A Background: CountSketch and TensorSketch

We start by describing the CountSketch transform Charikar et al. (2004). Let m be the target dimension. When applied to n-dimensional vectors, the transform is specified by a 2-wise independent hash function $h : [n] \to [m]$ and a 2-wise independent sign function $s : [n] \to \{-1, +1\}$. When applied to v, the value at coordinate i of the output, $i = 1, 2, \ldots, m$ is $\sum_{j | h(j)=i} s(j) v_j$. Note that CountSketch can be represented as an $m \times n$ matrix in which the j-th column contains a single non-zero entry $s(j)$ in the $h(j)$-th row.

We now describe the TensorSketch transform Pagh (2013). Suppose we are given points $v_i \in \mathbb{R}^{n_i}$, where $i = 1, \ldots, q$ and so $\phi(v_1, \ldots, v_q) = v_1 \otimes v_2 \otimes \cdots \otimes v_q \in \mathbb{R}^{n_1 n_2 \cdots n_q}$, and the target dimension is again m. The transform is specified using q 3-wise independent hash functions $h_i : [n_i] \to [m]$, and q 4-wise independent sign functions $s_i : [n_i] \to \{+1, -1\}$, where $i = 1, \ldots, q$. TensorSketch applied to $v_1, \ldots \otimes v_q$ is then CountSketch applied to $\phi(v_1, \ldots, v_q)$ with hash function $H : [n_1 n_2 \cdots n_q] \to [m]$ and sign function $S : [n_1 n_2 \cdots n_q] \to \{+1, -1\}$ defined as follows:

$$H(i_1, \ldots, i_q) = h_1(i_1) + h_2(i_2) + \cdots + h_q(i_q) \bmod m,$$

and

$$S(i_1, \ldots, i_q) = s_1(i_1) \cdot s_2(i_2) \cdots s_q(i_q),$$

where $i_j \in [n_j]$. It is well-known that if H is constructed this way, then it is 3-wise independent Carter and Wegman (1979); Patrascu and Thorup (2012). Unlike the work of Pham and Pagh Pham and Pagh (2013), which only used that H was 2-wise independent, our analysis needs this stronger property of H.

The TensorSketch transform can be applied to $v_1, \ldots, v_q$ without computing $\phi(v_1, \ldots, v_q)$ as follows. Let $v_j = (v_{j_\ell}) \in \mathbb{R}^{n_j}$. First, compute the polynomials

$$p_\ell(x) = \sum_{i=0}^{B-1} x^i \sum_{j_\ell | h_\ell(j_\ell)=i} v_{j_\ell} \cdot s_\ell(j_\ell),$$

for $\ell = 1, 2, \ldots, q$. A calculation Pagh (2013) shows

$$\prod_{\ell=1}^{q} p_\ell(x) \bmod (x^B - 1) = \sum_{i=0}^{B-1} x^i \sum_{(j_1, \ldots, j_q) | H(j_1, \ldots, j_q)=i} v_{j_1} \cdots v_{j_q} S(j_1, \ldots, j_q),$$

that is, the coefficients of the product of the q polynomials mod $(x^m - 1)$ form the value of TensorSketch$(v_1, \ldots, v_q)$. Pagh observed that this product of polynomials can be computed in $O(qm \log m)$ time using the Fast Fourier Transform. As it takes $O(q \max(\mathtt{nnz}(v_i)))$ time to form the q polynomials, the overall time to compute TensorSketch(v) is $O(q(\max(\mathtt{nnz}(v_i)) + m \log m))$.

B TensorSketch is an Oblivious Subspace Embedding (OSE)

Let S be the $m \times (n_1 n_2 \cdots n_q)$ matrix such that TensorSketch $(v_1, \ldots, v_q)$ is $S \cdot \phi(v_1, \ldots, v_q)$ for a randomly selected TensorSketch. Notice that S is a random matrix. In the rest of the paper, we refer to such a matrix as a TensorSketch matrix with an appropriate number of rows, i.e., the number of hash buckets. We will show that S is an oblivious subspace embedding for subspaces in $\mathbb{R}^{n_1 n_2 \cdots n_q}$ for appropriate values of m. Notice that S has exactly one non-zero entry per column. The index of the non-zero in the column $(i_1, \ldots, i_q)$ is $H(i_1, \ldots, i_q) = \sum_{j=1}^{q} h_j(i_j) \bmod m$. Let $\delta_{a,b}$ be the indicator random variable of whether $S_{a,b}$ is non-zero. The sign of the non-zero entry in column $(i_1, \ldots, i_q)$ is $S(i_1, \ldots, i_q) = \prod_{j=1}^{q} s_j(i_j)$. We show that the embedding matrix S of TensorSketch can be used to approximate matrix product and is an oblivious subspace embedding (OSE).

Theorem B.1. *Let S be the $m \times (n_1 n_2 \cdots n_q)$ matrix such that*

$$\text{TensorSketch}(v_1, \ldots, v_q)$$

is $S \cdot \phi(v_1, \ldots, v_q)$ for a randomly selected TensorSketch. *The matrix S satisfies the following two properties.*

1. *(*Approximate Matrix Product :*) Let A and B be matrices with $n_1n_2\cdots n_q$ rows. For $m \geq (2+3^q)/(\epsilon^2\delta)$, we have*
$$\Pr_S\left[\|A^\top S^\top SB - A^\top B\|_F^2 \leq \epsilon^2\|A\|_F^2\|B\|_F^2\right] \geq 1-\delta.$$

2. *(*Subspace Embedding :*) Consider a fixed k-dimensional subspace V. If $m \geq k^2(2+3^q)/(\epsilon^2\delta)$, then with probability at least $1-\delta$, $\|Sx\| = (1\pm\epsilon)\|x\|$ simultaneously for all $x \in V$.*

We establish the theorem via two lemmas as in Avron et al. (2016). The first lemma proves the approximate matrix product property via a careful second moment analysis.

Lemma B.2 (Approximate matrix product)**.** *Let A and B be matrices with $n_1n_2\cdots n_q$ rows. For $m \geq (2+3^q)/(\epsilon^2\delta)$, we have*

$$\Pr_S\left[\|A^\top S^\top SB - A^\top B\|_F^2 \leq \epsilon^2\|A\|_F^2\|B\|_F^2\right] \geq 1-\delta.$$

Proof. The proof follows that in Avron et al. (2016). Let $C = A^\top S^\top SB$. We have

$$C_{u,u'} = \sum_{t=1}^{m}\sum_{i,j\in[n_1n_2\cdots n_q]} S(i)S(j)\delta_{t,i}\delta_{t,j}A_{i,u}B_{j,u'} = \sum_{t=1}^{m}\sum_{i\neq j\in[n_1n_2\cdots n_q]} S(i)S(j)\delta_{t,i}\delta_{t,i}A_{i,u}B_{j,u'} + (A^\top B)_{u,u'}$$

Thus, $\mathbf{E}[C_{u,u'}] = (A^\top B)_{u,u'}$.

Next, we analyze $\mathbf{E}[((C - A^\top B)_{u,u'})^2]$. We have

$$((C-A^\top B)_{u,u'})^2 = \sum_{t_1,t_2=1}^{m}\sum_{i_1\neq j_1, i_2\neq j_2\in[n_1n_2\cdots n_q]} S(i_1)S(i_2)S(j_1)S(j_2)\cdot\delta_{t_1,i_1}\delta_{t_1,j_1}\delta_{t_2,i_2}\delta_{t_2,j_2}\cdot A_{i_1,u}A_{i_2,u}B_{j_1,u'}B_{j_2,u'}$$

For a term in the summation on the right hand side to have a non-zero expectation, it must be the case that $\mathbf{E}[S(i_1)S(i_2)S(j_1)S(j_2)] \neq 0$. Note that $S(i_1)S(i_2)S(j_1)S(j_2)$ is a product of random signs (possibly with multiplicities) where the random signs in different coordinates in $\{1,\ldots,q\}$ are independent and they are 4-wise independent within each coordinate. Thus, $\mathbf{E}[S(i_1)S(i_2)S(j_1)S(j_2)]$ is either 1 or 0. For the expectation to be 1, all random signs must appear with even multiplicities. In other words, in each of the q coordinates, the 4 coordinates of i_1, i_2, j_1, j_2 must be the same number appearing 4 times or 2 distinct numbers, each appearing twice. All the subsequent claims in the proof regarding i_1, i_2, j_1, j_2 agreeing on some coordinates follow from this property.

Let S_1 be the set of coordinates where i_1 and i_2 agree. Note that j_1 and j_2 must also agree in all coordinates in S_1 by the above argument. Let $S_2 \subset [q]\setminus S_1$ be the coordinates among the remaining where i_1 and j_1 agree. Finally, let $S_3 = [q]\setminus(S_1\cup S_2)$. All coordinates in S_3 of i_1 and j_2 must agree. Similarly as before, note that i_2 and j_2 agree on all coordinates in S_2 and i_2 and j_1 agree on all coordinates in S_3. We can rewrite $i_1 = (a,b,c), i_2 = (a,e,f), j_1 = (g,b,f), j_2 = (g,e,c)$ where $a = (a_\ell), g = (g_\ell)$ with $\ell \in S_1$, $b = (b_\ell), e = (e_\ell)$ with $\ell \in S_2$ and $c = (c_\ell), f = (f_\ell)$ with $\ell \in S_3$.

First we show that the contribution of the terms where $i_1 = i_2$ or $i_1 = j_2$ is bounded by $\frac{2\|A_u\|_2^2\|B_{u'}\|_2^2}{m}$, where A_u is the uth column of A and $B_{u'}$ is the u'th column of B. Indeed, consider the case $i_1 = i_2$. As observed before, we must have $j_1 = j_2$ to get a non-zero contribution. Note that if $t_1 \neq t_2$, we always have $\delta_{t_1,i_1}\delta_{t_2,i_2} = 0$ as $H(i_1)$ cannot be equal to both t_1 and t_2. Thus, for fixed $i_1 = i_2, j_1 = j_2$,

$$\begin{aligned}
&\mathbf{E}\left[\sum_{t_1,t_2=1}^{m} S(i_1)S(i_2)S(j_1)S(j_2)\cdot\delta_{t_1,i_1}\delta_{t_1,j_1}\delta_{t_2,i_2}\delta_{t_2,j_2}\cdot A_{i_1,u}A_{i_2,u}B_{j_1,u'}B_{j_2,u'}\right]\\
&= \mathbf{E}\left[\sum_{t_1=1}^{m}\delta_{i_1,t_1}^2\delta_{j_1,t_1}^2A_{i_1,u}^2B_{j_1,u'}^2\right]\\
&= \frac{A_{i_1,u}^2B_{j_1,u'}^2}{m}
\end{aligned}$$

Summing over all possible values of i_1, j_1, we get the desired bound of $\frac{\|A_u\|_2^2\|B_{u'}\|_2^2}{m}$. The case $i_1 = j_2$ is analogous.

Next we compute the contribution of the terms where $i_1 \neq i_2, j_1, j_2$ i.e., there are at least 3 distinct numbers among i_1, i_2, j_1, j_2. Notice that $\mathbf{E}[\delta_{t_1,i_1}\delta_{t_1,j_1}\delta_{t_2,i_2}\delta_{t_2,j_2}] \leq \frac{1}{m^3}$ because the $\delta_{t,i}$'s are 3-wise independent. For fixed i_1, j_1, i_2, j_2, there are m^2 choices of t_1, t_2 so the total contribution to the expectation from terms with the same i_1, j_1, i_2, j_2 is bounded by $m^2 \cdot \frac{1}{m^3} \cdot |A_{i_1,u}A_{i_2,u}B_{j_1,u'}B_{j_2,u'}| = \frac{1}{m}|A_{i_1,u}A_{i_2,u}B_{j_1,u'}B_{j_2,u'}|$.

Therefore,

$$
\begin{aligned}
&\mathbf{E}[((C - A^\top B)_{u,u'})^2] \\
\leq\ & \frac{2\|A_u\|_2^2\|B_{u'}\|_2^2}{m} + \frac{1}{m}\sum_{\text{partition } S_1,S_2,S_3}\ \sum_{a,g,b,e,c,f} |A_{(a,b,c),u}B_{(g,b,f),u'}A_{(a,e,f),u}B_{(g,e,c),u'}| \\
\leq\ & \frac{2\|A_u\|_2^2\|B_{u'}\|_2^2}{m} + \frac{3^q}{m}\sum_{a,b,c,g,e,f} |A_{(a,b,c),u}B_{(g,b,f),u'}A_{(a,e,f),u}B_{(g,e,c),u'}| \\
\leq\ & \frac{2\|A_u\|_2^2\|B_{u'}\|_2^2}{m} + \frac{3^q}{m}\sum_{g,e,f}\left(\sum_{a,b,c} A^2_{(a,b,c),u}\right)^{1/2}\left(\sum_{a,b,c} B^2_{(g,b,f),u'}A^2_{(a,e,f),u}B^2_{(g,e,c),u'}\right)^{1/2} \\
=\ & \frac{2\|A_u\|_2^2\|B_{u'}\|_2^2}{m} + \frac{3^q\|A_u\|}{m}\sum_{g,e,f}\left(\sum_{b} B^2_{(g,b,f),u'}\right)^{1/2}\left(\sum_{a,c} A^2_{(a,e,f),u}B^2_{(g,e,c),u'}\right)^{1/2} \\
\leq\ & \frac{2\|A_u\|_2^2\|B_{u'}\|_2^2}{m} + \frac{3^q\|A_u\|}{m}\sum_{e}\left(\sum_{b,g,f} B^2_{(g,b,f),u'}\right)^{1/2}\left(\sum_{a,c,g,f} A^2_{(a,e,f),u}B^2_{(g,e,c),u'}\right)^{1/2} \\
=\ & \frac{2\|A_u\|_2^2\|B_{u'}\|_2^2}{m} + \frac{3^q\|A_u\|\cdot\|B_{u'}\|}{m}\sum_{e}\left(\sum_{a,f} A^2_{(a,e,f),u}\right)^{1/2}\left(\sum_{g,c} B^2_{(g,e,c),u'}\right)^{1/2} \\
\leq\ & \frac{2\|A_u\|_2^2\|B_{u'}\|_2^2}{m} + \frac{3^q\|A_u\|\cdot\|B_{u'}\|}{m}\left(\sum_{a,e,f} A^2_{(a,e,f),u}\right)^{1/2}\left(\sum_{g,e,c} B^2_{(g,e,c),u'}\right)^{1/2} \\
=\ & \frac{(2+3^q)\|A_u\|_2^2\|B_{u'}\|_2^2}{m},
\end{aligned}
$$

where the second inequality follows from the fact that there are at most 3^q partitions of $[q]$ into 3 sets. The other inequalities are from Cauchy-Schwarz.

Combining the above bounds, we have $\mathbf{E}[((C - A^\top B)_{u,u'})^2] \leq \frac{(2+3^q)\|A_u\|_2^2\|B_{u'}\|_2^2}{m}$. For $m \geq (2+3^q)/(\epsilon^2\delta)$, by the Markov inequality, $\|A^\top S^\top SB - A^\top B\|_F^2 \leq \epsilon^2\|A\|_F^2\|B\|_F^2$ with probability $1-\delta$. □ □

The second lemma proves that the subspace embedding property follows from the approximate matrix product property.

Lemma B.3 (Oblivious subspace embeddings)**.** *Consider a fixed k-dimensional subspace $V \subset \mathbb{R}^{n_1n_2\cdots n_q}$. If $m \geq k^2(2+3^q)/(\epsilon^2\delta)$, then with probability at least $1-\delta$, $\|Sx\|_2 = (1\pm\epsilon)\|x\|_2$ simultaneously for all $x \in V$.*

Proof. Let B be a $(n_1n_2\cdots n_q)\times k$ matrix whose columns form an orthonormal basis of V. Thus, we have $B^\top B = I_k$ and $\|B\|_F^2 = k$. The condition that $\|Sx\|_2 = (1\pm\epsilon)\|x\|_2$ simultaneously for all $x \in V$ is equivalent to the condition that the singular values of SB are bounded by $1\pm\epsilon$. By Lemma B.2, for $m \geq (2+3^q)/((\epsilon/k)^2\delta)$, with probability at least $1-\delta$, we have

$$\|B^\top S^\top SB - B^\top B\|_F^2 \leq (\epsilon/k)^2\|B\|_F^4 = \epsilon^2$$

Thus, we have $\|B^\top S^\top SB - I_k\|_2 \leq \|B^\top S^\top SB - I_k\|_F \leq \epsilon$. In other words, the squared singular values of SB are bounded by $1\pm\epsilon$, implying that the singular values of SB are also bounded by $1\pm\epsilon$. Note that $\|A\|_2$ for a matrix A denotes its operator norm. □ □

C Missing Proofs

C.1 Proofs for Tensor Product Least Square Regression

Theorem 3.1. (Tensor regression) Suppose $\widetilde{x}$ is the output of Algorithm 1 with TENSORSKETCH $S \in \mathbb{R}^{m\times n}$, where $m = 8(d_1 d_2 \cdots d_q + 1)^2(2+3^q)/(\epsilon^2\delta)$. Then the following approximation $\|(A_1 \otimes A_2 \otimes \cdots \otimes A_q)\widetilde{x} - b\|_2 \leq (1+\epsilon)\,\mathrm{OPT}$, holds with probability at least $1-\delta$.

Proof. It is easy to see that

$$\|(A_1 \otimes A_2 \otimes \cdots \otimes A_q)x - b\|_2 = \left\| \begin{bmatrix} (A_1 \otimes A_2 \otimes \cdots \otimes A_q) & b \end{bmatrix} \begin{bmatrix} x \\ -1 \end{bmatrix} \right\|_2,$$

and identifying

$$y = \begin{bmatrix} (A_1 \otimes A_2 \otimes \cdots \otimes A_q) & b \end{bmatrix} \begin{bmatrix} x \\ -1 \end{bmatrix} \in \mathbb{R}^{n_1 n_2 \cdots n_q}$$

and y is a vector of a subspace $V \subset \mathbb{R}^{n_1 n_2 \cdots n_q}$ with dimension at most $d_1 d_2 \cdots d_q + 1$, we can use Lemma B.3 to conclude that

$$\Pr\left[\left| \|Sy\|_2 - \|y\|_2 \right| \leq \epsilon \|y\|_2 \right] \geq 1 - \delta$$

when $m = (d_1 d_2 \cdots d_q + 1)^2(2+3^q)/(\epsilon^2\delta)$.

Thus we have

$$\|(A_1 \otimes A_2 \otimes \cdots \otimes A_q)\widetilde{x} - b\|_2 \leq \frac{1}{1-\epsilon}\|S(A_1 \otimes A_2 \otimes \cdots \otimes A_q)\widetilde{x} - Sb\|_2$$

and

$$\|S(A_1 \otimes A_2 \otimes \cdots \otimes A_q)x - Sb\|_2 \leq (1+\epsilon)\|(A_1 \otimes A_2 \otimes \cdots \otimes A_q)x - b\|_2$$

hold with probability at least $1-\delta$. Then using a union bound, we have

$$\begin{aligned} & \|(A_1 \otimes A_2 \otimes \cdots \otimes A_q)\widetilde{x} - b\|_2 \\ \leq & \frac{1}{1-\epsilon}\|S(A_1 \otimes A_2 \otimes \cdots \otimes A_q)\widetilde{x} - Sb\|_2 \\ \leq & \frac{1}{1-\epsilon}\|S(A_1 \otimes A_2 \otimes \cdots \otimes A_q)x - Sb\|_2 \\ \leq & \frac{1+\epsilon}{1-\epsilon}\|(A_1 \otimes A_2 \otimes \cdots \otimes A_q)x - b\|_2 \end{aligned}$$

holds with probability at least $1-2\delta$. □

Corollary 3.2. (Sketch for tensor nonnegative regression) Suppose $\tilde{x} = \min_{x\geq 0}\|S\mathcal{A}x - Sb\|_2$ with TENSORSKETCH $S \in \mathbb{R}^{m\times n}$, where $m = 8(d_1 d_2 \cdots d_q + 1)^2(2+3^q)/(\epsilon^2\delta)$. Then the following approximation $\|(A_1 \otimes A_2 \otimes \cdots \otimes A_q)\widetilde{x} - b\|_2 \leq (1+\epsilon)\,\mathrm{OPT}$ holds with probability at least $1-\delta$, where $\mathrm{OPT} = \min_{x\geq 0}\|(A_1 \otimes A_2 \otimes \cdots \otimes A_q)x - b\|_2$.

Proof. The proof of Theorem. 3.2 is similar to the proof of theorem 3.1. Denote $\tilde{x} = \min_{x\geq 0}\|S\mathcal{A}x - Sb\|_2$ and $x^* = \min_{x\geq 0}\|\mathcal{A}x - b\|_2$. Using Lemma. B.3, we have:

$$\|\mathcal{A}\tilde{x} - b\|_2 \leq \frac{1}{1-\epsilon}\|S\mathcal{A}\tilde{x} - Sb\|_2, \tag{6}$$

with probability at least $1-\delta$, and

$$\|S\mathcal{A}x^* - Sb\|_2 \leq (1+\epsilon)\|\mathcal{A}x^* - b\|_2, \tag{7}$$

with probability at least $1-\delta$. Hence applying a union bound we have:

$$\begin{aligned} & \|\mathcal{A}\tilde{x}-b\|_2 && (8)\\ \le & \frac{1}{1-\epsilon}\|S\mathcal{A}\tilde{x}-Sb\|_2 \\ \le & \frac{1}{1-\epsilon}\|S\mathcal{A}x^*-Sb\|_2 \\ \le & \frac{1+\epsilon}{1-\epsilon}\|\mathcal{A}x^*-b\|_2, && (9)\end{aligned}$$

with probability at least $1-2\delta$. □

C.2 Proofs for P-Splines

Lemma 4.1. *Let $x^* \in \mathbb{R}^d$, $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$ as above. Let $U_1 \in \mathbb{R}^{n\times d}$ denote the first n rows of an orthogonal basis for $\begin{bmatrix} A \\ \sqrt{\lambda}L \end{bmatrix} \in \mathbb{R}^{(n+p)\times d}$. Let sketching matrix $S \in \mathbb{R}^{m\times n}$ have a distribution such that with constant probability*

$$\text{(I)}\ \|U_1^\top S^\top S U_1 - U_1^\top U_1\|_2 \le 1/4,$$

and

$$\text{(II)}\ \|U_1^\top (S^\top S - I)(b - Ax^*)\|_2 \le \sqrt{\epsilon \operatorname{OPT}/2}.$$

Let $\widetilde{x}$ denote $\operatorname{argmin}_{x\in\mathbb{R}^d}\|S(Ax-b)\|_2^2 + \lambda\|Lx\|_2^2$. *Then with probability at least $9/10$,*

$$\|A\widetilde{x}-b\|_2^2 + \lambda\|L\widetilde{x}\|_2^2 \le (1+\epsilon)\operatorname{OPT}.$$

Proof. Let $\hat{A} \in \mathbb{R}^{(n+d)\times d}$ have orthonormal columns with $\operatorname{range}(\hat{A}) = \operatorname{range}(\begin{bmatrix} A \\ \sqrt{\lambda}L \end{bmatrix})$. (An explicit expression for one such $\hat{A}$ is given below.) Let $\hat{b} \equiv \begin{bmatrix} b \\ 0_d \end{bmatrix}$. We have

$$\min_{y\in\mathbb{R}^d}\|\hat{A}y-\hat{b}\|_2 \tag{10}$$

equivalent to $\|b-Ax\|_2^2+\lambda\|Lx\|_2^2$, , in the sense that for any $\hat{A}y \in \operatorname{range}(\hat{A})$, there is $x \in \mathbb{R}^d$ with $\hat{A}y = \begin{bmatrix} A \\ \sqrt{\lambda}L \end{bmatrix} x$, so that $\|\hat{A}y-\hat{b}\|_2^2 = \|\begin{bmatrix} A \\ \sqrt{\lambda}L \end{bmatrix} x-\hat{b}\|_2^2 = \|b-Ax\|_2^2+\lambda\|Lx\|_2^2$. Let $y^* = \operatorname{argmin}_{y\in\mathbb{R}^d}\|\hat{A}y-\hat{b}\|_2$, so that $\hat{A}y^* = \begin{bmatrix} Ax^* \\ \sqrt{\lambda}Lx^* \end{bmatrix}$.

Let $\hat{A} = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix}$, where $U_1 \in \mathbb{R}^{n\times d}$ and $U_2 \in \mathbb{R}^{d\times d}$, so that U_1 is as in the lemma statement.

We define $\hat{S}$ to be $\begin{bmatrix} S & 0_{m\times d} \\ 0_{d\times n} & I_d \end{bmatrix}$ and $\hat{S}$ satisfies Property (I) and (II) of Lemma 4.1.

Using $\|U_1^\top S^\top S U_1 - U_1^\top U_1\|_2 \le 1/4$, with constant probability

$$\|\hat{A}^\top\hat{S}^\top\hat{S}\hat{A} - I_d\|_2 = \|U_1^\top S^\top S U_1 + U_2^\top U_2 - I_d\|_2 = \|U_1^\top S^\top S U_1 - U_1^\top U_1\|_2 \le 1/4. \tag{11}$$

Using the normal equations for Eq. (10), we have

$$0 = \hat{A}^\top(\hat{b}-\hat{A}y^*) = U_1^\top(b-Ax^*) - \sqrt{\lambda}U_2^\top x^*,$$

and so

$$\hat{A}^\top\hat{S}^\top\hat{S}(\hat{b}-\hat{A}y^*) = U_1^\top S^\top S(b-Ax^*) - \sqrt{\lambda}U_2^\top x^* = U_1^\top S^\top S(b-Ax^*) - U_1^\top(b-Ax^*).$$

Using Property (II) of Lemma 4.1, with constant probability

$$\begin{aligned} & \|\hat{A}^\top\hat{S}^\top\hat{S}(\hat{b}-\hat{A}y^*)\|_2 \\ = & \|U_1^\top S^\top S(b-Ax^*) - U_1^\top(b-Ax^*)\|_2 \\ \le & \sqrt{\epsilon \operatorname{OPT}/2} \\ = & \sqrt{\epsilon/2}\|\hat{b}-\hat{A}y^*\|_2. && (12)\end{aligned}$$

It follows by a standard result from (11) and (12) that the solution $\tilde{y} \equiv \text{argmin}_{y\in\mathbb{R}^d} \|\hat{S}(\hat{A}y - \hat{b})\|_2$ has $\|\hat{A}\tilde{y} - \hat{b}\|_2 \le (1+\epsilon)\min_{y\in\mathbb{R}^d} \|\hat{A}y - \hat{b}\|_2$, and therefore that $\tilde{x}$ satisfies the claim of the theorem.

For convenience we give the proof of the standard result: (11) implies that $\hat{A}^\top \hat{S}^\top \hat{S}\hat{A}$ has smallest singular value at least $3/4$. The normal equations for the unsketched and sketched problems are

$$\hat{A}^\top(\hat{b} - \hat{A}y^*) = 0 = \hat{A}^\top \hat{S}^\top \hat{S}(\hat{b} - \hat{A}\tilde{y}).$$

The normal equations for the unsketched case imply $\|\hat{A}\tilde{y} - \hat{b}\|_2^2 = \|\hat{A}(\tilde{y} - y^*)\|_2^2 + \|\hat{b} - \hat{A}y^*\|_2^2$, so it is enough to show that $\|\hat{A}(\tilde{y} - y^*)\|_2^2 = \|\tilde{y} - y^*\|_2^2 \le \epsilon\,\text{OPT}$. We have

$$\begin{aligned}
(3/4)\|\tilde{y} - y^*\|_2 &\le \|\hat{A}^\top \hat{S}^\top \hat{S}\hat{A}(\tilde{y} - y^*)\|_2 && \text{by Eq. (11)}\\
&= \|\hat{A}^\top \hat{S}^\top \hat{S}\hat{A}(\tilde{y} - y^*) - \hat{A}^\top \hat{S}^\top \hat{S}(\hat{b} - \hat{A}\tilde{y})\|_2 && \text{by Normal Equation}\\
&= \|\hat{A}^\top \hat{S}^\top \hat{S}(\hat{b} - \hat{A}y^*)\|_2\\
&\le \sqrt{\epsilon\,\text{OPT}/2} && \text{by Eq. (12)},
\end{aligned}$$

so that $\|\tilde{y} - y^*\|_2^2 \le (4/3)^2 \epsilon\,\text{OPT}/2 \le \epsilon\,\text{OPT}$. The lemma follows.

□

The following lemma computes the statistical dimension $\texttt{sd}_\lambda(A, L)$ that will be used for computing the number of rows of sketching matrix S.

Lemma C.1. *For U_1 as in Lemma 4.1, $\|U_1\|_F^2 = \texttt{sd}_\lambda(A, L) = \sum_i 1/(1 + \lambda/\gamma_i^2) + d - p$, where A has singular values σ_i. Also $\|U_1\|_2 = \max\{1/\sqrt{1 + \lambda/\gamma_1^2},\, 1\}$.*

Proof. Suppose we have the GSVD of $(A,\, L)$. Let

$$D \equiv \begin{bmatrix} \Sigma^\top\Sigma + \lambda\Omega^\top\Omega & 0_{p\times(n-p)} \\ 0_{(n-p)\times p} & I_{d-p} \end{bmatrix}^{-1/2}.$$

Then

$$\hat{A} = \begin{bmatrix} U\begin{bmatrix} \Sigma & 0_{p\times(n-p)} \\ 0_{(n-p)\times p} & I_{d-p} \end{bmatrix} D \\ \sqrt{\lambda}V\begin{bmatrix} \Omega & 0_{p\times(n-p)} \end{bmatrix} D \end{bmatrix}$$

has $\hat{A}^\top\hat{A} = I_d$, and for given x, there is $y = D^{-1}RQ^\top x$ with $\hat{A}y = \begin{bmatrix} A \\ \sqrt{\lambda}L \end{bmatrix} x$. We have $\|U_1\|_F^2 = \left\| U \begin{bmatrix} \Sigma & 0_{p\times(n-p)} \\ 0_{(n-p)\times p} & I_{d-p} \end{bmatrix} D \right\|_F^2 = \left\| \begin{bmatrix} \Sigma & 0_{p\times(n-p)} \\ 0_{(n-p)\times p} & I_{d-p} \end{bmatrix} D \right\|_F^2 = \sum_{i=1}^p 1/(1 + \lambda/\gamma_i^2) + d - p$ as claimed. □

Theorem 4.3. (P-Spline regression) There is a constant $K > 0$ such that for $m \ge K(\epsilon^{-1}\texttt{sd}_\lambda(A, L) + \texttt{sd}_\lambda(A, L)^2)$ and $S \in \mathbb{R}^{m\times n}$ a sparse embedding matrix (e.g., CountSketch) with SA computable in $O(\texttt{nnz}(A))$ time, Property (I) and (II) of Lemma 4.1 apply, and with constant probability the corresponding $\tilde{x} = \text{argmin}_{x\in\mathbb{R}^d} \|S(Ax - b)\|_2 + \lambda\|Lx\|_2^2$ is an ϵ-approximate solution to $\min_{x\in\mathbb{R}^d} \|b - Ax\|_2^2 + \lambda\|Lx\|_2^2$.

Proof. Recall that $\texttt{sd}_\lambda(A, L) = \|U_1\|_F^2$. Sparse embedding distributions satisfy the bound for approximate matrix multiplication

$$\|W^\top S^\top SH - W^\top H\|_F \le C\|W\|_F\|H\|_F/\sqrt{m},$$

for a constant C (Clarkson and Woodruff, 2013; Meng and Mahoney, 2013; Nelson and Nguyên, 2013); this is also true of OSE matrices. We set $W = H = U_1$ and use $\|X\|_2 \le \|X\|_F$ for all X and $m \ge K\|U_1\|_F^4$ to obtain Property (I) of Lemma 4.1, and set $W = U_1$, $H = b - Ax^*$ and use $m \ge K\|U_1\|_F^2/\epsilon$ to obtain Property (II) of Lemma 4.1. (Here the bound is slightly stronger than Property (II), holding for $\lambda = 0$.) With Property (I) and Property (II), the claim for $\tilde{x}$ from a sparse embedding follows using Lemma 4.1. □

C.3 Proofs for Tensor Product ℓ_1 Regression

Lemma 5.3. *For any $p \geq 1$.* $\mathsf{Condition}(\mathcal{A})$ *computes $\mathcal{A}U/(d\gamma_p)$ which is an $(\alpha, \beta\sqrt{3}d(tw)^{|1/p-1/2|}, p)$-well-conditioned basis of $\mathcal{A}$, with probability at least $1 - \prod_{i=1}^q (n_i/w_i)\delta$.*

Proof. This lemma is similar to arguments in Clarkson et al. (2013), we simply adjust notation and parameters for completeness. Applying Theorem 5.2, we have that with probability at least $1 - \prod_{i=1}^q (n_i/w_i)\delta$, for all $x \in \mathbb{R}^r$, if we consider $y = \mathcal{A}x$ and write $y^\top = [z_1^\top, z_2^\top, \ldots, z_{\prod_{i=1}^q n_i/w_i}]^\top$, then for all $i \in [\prod_{i=1}^q n_i/w_i]$,

$$\sqrt{\tfrac{1}{2}}\|z_i\|_2 \leq \|S_i z_i\|_2 \leq \sqrt{\tfrac{3}{2}}\|z_i\|_2,$$

where $S_i \in \mathbb{R}^{m_i \times \prod_{i=1}^q w_i}$. In the following, suppose $m_i = t$. By relating the 2-norm and the p-norm, for $1 \leq p \leq 2$, we have

$$\|S_i z_i\|_p \leq t^{1/p-1/2}\|S z_i\|_2 \leq t^{1/p-1/2}\sqrt{\tfrac{3}{2}}\|z_i\|_2 \leq t^{1/p-1/2}\sqrt{\tfrac{3}{2}}\|z_i\|_p,$$

and similarly,

$$\|S_i z_i\|_p \geq \|S_i z_i\|_2 \geq \sqrt{\tfrac{1}{2}}\|z_i\|_2 \geq \sqrt{\tfrac{1}{2}}w^{1/2-1/p}\|z_i\|_p, \; w = \prod_{j=1}^q w_j.$$

If $p > 2$, then

$$\|S_i z_i\|_p \leq \|S_i z_i\|_2 \leq \sqrt{\tfrac{3}{2}}\|z_i\|_2 \leq \sqrt{\tfrac{3}{2}}w^{1/2-1/p}\|z_i\|_p,$$

and similarly,

$$\|S_i z_i\|_p \geq t^{1/p-1/2}\|S_i z_i\|_2 \geq t^{1/p-1/2}\sqrt{\tfrac{1}{2}}\|z_i\|_2 \geq t^{1/p-1/2}\sqrt{\tfrac{1}{2}}\|z_i\|_p.$$

Since $\|\mathcal{A}x\|_p^p = \|y\|_p^p = \sum_i \|z_i\|_p^p$ and $\|S\mathcal{A}x\|_p^p = \sum_i \|S_i z_i\|_p^p$, for $p \in [1, 2]$ we have with probability $1 - \prod_{i=1}^q (n_i/w_i)\delta$

$$\sqrt{\tfrac{1}{2}}w^{1/2-1/p}\|\mathcal{A}x\|_p \leq \|S\mathcal{A}x\|_p \leq \sqrt{\tfrac{3}{2}}t^{1/p-1/2}\|\mathcal{A}x\|_p,$$

and for $p \in [2, \infty)$ with probability $1 - \prod_{i=1}^q (n_i/w_i)\delta$

$$\sqrt{\tfrac{1}{2}}t^{1/p-1/2}\|\mathcal{A}x\|_p \leq \|S\mathcal{A}x\|_p \leq \sqrt{\tfrac{3}{2}}w^{1/2-1/p}\|\mathcal{A}x\|_p.$$

In either case,

$$\|\mathcal{A}x\|_p \leq \gamma_p\|S\mathcal{A}x\|_p \leq \sqrt{3}(tw)^{|1/p-1/2|}\|\mathcal{A}x\|_p. \tag{13}$$

We have, from the definition of an (α, β, p)-well-conditioned basis, that

$$\|S\mathcal{A}U\|_p \leq \alpha \tag{14}$$

and for all $x \in \mathbb{R}^d$,

$$\|x\|_q \leq \beta\|S\mathcal{A}Ux\|_p. \tag{15}$$

Combining (13) and (14), we have that with probability at least $1 - \prod_{i=1}^q (n_i/w_i)\delta$,

$$\|\mathcal{A}U/(r\gamma_p)\|_p \leq \sum_i \|\mathcal{A}U_i/r\gamma_p\|_p \leq \sum_i \|S\mathcal{A}U_i/r\|_p \leq \alpha.$$

Combining (13) and (15), we have that with probability at least $1 - \prod_{i=1}^q (n_i/w_i)\delta$, for all $x \in \mathbb{R}^r$,

$$\|x\|_q \leq \beta\|S\mathcal{A}Ux\|_p \leq \beta\sqrt{3}r(tw)^{|1/p-1/2|}\|\mathcal{A}U\frac{1}{r\gamma_p}x\|_p.$$

Hence $\mathcal{A}U/(r\gamma_p)$ is an $(\alpha, \beta\sqrt{3}r(tw)^{|1/p-1/2|}, p)$-well-conditioned basis. □

Theorem 5.4. (Main result) Given $\epsilon \in (0,1)$, $\mathcal{A} \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$, Alg. 3 computes $\widehat{x}$ such that with probability at least $1/2$, $\|\mathcal{A}\hat{x} - b\|_1 \leq (1+\epsilon)\min_{x\in\mathbb{R}^d} \|\mathcal{A}x - b\|_1$. For the special case when $q = 2$, $n_1 = n_2$, the algorithm's running time is $O({n_1}^{3/2}\operatorname{poly}(\prod_{i=1}^2 d_i/\epsilon))$.

Proof. For notational simplicity, let us denote $n_{[q_1]} = \prod_{i=1}^{q_1} n_i$, $n_{[q]\setminus[q_1]} = \prod_{i=q_1+1}^{q} n_1$, $d_{[q_1]} = \prod_{i=1}^{q_1} d_i$, and $d_{[q]\setminus[q_1]} = \prod_{i=q_1+1}^{q} d_i$. For any row-block $A_{i_1}^1 \otimes \ldots \otimes A_{i_q}^{(q)}$, computing $S_{i1i2\ldots i_q}(A_{i_1}^1 \otimes \ldots \otimes A_{i_q}^{(q-1)})$ takes $O(d(\sum_{k=1}^q nnz(A_{i_k}^{(k)})) + dqm\log(m))$ (see Sec 2). Hence for $S\mathcal{A}$, it takes:

$$\left(d\sum_{k=1}^{q}\mathtt{nnz}(A_k)\prod_{i\in[q]\setminus\{k\}}^{q} n_i/w_i\right) + \left(dqm\log(m)\prod_{i=1}^{q} n_i/w_i\right)$$

where $S \in \mathbb{R}^{(m\prod_{i=1}^q (n_i/w_i))\times\prod_{i=1}^q w_i}$ and $m \geq 100\prod_{i=1}^q d_i^2(2+3^q)/\epsilon^2 = O(\operatorname{poly}(d/\epsilon))$. We need to compute an orthogonal factorization $S\mathcal{A} = QR_{\mathcal{A}}$ in $O(qmd^2)$ and then compute $U = R_{\mathcal{A}}^{-1}$ in $O(d^3)$ time. Hence the total running time of Algorithm $\mathsf{Condition}(\mathcal{A})$ is $O(qmd^2 + d^3)$. Thus the total running time of computing $S\mathcal{A}$ and Condition(A) is

$$O\left(\left(\sum_{k=1}^{q}\mathtt{nnz}(A_k)\prod_{i\in[q]\setminus\{k\}}^{q} n_i/w_i\right) + \left(\prod_{i=1}^{q} n_i/w_i\right)\operatorname{poly}(d/\epsilon) + qmd^2 + d^3\right),$$

We will compute UG in $O(d^2\log n)$ time. We compute $\widetilde{E} = E(A_{q_1+1}\otimes\ldots\otimes A_q)^T$ in $O(dn_{[q]\setminus[q_1]})$ time.

Then we can compute $R(A_1\otimes\cdots\otimes A_{q_1})\widetilde{E}_j$ in $O(n_{[q_1]}d_{[q_1]}\log n + d_{[q_1]}n_{[q]\setminus[q_1]}\log n)$ time.

Since computation of the median λ_i takes $O(\log n)$ time, computing all λ_i and then λ_e takes $O(n_{[q]\setminus[q_1]}\log n)$ time.

As $\mathcal{A}UG$ has $O(\log n)$ columns, we need to compute λ_e for each $\mathcal{A}UG$ using the above procedure and hence it takes in total $O(d(n_{[q_1]} + n_{[q]\setminus[q_1]})\log^2 n)$ time.

Sampling a column of $\mathcal{A}UG$ using λ_e takes $O(\log n)$ time, sampling an entry in M takes in total $O(n_{[q_1]} + n_{[q]\setminus[q_1]})$ time.

Since we need $\sqrt{\prod_{k=1}^q w_k}\operatorname{poly}(r)$ samples to select rows, the running time is $d(n_{[q_1]} + n_{[q]\setminus[q_1]})\log^2 n \cdot \sqrt{\prod_{k=1}^q w_k}\operatorname{poly}(r)$.

Now for simplicity, we set $q = 2$, $n_i = n_0$ for $i \in [2]$. Note that it is optimal to choose $w_i = w$ for $i \in [2]$. Substituting $q = 2, n_i = n_0$ and $w_i = w$, we that the total running time of Alg. 3:

$$O\left(dw^{-1}n_0(\mathtt{nnz}(A_1) + \mathtt{nnz}(A_2)) + w^{-2}n_0^2\operatorname{poly}(d/\epsilon) + wn_0\operatorname{poly}(d)\log(n)\right).$$

For dense A_1 and A_2, $\mathtt{nnz}(A_1) + \mathtt{nnz}(A_2) = O(n_0)$ time, and so ignoring poly and log terms that do not depend on n_0, the total running time can be simplified to:

$$O(w^{-1}n_0^2 + wn_0).$$

Setting $w = \sqrt{n_0}$, we can minimize the above running time to $O(n_0^{3/2})$, which is faster than the n_0^2 time for solving the problem by forming $A_1 \otimes A_2$. $\square$