Supplementary Material for  
International Conference on Machine Learning (ICML 2018)

Anonymous Authors

Lemma 1 (covariance error to projection error (modified)).
\[ \|A - \pi_B^k(A)\|_F^2 \leq \|A - [A]_k\|_F^2 + 2k \cdot \|A^T A - B^T B\|_2. \]

Proof. For any \( x \) with \( \|x\| = 1 \), we have
\[ \|Ax\|^2 - \|Bx\|^2 = \|x^T (A^T A - B^T B) x\| \leq \|A^T A - B^T B\|_2 \]
Let \( u_i \) and \( w_i \) be the \( i \)th right singular vector of \( B \) and \( A \) respectively
\[ \|A - \pi_B^k(A)\|_F^2 = \|A\|_F^2 - \|\pi_B^k(A)\|_F^2 \]
\[ = \|A\|_F^2 - \sum_{i=1}^k \|Au_i\|^2 \quad \text{Pathagorean theorem} \]
\[ \leq \|A\|_F^2 - \sum_{i=1}^k \|Bu_i\|^2 + k \cdot \|A^T A - B^T B\|_2 \]
by Eq. (1)
\[ \leq \|A\|_F^2 - \sum_{i=1}^k \|Bu_i\|^2 + k \cdot \|A^T A - B^T B\|_2 \]
because \( \sum_{i=1}^k \|Bu_i\|^2 \leq \sum_{i=1}^k \|Bu_i\|^2 \)
\[ \leq \|A\|_F^2 - \sum_{i=1}^k \|Au_i\|^2 + 2k \cdot \|A^T A - B^T B\|_2 \]
by Eq. (1)
\[ = \|A - [A]_k\|_F^2 + 2k \cdot \|A^T A - B^T B\|_2. \]
\[ \square \]

Row sampling.

Theorem 1. For any \( A \in \mathbb{R}^{n \times d} \) and \( F > 0 \), we sample each row \( A_i \) with probability \( p_i = \frac{\|A_i\|_F^2}{\alpha F^2} \); if it is sampled, scale it by \( 1/\sqrt{p_i} \). Let \( B \) be the (rescaled) sampled rows, then w.p. \( 0.99 \), \( \|A^T A - B^T B\|_2 \leq 10\alpha \sqrt{F} \|A\|_F \), and \( \|B\|_F \leq 10 \|A\|_F \). The expected number of rows sampled is \( O(\|A\|_F^2) \).

Proof. Since spectral norm is no larger than the Frobenius norm, it is sufficient to prove \( \|A^T A - B^T B\|_2 \leq 10\alpha \sqrt{F} \|A\|_F \).
For each \( j \in [n] \), let
\[ x_j = \begin{cases} 1 & \text{if the } j \text{th rows of } A \text{ is sampled} \\ 0 & \text{otherwise.} \end{cases} \]
We have \( (A^T A)_{i,j} = \sum_{t=1}^n a_{t,i}a_{t,j} \), while
\[ (B^T B)_{i,j} = \sum_{t=1}^n \frac{x_t^2 \cdot a_{t,i}a_{t,j}}{p_t}. \]
So \( \mathbb{E}[(B^T B)_{i,j}] = (A^T A)_{i,j} \). We also have
\[ \mathbb{V}[r] = \mathbb{V} \left[ \sum_{t=1}^n \frac{x_t^2 \cdot a_{t,i}a_{t,j}}{p_t} \right] \]
\[ = \sum_{t=1}^n \frac{a_{t,i}^2 a_{t,j}^2 \cdot \mathbb{V}[x_t^2]}{p_t^2} \]
\[ \leq \sum_{t=1}^n \frac{a_{t,i}^2 a_{t,j}^2}{p_t}. \]
where we use the fact \( \mathbb{V}[x_t^2] = p_t (1 - p_t) \leq p_t \). So we have
\[ \mathbb{E} \left[ ((A^T A)_{i,j} - (B^T B)_{i,j})^2 \right] = \mathbb{V}[(B^T B)_{i,j}] \]
\[ \leq \sum_{t=1}^n \frac{a_{t,i}^2 a_{t,j}^2}{p_t}. \]
Therefore,
\[
\mathbb{E} \left[ \| A^T A - B^T B \|_F^2 \right] = \sum_{i,j} \mathbb{E} \left[ \left( (A^T A)_{i,j} - (B^T B)_{i,j} \right)^2 \right]
\]
\[
\leq \sum_{i,j} \sum_{t=1}^n \frac{a^2_t \| A_t \|^2}{p_t}
\]
\[
= \sum_{i=1}^n \| A_i \|^2 \| A_i \|^2
\]
\[
= \sum_{i=1}^n \alpha^2 F \| A_i \|^2 = \alpha^2 F \| A \|^2.
\]

We adjust \( \alpha \) by a constant, and using Markov’s inequality
\[
\Pr \left[ \| A^T A - B^T B \|_F^2 \geq 100 \alpha^2 F \| A \|^2 \right] \leq 0.01,
\]
which is equivalent to
\[
\Pr \left[ \| A^T A - B^T B \|_F \geq 10 \sqrt{F} \| A \|_F \right] \leq 0.01.
\]

The success probability can be boosted by a similar argument as in (2) via McDiarmid’s inequality (see e.g. (2)). It is not hard to verify that
\[
\mathbb{E} \left[ \| B \|_F^2 \right] = \| A \|_F^2.
\]
So by another Markov inequality, we prove the second part. \( \square \)

Input-sparsity time lower rank approximation algorithm.

**Theorem 2** (weak low rank approximation). For any integers \( \ell, d \), given \( A \in \mathbb{R}^{\ell \times d} \), there is an algorithm that uses \( O(\text{nnz}(A) \log(1/\delta)) + O(\ell k^3) \) time and \( O(\ell (k^2 + \log 1/\delta)) \) space, and outputs a matrix \( Z \in \mathbb{R}^{\ell \times k} \) with orthonormal rows such that with probability \( 1 - \delta \), \( \| A - Z^T Z A \|_F^2 \leq O(1) \| A - [A]_k \|_F^2 \).

**Proof.** Let \( J \) be a \( O(k) \times t_1 \) matrix with iid Gaussian random variables, and \( C \) be a \( t_1 \times \ell \) sparse subspace embedding matrix (see (2) for details), with \( t_1 = O(k^2) \). It was proved that, with constant probability, the column space of \( S = AC^T J^T \) contains a \( O(1) \) rank-\( k \) approximation to \( A \) (see e.g., Lemma 4.2 and Remark 4.1 in (2)), moreover \( S = AC^T J^T \) can be computed in time \( O(\text{nnz}(A)) + O(\ell k^3) \). In particular, let \( z_1, \cdots, z_{O(k)} \) be the an orthonormal basis of the column space of \( S \), then there exists \( X \) with \( \text{rank}(X) \leq k \) such that
\[
\| A - Z^T X \|_F^2 \leq O(1) \| A - [A]_k \|_F^2,
\]
where \( Z \) is the matrix whose rows are \( z_1, \cdots, z_{O(k)} \). Hence,
\[
\| A - Z^T Z A \|_F^2 \leq \| A - Z^T Z A \|_F^2 + \| Z^T Z A - Z^T X \|_F^2
\]
\[
= \| A - Z^T X \|_F^2 \leq O(1) \| A - [A]_k \|_F^2.
\]

Note that \( Z \) can be computed from \( S \) with \( O(\ell k^2) \) time. To boost the success probability to \( 1 - \delta \), we repeat the algorithm \( \gamma = \log(1/\delta) \) times, which compute \( Z^{(1)}, \cdots, Z^{(\gamma)} \) and pick the best one. This needs \( O(\text{nnz}(A) \log 1/\delta) + O(\ell k^3) \) time. However, it will take too much time to compute \( A - Z^T Z A \|_F^2 \). To avoid this, we instead just compute a constant approximation using Johnson-Lindenstrauss Transform. Let \( \Phi \in \mathbb{R}^{\ell \times d} \) be a Johnson-Lindenstrauss matrix, where \( t = O(\log(1/\delta)) \). We have \( \Pr[\| \Phi x \| = O(1) \cdot \| x \| \geq 1 - \frac{\delta}{\ell} \) for any fixed \( x \). By union bound, with probability at least \( 1 - \delta \), it holds simultaneously for all \( i \) that
\[
\|(I - Z^{(i)^T} Z^{(i)}) A \Phi^T \|_F^2 = O(1) \| A - Z^{(i)^T} Z^{(i)} A \|_F^2.
\]

Note that \( A \Phi^T \) can be computed in \( O(\text{nnz}(A) \log 1/\delta) \) time. Given this, each \( \|(I - Z^{(i)^T} Z^{(i)}) \Phi^T \|_F^2 \) can be computed in \( O(kt \log 1/\delta) \) time. So the total running time is \( O(\text{nnz}(A) \log 1/\delta) + O(\ell k^3) \). Since each \( Z^{(i)} \) is good with constant probability, with probability at least \( 1 - \delta \), there exists an \( i' \) such that
\[
\| A - Z^{(i')^T} Z^{(i')} A \|_F^2 \leq O(1) \| A - [A]_k \|_F^2.
\]
Hence,
\[
\|(I - Z^{(i')^T} Z^{(i')}) \Phi^T \|_F^2 = O(1) \| A - [A]_k \|_F^2.
\]
Because we pick \( Z^{(j)} \) minimizing
\[
\|(I - Z^{(j)^T} Z^{(j)}) \Phi^T \|_F^2,
\]
with probability \( 1 - \delta \), then
\[
\|(I - Z^{(j)^T} Z^{(j)}) A \|_F^2 = O(1) \|(I - Z^{(j)^T} Z^{(j)}) \Phi^T \|_F^2
\]
\[
\leq \|(I - Z^{(i')^T} Z^{(i')}) \Phi^T \|_F^2
\]
\[
= O(1) \| A - [A]_k \|_F^2,
\]
which proves the correctness.

For space, computing each \( S = AC^T J^T \) and \( Z \) needs \( O(\ell k^2) \) space. We do not store all \( Z^{(i)} \), but compute one at a time. We only need to store the current best at any time, so this does not increase space. We also need to store \( A \Phi \), which takes \( O(\ell \log 1/\delta) \) space. \( \square \)