3 *With the notations of Lemma 1.1, for* $X = \mathrm{tr}\,\boldsymbol{\Sigma}(\mathbf{P}_{-k} - \mathbb{E}[\mathbf{P}_{-k}])$ *and* $\|\boldsymbol{\Sigma}\| = 1$, *we have*

$$\mathbb{E}[X^2] \leq Ck \quad \textit{and} \quad \Pr\{|X| \geq t\} \leq 2\mathrm{e}^{-\frac{t^2}{ck}}.$$

for some universal constant $C, c > 0$.

Proof To simplify notations, we work on $\mathbf{P}$ instead of $\mathbf{P}_{-k}$, the same line of argument applies to $\mathbf{P}_{-k}$ by changing the sample size k to $k-1$.

First note that

$$\begin{aligned} X &= \mathrm{tr}\boldsymbol{\Sigma}(\mathbf{P} - \mathbb{E}\mathbf{P}) = \mathbb{E}_k[\mathrm{tr}\boldsymbol{\Sigma}\mathbf{P}] - \mathbb{E}_0[\mathrm{tr}\boldsymbol{\Sigma}\mathbf{P}] \\ &= \sum_{i=1}^{k} (\mathbb{E}_i[\mathrm{tr}\boldsymbol{\Sigma}\mathbf{P}] - \mathbb{E}_{i-1}[\mathrm{tr}\boldsymbol{\Sigma}\mathbf{P}]) = \sum_{i=1}^{k} (\mathbb{E}_i - \mathbb{E}_{i-1})\mathrm{tr}\boldsymbol{\Sigma}(\mathbf{P} - \mathbf{P}_{-i}) \end{aligned}$$

where we used the fact that $\mathbb{E}_i[\mathrm{tr}\boldsymbol{\Sigma}\mathbf{P}_{-i}] = \mathbb{E}_{i-1}[\mathrm{tr}\boldsymbol{\Sigma}\mathbf{P}_{-i}]$, for $\mathbb{E}_i[\cdot]$ the conditional expectation with respect to $\mathcal{F}_i$ the σ-field generating the rows $\mathbf{x}_1 \ldots, \mathbf{x}_i$ of $\mathbf{X}$. This forms a martingale difference sequence (it is a difference sequence of the Doob martingale for $\mathrm{tr}\boldsymbol{\Sigma}(\mathbf{P} - \mathbf{P}_{-i})$ with respect to filtration $\mathcal{F}_i$) hence it falls within the scope of the Burkholder inequality [**burkholder1973distribution**], recalled as follows.

Lemma 1.4 *For* $\{x_i\}_{i=1}^k$ *a real martingale difference sequence with respect to the increasing* σ *field* $\mathcal{F}_i$, *we have, for* $L > 1$, *there exists* $C_L > 0$ *such that*

$$\mathbb{E}\Big[\Big|\sum_{i=1}^{k} x_i\Big|^L\Big] \leq C_L\mathbb{E}\Big[\Big(\sum_{i=1}^{k} |x_i|^2\Big)^{L/2}\Big].$$

From Lemma 1.1, $\mathbf{P}-\mathbf{P}_{-i}=\frac{(\mathbf{I}-\mathbf{P}_{-i})\mathbf{x}_i\mathbf{x}_i^\top(\mathbf{I}-\mathbf{P}_{-i})}{\mathbf{x}_i^\top(\mathbf{I}-\mathbf{P}_{-i})\mathbf{x}_i}$ is positive semi-definite, we have $\mathrm{tr}\boldsymbol{\Sigma}(\mathbf{P}-\mathbf{P}_{-i})\leq\|\boldsymbol{\Sigma}\|=1$ so that with Lemma 1.4 we obtain with $x_i=(\mathbb{E}_i-\mathbb{E}_{i-1})\mathrm{tr}\boldsymbol{\Sigma}(\mathbf{P}-\mathbf{P}_{-i})$ that, for $L>1$

$$\mathbb{E}|X|^L\leq C_Lk^{L/2}.$$

In particular, for $L=2$, we obtain $\mathbb{E}|X|^2\leq Ck$.

For the second result, since we have almost surely bounded martingale differences ($|x_i|\leq 2$), by the Azuma-Hoeffding inequality

$$\Pr\{|X|\geq t\}\leq 2\mathrm{e}^{\frac{-t^2}{8k}}$$

as desired. ■

Complete proof of Theorem 1.2

Equipped with the lemmas above, we are ready to prove Theorem 1.2. First note that:

1. Since $\mathbf{X}^\dagger\mathbf{X}\overset{d}{=}(\alpha\mathbf{X})^\dagger(\alpha\mathbf{X})$ for any $\alpha\in\mathbb{R}\setminus\{0\}$, we can assume without loss of generality (after rescaling $\bar{\mathbf{P}}_\perp$ correspondingly) that $\|\boldsymbol{\Sigma}\|=1$.

2. According to the definition of $\bar{\mathbf{P}}_\perp$ and γ, the following bounds hold

$$\frac{1}{\gamma+1}\mathbf{I}\preceq\bar{\mathbf{P}}_\perp\preceq\mathbf{I},\quad\gamma\leq\frac{k}{r-k}=\frac{1}{\rho-1}\tag{1.6}$$

   for $r\equiv\frac{\mathrm{tr}\boldsymbol{\Sigma}}{\|\boldsymbol{\Sigma}\|}=\mathrm{tr}\boldsymbol{\Sigma}$ and $\rho\equiv\frac{r}{k}>1$, where we used the fact that

$$k=n-\mathrm{tr}\,\bar{\mathbf{P}}_\perp=\mathrm{tr}\,\bar{\mathbf{P}}_\perp(\gamma\boldsymbol{\Sigma}+\mathbf{I})-\mathrm{tr}\,\bar{\mathbf{P}}_\perp=\gamma\mathrm{tr}\,\bar{\mathbf{P}}_\perp\boldsymbol{\Sigma}\geq\frac{\gamma}{\gamma+1}\mathrm{tr}\,\boldsymbol{\Sigma},$$

   so that $r=\mathrm{tr}\boldsymbol{\Sigma}\leq k\cdot\frac{\gamma+1}{\gamma}$.

3. As already discussed in Section 1.3, to obtain the lower and upper bound for $\mathbb{E}[\mathbf{P}_\perp]$ in the sense of symmetric matrix as in Theorem 1.2, it suffices to bound the following spectral norm

$$\|\mathbf{I}-\mathbb{E}[\mathbf{P}_\perp]\bar{\mathbf{P}}_\perp^{-1}\|\leq\frac{C_\rho}{\sqrt{r}},\tag{1.7}$$

   so that, with $\frac{\rho-1}{\rho}\mathbf{I}\preceq\bar{\mathbf{P}}_\perp\preceq\mathbf{I}$ from (1.6), we have

$$\|\mathbf{I}-\bar{\mathbf{P}}_\perp^{-\frac{1}{2}}\mathbb{E}[\mathbf{P}_\perp]\bar{\mathbf{P}}_\perp^{-\frac{1}{2}}\|=\|\bar{\mathbf{P}}_\perp^{-\frac{1}{2}}(\mathbf{I}-\mathbb{E}[\mathbf{P}_\perp]\bar{\mathbf{P}}_\perp^{-1})\bar{\mathbf{P}}_\perp^{\frac{1}{2}}\|\leq\frac{C_\rho}{\sqrt{r}}\sqrt{\frac{\rho}{\rho-1}}.$$

Defining $\epsilon = \frac{C_\rho}{\sqrt{r}}\sqrt{\frac{\rho}{\rho-1}}$, this means that all eigenvalues of the p.s.d. matrix $\bar{\mathbf{P}}_\perp^{-\frac{1}{2}}\mathbb{E}[\mathbf{P}_\perp]\bar{\mathbf{P}}_\perp^{-\frac{1}{2}}$ lie in the interval $[1-\epsilon, 1+\epsilon]$, and

$$(1-\epsilon)\mathbf{I} \preceq \bar{\mathbf{P}}_\perp^{-\frac{1}{2}}\mathbb{E}[\mathbf{P}_\perp]\bar{\mathbf{P}}_\perp^{-\frac{1}{2}} \preceq (1+\epsilon)\mathbf{I}.$$

so that by multiplying $\bar{\mathbf{P}}_\perp^{\frac{1}{2}}$ on both sides, we obtain the desired bound.

As a consequence of the above observations, we only need to prove (1.7) under the setting $\|\mathbf{\Sigma}\| = 1$. The proof comes in the following two steps:

1. For $\mathbf{P}_{-i} = \mathbf{X}_{-i}^\dagger \mathbf{X}_{-i}$, with $\mathbf{X}_{-i} \in \mathbb{R}^{(k-1)\times n}$ the matrix $\mathbf{X}$ without its i-th row, we define, for $i \in \{1, \ldots, k\}$, the following events

$$E_i : \left\{ \left| \frac{\operatorname{tr}(\mathbf{I}-\mathbf{P}_{-i})\mathbf{\Sigma}}{\mathbf{x}_i^\top(\mathbf{I}-\mathbf{P}_{-i})\mathbf{x}_i} - 1 \right| \leq \frac{1}{2} \right\}, \tag{1.8}$$

   where we recall $\mathbf{x}_i \in \mathbb{R}^n$ is the i-th row of $\mathbf{X}$ so that $\mathbb{E}[\mathbf{x}_i] = 0$ and $\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\top] = \mathbf{\Sigma}$. With Lemma 1.2, we can bound the probability of $\neg E_i$, and consequently that of $\neg E$ for $E = \bigwedge_{i=1}^k E_i$;

2. We then bound, conditioned on E and $\neg E$ respectively, the spectral norm $\|\mathbf{I} - \mathbb{E}[\mathbf{P}_\perp]\bar{\mathbf{P}}_\perp^{-1}\|$. More precisely, since

$$\begin{aligned}
\mathbf{I} - \mathbb{E}[\mathbf{P}_\perp]\bar{\mathbf{P}}_\perp^{-1} &= \mathbb{E}[\mathbf{P}] - \gamma\mathbb{E}[\mathbf{P}_\perp]\mathbf{\Sigma} \\
&= \mathbb{E}[\mathbf{P}\cdot\mathbf{1}_E] + \mathbb{E}[\mathbf{P}\cdot\mathbf{1}_{\neg E}] - \gamma\mathbb{E}[\mathbf{P}_\perp]\mathbf{\Sigma} \\
&= k\,\mathbb{E}\left[\frac{(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k\mathbf{x}_k^\top}{\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k}\cdot\mathbf{1}_E\right] - \gamma\mathbb{E}[\mathbf{P}_\perp]\mathbf{\Sigma} \;+\; \mathbb{E}[\mathbf{P}\cdot\mathbf{1}_{\neg E}] \\
&= \gamma\,\underbrace{\mathbb{E}\left[(\bar{s}-\hat{s})\cdot\frac{(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k\mathbf{x}_k^\top}{\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k}\cdot\mathbf{1}_E\right]}_{\mathbf{T}_1} - \gamma\underbrace{\mathbb{E}[(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k\mathbf{x}_k^\top\cdot\mathbf{1}_{\neg E}]}_{\mathbf{T}_2} \\
&\quad + \gamma\underbrace{\mathbb{E}[\mathbf{P}-\mathbf{P}_{-k}]\mathbf{\Sigma}}_{\mathbf{T}_3} + \underbrace{\mathbb{E}[\mathbf{P}\cdot\mathbf{1}_{\neg E}]}_{\mathbf{T}_4},
\end{aligned}$$

   where we used Lemma 1.1 for the third equality and denote $\hat{s} = \mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k$ as well as $\bar{s} = \operatorname{tr}\bar{\mathbf{P}}_\perp\mathbf{\Sigma} = k/\gamma$. It then remains to bound the spectral norms of $\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3, \mathbf{T}_4$ to reach the conclusion.

Another important relation that will be constantly used throughout the proof is

$$\operatorname{tr}(\mathbf{I}-\mathbf{P}_{-k})\mathbf{\Sigma} = \operatorname{tr}\mathbf{\Sigma}^{\frac{1}{2}}(\mathbf{I}-\mathbf{P}_{-k})^2\mathbf{\Sigma}^{\frac{1}{2}} = \|\mathbf{\Sigma}^{\frac{1}{2}} - \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{X}_{-k}^\dagger\mathbf{X}_{-k}\|_F^2 \geq \sum_{i\geq k}\lambda_i(\mathbf{\Sigma}) \geq r-k \tag{1.9}$$

where we used the fact that $\mathrm{rank}(\mathbf{X}_{-k}^{\dagger}\mathbf{X}_{-k}) \leq \mathrm{rank}(\mathbf{X}_{-k}) \leq k-1$ and arranged the eigenvalues $1 = \lambda_1(\mathbf{\Sigma}) \geq \ldots \geq \lambda_n(\mathbf{\Sigma})$ in a non-increasing order. As a consequence, we also have

$$\frac{\mathrm{tr}(\mathbf{I}-\mathbf{P}_{-k})\mathbf{\Sigma}}{\|(\mathbf{I}-\mathbf{P}_{-k})\mathbf{\Sigma}\|} \geq \mathrm{tr}(\mathbf{I}-\mathbf{P}_{-k})\mathbf{\Sigma} \geq r-k. \tag{1.10}$$

For the first step, we have, with Lemma 1.2 and (1.10) that

$$\begin{aligned}\Pr(\neg E_i) &\leq \Pr\left\{|\mathbf{x}_i^{\top}(\mathbf{I}-\mathbf{P}_{-i})\mathbf{x}_i - \mathrm{tr}\mathbf{\Sigma}(\mathbf{I}-\mathbf{P}_{-i})| \geq \frac{1}{3}\mathrm{tr}\mathbf{\Sigma}(\mathbf{I}-\mathbf{P}_{-i})\right\} \\ &\leq 2\mathrm{e}^{-\min\left\{\frac{r-k}{9C^2K^4}, \frac{\sqrt{r-k}}{3CK^2}\right\}}.\end{aligned}$$

so that with the union bound we obtain

$$\Pr(\neg E) \leq 2k\mathrm{e}^{-\min\left\{\frac{r-k}{9C^2K^4}, \frac{\sqrt{r-k}}{3CK^2}\right\}} \leq \frac{k}{(r-k)^2}\cdot 2(r-k)^2\mathrm{e}^{-\min\left\{\frac{r-k}{9C^2K^4}, \frac{\sqrt{r-k}}{3CK^2}\right\}} \leq \frac{C_\rho}{r-k} \tag{1.11}$$

where we used the fact that, for $\alpha > 0$, $x^2\mathrm{e}^{-\alpha x} \leq \frac{4\mathrm{e}^{-2}}{\alpha^2}$ and $x^4\mathrm{e}^{-\alpha x} \leq \frac{256\mathrm{e}^{-4}}{\alpha^4}$ on $x > 0$. Also, denote $c_\rho = \frac{r-k}{r} = \frac{\rho-1}{\rho} > 0$, we have

$$\Pr(\neg E) \leq \frac{C_\rho}{r-k} = \frac{C_\rho}{c_\rho r} = \frac{C'_\rho}{r} \tag{1.12}$$

for some $C'_\rho > 0$ that depends on $\rho = r/k > 1$ and the sub-gaussian norm K.

At this point, note that, conditioned on the event E, we have for $i \in \{1, \ldots, k\}$

$$\frac{1}{2}\frac{1}{\mathrm{tr}(\mathbf{I}-\mathbf{P}_{-i})\mathbf{\Sigma}} \leq \frac{1}{\mathbf{x}_i^{\top}(\mathbf{I}-\mathbf{P}_{-i})\mathbf{x}_i} \leq \frac{3}{2}\frac{1}{\mathrm{tr}(\mathbf{I}-\mathbf{P}_{-i})\mathbf{\Sigma}}, \tag{1.13}$$

Also, with (1.12) and the fact that $\|\mathbf{P}\| \leq 1$, we have $\|\mathbf{T}_4\| \leq \frac{C_\rho}{r}$ for some $C_\rho > 0$ that depends on ρ and K. To handle non-symmetric matrix $\mathbf{T}_2$, note that $\mathbf{T}_2 + \mathbf{T}_2^{\top}$ is symmetric and

$$-\mathbb{E}[(\mathbf{I}-\mathbf{P}_{-k})\cdot\mathbf{1}_{\neg E}] - \mathbb{E}[(\mathbf{x}_k^{\top}\mathbf{x}_k)\mathbf{x}_k\mathbf{x}_k^{\top}\cdot\mathbf{1}_{\neg E}] \preceq \mathbf{T}_2 + \mathbf{T}_2^{\top} \preceq \mathbb{E}[(\mathbf{I}-\mathbf{P}_{-k})\cdot\mathbf{1}_{\neg E}] + \mathbb{E}[(\mathbf{x}_k^{\top}\mathbf{x}_k)\mathbf{x}_k\mathbf{x}_k^{\top}\cdot\mathbf{1}_{\neg E}] \tag{1.14}$$

with $-(\mathbf{A}\mathbf{A}^{\top} + \mathbf{B}\mathbf{B}^{\top}) \preceq \mathbf{A}\mathbf{B}^{\top} + \mathbf{B}\mathbf{A}^{\top} \preceq \mathbf{A}\mathbf{A}^{\top} + \mathbf{B}\mathbf{B}^{\top}$. To obtain an upper bound for operator norm of $\mathbb{E}[(\mathbf{x}_k^{\top}\mathbf{x}_k)\mathbf{x}_k\mathbf{x}_k^{\top}\cdot\mathbf{1}_{\neg E}]$, note that

$$\begin{aligned}\|\mathbb{E}[(\mathbf{x}_k^{\top}\mathbf{x}_k)\mathbf{x}_k\mathbf{x}_k^{\top}\cdot\mathbf{1}_{\neg E}]\| &\leq \mathbb{E}[(\mathbf{x}_k^{\top}\mathbf{x}_k)^2\cdot\mathbf{1}_{\neg E}] = \int_0^{\infty}\Pr(\mathbf{x}^{\top}\mathbf{x}\cdot\mathbf{1}_{\neg E} \geq \sqrt{t})dt \\ &\leq \int_0^{X}\Pr(\mathbf{x}^{\top}\mathbf{x}\cdot\mathbf{1}_{\neg E} \geq \sqrt{t})dt + \int_X^{\infty}\Pr(\mathbf{x}^{\top}\mathbf{x} \geq \sqrt{t})dt \\ &\leq X\cdot\Pr(\neg E) + \int_X^{\infty}\mathrm{e}^{-\min\left\{\frac{t}{C^2K^4r}, \frac{\sqrt{t}}{CK^2\sqrt{r}}\right\}}dt \leq \frac{c_\rho}{r}\end{aligned}$$

where we recall $\mathbb{E}[\mathbf{x}^\top\mathbf{x}] = \mathrm{tr}\boldsymbol{\Sigma} = r$ and take $X \geq C^2K^4r$, the third line follows from the proof of Lemma 1.2 and the forth line from the same argument as in (1.11). Moreover, since $\|\mathbf{T}_2\| \leq \|\mathbf{T}_2 + \mathbf{T}_2^\top\|$ (see for example [**serre2010matrices**]), we conclude that $\|\mathbf{T}_2\| \leq \frac{C_\rho}{r}$.

And it thus remains to handle the terms $\mathbf{T}_1$ and $\mathbf{T}_3$ to obtain a bound on $\|\mathbf{I} - \mathbb{E}[\mathbf{P}_\perp]\widehat{\mathbf{P}}_\perp^{-1}\|$.

To bound $\mathbf{T}_3$, with $\mathbf{P} - \mathbf{P}_{-k} = \frac{(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})}{\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k}$ in Lemma 1.1, we have

$$\begin{aligned}\|\mathbf{T}_3\| &\leq \left\|\mathbb{E}\left[\frac{(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})}{\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k}\cdot\mathbf{1}_E\right]\right\| + \|\mathbb{E}[(\mathbf{P}-\mathbf{P}_{-k})\cdot\mathbf{1}_{\neg E}]\| \\ &\leq \frac{3}{2}\mathbb{E}\left[\frac{1}{\mathrm{tr}(\mathbf{I}-\mathbf{P}_{-k})\boldsymbol{\Sigma}}\right] + \frac{c_\rho}{r-k} \leq \frac{C_\rho}{r-k} = \frac{C'_\rho}{r}\end{aligned}$$

where we used the fact that $\mathrm{tr}\,(\mathbf{I}-\mathbf{P}_{-k})\boldsymbol{\Sigma} \geq r-k$ from (1.9) and recall $\rho \equiv r/k > 1$.

For $\mathbf{T}_1$ we write

$$\begin{aligned}\|\mathbf{T}_1\| &\leq \mathbb{E}\left[\|\mathbf{I}-\mathbf{P}_{-k}\|\cdot\left\|\mathbb{E}\Big[|\bar{s}-\hat{s}|\cdot\frac{\mathbf{x}_k\mathbf{x}_k^\top}{\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k}\cdot\mathbf{1}_E \mid \mathbf{P}_{-k}\Big]\right\|\right] \\ &\leq \frac{3}{2}\frac{1}{r-k}\cdot\mathbb{E}\left[\sup_{\|\mathbf{v}\|=1}\mathbb{E}\Big[|\bar{s}-\hat{s}|\cdot\mathbf{v}^\top\mathbf{x}_k\mathbf{x}_k^\top\mathbf{v}\cdot\mathbf{1}_E \mid \mathbf{P}_{-k}\Big]\right] \\ &\leq \frac{C_\rho}{r}\cdot\mathbb{E}\Bigg[\underbrace{\sqrt{\mathbb{E}\big[(\bar{s}-\hat{s})^2\cdot\mathbf{1}_E \mid \mathbf{P}_{-k}\big]}}_{T_{1,1}}\cdot\underbrace{\sup_{\|\mathbf{v}\|=1}\sqrt{\mathbb{E}\big[(\mathbf{v}^\top\mathbf{x}_k)^4\big]}}_{T_{1,2}}\Bigg]\end{aligned}$$

where we used Jensen's inequality for the first inequality, the relation in (1.9) for the second inequality, and Cauchy–Schwarz for the third inequality.

We first bound $T_{1,2}$ by definition of sub-gaussian random vectors. We have for $\mathbf{x}_k$ a K-sub-gaussian and $\|\mathbf{v}\| = 1$ that, $\mathbf{v}^\top\mathbf{x}_k$ is a sub-gaussian random variable with $\|\mathbf{v}^\top\mathbf{a}\|_{\psi_2} \leq K$. As such, $T_{1,2} \leq CK^2$ for some absolute constant $C > 0$, see for example [**vershynin2018high**].

For $T_{1,1}$ we have

$$\sqrt{\mathbb{E}[(\bar{s}-\hat{s})^2\cdot\mathbf{1}_E \mid \mathbf{P}_{-k}]} = \sqrt{(\bar{s}-s)^2 + \mathbb{E}[(s-\hat{s})^2\cdot\mathbf{1}_E]}$$

where we denote $s = \mathbb{E}[\hat{s}] = \mathrm{tr}\,\mathbb{E}[\mathbf{I}-\mathbf{P}_{-k}]\boldsymbol{\Sigma}$. Note that

$$\begin{aligned}\mathbb{E}\big[(s-\hat{s})^2\big] &= \mathbb{E}\big[\big(\mathrm{tr}\,\boldsymbol{\Sigma}(\mathbf{P}_{-k}-\mathbb{E}[\mathbf{P}_{-k}])\big)^2\big] + \mathbb{E}\big[(\mathrm{tr}\,(\mathbf{I}-\mathbf{P}_{-k})\boldsymbol{\Sigma} - \mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k)^2\big] \\ &\leq C_1k + C_2\mathbb{E}\big[\mathrm{tr}\,(\boldsymbol{\Sigma}-\mathbf{P}_{-k}\boldsymbol{\Sigma})^2\big] \leq C(k+s) \leq C\big(k+\bar{s}+|s-\bar{s}|\big)\end{aligned}$$

where we used Lemma 1.3 and Lemma 1.2. Recall that $\bar{s} = \mathrm{tr}\bar{\mathbf{P}}_\perp\boldsymbol{\Sigma} \leq \mathrm{tr}\boldsymbol{\Sigma} = r$ and $k < r$, we have

$$T_{1,1} \leq \sqrt{(\bar{s}-s)^2 + C(|\bar{s}-s| + 2r)} \tag{1.15}$$

It remains to bound $|\bar{s} - s|$. Note that $\mathbf{P} = (\mathbf{X}^\top\mathbf{X})^\dagger\mathbf{X}^\top\mathbf{X} = \mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^\dagger$ and is symmetric, so

$$\begin{aligned}
&\mathbf{I} - \mathbb{E}[\mathbf{P}_\perp]\bar{\mathbf{P}}_\perp^{-1} + \mathbf{I} - \bar{\mathbf{P}}_\perp^{-1}\mathbb{E}[\mathbf{P}_\perp] = 2\mathbb{E}[\mathbf{P}] - \mathbb{E}[\gamma\mathbf{P}_\perp\boldsymbol{\Sigma}] - \mathbb{E}[\gamma\boldsymbol{\Sigma}\mathbf{P}_\perp] \\
&= \sum_{i=1}^{k} \mathbb{E}\big[(\mathbf{X}^\top\mathbf{X})^\dagger\mathbf{x}_i\mathbf{x}_i^\top + \mathbf{x}_i\mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^\dagger\big] - \gamma(\mathbb{E}[\mathbf{P}_\perp]\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\mathbb{E}[\mathbf{P}_\perp]) \\
&= \gamma\,\mathbb{E}\left[\bar{s}\cdot\frac{(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k\mathbf{x}_k^\top + \mathbf{x}_k\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})}{\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k}\right] - \gamma\,\mathbb{E}\left[\hat{s}\cdot\frac{(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k\mathbf{x}_k^\top + \mathbf{x}_k\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})}{\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k}\right] \\
&+ \gamma\,(\mathbb{E}[(\mathbf{I}-\mathbf{P}_{-k})\boldsymbol{\Sigma}] + \mathbb{E}[\boldsymbol{\Sigma}(\mathbf{I}-\mathbf{P}_{-k})]) - \gamma(\mathbb{E}[\mathbf{P}_\perp]\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\mathbb{E}[\mathbf{P}_\perp]) \\
&= \gamma\,\mathbb{E}\left[(\bar{s}-\hat{s})\cdot\frac{(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k\mathbf{x}_k^\top + \mathbf{x}_k\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})}{\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k}\right] + \gamma(\mathbb{E}[\mathbf{P}-\mathbf{P}_{-k}]\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\mathbb{E}[\mathbf{P}-\mathbf{P}_{-k}]).
\end{aligned}$$

Moreover, using the fact that $\bar{\mathbf{P}}_\perp\boldsymbol{\Sigma} \preceq \frac{1}{\gamma+1}\mathbf{I}$ and $\bar{\mathbf{P}}_\perp\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\bar{\mathbf{P}}_\perp$, we obtain that

$$\begin{aligned}
&|\bar{s} - s| = |\mathrm{tr}(\bar{\mathbf{P}}_\perp - \mathbb{E}[\mathbf{I}-\mathbf{P}_{-k}])\boldsymbol{\Sigma}| \le |\mathrm{tr}(\bar{\mathbf{P}}_\perp - \mathbb{E}[\mathbf{P}_\perp])\boldsymbol{\Sigma}| + |\mathrm{tr}\mathbb{E}[\mathbf{P}-\mathbf{P}_{-k}]\boldsymbol{\Sigma}| \\
&= \frac{1}{2}\big|\mathrm{tr}\,(\mathbf{I}-\mathbb{E}[\mathbf{P}_\perp]\bar{\mathbf{P}}_\perp^{-1})\bar{\mathbf{P}}_\perp\boldsymbol{\Sigma} + \mathrm{tr}\,\bar{\mathbf{P}}_\perp(\mathbf{I}-\bar{\mathbf{P}}_\perp^{-1}\mathbb{E}[\mathbf{P}_\perp])\boldsymbol{\Sigma}\big| + \mathrm{tr}\,\mathbb{E}\left[\frac{(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})}{\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k}\right]\boldsymbol{\Sigma} \\
&\le \frac{1}{2}\big|\mathrm{tr}\,(\mathbf{I}-\mathbb{E}[\mathbf{P}_\perp]\bar{\mathbf{P}}_\perp^{-1} + \mathbf{I} - \bar{\mathbf{P}}_\perp^{-1}\mathbb{E}[\mathbf{P}_\perp])\bar{\mathbf{P}}_\perp\boldsymbol{\Sigma}\big| + 1 \\
&\le \frac{\gamma}{2}\,\mathbb{E}\left[|\bar{s}-\hat{s}|\cdot\frac{\mathrm{tr}\,((\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k\mathbf{x}_k^\top + \mathbf{x}_k\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k}))\bar{\mathbf{P}}_\perp\boldsymbol{\Sigma}}{\mathrm{tr}\,(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k\mathbf{x}_k^\top}\right] \\
&+ \gamma\,\mathbb{E}\left[\frac{\mathrm{tr}\,(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\bar{\mathbf{P}}_\perp\boldsymbol{\Sigma}}{\mathrm{tr}\,(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k\mathbf{x}_k^\top}\right] + 1 \\
&\le \frac{\gamma}{\gamma+1}\left(\mathbb{E}\left[|\bar{s}-\hat{s}|\cdot\frac{\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k}{\mathbf{x}_k^\top(\mathbf{I}-\mathbf{P}_{-k})\mathbf{x}_k}\right] + 1\right) + 1 \le \frac{\gamma}{\gamma+1}\Big(|\bar{s}-s| + \mathbb{E}\big[|s-\hat{s}|\big] + 1\Big) + 1 \\
&\le \frac{\gamma}{\gamma+1}\Big(|\bar{s}-s| + C\sqrt{|\bar{s}-s|} + C\sqrt{2r} + 1\Big) + 1.
\end{aligned}$$

Solving for $|\bar{s} - s|$, we deduce that

$$|\bar{s} - s| \le C_1\sqrt{r} + C_2,$$

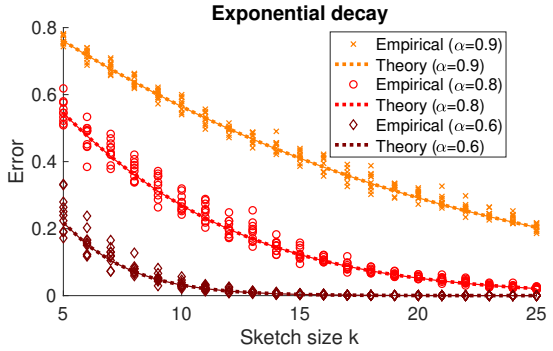
so plugging back to (1.15) we get $T_{1,1} \le C\sqrt{r}$ and $\|\mathbf{T}_1\| \le \frac{C_\rho}{\sqrt{r}}$, thus completing the proof.

1.5 Explicit formulas under known spectral decay

The expression we give for the expected residual projection, $\mathbb{E}[\mathbf{P}_\perp] \simeq (\gamma\mathbf{A}^\top\mathbf{A} + \mathbf{I})^{-1}$, is implicit in that it depends on the parameter γ which is the solution of the following equation:

$$\sum_{i\ge 1}\frac{\gamma\sigma_i^2}{\gamma\sigma_i^2 + 1} = k, \qquad \text{where } \sigma_i \text{ are the singular values of } \mathbf{A}. \tag{1.16}$$

(a) Singular values are given by $\sigma_i^2 = C \cdot \alpha^{i-1}$.



(b) Singular values are given by $\sigma_i^2 = C \cdot i^{-\beta}$.

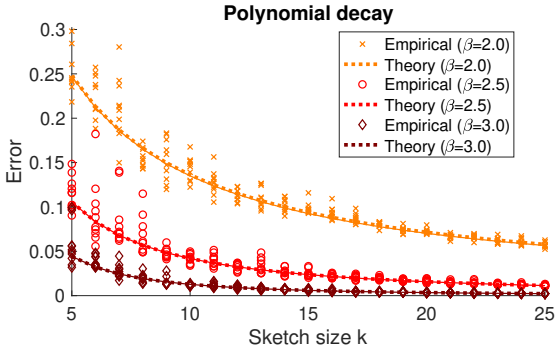


Figure 1.1: Theoretical predictions of low-rank approximation error of a Gaussian sketch under known spectral decays, compared to the empirical results. The constant C is scaled so that $\|\mathbf{A}\|_F^2 = 1$ and we let $n = m = 1000$. For the theory, we plot the explicit formulas (1.17) and (1.18) (dashed lines), as well as the implicit expression from Corollary 1.1 (thin solid lines) obtained by numerically solving (1.16). Observe that the explicit and implicit predictions are nearly (but not exactly) identical.

In general, it is impossible to solve this equation analytically, i.e., to write γ as an explicit formula of n, k and the singular values of $\mathbf{A}$. However, we show that when the singular values exhibit a known rate of decay, then it is possible to obtain explicit formulas for γ. In particular, this allows us to provide precise and easily interpretable rates of decay for the low-rank approximation error of a sub-gaussian sketch.

Matrices that have known spectral decay, most commonly with either exponential or polynomial rate, arise in many machine learning problems [**randomized-newton**]. Such behavior can be naturally occurring in data, or it can be induced by feature expansion using, say, the RBF kernel (for exponential decay) or Matérn (for polynomial decay) kernels [**Santa97gaussianregression**; **RasmussenWilliams06**]. Understanding these two classes of decay plays an important role in distinguishing the properties of light-tailed and heavy-tailed data distributions. Note that in the kernel setting we may often represent our data via the $m \times m$ kernel matrix $\mathbf{K}$, instead of the $m \times n$ data matrix $\mathbf{A}$, and study the sketched Nyström method [**revisiting-nystrom**] for low-rank approximation. To handle the kernel setting in our analysis, it suffices to replace the squared singular values σ_i^2 of $\mathbf{A}$ with the eigenvalues of $\mathbf{K}$.

Exponential spectral decay

Suppose that the squared singular values of $\mathbf{A}$ exhibit exponential decay, i.e. $\sigma_i^2 = C \cdot \alpha^{i-1}$, where C is a constant and $\alpha \in (0, 1)$. For simplicity of presentation, we will let $m, n \to \infty$. Under this spectral decay, we can approximate the sum in (1.16) by the analytically

computable integral $\int_y^\infty \frac{1}{1+(C\gamma)^{-1}\alpha^{-x}}dx$, obtaining $\gamma \approx (\alpha^{-k}-1)\sqrt{\alpha}/C$. Applying this to the formula from Corollary 1.1, we can express the low-rank approximation error for a sketch of size k as follows:

$$\mathbb{E}\big[\|\mathbf{A}-\mathbf{AP}\|_F^2\big] \approx \frac{C}{\sqrt{\alpha}} \cdot \frac{k}{\alpha^{-k}-1}, \quad \text{when} \quad \sigma_i^2 = C \cdot \alpha^{i-1} \ \text{ for all } i. \tag{1.17}$$

In Figure 1.1a, we plot the above formula against the numerically obtained implicit expression from Corollary 1.1, as well as empirical results for a Gaussian sketch. First, we observe that the theoretical predictions closely align with empirical values even after the sketch size crosses the stable rank $r \approx \frac{1}{1-\alpha}$, suggesting that Theorem 1.1 can be extended to this regime. Second, while it is not surprising that the error decays at a similar rate as the singular values, our predictions offer a much more precise description, down to lower order effects and even constant factors. For instance, we observe that the error (normalized by $\|\mathbf{A}\|_F^2$, as in the figure) only starts decaying exponentially after k crosses the stable rank, and until that point it decreases at a linear rate with slope $-\frac{1-\alpha}{2\sqrt{\alpha}}$.

Polynomial spectral decay

We now turn to polynomial spectral decay, which is a natural model for analyzing heavy-tailed data distributions. Let $\mathbf{A}$ have squared singular values $\sigma_i^2 = C \cdot i^{-\beta}$ for some $\beta \geq 2$, and let $m, n \to \infty$. As in the case of exponential decay, we use the integral $\int_y^\infty \frac{1}{1+(C\gamma)^{-1}x^{-\beta}}dx$ to approximate the sum in (1.16), and solve for γ, obtaining $\gamma \approx \big((k+\frac{1}{2})\frac{\beta}{\pi}\sin(\frac{\pi}{\beta})\big)^\beta$. Combining this with Corollary 1.1 we get:

$$\mathbb{E}\big[\|\mathbf{A}-\mathbf{AP}\|_F^2\big] \approx C \cdot \frac{k}{(k+\frac{1}{2})^\beta}\left(\frac{\pi/\beta}{\sin(\pi/\beta)}\right)^\beta, \quad \text{when} \quad \sigma_i^2 = C \cdot i^{-\beta} \ \text{ for all } i. \tag{1.18}$$

Figure 1.1b compares our predictions to the empirical results for several values of β. In all of these cases, the stable rank is close to 1, and yet the theoretical predictions align very well with the empirical results. Overall, the asymptotic rate of decay of the error is $k^{1-\beta}$. However it is easy to verify that the lower order effect of $(k+\frac{1}{2})^\beta$ appearing instead of k^β in (1.18) significantly changes the trajectory for small values of k. Also, note that as β grows large, the constant $\big(\frac{\pi/\beta}{\sin(\pi/\beta)}\big)^\beta$ goes to 1, but it plays a significant role for $\beta = 2$ or 3 (roughly, scaling the expression by a factor of 2). Finally, we remark that for $\beta \in (1,2)$, our integral approximation of (1.16) becomes less accurate. We expect that a corrected expression is possible, but likely more complicated and less interpretable.

1.6 Empirical results

In this section, we numerically verify the accuracy of our theoretical predictions for the low-rank approximation error of sketching on benchmark datasets from the libsvm repository

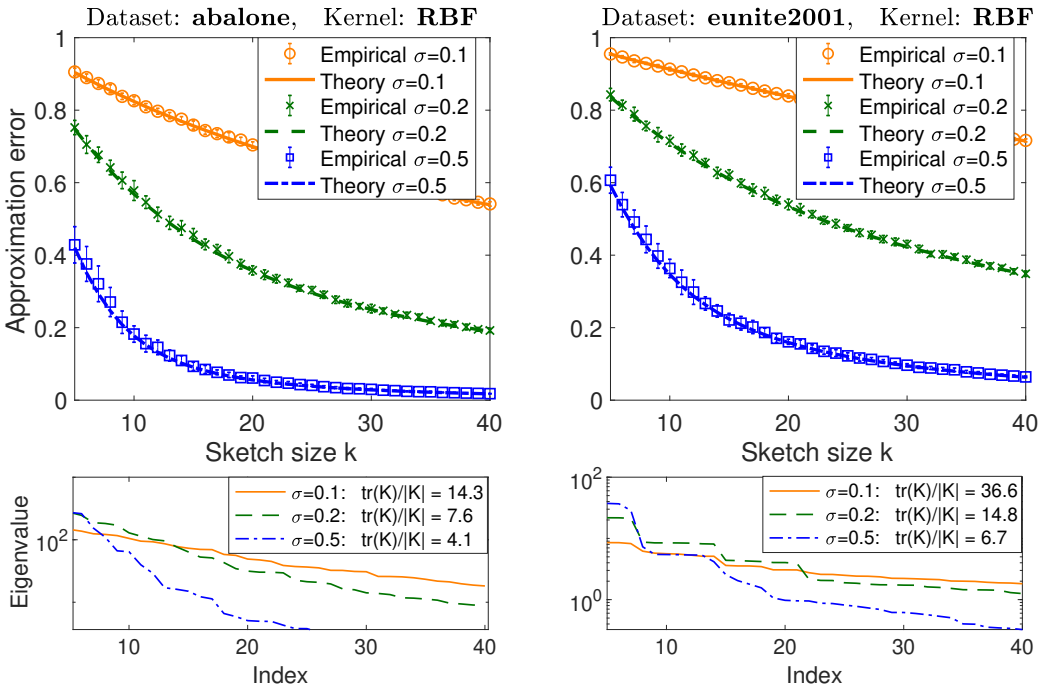


Figure 1.2: Theoretical predictions versus approximation error for the sketched Nyström with the RBF kernel (spectral decay shown at the bottom).

[**libsvm**] (further numerical results are in Appendix 1.6). We repeated every experiment 10 times, and plot both the average and standard deviation of the results. We use the following $k \times m$ sketching matrices $\mathbf{S}$:

1. *Gaussian sketch:* with i.i.d. standard normal entries;

2. *Rademacher sketch:* with i.i.d. entries equal 1 with probability 0.5 and -1 otherwise.

Varying spectral decay. To demonstrate the role of spectral decay and the stable rank on the approximation error, we performed feature expansion using the radial basis function (RBF) kernel $k(\mathbf{a}_i, \mathbf{a}_j) = \exp(-\|\mathbf{a}_i - \mathbf{a}_j\|^2/(2\sigma^2))$, obtaining an $m \times m$ kernel matrix $\mathbf{K}$. We used the sketched Nyström method to construct a low-rank approximation $\tilde{\mathbf{K}} = \mathbf{K}\mathbf{S}^\top(\mathbf{S}\mathbf{K}\mathbf{S}^\top)^\dagger\mathbf{S}\mathbf{K}$, and computed the normalized trace norm error $\|\mathbf{K} - \tilde{\mathbf{K}}\|_*/\|\mathbf{K}\|_*$. The theoretical predictions are coming from (1.2), which in turn uses Theorem 1.1. Following [**revisiting-nystrom**], we use the RBF kernel because varying the scale parameter σ allows us to observe the approximation error under qualitatively different spectral decay profiles of the kernel. In Figure 1.2, we present the results for the Gaussian sketch on two datasets, with three values of σ, and in all cases our theory aligns with the empirical results. Furthermore, as smaller σ leads to slower spectral decay and larger stable rank, it also makes the approximation error decay

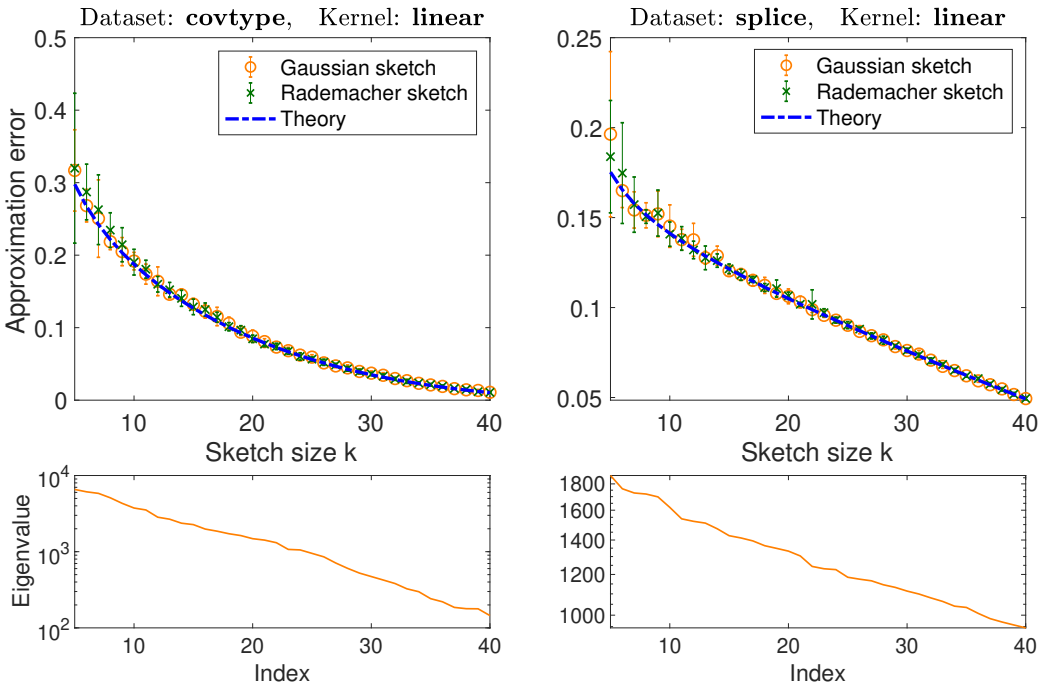


Figure 1.3: Theoretical predictions versus approximation error for the Gaussian and Rademacher sketches (spectral decay shown at the bottom).

more linearly for small sketch sizes. This behavior is predicted by our explicit expressions (1.17) for the error under exponential spectral decay from Section 1.5. Once the sketch sizes are sufficiently larger than the stable rank of $\mathbf{K}^{\frac{1}{2}}$, the error starts decaying at an exponential rate. Note that Theorem 1.1 only guarantees accuracy of our expressions for sketch sizes below the stable rank, however the predictions are accurate regardless of this constraint.

Varying sketch type. In the next set of empirical results, we compare the performance of Gaussian and Rademacher sketches, and also verify the theory when sketching the data matrix $\mathbf{A}$ without kernel expansion, plotting $\|\mathbf{A} - \mathbf{A}(\mathbf{SA})^{\dagger}\mathbf{SA}\|_F^2/\|\mathbf{A}\|_F^2$. Since both of the sketching methods have sub-gaussian entries, Corollary 1.1 predicts that they should have comparable performance in this task and match our expressions. This is exactly what we observe in Figure 1.3 for two datasets and a range of sketching sizes, as well as in other empirical results shown in Section 1.6.

Additional empirical results on libsvm datasets

We complement the results of Section 1.6 with empirical results on four additional libsvm datasets [**libsvm**] (bringing the total number of benchmark datasets to eight), which further establish the accuracy of our surrogate expressions for the low-rank approximation error.

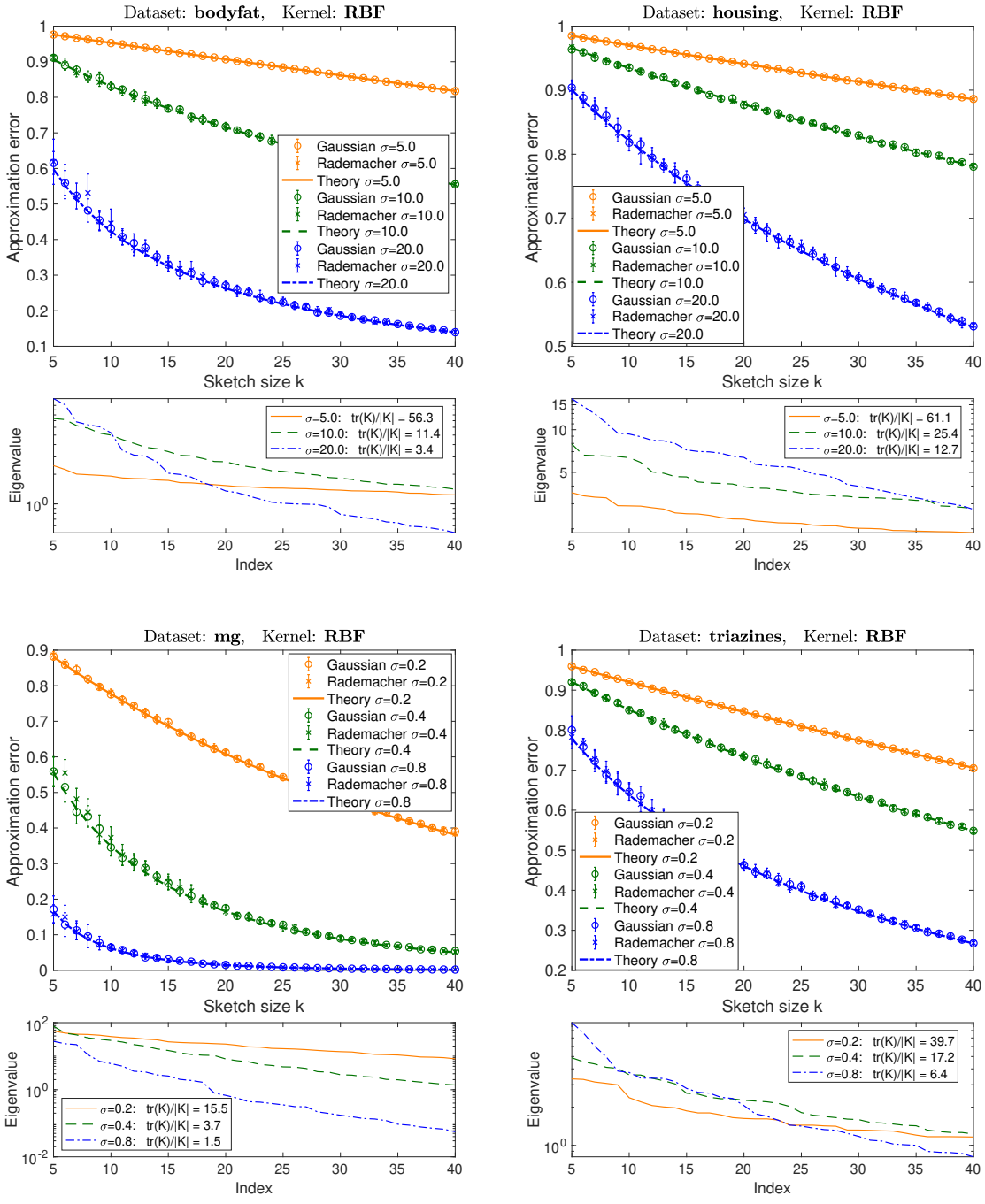


Figure 1.4: Theoretical predictions versus approximation error for the sketched Nyström with the RBF kernel, using Gaussian and Rademacher sketches (spectral decay shown at the bottom).

Similary as in Figure 1.2, we use the sketched Nyström method [**revisiting-nystrom**] with the RBF kernel $k(\mathbf{a}_i, \mathbf{a}_j) = \exp(-\|\mathbf{a}_i - \mathbf{a}_j\|^2/(2\sigma^2))$, for several values of the parameter σ. The values of σ were chosen so as to demonstrate the effectiveness of our theoretical predictions both when the stable rank is moderately large and when it is very small.

In Figure 1.4 we show the results for both Gaussian and Rademacher sketches. These results reinforce the conclusions we made in Section 1.6: our theoretical estimates are very accurate in all cases, for both sketching methods, and even when the stable rank is close to 1 (a regime that is not supported by the current theory).

1.7 Conclusions

We derived the first theoretically supported precise expressions for the expected residual projection matrix, which is a central component in the analysis of RandNLA dimensionality reduction via sketching. Our analysis provides a new understanding of low-rank approximation, the Nyström method, and the convergence properties of many randomized iterative algorithms. As a direction for future work, we conjecture that our main result can be extended to sketch sizes larger than the stable rank of the data matrix.