Leverage Score Sampling

Federico Betti
MATH-403 Low-Rank Approximation Techniques, EPFL Lausanne, Switzerland

December 19, 2022

Abstract

Given a fixed matrix $A \in \mathbb{R}^{m\times n}$, a sketching algorithm for column subset selection aims at the selection of a column submatrix $C \in \mathbb{R}^{m\times k}$ of A that has favorable spectral properties. In randomized linear algebra, the column subset $J = [j_1, \dots, j_k]$ is obtained by sampling each $j_i \in [n]$ from a suitable probability distribution defined on the columns of A. In this project, we present the properties of the standard leverage scores as one such sampling probability. We present some of its advantages and show its poor performance in the case of full column rank matrices. Motivated by this observation, we follow the idea of Cohen et al. [2017] and we introduce the ridge leverage scores as a stable and natural solution to filter out small principal components. In the final section, we show the results of the numerical experiments supporting our claims on standard benchmark matrices of reduced size, and we draw conclusions about the work. The code is available at `https://github.com/federicobetti99/Low-Rank-Approximation-Techniques`: the obtained results are reproducible by executing the scripts provided thereby.

1 Introduction

Introduction to the problem

Let us consider an arbitrary matrix $A \in \mathbb{R}^{m\times n}$ with columns $[a_1, \dots, a_n]$. The problem of column subset selection requests a columns index set $J = \{j_1, \cdots, j_k\}$ such that the column sub-matrix $C = [a_{j_1}, \dots, a_{j_k}] \in \mathbb{R}^{m\times k}$ has favorable spectral properties and approximates well enough the range of A.

Definition 1.1. *Let $A \in \mathbb{R}^{m\times n}$ and $C = [a_{j_1}, \dots, a_{j_k}] \in \mathbb{R}^{m\times k}$ a column submatrix of A. Moreover, let the columns of $Q \in \mathbb{R}^{m\times k}$ be an orthogonal basis for the range of C. Then, we call the subset $J = \{j_1, \cdots, j_k\}$ a $(1+\epsilon)$-factor column subset if*

$$\|A - QQ^TA\|_2^2 \leq (1+\epsilon)\|A - \mathcal{T}_k(A)\|_2^2, \tag{1}$$

where $\mathcal{T}_k(A)$ denotes the best rank-k approximation of A, obtained by projecting onto the k principal left singular vectors of A.

According to the definition above, the goal is to find a $(1+\epsilon)$ factor column subset. We recall that for a generic matrix C (not necessarily a column subset of A), it holds that

$$\|A - QQ^TA\|_2^2 \leq \|A - \mathcal{T}_k(A)\|_2^2 + 2\|AA^T - CC^T\|_2. \tag{2}$$

Therefore, the task at hand for sketching can be reduced to finding a column subset J for which the corresponding matrix C approximates well enough the range of the original matrix A, which eventually means that the additional term $\|AA^T - CC^T\|_2$ is not too large.

Literature review

Recently, *random column sampling* have been shown to perform well for the sketching task. The main task in this setting is the definition of a suitable probability distribution p on the columns of A such that $p_j \geq 0$ for $j \in [n]$ and $\sum_{j=1}^n p_j = 1$. Many sampling probabilities have been studied in the literature. In our opinion, Tropp [2009] provides an exhaustive summary. A naive approach would be to pick uniformly at random columns of A, each of them with probability $1/n$: this is clearly a sub-optimal choice, as one can diminish the importance sampling probability of the main principal components by simply adding an arbitrary number of zero columns to A which are not contributing in forming the column space of A. In this work, we mainly consider the latter as a baseline to gain further insight on the advantage of the proposed sampling distribution. On the other hand, standard results in low-rank approximation have shown that sampling proportionally to the column norms of A, namely

$$p_l = \frac{\|a_l\|_2^2}{\|A\|_F^2}, \tag{3}$$

minimizes the quantity $\mathbb{E}\left[\|AA^T - CC^T\|_F^2\right]$ over all possible choices of sampling probabilities $\mathbf{p}$. Note that the distribution (3) remains unchanged under the addition of an arbitrary number of zero columns (because the corresponding column norm is zero), and thus improves with respect to uniform sampling. However, the main drawback for such a sampling strategy is that the importance of a column is based only on its vector norm, and this may be unnecessarily restrictive. As a consequence, one can easily construct counterexamples for which this sampling algorithm gets tricked into selecting a sub-optimal index set J: consider for example a $n \times 2$ matrix $A = [Q_1, Q_2 \mathbb{1}_{n-1}]$ where Q is the orthogonal factor of the QR decomposition of a random matrix and $\mathbb{1}_{n-1}$ denotes the vector of length $n-1$ with all components equal to one.

The disadvantage in sampling according to (3) which we introduced in the previous paragraph leads to the main focus of this work: we discuss how sketching proportionally to the **leverage scores** Cohen et al. [2017] will mitigate this issue, while providing additional desirable properties for the underlying probability distribution. In the first section, we define the **classical leverage scores** associated to the columns of A and we prove their main properties and advantages for their application to sketching algorithms: we show that maximum sampling probability is given to columns which are orthogonal to all the others and that the underlying sampling distribution is invariant to the addition of zero columns. While these are desirable properties, they imply that the leverage scores are all equal to one for a full column rank matrix, and hence sketching proportionally to them reduces to uniform sampling, which performs poorly in practice. To mitigate this issue, we study instead the **ridge leverage scores** which arise as a stable and natural regularization of the classical leverage scores that gives proportionally less probability to small principal components. In the last section, we show empirically the advantage of the ridge leverage scores with respect to the other sketching distributions introduced in the above by testing the different sampling strategies on a Hilbert matrix of reduced size. Finally, we draw conclusions about the work and we remark possible improvements of the presented work.

Notation and preliminary results

For the rest of the work we consider a matrix $A \in \mathbb{R}^{m \times n} = [a_1, \dots, a_n]$ with $m \geq n$. Thus, a_i denotes the i-th column of A. We denote by $A = U\Sigma V^T$ the singular value decomposition (SVD) of A and $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$ the singular values of A. We always denote by r the rank of A, namely we assume $\sigma_r > 0$ and $\sigma_{r+1} = 0$. For the sake of completeness, we now recall the main properties of the pseudo-inverse which we use in the elaborate. The Moore-Penrose inverse of A is defined as the unique matrix $P \in \mathbb{R}^{n \times m}$ satisfying the Moore-Penrose conditions:

$$APA = A, \quad PAP = P, \quad (AP)^T = AP, \quad (PA)^T = PA. \tag{4}$$

Given an SVD of A, one can show that $A^\dagger = V\Sigma^\dagger U^T$, where $\Sigma^\dagger = \mathrm{diag}(\sigma_1^{-1}, \dots, \sigma_r^{-1}, 0, \dots, 0)$, where we denote $r = \mathrm{rank}(A)$. Moreover, the following useful properties hold:

$$(A^\dagger)^T = (A^T)^\dagger, \quad (A^\dagger)^\dagger = A, \quad (AA^T)^\dagger = (A^T)^\dagger A^\dagger. \tag{5}$$

For space reasons, we refer to Barata and Hussein [2012] for an exhaustive summary with corresponding proofs of the results above and of the definitions. The best rank-k approximation of A, corresponding to the orthogonal projection onto the k principal left singular vectors $U_k U_k^T A$, is denoted by $\mathcal{T}_k(A)$, where $\mathcal{T}$ denotes the truncation operator and U_k is the restriction of U to its first k columns. Finally, we denote by e_i the i-th canonical basis vector with $(e_i)_j = 1$ if $i = j$, 0 otherwise.

In the context of sampling algorithms, it is always implicitly assumed that the k sampled columns of A contained in C according to some probability distribution p defined for $i \in [n]$ $\{a_{j_i}\}_{i=1}^k$ are properly re-scaled according to $a_{j_i} = a_{j_i}/\sqrt{k p_{j_i}}$, as to guarantee the unbiased estimate $\mathbb{E}\left[CC^T\right] = AA^T$.

2 Leverage Scores

Leverage scores are used ubiquitously as importance sampling probabilities for matrix sketching. In the literature, they are often introduced as the solution to a least squares problem. We find that this interpretation gives a nice insight, hence this is also our starting point.

Definition 2.1. *(Leverage score). The leverage score associated with the i-th column of A, which we denote as $l_i(A)$ for the rest of the work, is the solution to the problem*

$$\min_y \|y\|_2^2 \quad \textit{such that } a_i = Ay. \tag{6}$$

Lemma 2.1. *For $i \in [n]$ it holds that*

$$l_i(A) = a_i^T (AA^T)^\dagger a_i. \tag{7}$$

Proof. The solution $\hat{y}$ to problem (6) satisfies

$$\hat{y} \in \text{argmin}_y \|Ay - a_i\|_2^2, \tag{8}$$

and is the minimum norm solution to such a problem, in the sense that for any $\tilde{y}$ for which $\|A\tilde{y} - a_i\|_2^2$ is minimized, we have $\|\hat{y}\|_2 \leq \|\tilde{y}\|_2$. It is a well known fact, see for example Barata and Hussein [2012], that the minimum norm solution to (8) is given by $\hat{y} = A^\dagger a_i$, for which

$$\|\hat{y}\|_2^2 = a_i^T (A^T)^\dagger A^\dagger a_i = a_i^T (AA^T)^\dagger a_i. \tag{9}$$

In the last equality of (9), we used the fact that for any matrix A, $B = A^T$ is a sufficient condition for $(AB)^\dagger = B^\dagger A^\dagger$ to hold true. Furthermore, the constraint $A\hat{y} = a_i$ is satisfied because we have

$$A\hat{y} = AA^\dagger a_i = AA^\dagger Ae_i \overset{(*)}{=} Ae_i = a_i,$$

where at $(*)$ we used the fact that by definition $AA^\dagger A = A$. □

Remark 2.1. *An alternative way to see that the constraint is satisfied is to notice that the residual $\|Ay - a_i\|_2 = 0$ when taking $y = e_i$, being e_i the i-th canonical basis vector. Hence, this must hold true also for the minimum norm solution to* (8) *$\hat{y}$, as $\|A\hat{y} - a_i\|_2 \leq \|Ae_i - a_i\|_2 = 0$.*

Definition 2.2. *Let $A \in \mathbb{R}^{m\times n}$. We define the **hat matrix** of A as $H = A^T(AA^T)^\dagger A$.*

Remark 2.2. *We can simplify the expression for the **hat matrix** by noticing that $A^T(AA^T)^\dagger = A^\dagger$. This can be verified by showing that $A^T(AA^T)^\dagger$ satisfies all the properties of the pseudoinverse. Details can be found in the appendix. Most importantly, we recognize the hat matrix $H = A^\dagger A$ to be the orthogonal projector onto* range(A^T)*, the row-space of A.*

Lemma 2.2. *The hat matrix H is idempotent, i.e. $H^2 = H$, and symmetric, i.e. $H = H^T$.*

Proof. To show that H is idempotent, we compute directly that

$$H^2 = A^\dagger AA^\dagger A = A^\dagger A = H,$$

where we use the fact that $A^\dagger AA^\dagger = A^\dagger$ by definition of the Moore-Penrose inverse. To show that H is symmetric, let us take $A = U\Sigma V^T$ the singular value decomposition of A. Then

$$\begin{aligned} H^T = (A^\dagger A)^T &= (V\Sigma^\dagger U^T U\Sigma V^T)^T \\ &= (V\Sigma^\dagger \Sigma V^T)^T && (U^T U = I_m) \\ &= V(\Sigma^\dagger \Sigma)^T V^T \\ &= V\Sigma^\dagger \Sigma V^T && (\Sigma^\dagger\Sigma \text{ is diagonal}) \\ &= V\Sigma^\dagger U^T U\Sigma V^T = A^\dagger A = H && (U^T U = I_m). \end{aligned}$$

Thus, the claim follows. □

Lemma 2.3. *For $i \in [n]$, it holds that $l_i(A) \leq 1$.*

Proof. Note that the leverage scores $\{l_i(A)\}_{i=1}^n$ are the diagonal entries of the hat matrix H, because we can write

$$l_i(A) = a_i^T (AA^T)^\dagger a_i = e_i^T A^T (AA^T)^\dagger Ae_i = e_i^T He_i = [H]_{ii}.$$

Therefore, we compute directly

$$l_i(A) = [H]_{ii} = e_i^T He_i \overset{CS}{\leq} \|e_i\|_2 \|He_i\|_2 \leq \|e_i\|_2 \|H\|_2 \|e_i\|_2 = \|H\|_2 \leq 1.$$

Thus, $0 \leq l_i(A) \leq 1$ for $i \in [n]$. In the first inequality, we used the fact that $l_i(A) \geq 0$ by construction and the Cauchy-Schwarz inequality. In the second inequality, we used the fact that $\|Hx\|_2 \leq \|H\|_2\|x\|_2$ for any $x \in \mathbb{R}^n$. In the last equality, we used the fact that $\|e_i\|_2 = 1$. In the last inequality, we used that for any orthogonal projector P, it holds that $\|P\|_2 \leq 1$ Golub and Van Loan [2013]. Indeed, we have

$$\|P\|_2^2 = \max_{v:\|v\|_2=1} |\langle Pv, Pv\rangle| = \max_{v:\|v\|_2=1} |v^T P^T Pv| \overset{(*)}{=} \max_{v:\|v\|_2=1} |v^T Pv| \overset{CS}{\leq} \|v\|_2 \|P\|_2 \|v\|_2 = \|P\|_2,$$

where at $(*)$ we used that $P^T P = P^2 = P$. Thus, $\|P\|_2^2 \leq \|P\|_2$ or equivalently $\|P\|_2 \leq 1$. □

Remark 2.3. *By construction of the problem* (6) *which is solved by the i-th leverage score, we can alternatively prove that* $l_i(A) \leq 1$ *by noting that*

$$l_i(A) = \underset{y: Ay = a_i}{\operatorname{argmin}} \ \|y\|_2^2 \leq \|e_i\|_2^2 = 1,$$

where e_i *is the i-th canonical basis vector. However, we preferred the previous proof as the representation of the leverage scores through the diagonal elements of the hat matrix will often turn out to be useful throughout the work.*

Similarly to the sampling distribution (3) for sketching, the sampling distribution induced by the leverage scores (7) remains unchanged under addition of zero columns, because the leverage score for the latter will be trivially zero. This also implies that the normalization constant $\sum_{i=1}^n l_i(A)$ should be invariant, as it happens for the Frobenius norm $\|A\|_F^2$ for the sketching distribution given by (3). We now show that we can compute directly the normalization constant, exploiting again the fact that H is a projection matrix.

Lemma 2.4. *It holds that* $\sum_{i=1}^n l_i(A) = rank(A) \leq n$.

Proof. Let r denote the rank of A. We notice first that $\sum_{i=1}^n l_i(A) = \mathrm{Tr}(H)$. Letting $A = U\Sigma V^T$ be the SVD of A, this gives

$$\begin{aligned}
\mathrm{Tr}(H) &= \mathrm{Tr}(A^\dagger A) \\
&= \mathrm{Tr}(V\Sigma^\dagger U^T U\Sigma V^T) && (\text{if } A = U\Sigma V^T \text{ then } A^\dagger = V\Sigma^\dagger U^T) \\
&= \mathrm{Tr}(V\Sigma^\dagger \Sigma V^T) && (U^T U = I_m) \\
&= \mathrm{Tr}(V^T V\Sigma^\dagger \Sigma) && (\text{trace invariance under cyclic permutations}) \\
&= \mathrm{Tr}(\Sigma^\dagger \Sigma) && (V^T V = I_n) \\
&= \mathrm{rank}(A) && (\Sigma^\dagger = \mathrm{diag}(\sigma_1^{-1}, \ldots, \sigma_r^{-1}, 0, \ldots, 0)).
\end{aligned}$$

Thus, the claim follows. □

With a slight abuse of notation, for the rest of the work we implicitly assume the probability distribution induced by $\{l_i(A)\}_{i=1}^n$ to be normalized by $r = \mathrm{rank}(A)$, unless specified otherwise.

Lemma 2.5. *Let* $A = U\Sigma V^T$ *be the singular value decomposition of A. Furthermore, denote by* V_k *be the restriction of* V *to its first k columns. If A has rank k, then* $l_i(A) = \|(V_k^T)_i\|_2^2$, *where* $(V_k^T)_i$ *denotes the i-th row of* V_k.

Proof. If A has rank k it coincides with its k-truncated SVD, namely $A = U_k\Sigma_k V_k^T$, where $U_k \in \mathbb{R}^{m\times k}, \Sigma_k \in \mathbb{R}^{k\times k}$ and $V_k \in \mathbb{R}^{n\times k}$. This is because $\sigma_{k+1} = \ldots = \sigma_n = 0$. Thus, we compute

$$\begin{aligned}
l_i(A) &= a_i^T (AA^T)^\dagger a_i \\
&= e_i^T A^\dagger A e_i && A^T(AA^T)^\dagger = A^\dagger \\
&= e_i^T V_k \Sigma_k^\dagger U_k^T U_k \Sigma_k V_k^T e_i && (\text{if } A = U_k\Sigma_k V_k^T \text{ then } A^\dagger = V_k\Sigma_k^\dagger U_k^T) \\
&= e_i^T V_k \Sigma_k^\dagger \Sigma_k V_k^T e_i && (U_k^T U_k = I_k) \\
&= (V_k^T e_i)^T (V_k^T e_i) = \|(V_k^T)_i\|_2^2. && (\Sigma_k^\dagger = \Sigma_k^{-1} \text{ because } A \text{ has rank } k)
\end{aligned}$$

Therefore, the claim follows. □

Remark 2.4. *Lemma 2.5 implies that, in the special case in which A has rank k, the sampling probability induced by* (7) *coincides with*

$$p_i = \|(V_k^T)_i\|_2^2/k, \quad i = 1, \ldots, n. \tag{10}$$

Drineas et al. [2008] have shown that sampling $\mathcal{O}(k^2 \log(1/\delta)\epsilon^{-2})$ *columns according to* (10) *gives*

$$\|A - QQ^T A\|_2 \leq (1+\epsilon)\|A - \mathcal{T}_k(A)\|_2 \tag{11}$$

with probability $1-\delta$, *where* $\mathcal{T}_k(A)$ *is the best rank-k approximation of A and Q contains an orthogonal basis for the column space of the sampled columns matrix.*

Lemma 2.6. *If a column* a_i *is orthogonal to all the other columns then* $l_i(A) = 1$.

Proof. Let us assume that $a_i \neq 0$, so that $\|a_i\|_2 \neq 0$. If $a_i = 0$, we trivially have $l_i(A) = 0$ which is inconsistent with the claim. Because a_i is orthogonal to all the other columns, $u_i = \frac{a_i}{\|a_i\|_2}$ is a left singular vector of A. To see this, we show equivalently that u_i is an eigenvector of AA^T with eigenvalue $\sigma_i^2 = \|a_i\|_2^2$. We have

$$AA^T u_i = AA^T \frac{a_i}{\|a_i\|_2} = A \begin{bmatrix} — & a_1 & — \\ & \dots & \\ — & a_i & — \\ & \dots & \\ — & a_n & — \end{bmatrix} \frac{a_i}{\|a_i\|_2} = A \begin{bmatrix} 0 \\ \dots \\ 0 \\ \frac{a_i^T a_i}{\|a_i\|_2} \\ 0 \\ \dots \\ 0 \end{bmatrix} = A \begin{bmatrix} 0 \\ \dots \\ 0 \\ \|a_i\|_2 \\ 0 \\ \dots \\ 0 \end{bmatrix}$$

where in the third equality we used the fact that a_i is orthogonal to all the other columns of A (and hence, rows of A^T). We conclude that

$$AA^T u_i = a_i \|a_i\|_2 = \|a_i\|_2^2 u_i,$$

which shows the claim. Up to permutations of the columns of A we can assume without loss of generality that $i = 1$. Thus, we write

$$A = \begin{bmatrix} | & \\ a_i & B \\ | & \end{bmatrix},$$

where $B = [a_2, \dots, a_n] \in \mathbb{R}^{m \times n-1}$. Now, if $\tilde{U}\tilde{\Sigma}\tilde{V}^T$ denotes the economy sized SVD of B, namely with $\tilde{U} \in \mathbb{R}^{m \times n-1}, \tilde{\Sigma} \in \mathbb{R}^{n-1 \times n-1}$ and $\tilde{V} \in \mathbb{R}^{n-1 \times n-1}$, we can derive an SVD for A as

$$\begin{aligned} A = \begin{bmatrix} | & \\ a_i & B \\ | & \end{bmatrix} &= \begin{bmatrix} | & \\ u_i \|a_i\|_2 & \tilde{U}\tilde{\Sigma}\tilde{V}^T \\ | & \end{bmatrix} = \begin{bmatrix} | & \\ u_i & \tilde{U} \\ | & \end{bmatrix} \begin{bmatrix} \|a_i\|_2 & 0 \\ & \\ 0 & \tilde{\Sigma}\tilde{V}^T \end{bmatrix} \\ &= \begin{bmatrix} | & \\ u_i & \tilde{U} \\ | & \end{bmatrix} \begin{bmatrix} \|a_i\|_2 & 0 \\ & \\ 0 & \tilde{\Sigma} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ & \\ 0 & \tilde{V}^T \end{bmatrix} = \begin{bmatrix} | & \\ u_i & \tilde{U} \\ | & \end{bmatrix} \begin{bmatrix} \|a_i\|_2 & 0 \\ & \\ 0 & \tilde{\Sigma} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ & \\ 0 & \tilde{V} \end{bmatrix}^T := U\Sigma V^T \end{aligned}$$

Let us denote $k = \text{rank}(A)$. To conclude, we note that the elements of the matrix Σ are not yet necessarily in non-increasing order. However, because we assumed that $a_i \neq 0$, we have $\|a_i\|_2 > 0$, which equivalently means that if we order the elements of Σ in non-increasing order, such a singular value will belong in the top k diagonal entries of Σ. Similarly, u_i will belong to the k-th left principal subspace of A and the first column of V to the k-th right principal subspace of A. Hence, when considering the truncated SVD of A given by $U_k \Sigma_k V_k^T$, the canonical vector e_1 will belong to the columns of V_k. Using Lemma 2.5, we conclude that $l_i(A) = \|(V_k^T)_i\|_2^2 = \|(1, 0, \dots, 0)\|_2^2 = 1$. Thus, the claim follows. □

Therefore, if a column a_i is orthogonal to all other columns (and not identically zero), its leverage score is maximum. Thus, the **classical leverage score** (7) is a measure of the importance of a_i in the range of A. This is a desirable feature for the sampling distribution we aim at defining because removing such a column would decrease the column rank of A, completely changing its column space.

Remark 2.5. *In the proof of Lemma 2.6, we assumed without loss of generality and for simplicity of exposition that a_i was corresponding to the first column of A. If this is not the case, one can repeat a similar proof by taking an SVD of the first $i-1$ columns of A given by $U_i^- \Sigma_i^- (V_i^-)^T$ and an SVD of the last $n-i-1$ columns of A given by $U_i^+ \Sigma_i^+ (V_i^+)^T$. Using this, one can derive an SVD for A as*

$$\begin{bmatrix} | & & \\ U_i^- & u_i & U_i^+ \\ | & & \end{bmatrix} \begin{bmatrix} \Sigma_i^- & 0 & 0 \\ 0 & \|a_i\|_2 & 0 \\ 0 & 0 & \Sigma_i^+ \end{bmatrix} \begin{bmatrix} V_i^- & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & V_i^+ \end{bmatrix}^T := U\Sigma V^T.$$

However, note that in principle the diagonal elements of Σ are not in decreasing order. In particular, because we assumed $a_i \neq 0$, then $\sigma_i = \|a_i\|_2 > 0$ which implies that for an ordered SVD u_i is in the top $k = \text{rank}(A)$ columns of U, $\|a_i\|_2 > 0$ is in the top k entries of Σ and, finally, $(V^T)_i$ is in the top k rows of V^T. Thus, $(V^T)_i$ belongs to the rows of V_k when applying to A the $\mathcal{T}_k(\cdot)$ truncation operator. Thus, we can conclude also in this case from Lemma 2.5 that $l_i(A) = 1$.

Remark 2.6. *In the introduction, we claimed that the main drawback of sampling proportionally to the columns norm* (3) *is the fact that the importance sampling probability for a column is just given by its norm, while ideally the sampling probability of a column should also measure its contribute in the column space of A. The previous result shows precisely that the leverage scores do satisfy this property; namely, the i-th leverage score takes into account the importance of the column a_i in composing the range of A, for $i \in [n]$. As a consequence of this, we naturally expect sketching proportionally to* (7) *to be superior to sampling from* (3)*.*

Definition 2.3. *The maximum leverage score $c(A) = \max_{i\in[n]} l_i(A)$ over the columns of A is defined as the* ***coherence*** *of A.*

Remark 2.7. *If there is a column a_i which is orthogonal to all the others, then A contains at least one column whose removal would significantly affect the composition of A's column space, and hence A is said to have maximum coherence. Thus, sampling accordingly to the column norms* (3) *may not capture such a feature of A, while the leverage scores provide additional information on the coherence of a matrix.*

3 Ridge Leverage Scores

Combining Lemma 2.3 and Lemma 2.4 from the previous section, we conclude that if A has full column rank n we have $0 \leq l_i(A) \leq 1$ for $i \in [n]$ and $\sum_{i=1}^{n} l_i(A) = n$, which can hold true only if $l_i(A) = 1$ for all i. Therefore, the leverage score sampling reduces to uniform sampling of the columns of A with probability $\frac{1}{n}$, which performs poorly in practice. As a first solution, inspired by the result of Lemma 2.5, one can compute rank-k subspace leverage scores McCurdy [2018].

Definition 3.1. *(Rank-k subspace leverage scores). Let $A \in \mathbb{R}^{m\times n}$ and $\mathcal{T}_k(A)$ its best rank k approximation obtained by projecting onto the k principal left singular vectors. The rank-k subspace leverage scores associated to the columns of A are defined for $i \in [n]$ as*

$$l_i^k(A) = a_i^T(\mathcal{T}_k(A)\mathcal{T}_k(A)^T)^\dagger a_i. \tag{12}$$

Remark 3.1. *By substituting A with its truncated SVD $\mathcal{T}_k(A)$ in* (7) *and repeating the same exact calculations carried out in Lemma* (2.5)*, we recover an expression for the rank-k subspace leverage scores given by $l_i^k(A) = \|(V_k^T)_i\|_2^2$.*

In (12), the classical leverage scores introduced in the previous section are modified to only capture how important each column a_i is in composing the top k singular directions of the range of A. This procedure aims at mitigating the small singular values components of AA^T in the classical leverage scores, which are selected with probability $1/n$ otherwise, by completely omitting them. However, note that in principle the best rank-k approximation $\mathcal{T}_k(A)$ may not even be well defined (for the Frobenius norm it is not unique if for example $\sigma_k = \sigma_{k+1}$). Furthermore, even if the latter $\mathcal{T}_k(A)$ is unique, it can be sensitive to matrix perturbations, so the scores can drastically change when A is slightly modified or when the algorithm is not given access to a full knowledge of A. As a consequence, Cohen et al. [2017] argue that the rank-k subspace leverage scores are not stable and, instead, propose to regularize the problem as a more natural solution.

Definition 3.2. *(Ridge leverage scores). Let $A \in \mathbb{R}^{m\times n}$ and let $\lambda > 0$ be a suitable regularization parameter. The ridge leverage scores associated to the columns of A are defined for $i \in [n]$ as*

$$l_{i,\lambda}(A) = a_i^T(AA^T + \lambda^2 I)^{-1} a_i. \tag{13}$$

Remark 3.2. *Note that in* (13)*, we can take the inverse of $(AA^T + \lambda^2 I)$ instead of the Moore-Penrose pseudoinverse of the latter. This is because by regularizing the problem, all the singular values of $AA^T + \lambda^2 I$ are at least λ^2. Furthermore, we recover the orthogonal projector onto* $\mathrm{range}(A^T)$ *in the limit as $\lambda \to 0$. Indeed note that, by definition*

$$A^T(AA^T)^\dagger = A^\dagger := \lim_{\lambda\to 0} A^T(AA^T + \lambda^2 I)^{-1}.$$

In particular, if we always denote $r = \mathrm{rank}(A)$, it holds that Golub and Van Loan [2013]

$$\|A^\dagger - A^T(AA^T + \lambda^2 I)^{-1}\|_2 \leq \frac{\lambda^2}{\sigma_r(\sigma_r^2 + \lambda^2)}. \tag{14}$$

Hence, we also have

$$\begin{aligned}
|l_{i,\lambda}(A) - l_i(A)| &= |e_i^T A^T(AA^T + \lambda^2 I)^{-1} A e_i - e_i^T A^\dagger A e_i| \\
&= |e_i^T[A^T(AA^T + \lambda^2 I)^{-1} A - A^\dagger A]e_i| \overset{CS}{\leq} \|e_i\|_2 \|A^T(AA^T + \lambda^2 I)^{-1} A - A^\dagger A\|_2 \|e_i\|_2 \\
&= \|A^T(AA^T + \lambda^2 I)^{-1} A - A^\dagger A\|_2 \leq \|A^T(AA^T + \lambda^2 I)^{-1} - A^\dagger\|_2 \|A\|_2 \\
&\leq \frac{\sigma_1 \lambda^2}{\sigma_r(\sigma_r^2 + \lambda^2)} \to 0 \quad \textit{as } \lambda \to 0,
\end{aligned}$$

where in the first inequality we used the Cauchy-Schwarz inequality and in the last inequality we used (14)*. Thus, the classical leverage scores* (7) *are recovered in the limit as* $\lambda \to 0$ *of the ridge leverage scores* (13)*. Note that, the factor* $\frac{\sigma_1}{\sigma_r}$ *corresponds to the condition number of the matrix. Thus, while the result holds true in general, for badly conditioned matrices the convergence may be slower.*

Remark 3.3. *Intuitively, sampling by ridge leverage scores* (13) *is equivalent to sampling by classical leverage scores* (7) *for the matrix* $[A, \lambda I_n]$*. However, samples are only taken among the columns of* A *and not among the ones of the identity matrix. One can show a derivation for* (13) *similar to the one proved for the classical leverage scores in Lemma 2.1 and involving a regularized version of the least squares problem* (8)*. For space reasons, we omit such a discussion, referring the interested reader to El Alaoui and Mahoney [2014].*

Ridge leverage scores take a different approach to mitigate the small singular values components of AA^T compared to the rank-k leverage scores (12): they only diminish the importance of small principle components through regularization, instead of completely omitting them. We give an intuition for this claim by deriving a simpler expression for the ridge leverage scores.

Lemma 3.1. *Let* $A = U\Sigma V^T$ *be the singular value decomposition of* A*. For* $i \in [n]$ *it holds that*

$$l_{i,\lambda}(A) = (V^T e_i)^T \operatorname{diag}\left(\frac{\sigma_1^2}{\sigma_1^2+\lambda^2}, \ldots, \frac{\sigma_n^2}{\sigma_n^2+\lambda^2}\right) V^T e_i. \tag{15}$$

Proof. We compute

$$\begin{aligned}
l_{i,\lambda}(A) &= a_i^T (AA^T + \lambda^2 I)^{-1} a_i \\
&= e_i^T V\Sigma^T U^T (U\Sigma\Sigma^T U^T + \lambda^2 I)^{-1} U\Sigma V^T e_i \\
&= e_i^T V\Sigma^T U^T (U\Sigma\Sigma^T U^T + U\lambda^2 I U^T)^{-1} U\Sigma V^T e_i && (UU^T = I_m) \\
&= e_i^T V\Sigma^T U^T U(\Sigma\Sigma^T + \lambda^2 I)^{-1} U^T U\Sigma V^T e_i && (U^T U = I_m) \\
&= e_i^T V\Sigma^T (\Sigma\Sigma^T + \lambda^2 I)^{-1} \Sigma V^T e_i && (U^T U = I_m) \\
&= (V^T e_i)^T \operatorname{diag}\left(\frac{\sigma_1^2}{\sigma_1^2+\lambda^2}, \ldots, \frac{\sigma_n^2}{\sigma_n^2+\lambda^2}\right) V^T e_i.
\end{aligned}$$

□

The perturbation of AA^T with $\lambda^2 I$ allows to regularize the problem and can mitigate the issue presented at the beginning of the section for full column rank matrices. We illustrate this claim by looking at the following example: consider the case of a $m \times n$ matrix A (always with $m \geq n$) with a large gap between the k-th and $(k+1)$-th singular values, and choose λ such that

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k \gg \lambda \gg \sigma_{k+1} \geq \ldots \geq \sigma_n > 0.$$

Because $\sigma_n > 0$, A has full column rank so we already argued previously that $l_i(A) = 1$ for all $i \in [n]$. Hence, sampling from the classical leverage scores (7) is equivalent to uniformly sample columns of A, which performs poorly. On the other hand, for such a regularization parameter λ, the diagonal entries of the matrix

$$\Sigma^{2,\lambda} = \operatorname{diag}\left(\frac{\sigma_1^2}{\sigma_1^2+\lambda^2}, \ldots, \frac{\sigma_n^2}{\sigma_n^2+\lambda^2}\right) \tag{16}$$

are satisfying $\Sigma_{ii}^{2,\lambda} \approx 0$ for $i \geq k+1$ and $\Sigma_{ii}^{2,\lambda} \approx 1$ for $i \leq k$. Therefore

$$l_{i,\lambda}(A) = \sum_{j=1}^{n} \frac{\sigma_j^2}{\sigma_j^2+\lambda^2} v_{ij}^2 \approx \sum_{j=1}^{k} v_{ij}^2 = \|(V_k^T)_i\|_2^2, \tag{17}$$

where V_k denotes the restriction of V to its first k columns. Therefore, sampling proportionally to the ridge leverage scores $l_{i,\lambda}(A)$ is approximately equal to using the distribution (10).

Remark 3.4. *Note that, as we discussed at the beginning of the section, for a similar matrix using the rank-k subspace leverage scores* (12) *will omit completely the small principal components successive to* σ_k*. Instead, with ridge leverage scores sampling we assign to these principal components a non-zero sampling probability, which is decreasing with* λ*.*

From this example, we see that the regularization parameter becomes to all effects a hyper-parameter which needs to be tuned, depending on the desired cutoff effect on the columns of A. A too large value of λ may drastically change the spectral properties of the middle factor AA^T, while a too small λ may not regularize sufficiently the problem, thus recovering the rank-k subspace leverage scores (12). To our knowledge, for a general matrix A, the choice of λ is still an unresolved task. In the pioneering work on ridge leverage scores sampling, Cohen et al. [2015] propose a wise choice of λ given by $\lambda^2 = \frac{\|A-\mathcal{T}_k(A)\|_F^2}{k}$.

Remark 3.5. *Note that while $\mathcal{T}_k(A)$, used to compute the rank-k subspace leverage scores* (12)*, may not be uniquely determined, $\lambda = \frac{\|A-\mathcal{T}_k(A)\|_F}{\sqrt{k}}$ is uniquely determined because by the Von-Neumann trace inequality we can lower bound the best squared rank-k approximation error in the Frobenius norm with $\sum_{j=k+1}^n \sigma_j^2$. In particular, while the ridge scores* (13) *with this choice of regularization parameter depend on the value of $\|A-\mathcal{T}_k(A)\|_F^2$, they do not depend on a specific low-rank approximation. This is sufficient for stability since by the Weyl's inequality Marcus [1965] $\|A-\mathcal{T}_k(A)\|_F^2$ changes predictably under matrix perturbations even when $\mathcal{T}_k(A)$ itself does not.*

Note that computing both the rank-k subspace leverage scores (12) and the ridge scores (13) with $\lambda^2 = \frac{\|A-\mathcal{T}_k(A)\|_F^2}{k}$ may be computationally too expensive or even simply impossible in practice. For the former case, having access to the best rank-k approximation of A $\mathcal{T}_k(A)$ is more costly than the rest of the procedure. For the latter case, computing such a value of λ requires knowledge (or an estimation) of the singular values of A $\sigma_{k+1}, \dots, \sigma_n$, which may not be feasible in many situations. Even for a generic choice of λ, computing the ridge leverage scores for all the columns of A may be unfeasible in some applications in which the algorithm does not have full access to the entries of A. Cohen et al. [2017] discuss an iterative algorithm to compute overestimates of the true ridge leverage scores by uniformly sampling at random columns of A. We now show that under the aforementioned choice of λ, the sum of the corresponding ridge scores is not too large.

Lemma 3.2. *Let $\lambda^2 = \frac{\|A-\mathcal{T}_k(A)\|_F^2}{k}$ in the ridge scores* (13) *of the columns of A. Then $\sum_{i=1}^n l_{i,\lambda}(A) \leq 2k$.*

Proof. We rewrite (13) using the SVD decomposition of $A = U\Sigma V^T$. This gives

$$l_{i,\lambda}(A) = a_i^T (U\bar{\Sigma}^{-2}U^T)a_i,$$

where $\bar{\Sigma}_{ii}^{-2} = (\sigma_i^2(A) + \lambda^2)^{-1}$. Then

$$\sum_{i=1}^n l_{i,\lambda}(A) = \mathrm{Tr}(\Sigma^2\bar{\Sigma}^{-2}) = \sum_{i=1}^n \frac{\sigma_i^2(A)}{\sigma_i^2(A) + \frac{\|A-\mathcal{T}_k(A)\|_F^2}{k}}.$$

Now, for $i \leq k$ we can bound the terms in the previous sum by 1. This gives eventually

$$\sum_{i=1}^n l_{i,\lambda}(A) \leq k + \sum_{i=k+1}^n \frac{\sigma_i^2(A)}{\sigma_i^2(A) + \frac{\|A-\mathcal{T}_k(A)\|_F^2}{k}} \leq k + \frac{1}{\frac{\|A-\mathcal{T}_k(A)\|_F^2}{k}} \sum_{i=k+1}^n \sigma_i^2(A) = 2k.$$

□

4 Numerical results

We conclude this survey on the leverage scores by presenting numerical experiments. We start this section by considering a Hilbert matrix of size $n = 100$. Recall that the Hilbert matrix is defined as follows:

$$A_{ij} = \frac{1}{i+j-1}, \quad i,j = 1,\dots n. \tag{18}$$

It is a well-known fact that the Hilbert matrix have exponentially fast decaying of the singular values, which implies that it is not a full column rank matrix and the leverage scores are not all trivially equal. Moreover, these feature makes it suitable for numerical experiments on sketching algorithms. In the sequel, we compare the performance of the three main sampling probabilities which we discussed in the work: uniform sampling, columns norm sampling (3) and ridge leverage scores (13) with $\lambda = 1e-4$. The value of λ was tuned by experimental evidence. However, the value of λ had a little influence on the performance.

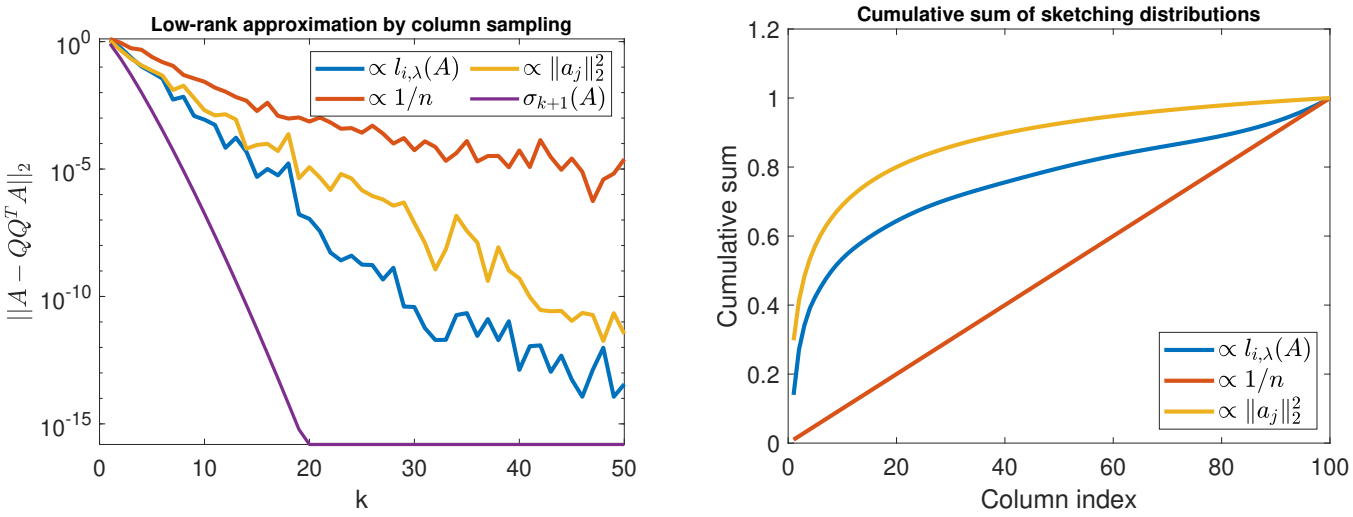

Figure 1: Performance of sketching strategies for the low-rank approximation of a Hilbert matrix of size $n = 100$. Approximations errors for all the sampling strategies (uniform sampling, columns norm sampling and ridge leverage score sampling) are computed as $\|A - QQ^TA\|_2$, where Q is the orthogonal factor of the reduced QR decomposition of the sampled columns matrix C, and compared with the gold standard for a rank k approximation given by $\sigma_{k+1}(A)$. Results are averaged over 20 runs. The figure on the right shows the cumulative sum of the sketching distributions: by construction, the Frobenius norm of A is concentrated in the very first columns, while the ridge leverage scores assign higher probabilities to the last columns.

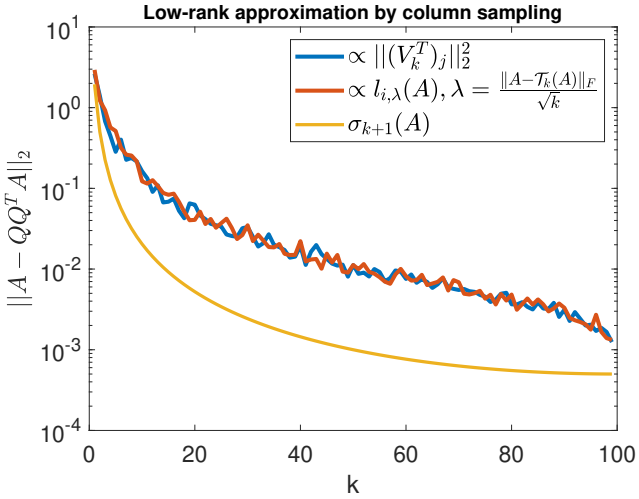

Figure 2: Comparison between sketching algorithms using the rank-k subspace leverage scores (12) and the ridge scores (13) with adaptive regularization parameter $\lambda^2 = \frac{\|A-\mathcal{T}_k(A)\|_F^2}{k}$ on the matrix (19) of size 100×100. Approximations errors for the two sampling strategies are computed as $\|A - QQ^TA\|_2$, where Q is the orthogonal factor of the reduced QR decomposition of the sampled columns matrix C, and compared with the gold standard for a rank k approximation given by $\sigma_{k+1}(A)$. Results are averaged over 20 runs.

We recall for the sake of clarity that the random column submatrix C is constructed by sampling k columns of A given by $[a_{j_1}, \dots, a_{j_k}]$ independently and with replacement according to a probability distribution p, and by successively rescaling the sampled columns with the factor $1/\sqrt{kp_{j_i}}$ as to guarantee that $\mathbb{E}\left[CC^T\right] = AA^T$.

Figure 1 (left) shows the low-rank approximation error using the three different strategies for ranks $k = 1, \dots, 50$, and their comparison to the gold standard given by the $(k+1)$-th singular value of A $\sigma_{k+1}(A)$. As these strategies are randomized, the results are averaged over 20 runs for all considered ranks. The projection error in the spectral norm is measured using QQ^T, the orthogonal projector onto the column space of C, where Q is computed from the reduced QR factorization of C. We see that ridge leverage score sampling (13) is outperforming both columns norm sampling (3) and uniform column sampling and the hierarchy in the quality of the approximation follows the outline of the work: uniform sampling performs poorly and remains far away from the gold standard, because all the columns of A are given the same probability. Columns norm sampling improves majorly the results, but ridge

leverage scores are always yielding a better column subset for approximating the range of A.

To understand the gap in the performance between columns norm sampling and ridge leverage scores, we show in Figure 1 (right) the cumulative sum of the scores obtained with the two strategies. First, we note that they both deviate quite clearly from the uniform distribution, lying on the bisector of the plot. This explains the poor performance of uniform sampling for the Hilbert matrix we take under consideration, and suggests that the Hilbert matrix has high coherence. However, we notice that, coherently with the structure of (18), the cumulative sum of the column norms is saturated in the very first columns, and almost zero sampling probability is given to the last columns of A. On the other hand, the ridge leverage scores are assigning non-negligible sampling probabilities to such columns. This difference in the growth of the cumulative distribution function of the two distributions gives an additional insight on the difference in performance which we observe in Figure 1.

Figure 2 presents evidence for the similarity between the truncated rank-k subspace leverage scores (which are equivalent to the rows norm of V_k (10)) and the ridge leverage scores (13) with the adaptive regularization parameter $\lambda^2 = \frac{\|A-\mathcal{T}_k(A)\|_F^2}{k}$. For this experiment, we picked instead a matrix A of size 100 with entries defined by

$$A_{ij} = \exp(-|i-j|/1000), \tag{19}$$

which has a very slow singular value decay. As this matrix has full column rank, all its columns have leverage score equal to one. The figure shows that the approximation error induced by the two sampling strategies is on average the same for all ranks $k = 1, \dots, 100$.

5 Conclusions

In this work, we started from the main purpose of defining a powerful sampling probability for the random column subset selection task of a generic matrix $A \in \mathbb{R}^{m\times n}$. To this aim, we introduced the classical leverage scores (7) associated to the columns of A as a measure of the importance of each column in composing the column space of A. We argued that the leverage scores are superior with respect to the columns norms for sketching because they maintains some nice features of the latter while doing coherence-revealing for the matrix at hand, because the leverage score is maximum for a column which is orthogonal to all the others. Motivated by their poor applicability in this last case (as sketching reduces to uniform sampling), we introduced the ridge leverage scores (13) as the classical leverage scores resulting from a regularization of the original problem. We argue that they are able to assign low sampling probabilities to small principal components, while providing sketching distributions which are stable under matrix perturbations. The numerical results support our claim and show their superior behaviour in the low-rank approximation of a Hilbert matrix of size 100.

References

J. C. A. Barata and M. S. Hussein. The Moore–Penrose pseudoinverse: A tutorial review of the theory. *Brazilian Journal of Physics*, 42(1):146–165, 2012.

M. B. Cohen, Y. T. Lee, C. Musco, C. Musco, R. Peng, and A. Sidford. Uniform sampling for matrix approximation. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 181–190, 2015.

M. B. Cohen, C. Musco, and C. Musco. Input sparsity time low-rank approximation via ridge leverage score sampling. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1758–1777. SIAM, 2017.

P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Relative-error cur matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, 2008.

A. El Alaoui and M. W. Mahoney. Fast randomized kernel methods with statistical guarantees. *stat*, 1050:2, 2014.

G. H. Golub and C. F. Van Loan. *Matrix computations*. JHU press, 2013.

M. Marcus. Harnack's and weyl's inequalities. *Proceedings of the American Mathematical Society*, 16(5):864–866, 1965.

S. McCurdy. Ridge regression and provable deterministic ridge leverage score sampling. *Advances in Neural Information Processing Systems*, 31, 2018.

J. A. Tropp. Column subset selection, matrix factorization, and eigenvalue optimization. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pages 978–986. SIAM, 2009.