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

Anonymous Authors[1]

Lemma 1 (covariance error to projection error (modified))**.**

$$\|A - \pi_B^k(A)\|_F^2 \le \|A - [A]_k\|_F^2 + 2k \cdot \|A^TA - B^TB\|_2.$$

Proof. For any x with $\|x\| = 1$, we have

$$\begin{aligned} \big|\ \|Ax\|^2 - \|Bx\|^2\ \big| &= \big|\ x^T(A^TA - B^TB)x\ \big| \\ &\le \|A^TA - B^TB\|_2 \end{aligned} \quad (1)$$

Let u_i and w_i be the ith right singular vector of B and A respectively

$$\begin{aligned} &\|A - \pi_B^k(A)\|_F^2 = \|A\|_F^2 - \|\pi_B^k(A)\|_F^2 \\ &= \|A\|_F^2 - \sum_{i=1}^k \|Au_i\|^2 \text{Pathagorean theorem} \\ &\le \|A\|_F^2 - \sum_{i=1}^k \|Bu_i\|^2 + k \cdot \|A^TA - B^TB\|_2 \\ &\qquad \text{by Eq. (1)} \\ &\le \|A\|_F^2 - \sum_{i=1}^k \|Bw_i\|^2 + k \cdot \|A^TA - B^TB\|_2 \\ &\qquad \text{because } \sum_{i=1}^k \|Bw_i\|^2 \le \sum_{i=1}^k \|Bu_i\|^2 \\ &\le \|A\|_F^2 - \sum_{i=1}^k \|Aw_i\|^2 + 2k \cdot \|A^TA - B^TB\|_2 \\ &\qquad \text{by Eq. (1)} \\ &= \|A - [A]_k\|_F^2 + 2k \cdot \|A^TA - B^TB\|_2. \end{aligned}$$

□

Row sampling.

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

Proof. Since spectral norm is no larger than the Frobenius norm, it is sufficient to prove $\|A^TA - B^TB\|_F \le 10\alpha\sqrt{F}\|A\|_F$.

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 $(A^TA)_{i,j} = \sum_{t=1}^n a_{t,i}a_{t,j}$, while

$$(B^TB)_{i,j} = \sum_{t=1}^n \frac{x_t^2 \cdot a_{t,i}a_{t,j}}{p_t}.$$

So $\mathsf{E}[(B^TB)_{i,j}] = (A^TA)_{i,j}$. We also have

$$\begin{aligned} \mathsf{Var}[(B^TB)_{i,j}] &= \mathsf{Var}\left[\sum_{t=1}^n \frac{x_t^2 \cdot a_{t,i}a_{t,j}}{p_t}\right] \\ &= \sum_{t=1}^n \frac{a_{t,i}^2 a_{t,j}^2 \cdot \mathsf{Var}\left[x_t^2\right]}{p_t^2} \\ &\le \sum_{t=1}^n \frac{a_{t,i}^2 a_{t,j}^2}{p_t}, \end{aligned}$$

where we use the fact $\mathsf{Var}[x_t^2] = p_t(1-p_t) \le p_t$. So we have

$$\begin{aligned} \mathsf{E}\left[\left((A^TA)_{i,j} - (B^TB)_{i,j}\right)^2\right] &= \mathsf{Var}[(B^TB)_{i,j}] \\ &\le \sum_{t=1}^n \frac{a_{t,i}^2 a_{t,j}^2}{p_t}. \end{aligned}$$

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Therefore,

$$\begin{aligned}\mathsf{E}\left[\|A^TA - B^TB\|_F^2\right] &= \sum_{i,j} \mathsf{E}\left[\left((A^TA)_{i,j} - (B^TB)_{i,j}\right)^2\right] \\ &\le \sum_{i,j}\sum_{t=1}^{n} \frac{a_{t,i}^2 a_{t,j}^2}{p_t} \\ &= \sum_{t=1}^{n} \frac{\|A_t\|^2\|A_t\|^2}{p_t} \\ &= \sum_{t=1}^{n} \alpha^2 F \|A_t\|^2 = \alpha^2 F \|A\|_F^2.\end{aligned}$$

We adjust α by a constant, and using Markov's inequality

$$\mathsf{Pr}\left[\|A^TA - B^TB\|_F^2 \ge 100\alpha^2 F\|A\|_F^2\right] \le 0.01,$$

which is equivalent to

$$\mathsf{Pr}\left[\|A^TA - B^TB\|_F \ge 10\alpha\sqrt{F}\|A\|_F\right] \le 0.01.$$

The success probability can be boosted by a similar argument as in (**?**) via McDiarmid's inequality (see e.g. (**?**)).

It is not hard to verify that

$$\mathsf{E}\left[\|B\|_F^2\right] = \|A\|_F^2.$$

So by another Markov inequality, we prove the second part. □

Input-sparsity time lower rank approximation algorithm.

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

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 (**?**) for details), with $t_1 = O(k^2)$. It was proved that, with constant probability, the column space of $S = AC^TJ^T$ contains a $O(1)$ rank-k approximation to A (see e.g., Lemma 4.2 and Remark 4.1 in (**?**)), moreover $S = AC^TJ^T$ can be computed in time $O(\mathsf{nnz}(A)) + O(\ell k^3)$. 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 $\mathsf{rank}(X) \le k$ such that

$$\|A - Z^TX\|_F^2 \le O(1)\|A - [A]_k\|_F^2,$$

where Z is the matrix whose rows are $z_1, \cdots, z_{O(k)}$. Hence,

$$\begin{aligned}\|A - Z^TZA\|_F^2 &\le \|A - Z^TZA\|_F^2 + \|Z^TZA - Z^TX\|_F^2 \\ &= \|A - Z^TX\|_F^2 \\ &\le O(1)\|A - [A]_k\|_F^2.\end{aligned}$$

Note that Z can be computed from S with $O(\ell k^2)$.

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 $O(\mathsf{nnz}(A)\log\frac{1}{\delta}) + \tilde{O}(\ell k^3)$ time. However, it will take too much time to compute $\|A - Z^TZA\|_F^2$. To avoid this, we instead just compute a constant approximation using Johnson-Lindenstrauss Transform. Let $\Phi \in \mathbb{R}^{t\times d}$ be a Johnson-Lindenstrauss matrix, where $t = O(\log(\frac{d}{\delta^2}))$. We have $\mathsf{Pr}[\|\Phi x\| = O(1)\cdot\|x\|] \ge 1 - \frac{\delta^2}{d}$ 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 $O(\mathsf{nnz}(A)\log\frac{d}{\delta})$ time. Given this, each $\|(I - Z^{(i)T}Z^{(i)})A\Phi^T\|_F^2$ can be computed in $O(k\ell\log\frac{d}{\delta})$ time. So the total running time is $O(\mathsf{nnz}(A)\log\frac{d}{\delta}) + \tilde{O}(\ell k^3)$. Since each $Z^{(i)}$ is good with constant probability, with probability at least $1-\delta$, there exists an i' such that

$$\|A - Z^{(i')T}Z^{(i')}A\|_F^2 \le O(1)\|A - [A]_k\|_F^2.$$

Hence,

$$\|(I - Z^{(i')T}Z^{(i')})A\Phi^T\|_F^2 = O(1)\|A - [A]_k\|_F^2.$$

Because we pick $Z^{(j)}$ minimizing

$$\|(I - Z^{(j)T}Z^{(j)})A\Phi^T\|_F^2,$$

with probability $1-\delta$, then

$$\begin{aligned}\|(I - Z^{(j)T}Z^{(j)})A\|_F^2 &= O(1)\|(I - Z^{(j)T}Z^{(j)})A\Phi^T\|_F^2 \\ &\le \|(I - Z^{(i')T}Z^{(i')})A\Phi^T\|_F^2 \\ &= O(1)\|A - [A]_k\|_F^2,\end{aligned}$$

which proves the correctness.

For space, computing each $S = AC^TJ^T$ 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 $O(\ell\log\frac{d}{\delta})$ space.

□