Faster Subspace Embeddings

Recall that a Countsketch matrix is a $k \times n$ matrix $S$ where $k = O(d^2/\varepsilon^2)$ and is constructed as follows: Take a single entry per column of $S$ at random and write $\pm 1$ with equal probability, leave the rest 0.

Let $\text{nnz}(A)$ denote the number of nonzero entries of the (sparse) matrix $A$. We claim that $SA$ can be computed in $O(\text{nnz}(A))$ time with the use of a hash function $h$ that maps the $n$ columns onto the $k$ rows that contain their nonzero entries: indeed, iterating trough nonzero entries of $A$ we note that the contribution of $A_{ij}$ to the product $SA$ is adding $S_{ki}A_{ij}$ to the $(SA)_{kj}$ for all $k,j$, and since $S_{ki} = 0 \iff k \neq h(i)$ it follows that in fact $A_{ij}$ only contributes to $(SA)_{h(i)j}$.

Simple Proof [Nguyen]

We will show that $S$ is a good sketching matrix with high probability. As before we may assume that the columns of $A$ are orthogonal. It suffices to show $\|SAx\|_2 = 1 \pm \varepsilon$ for all unit vectors $x$. For the regression problem we may apply $S$ to $[A, b]$. Write $SA$ as a $6d^2/(\delta \varepsilon^2) \times d$ matrix for some $\delta$.

It suffices to show $\|A^T S^T S A - I\|_F \leq \varepsilon$ as $\|A^T S^T S A - I\|_2 \leq \|A^T S^T S A - I\|_F$, where $\| \cdot \|_F$ denotes the Frobenius norm (square root of sum of squares of entries).

To this end we will show that in fact for any CountSketch matrix $S$,

$$\mathbb{P}\left(\left\|CS^T SD - CD\right\|_F^2 \leq \frac{6}{\delta \times \# \text{ rows in } S} \cdot \left\|C\right\|_F^2 \left\|D\right\|_F^2\right) \geq 1 - \delta. (*)$$

Setting $C = A^T$ and $D = A$, then since $A$ is orthonormal we have $\|A\|_F^2 = d$ and $A^T A = I_d$ whence the above becomes

$$\mathbb{P}\left(\|A^T S^T S A - A^T A\|_F^2 \leq \frac{6}{\delta \cdot 6d^2/\delta \varepsilon^2} d^2 - \varepsilon^2\right) \geq 1 - \delta$$

$$\iff \mathbb{P}\left(\|A^T S^T S A - I\|_F \leq \varepsilon\right) \geq 1 - \delta.$$

It remains to show $(*)$.

We say that a distribution on matrices $S \in \mathbb{R}^{k \times n}$ has the $(\varepsilon, \delta, \ell)$-JL moment property if for all $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$ (i.e., $x$ is a unit vector),

$$E_S[\|S x\|_2^\ell - 1] \leq \varepsilon^\ell \cdot \delta.$$  

Lemma [From vectors to matrices]: For $\varepsilon, \delta \in (0, 1/2)$, let $D$ be a distribution on matrices $S$ with $k$ rows and $n$ columns that satisfies the $(\varepsilon, \delta, \ell)$-JL moment property for some $\ell \geq 2$ then for matrices $A, B$ with $n$ rows

$$\mathbb{P}\left(\|A^T S^T S B - A^T B\|_F \geq 3\varepsilon\|A\|_F\|B\|_F\right) \leq \delta.$$
Proof. For a random variable $X$, we use the notation $\|X\|_\ell$ to denote $E[|X|^{\ell}]^{1/\ell}$. We will later show that $\| \cdot \|_\ell$ satisfies the triangle inequality which we will use in the following proof. Now, for arbitrary unit vectors $x, y$ we have

$$\|\langle Sx, Sy \rangle - \langle x, y \rangle\|_\ell = \frac{1}{2}\|\|Sx\|_2^2 - 1\| + (\|Sy\|_2^2 - 1) - (\|S(x - y)\|_2^2 - \|x - y\|_2^2)\|_\ell$$

$$\leq \frac{1}{2}\|\|Sx\|_2^2 - 1\| + \|Sy\|_2^2 - 1\| + \|S(x - y)\|_2^2 - \|x - y\|_2^2\|_\ell$$

$$\leq \frac{1}{2}(\varepsilon \delta^2 + \varepsilon \delta^2 + \|x - y\|_2^2 \varepsilon \delta^2)$$

$$\leq 3\varepsilon \delta^2.$$

By linearity, for arbitrary $x, y$ we have

$$\frac{\|\langle Sx, Sy \rangle - \langle x, y \rangle\|_\ell}{\|x\|_2 \|y\|_2} \leq 3\varepsilon \delta^2.$$

Suppose $A$ has $d$ columns and $B$ has $e$ columns, let the columns of $A$ be $A_1, \ldots, A_d$ and similarly for $B$.

Define $X_{i,j} = \frac{1}{\|A_i\|_2 \|B_j\|_2} \cdot (\langle SA_i, SB_j \rangle - \langle A_i, B_j \rangle)$ then

$$\|A^T S^T SB - A^T B\|_F^2 = \sum_{i,j} (\langle SA_i, SB_j \rangle - \langle A_i, B_j \rangle)^2 = \sum_{i,j} \|A_i\|_2^2 \|B_j\|_2^2 \|X_{i,j}\|_F^2$$

We calculate

$$\|A^T S^T SB - A^T B\|_F^2 \leq \sum_{i,j} \|A_i\|_2^2 \|B_j\|_2^2 \|X_{i,j}\|_F^2 \leq (3\varepsilon \delta^2)^2,$$

and observe that

$$\|X_{i,j}\|_F^2 = \|X_{i,j}\|_\ell^2 \leq (3\varepsilon \delta^2)^2,$$

replacing $X_{i,j}$ term on the RHS we obtain

$$\text{LHS} \leq (3\varepsilon \delta^2)^2 \sum_{i,j} \|A_i\|_2^2 \|A_j\|_2^2 = (3\varepsilon \delta^2)^2 \|A\|_F^2 \|B\|_F^2.$$

Considering the fact that

$$E[\|A^T S^T SB - A^T B\|_F^2] = \|A^T S^T SB - A^T B\|_F^2 \ell/2$$

we obtain by Markov’s Inequality,

$$\mathbb{P}(\|A^T S^T SB - A^T B\|_F > 3\varepsilon \|A\|_F \|B\|_F) \leq \frac{1}{(3\varepsilon \|A\|_F \|B\|_F)^\ell} E[\|A^T S^T SB - A^T B\|_F^\ell] \leq \delta.$$

Results for Vectors

It suffices to show that the CountSketch matrix $S$ satisfies JL-moment property and bound the number $k$ of rows.

We show this property holds with $\ell = 2$. First let’s consider $E_S[\|Sx\|_2^2]$.

For CountSketch matrix $S$ let
• \( h : [n] \to [k] \) be a 2-wise independent hash function.
• \( \sigma : [n] \to [-1,1] \) be a 4-wise independent hash function.

*Formally, \( h \) is \( k \)-wise independent if for all \( k \)-tuples \((x_1, \ldots, x_k)\) and \((y_1, \ldots, y_k)\) in the domain and range respectively the probability that \( h(x_i) = y_i \) for all \( i \) is \( 1/\text{size of hash output set}^k \).

For statements \( S \) define \([S] = \begin{cases} 1 & \text{if } S \text{ true} \\ 0 & \text{else} \end{cases} \).

We have
\[
E[||Sx||_2^2] = \sum_{j \in [k]} E[(\sum_{i \in [n]} [h(i) = j]\sigma_i x_i)^2]
\]

Expanding the sum on the right side and pulling out terms we obtain
\[
\text{RHS} = \sum_{i,j} x_i^2 E[[h(i) = j]^2] = ||x||^2
\]

We also compute by expansion
\[
E[||Sx||_2^2] = \sum_{i,j,k,l} ( [h(i) = j]\sigma_i x_i)^2 ( [h(k) = l]\sigma_k x_k)^2 \in [||x||_2^4,||x||_2^4(1 + \frac{2}{k})]
\]
similarly where we use the fact that
\[
\sum_{j} \frac{1}{k^2} \sum_{i_1,i_2} x_{i_1}^2 x_{i_2}^2 \leq \frac{1}{k} ||x||_2^4
\]
and symmetric inequalities. Since \( ||x||_2 = 1 \) we have
\[
ES[[||Sx||_2^2 - 1]^2] \leq \left( 1 + \frac{2}{k} \right) - 2 + 1 = \frac{2}{k}
\]
setting \( k = \frac{2}{\varepsilon^2} \) finishes the proof.

Where are we?

We just showed CountSketch has the JL property with \( \ell = 2 \) and \( k = \frac{2}{\varepsilon^2} \). The matrix product result (*) we wanted was
\[
P(||CSTSD - CD||_F^2 \leq (6/(\delta k)) \cdot ||C||_F^2 \cdot ||D||_F^2) \geq 1 - \delta
\]
We are now done with the proof that CountSketch is a subspace embedding.

Addendum: Minkowski Inequality and \( L^p \) Norm

For random scalar \( X \), let the \( L^p \)-norm of \( X \) be \( ||X||_p := (E|X|^p)^{1/p} \). Sometimes we will use \( X = ||T||_F \) for a random matrix \( T \).

We now show that \( || \cdot ||_p \) is a norm if \( p \geq 1 \) by Minkowski Inequality – the other properties of the norm are easy to check. It remains to show triangle inequality \( ||X+Y||_p \leq ||X||_p + ||Y||_p \) conditioned on \( ||X||_p, ||Y||_p < \infty \).
Proof of Minkowski’s Inequality. If \(\|X\|_p, \|Y\|_p < \infty\) then \(\|X + Y\|_p \leq 2^{p-1}(\|X\|_p + \|Y\|_p)\). To see this, note that \(f(x) = |x|^p\) is convex for \(p \geq 1\) so for any fixed \(x, y\),

\[
|0.5x + 0.5y|^p \leq |0.5|x| + 0.5|y|^p \leq (0.5)(|x|^p + |y|^p)
\]

by Jensen’s Inequality, from which we have \(|x + y|^p \leq 2^{p-1}(|x|^p + |y|^p)\). From the previous inequality we deduce that \(\|X + Y\|_p < \infty\).

Let \(\mu\) be the joint distribution of the random variables \(X\) and \(Y\). We have by definition,

\[
\|X + Y\|_p^p = \int \int |x + y|^p d\mu \leq \int \int |x + y|^{p-1} d\mu.
\]

We estimate using Hölder’s inequality that that

\[
\left(\int \int |x + y|^{p-1} d\mu\right)^p \leq \left(\int |x|^p d\mu\right)\left(\int |y|^p d\mu\right)^{p-1}.
\]

Taking \(1/p\)-th root of both sides and substituting into the RHS of the above we obtain the desired inequality.