A. Missing proofs from Section 3

Theorem A.1 (A version of Lemma 14 of (Sohler and Woodruff, 2018)). *Let P be an r dimensional subspace of $\mathbb{R}^d$ such that*

$$\sum_i \text{dist}(a_i, P) - \sum_i \text{dist}(a_i, \text{span}(P \cup H)) \leq \frac{\varepsilon^2}{80}\text{SubApx}_{k,1}(A)$$

for all k-dimensional subspaces H. Let $B \in \mathbb{R}^{d\times r}$ be an orthonormal basis for the subspace P. For each a_i, let $a_i^B \in \mathbb{R}^r$ be such that $\text{dist}(a_i, Ba_i^B) \leq (1+\varepsilon_c)\text{dist}(a_i, P)$ and let $(1-\varepsilon_c)\text{dist}(a_i, P) \leq \text{apx}_i \leq (1+\varepsilon_c)\text{dist}(a_i, P)$ for $\varepsilon_c = \varepsilon^2/6$. Then for any k dimensional shape S,

$$\sum_i \sqrt{\text{dist}(Ba_i^B, S)^2 + \text{apx}_i^2} = (1 \pm 5\varepsilon)\sum_i \text{dist}(a_i, S)$$

Proof. We have by the Pythagorean theorem that $\text{dist}(Ba_i^B, a_i)^2 = \text{dist}(Ba_i^B, \mathbb{P}_P a_i)^2 + \text{dist}(a_i, P)^2 \leq (1+3\varepsilon_c)\text{dist}(a_i, P)^2$ which implies that $\text{dist}(Ba_i^B, \mathbb{P}_P a_i)^2 \leq (3\varepsilon_c)\text{dist}(a_i, P)^2$.

Given a shape S, we partition $[n]$ into two sets *small* and *large*. We say $i \in [n]$ is *small* if $\text{dist}(\mathbb{P}_P a_i, S) \leq \text{dist}(\mathbb{P}_P a_i, Ba_i^B)$. In that case, $\text{dist}(Ba_i^B, S)^2 \leq 4\text{dist}(\mathbb{P}_P a_i, Ba_i^B)^2 \leq 12\varepsilon_c\text{dist}(a_i, P)^2$ by the triangle inequality and $\sqrt{\text{dist}(Ba_i^B, S)^2 + \text{apx}_i^2} \leq \sqrt{1+15\varepsilon_c}\text{dist}(a_i, P) \leq \sqrt{1+15\varepsilon_c}\sqrt{\text{dist}(a_i, P)^2 + \text{dist}(\mathbb{P}_P a_i, S)^2}$. Similarly, $\sqrt{\text{dist}(Ba_i^B, S)^2 + \text{apx}_i^2} \geq \text{apx}_i \geq (1-\varepsilon_c)\text{dist}(a_i, P) \geq (1-4\varepsilon_c)\sqrt{\text{dist}(\mathbb{P}_P a_i, S)^2 + \text{dist}(a_i, P)^2}$ by using the fact that $\text{dist}(\mathbb{P}_P a_i, S)^2 \leq 3\varepsilon_c\text{dist}(a_i, P)^2$.

We say that any $i \in [n]$ that is not *small*, is *large*. By the triangle inequality, we obtain that

$$\text{dist}(\mathbb{P}_P a_i, S) - \text{dist}(\mathbb{P}_P a_i, Ba_i^B) \leq \text{dist}(Ba_i^B, S) \leq \text{dist}(\mathbb{P}_P a_i, S) + \text{dist}(Ba_i^B, \mathbb{P}_P a_i). \tag{2}$$

As i is *large*, $\text{dist}(\mathbb{P}_P a_i, S) - \text{dist}(\mathbb{P}_P a_i, Ba_i^B) > 0$ and therefore by the AM-GM inequality, we obtain that

$$\text{dist}(Ba_i^B, S)^2 = (1 \pm \varepsilon)\text{dist}(\mathbb{P}_P a_i, S)^2 + \left(1 \pm \frac{1}{\varepsilon}\right)\text{dist}(Ba_i^B, \mathbb{P}_P a_i)^2.$$

Thus, $\text{dist}(Ba_i^B, S)^2 \leq (1+\varepsilon)\text{dist}(\mathbb{P}_P a_i, S)^2 + (2/\varepsilon)(3\varepsilon_c)\text{dist}(a_i, P)^2$ and $\text{dist}(Ba_i^B, S)^2 \geq (1-\varepsilon)\text{dist}(\mathbb{P}_P a_i, S)^2 - (1/\varepsilon)(3\varepsilon_c)\text{dist}(a_i, P)^2$. Letting $\varepsilon_c = \varepsilon^2/6$, we finally have

$$\text{dist}(Ba_i^B, S)^2 + \text{apx}_i^2 \leq (1+\varepsilon)\text{dist}(\mathbb{P}_P a_i, S)^2 + (1+2\varepsilon)\text{dist}(a_i, P)^2$$

and

$$\text{dist}(Ba_i^B, S)^2 + \text{apx}_i^2 \geq (1-\varepsilon)\text{dist}(\mathbb{P}_P a_i, S)^2 + (1-3\varepsilon)\text{dist}(a_i, P)^2.$$

Therefore, by combining both *small* and *large* indices,

$$\sum_i \sqrt{\text{dist}(Ba_i^B, S)^2 + \text{apx}_i^2} \leq \sqrt{1+O(\varepsilon)}\sum_i \sqrt{\text{dist}(\mathbb{P}_P a_i, S)^2 + \text{dist}(a_i, P)^2}$$

and

$$\sum_i \sqrt{\text{dist}(Ba_i^B, S)^2 + \text{apx}_i^2} \geq \sqrt{1-O(\varepsilon)}\sum_i \sqrt{\text{dist}(\mathbb{P}_P a_i, S)^2 + \text{dist}(a_i, P)^2}.$$

The theorem now follows from Theorem 8 of (Sohler and Woodruff, 2018). □

B. Missing Proofs from Section 4

B.1. Lopsided Embeddings and Gaussian Matrices

Recall $\|\cdot\|_h$ is defined as $\|A\|_h = \sum_j \|A_{*j}\|_2$. Note that $\|A\|_h = \|A^{\mathsf{T}}\|_{1,2}$ for all matrices A. The following lemma shows that lopsided-ε embeddings for certain matrices w.r.t. the norm $\|\cdot\|_h$ imply a dimension reduction for $\|\cdot\|_{1,2}$ subspace approximation.

Lemma B.1. *Given a matrix $A \in \mathbb{R}^{n\times d}$ and a parameter $k \in \mathbb{Z}_{>0}$, let $U_k \in \mathbb{R}^{n\times k}$ and $V_k^\mathsf{T} \in \mathbb{R}^{k\times d}$ be matrices such that*

$$\|U_k V_k^\mathsf{T} - A\|_{1,2} = \min_{\textit{rank-}k\ X} \|A(I - X)\|_{1,2}.$$

If S is a lopsided ε-embedding for (V_k, A^T) with respect to the norm $\|\cdot\|_h$, then

$$\min_{\textit{rank-}k\ X} \|AS^\mathsf{T}X - A\|_{1,2} \le (1 + O(\varepsilon)) \min_{\textit{rank-}k\ X} \|A(I - X)\|_{1,2}.$$

Proof. Note that $\|V_k U_k^\mathsf{T} - A^\mathsf{T}\|_h = \min_Y \|V_k Y^\mathsf{T} - A^\mathsf{T}\|_h$. By definition of a lopsided embedding, we have the following for any matrix Y:

$$\|YV_k^\mathsf{T}S^\mathsf{T} - AS^\mathsf{T}\|_{1,2} = \|SV_kY^\mathsf{T} - SA^\mathsf{T}\|_h \ge (1-\varepsilon)\|V_kY^\mathsf{T} - A^\mathsf{T}\|_h = (1-\varepsilon)\|YV_k^\mathsf{T} - A\|_{1,2}$$

and also that

$$\|U_kV_k^\mathsf{T}S^\mathsf{T} - AS^\mathsf{T}\|_{1,2} = \|SV_kU_k^\mathsf{T} - SA^\mathsf{T}\|_h \le (1+\varepsilon)\|V_kU_k^\mathsf{T} - A^\mathsf{T}\|_h = (1+\varepsilon)\|U_kV_k^\mathsf{T} - A\|_{1,2}.$$

Using these guarantees we now show that the column span of the matrix AS^T contains a good solution to the subspace approximation problem. First consider the minimization problem

$$\min_Y \|YV_k^\mathsf{T} - A\|_{1,2}.$$

Clearly, U_k is the optimal solution to the problem. Now consider the optimal solution $\widetilde{Y}$ to the sketched version of the above problem

$$\widetilde{Y} = \arg\min_Y \|YV_k^\mathsf{T}S^\mathsf{T} - AS^\mathsf{T}\|_{1,2}.$$

We can see that $\widetilde{Y} = (AS^\mathsf{T})(V_k^\mathsf{T}S^\mathsf{T})^+$. Now

$$\|\widetilde{Y}V_k^\mathsf{T} - A\|_{1,2} \le \frac{1}{1-\varepsilon}\|\widetilde{Y}V_k^\mathsf{T}S^\mathsf{T} - AS^\mathsf{T}\|_{1,2} \le \frac{1}{1-\varepsilon}\|U_KV_k^\mathsf{T}S^\mathsf{T} - AS^\mathsf{T}\| \le \frac{1+\varepsilon}{1-\varepsilon}\|U_kV_k^\mathsf{T} - A\|_{1,2}.$$

Therefore,

$$\min_{\text{rank-k}\ X} \|AS^\mathsf{T}X - A\|_{1,2} \le \|AS^\mathsf{T}(V_k^\mathsf{T}S^\mathsf{T})^+(V_k^\mathsf{T}) - A\|_{1,2} \le \frac{1+\varepsilon}{1-\varepsilon}\|U_kV_k^\mathsf{T} - A\|_{1,2} \le (1+3\varepsilon)\min_{\text{rank-}k\ X}\|A(I-X)\|_{1,2}.$$

Thus, if the number of rows of S is less than d, we obtain a dimension reduction for $\|\cdot\|_{1,2}$ subspace approximation. $\square$

Clarkson and Woodruff (2015) give the following sufficient conditions for a distribution of matrices to be an ε-lopsided embedding for (A, B). For the sake of completeness we reproduce their proof here.

Lemma B.2 (Sufficient Conditions)**.** *Given matrices (A, B), let $\mathbf{S}$ be a matrix drawn from a distribution such that*

1. *the matrix $\mathbf{S}$ is a subspace ε-contraction for A with respect to $\|\cdot\|_2$, i.e., simultaneously for all vectors x*

$$\|\mathbf{S}Ax\|_2 \ge (1-\varepsilon)\|Ax\|_2$$

   with probability $1 - \delta/3$,

2. *for all $i \in [d']$, with probability at least $1 - \delta\varepsilon^2/3$ the matrix $\mathbf{S}$ is a subspace ε^2-contraction for $[A\ B_{*i}]$ with respect to $\|\cdot\|_2$, i.e., for all vectors x,*

$$\|\mathbf{S}Ax - \mathbf{S}B_{*i}\|_2 \ge (1-\varepsilon^2)\|Ax - B_{*i}\|_2,$$

   and

3. *the matrix $\mathbf{S}$ is an ε^2-dilation for B^* with respect to $\|\cdot\|_h$, i.e., $\|\mathbf{S}B^*\|_h \le (1+\varepsilon^2)\|B^*\|_h$ with probability $\ge 1 - \delta/3$.*

In the Condition 3 above, $B^* = AX^* - B$ where $X^* = \arg\min_X \|AX - B\|_h$. With failure probability at most δ, the matrix $\mathbf{S}$ is an affine 6ε-contraction for (A, B) with respect to $\|\cdot\|_h$, i.e., for all matrices X,

$$\|\mathbf{S}(AX - B)\|_h \geq (1 - 6\varepsilon)\|AX - B\|_h$$

and therefore a lopsided 6ε-embedding for (A, B) with respect to $\|\cdot\|_h$.

Importantly, note that Condition 2 in the lemma is about the probability of $\mathbf{S}$ being a subspace contraction for $[A\, B_{*i}]$ separately for each i and *not* the probability of $\mathbf{S}$ being *simultaneously* a subspace contraction for $[A\, B_{*i}]$ for all $i \in [d']$.

Proof. Condition on the event that 1 and 3 hold. For $i \in [d']$, let $\mathbf{Z}_i$ be an indicator random variable where $\mathbf{Z}_i = 0$ if the matrix $\mathbf{S}$ is a subspace ε^2-contraction for $[A\ B_{*i}]$ and $\mathbf{Z}_i = 1$ otherwise. From the properties of $\mathbf{S}$, we have that $\Pr[\mathbf{Z}_i = 1] \leq \delta\varepsilon^2/3$ for all i. If $\mathbf{Z}_i = 1$, we call i *bad* and if $\mathbf{Z}_i = 0$, we call i *good*.

Consider an arbitrary matrix X. Say a *bad* i is *large* if $\|(AX - B)_{*i}\|_2 \geq (1/\varepsilon)(\|B_{*i}\|_2 + \|\mathbf{S}B_{*i}\|_2)$, otherwise a *bad* i is *small*. We have

$$\sum_{\textit{small } i} \|(AX - B)_{*i}\|_2 \leq (1/\varepsilon) \sum_{\textit{small } i} \|B_{*i}\|_2 + \|\mathbf{S}B_{*i}\|_2 \leq (1/\varepsilon) \sum_{\textit{bad } i} \|B_{*i}\|_2 + \|\mathbf{S}B_{*i}\|_2. \tag{3}$$

Using condition 2, we obtain that $\mathbb{E}[\sum_{\textit{bad } i} \|B^*_{*i}\|_2] \leq (\delta\varepsilon^2/3) \sum_i \|B^*_{*i}\|_2 \leq (\delta\varepsilon^2/3)\Delta^*$. By a Markov bound, we have that with probability $\geq 1 - \delta/3$, $\sum_{\textit{bad } i} \|B^*_{*i}\| \leq \varepsilon^2\Delta^*$. Assume that this event holds. Similarly,

$$\begin{aligned}
\sum_{\textit{bad } i} \|\mathbf{S}B^*_{*i}\|_2 &= \|\mathbf{S}B^*\|_h - \sum_{\textit{good } i} \|\mathbf{S}B^*_{*i}\|_2 \\
&\leq (1 + \varepsilon^2)\Delta^* - (1 - \varepsilon^2) \sum_{\textit{good } i} \|B^*_{*i}\|_2 \\
&\leq (1 + \varepsilon^2)\Delta^* - (1 - \varepsilon^2)(\Delta^* - \varepsilon^2\Delta^*) \\
&\leq 3\varepsilon^2\Delta^*.
\end{aligned}$$

Thus, we can bound the RHS of (3) and obtain

$$\sum_{\textit{small } i} \|(AX - B)_{*i}\|_2 \leq (1/\varepsilon)(\varepsilon^2\Delta^* + 3\varepsilon^2\Delta^*) \leq 4\varepsilon\Delta^*.$$

Now we lower bound $\sum_{\textit{bad } i} \|\mathbf{S}(AX - B)_{*i}\|_2$.

$$\begin{aligned}
\sum_{\textit{bad } i} \|\mathbf{S}(AX - B)_{*i}\|_2 &\geq \sum_{\textit{large } i} \|\mathbf{S}(AX - B)_{*i}\|_2 \\
&\geq \sum_{\textit{large } i} \|\mathbf{S}(AX - AX^*)_{*i}\|_2 - \|\mathbf{S}B^*_{*i}\|_2 \\
&\geq \sum_{\textit{large } i} (1 - \varepsilon)\|(AX - AX^*)_{*i}\|_2 - \|\mathbf{S}B^*_{*i}\|_2 \\
&\geq \sum_{\textit{large } i} (1 - \varepsilon)\|(AX - B)_{*i}\|_2 - (1 - \varepsilon)\|B^*_{*i}\|_2 - \|\mathbf{S}B^*_{*i}\|_2 \\
&\geq \sum_{\textit{large } i} (1 - \varepsilon)\|(AX - B)_{*i}\|_2 - \varepsilon\|(AX - B)_{*i}\|_2 \\
&\geq (1 - 2\varepsilon) \sum_{\textit{large } i} \|(AX - B)_{*i}\|_2.
\end{aligned}$$

In the above, we repeatedly used the triangle inequality for the $\|\cdot\|_2$ norm, and that $\mathbf{S}$ is a subspace ε-embedding for matrix A and for large i, we upper bound $(1 - \varepsilon)\|B^*_{*i}\|_2 + \|\mathbf{S}B^*_{*i}\|_2$ by $\varepsilon\|(AX - B)_{*i}\|_2$. We can finally lower bound

$\|\mathbf{S}(AX - B)\|_h$.

$$\begin{aligned}\|\mathbf{S}(AX - B)\|_h &= \sum_{\textit{good } i} \|\mathbf{S}(AX - B)_{*i}\|_2 + \sum_{\textit{bad } i} \|\mathbf{S}(AX - B)_{*i}\|_2 \\ &\geq (1-\varepsilon^2) \sum_{\textit{good } i} \|(AX - B)_{*i}\|_2 + (1-2\varepsilon) \sum_{\textit{large } i} \|(AX - B)_{*i}\|_2 \\ &\geq (1-\varepsilon^2) \sum_{\textit{good } i} \|(AX - B)_{*i}\|_2 + (1-2\varepsilon) \sum_{\textit{bad } i} \|(AX - B)_{*i}\|_2 \\ &\quad - (1-2\varepsilon) \sum_{\textit{small } i} \|(AX - B)_{*i}\|_2 \\ &\geq (1-2\varepsilon)\|AX - B\|_h - (1-2\varepsilon)4\varepsilon\Delta^* \\ &\geq (1-6\varepsilon)\|AX - B\|_h.\end{aligned}$$

Thus, by a union bound, with failure probability $\leq \delta$, $\mathbf{S}$ is an affine 6ε-contraction for (A, B) with respect to $\|\cdot\|_h$. □

Lemma B.3 (Gaussian Matrices are Lopsided Embeddings)**.** *Given arbitrary matrices A of rank k and B of* any *rank, a Gaussian matrix $\mathbf{S}$ with $\widetilde{O}(k/\varepsilon^4 + 1/\varepsilon^4\delta^2)$ rows is an ε-lopsided embedding for (A, B) with probability $\geq 1 - \delta$.*

Proof. We now show that a Gaussian matrix, with small dimension equal to $\widetilde{O}(k/\varepsilon^4 + 1/\varepsilon^4\delta^2)$, satisfies all of the sufficient conditions of Lemma B.2. Clearly, a Gaussian matrix with $O((k + \log(1/\delta))/\varepsilon^2)$ rows satisfies condition 1 and a Gaussian matrix with $O((k + \log(1/\delta\varepsilon))/\varepsilon^4)$ rows satisfies condition 2 (Woodruff, 2014).

We now show that a Gaussian matrix with at least $O(1/\varepsilon^4)$ rows satisfies

$$\mathbb{E}[(\|\mathbf{S}y\|_2^2 - 1)^2] \leq \varepsilon^4$$

for any given unit vector y. If $\mathbf{S}$ is a Gaussian matrix of t rows with each entry drawn i.i.d. from $N(0, 1/t)$, then the entries of Sy are each drawn i.i.d. from $N(0, \|y\|_2^2/t) = N(0, 1/t)$. Therefore, $\|\mathbf{S}y\|_2^2 = \mathbf{Y}_1^2 + \ldots + \mathbf{Y}_t^2$, where $\mathbf{Y}_i \sim N(0, 1/t)$, which gives

$$\begin{aligned}\mathbb{E}[(\|\mathbf{S}y\|_2^2 - 1)^2] &= \mathbb{E}[(\mathbf{Y}_1^2 + \ldots + \mathbf{Y}_t^2 - 1)^2] \\ &= t\mathbb{E}[\mathbf{Y}_1^4] + 1 + 2\binom{t}{2}\mathbb{E}[\mathbf{Y}_1^2\mathbf{Y}_2^2] - 2t\mathbb{E}[\mathbf{Y}_1^2] = t\frac{3}{t^2} + 1 + 2\binom{t}{2}\frac{1}{t^2} - 2t\frac{1}{t} \\ &= 2/t.\end{aligned}$$

Thus, with $t \geq 1/\varepsilon^4$, we have that $\mathbb{E}[(\|\mathbf{S}y\|_2^2 - 1)^2] \leq \varepsilon^4$. By Lemma 28 of (Clarkson and Woodruff, 2015), we obtain that $\mathbb{E}[\max(\|\mathbf{S}y\|_2^4, 1)] \leq (1 + \varepsilon^2)^2 \leq 1 + 3\varepsilon^2$. Now, by Holder's inequality,

$$\mathbb{E}[\max(\|\mathbf{S}y\|_2, 1)] \leq \mathbb{E}[\max(\|\mathbf{S}y\|_2, 1)^4]^{1/4} \leq (1 + 3\varepsilon^2)^{1/4} \leq 1 + (3/4)\varepsilon^2.$$

As $(\|\mathbf{S}y\|_2 - 1)_+ = \max(\|\mathbf{S}y\|_2, 1) - 1$, we obtain that $\mathbb{E}[(\|\mathbf{S}y\|_2 - 1)_+] \leq (3/4)\varepsilon^2$, which implies by scaling that for an arbitrary vector y,

$$\mathbb{E}[(\|\mathbf{S}y\|_2 - \|y\|_2)_+] \leq (3/4)\varepsilon^2\|y\|_2$$

which gives

$$\mathbb{E}[(\|\mathbf{S}B^*\|_h - \|B^*\|_h)_+] \leq (3/4)\varepsilon^2\|B^*\|_h.$$

By Markov's inequality, with probability $\geq 1 - \delta/3$, $(\|\mathbf{S}B^*\|_h - \|B^*\|_h)_+ \leq (9/4)(\varepsilon^2/\delta)\|B^*\|_h$ and hence, with probability $\geq 1 - \delta/3$, $\|\mathbf{S}B^*\|_h \leq (1 + (9/4)(\varepsilon^2/\delta))\|B^*\|_h$. Thus, a Gaussian matrix with $m = O(1/\varepsilon^4\delta^2)$ rows satisfies that with probability $\geq 1 - \delta/3$ that

$$\|\mathbf{S}B^*\|_h \leq (1 + \varepsilon^2)\|B^*\|_h.$$

□

B.2. Utilizing Sampling based ℓ_1 embeddings

Let A be a matrix that has r columns. Suppose $\mathbf{L}$ is a random matrix such that with probability $\geq 9/10$, simultaneously for all vectors y,

$$\alpha\|Ay\|_1 \leq \|\mathbf{L}Ay\|_1 \leq \beta\|Ay\|_1.$$

Assume the above event holds. Let X be an arbitrary matrix with t columns. We have that for a suitably scaled Gaussian matrix $\mathbf{G}$ with $\widetilde{O}(t/\varepsilon^2)$ columns, with probability $\geq 9/10$, simultaneously for all vectors $x \in \mathbb{R}^t$, $\|x^\mathsf{T}\mathbf{G}\|_1 = (1 \pm \varepsilon)\|x\|_2$ (Matoušek, 2013). Thus there exists a matrix M with $\widetilde{O}(t/\varepsilon^2)$ columns such that for all vectors $x \in \mathbb{R}^t$,

$$\|x^\mathsf{T}M\|_1 = (1 \pm \varepsilon)\|x\|_2.$$

Therefore,

$$\frac{1}{1+\varepsilon}\|AXM\|_{1,1} \leq \|AX\|_{1,2} = \frac{1}{1-\varepsilon}\|AXM\|_{1,1}$$

and

$$\frac{1}{1+\varepsilon}\|\mathbf{L}AXM\|_{1,1} \leq \|\mathbf{L}AX\|_{1,2} \leq \frac{1}{1-\varepsilon}\|\mathbf{L}AXM\|_{1,1}$$

Now we upper bound $\|\mathbf{L}AX\|_{1,2}$.

$$\begin{aligned}\|\mathbf{L}AX\|_{1,2} &\leq \frac{1}{1-\varepsilon}\|\mathbf{L}AXM\|_{1,1} \leq \frac{1}{1-\varepsilon}\sum_j \|\mathbf{L}A(XM)_{*j}\|_1 \\ &\leq \frac{\beta}{1-\varepsilon}\sum_j \|A(XM)_{*j}\|_1 = \frac{\beta}{1-\varepsilon}\|AXM\|_{1,1} \leq \beta\frac{1+\varepsilon}{1-\varepsilon}\|AX\|_{1,2}.\end{aligned}$$

We now lower bound $\|\mathbf{L}AX\|_{1,2}$ similarly.

$$\begin{aligned}\|\mathbf{L}AX\|_{1,2} &\geq \frac{1}{1+\varepsilon}\|\mathbf{L}AXM\|_{1,1} = \frac{1}{1+\varepsilon}\sum_j \|\mathbf{L}A(XM)_{*j}\|_1 \\ &\geq \frac{\alpha}{1+\varepsilon}\sum_j \|A(XM)_{*j}\|_1 = \frac{\alpha}{1+\varepsilon}\|AXM\|_{1,1} \geq \alpha\frac{1-\varepsilon}{1+\varepsilon}\|AX\|_{1,2}.\end{aligned}$$

By picking appropriate ε, we conclude that for any matrix X,

$$\frac{\alpha}{2}\|AX\|_{1,2} \leq \|\mathbf{L}AX\|_{1,2} \leq 2\beta\|AX\|_{1,2}. \tag{4}$$

Lemma B.4. *If $\mathbf{S}^\mathsf{T}$ is a random Gaussian matrix with $O(k)$ columns such that with probability $\geq 9/10$,*

$$\min_{\text{rank-}k\ X}\|A\mathbf{S}^\mathsf{T}X - A\|_{1,2} \leq (3/2)\min_{\text{rank-}k\ X}\|AX - A\|_{1,2},$$

and if $\mathbf{L}$ is a random matrix drawn from a distribution such that with probability $\geq 9/10$ over the draw of matrix $\mathbf{L}$,

$$\alpha\|A\mathbf{S}^\mathsf{T}y\|_1 \leq \|\mathbf{L}A\mathbf{S}^\mathsf{T}y\|_1 \leq \beta\|A\mathbf{S}^\mathsf{T}y\|_1$$

for all vectors y and

$$\mathbb{E}_{\mathbf{L}}[\|\mathbf{L}M\|_{1,2}] = \|M\|_{1,2}$$

for any matrix M, then with probability $\geq 3/5$, all matrices X such that $\|\mathbf{L}A\mathbf{S}^\mathsf{T}X - \mathbf{L}A\|_{1,2} \leq 10 \cdot \mathrm{SubApx}_{k,1}(A)$ satisfy

$$\|A\mathbf{S}^\mathsf{T}X - A\|_{1,2} \leq \big(2 + 40/\alpha\big)\,\mathrm{SubApx}_{k,1}(A).$$

Proof. Let $X_1 = \arg\min_{\text{rank-}k\ X} \|A\mathbf{S}^\mathsf{T}X - A\|_{1,2}$. With probability $\geq 9/10$, we have that $\|A\mathbf{S}^\mathsf{T}X_1 - A\|_{1,2} \leq (3/2)\text{SubApx}_{k,1}(A)$. By a Markov bound, we obtain that with probability $\geq 4/5$, $\|\mathbf{L}A\mathbf{S}^\mathsf{T}X_1 - \mathbf{L}A\|_{1,2} \leq 10\text{SubApx}_{k,1}(A)$. Assume this event holds. For any matrix X,

$$\|\mathbf{L}A\mathbf{S}^\mathsf{T}X - \mathbf{L}A\|_{1,2} \geq \|\mathbf{L}A\mathbf{S}^\mathsf{T}X - \mathbf{L}A\mathbf{S}^\mathsf{T}X_1\|_{1,2} - \|\mathbf{L}A\mathbf{S}^\mathsf{T}X_1 - \mathbf{L}A\|_{1,2}.$$

We have

$$\|\mathbf{L}A\mathbf{S}^\mathsf{T}X - \mathbf{L}A\|_{1,2} \geq \|\mathbf{L}A\mathbf{S}^\mathsf{T}X - \mathbf{L}A\mathbf{S}^\mathsf{T}X_1\|_{1,2} - 10 \cdot \text{SubApx}_{k,1}(A).$$

From (4), we have

$$\begin{aligned}\|\mathbf{L}A\mathbf{S}^\mathsf{T}X - \mathbf{L}A\|_{1,2} &\geq \frac{\alpha}{2}\|A\mathbf{S}^\mathsf{T}X - A\mathbf{S}^\mathsf{T}X_1\|_{1,2} - 10 \cdot \text{SubApx}_{k,1}(A)\\ &\geq \frac{\alpha}{2}\|A\mathbf{S}^\mathsf{T}X - A\|_{1,2} - \frac{\alpha}{2}\|A\mathbf{S}^\mathsf{T}X_1 - A\|_{1,2} - 10 \cdot \text{SubApx}_{k,1}(A)\\ &\geq \frac{\alpha}{2}\|A\mathbf{S}^\mathsf{T}X - A\|_{1,2} - (3\alpha/4 + 10) \cdot \text{SubApx}_{k,1}(A).\end{aligned}$$

Thus, for any matrix X of rank r, if $\|A\mathbf{S}^\mathsf{T}X - A\|_{1,2} > (2/\alpha)(20 + 3\alpha/4) \cdot \text{SubApx}_{k,1}(A)$, then $\|\mathbf{L}A\mathbf{S}^\mathsf{T}X - \mathbf{L}A\|_{1,2} > 10 \cdot \text{SubApx}_{k,1}(A)$. $\square$

B.3. Main Theorem for constructing an $(O(1), \widetilde{O}(k))$-bicriteria solution

Theorem B.1. *Given any matrix $A \in \mathbb{R}^{n\times d}$ and a matrix $B \in \mathbb{R}^{d\times c_1}$ with orthonormal columns, Algorithm 1 returns a matrix $\widehat{X}$ with $\widetilde{O}(k)$ orthonormal columns that with probability $1-\delta$ satisfies*

$$\|A(I - BB^\mathsf{T})(I - \widehat{X}\widehat{X}^\mathsf{T})\|_{1,2} \leq O(1) \cdot \text{SubApx}_{k,1}(A(I - BB^\mathsf{T})),$$

in time $\widetilde{O}((\mathsf{nnz}(A) + d\mathsf{poly}(k/\varepsilon))\log(1/\delta))$.

Proof. It is shown in Lemma B.3 that a Gaussian matrix with $O(k)$ rows is a $1/6$-lopsided embedding for (V_k, A^T) with probability $\geq 9/10$. Thus by Lemma B.1, we obtain that

$$\min_{\text{rank-}k\ X} \|A(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T}X - A(I - BB^\mathsf{T})\|_{1,2} \leq (3/2)\text{SubApx}_{k,1}(A(I - BB^\mathsf{T}))$$

with probability $\geq 9/10$. (Cohen and Peng, 2015) show that a sampling matrix $\mathbf{L}$ obtained using Lewis weights has $\widetilde{O}(k)$ rows and is a $(1/2, 3/2)$ ℓ_1 subspace embedding for the matrix $A(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T}$. Thus, the matrices $\mathbf{S}^\mathsf{T}$ and $\mathbf{L}$ constructed in Algorithm 1 satisfy the conditions of Lemma 4.1. Therefore from Lemma B.4, with probability $\geq 3/5$, if a matrix X satisfies $\|\mathbf{L}A(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T}X - \mathbf{L}A(I - BB^\mathsf{T})\|_{1,2} \leq 10 \cdot \text{SubApx}_{k,1}(A(I - BB^\mathsf{T}))$, then $\|A(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T}X - A(I - BB^\mathsf{T})\|_{1,2} \leq 82 \cdot \text{SubApx}_{k,1}(A(I - BB^\mathsf{T}))$.

Let $\widetilde{X} = \arg\min_{\text{rank-}k\ X} \|A(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T}X - A(I - BB^\mathsf{T})\|_{1,2}$. We have $\|A(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T}\widetilde{X} - A(I - BB^\mathsf{T})\|_{1,2} \leq (3/2)\text{SubApx}_{k,1}(A(I - BB^\mathsf{T}))$. By Markov's bound, with probability $\geq 3/4$, $\|\mathbf{L}A(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T}\widetilde{X} - \mathbf{L}A(I - BB^\mathsf{T})\|_{1,2} \leq 10 \cdot \text{SubApx}_{k,1}(A(I - BB^\mathsf{T}))$. We now have the following:

$$\|\mathbf{L}A(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T}\widetilde{X}(\mathbf{L}A(I - BB^\mathsf{T}))^+\mathbf{L}A(I - BB^\mathsf{T}) - \mathbf{L}A(I - BB^\mathsf{T})\|_{1,2} \leq 10 \cdot \text{SubApx}_{k,1}(A(I - BB^\mathsf{T})).$$

Thus $\|A(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T}\widetilde{X}(\mathbf{L}A(I - BB^\mathsf{T}))^+\mathbf{L}A(I - BB^\mathsf{T}) - A(I - BB^\mathsf{T})\|_{1,2} \leq 82 \cdot \text{SubApx}_{k,1}(A(I - BB^\mathsf{T}))$. Finally,

$$\begin{aligned}&\|A(I - BB^\mathsf{T})(\mathbf{L}A(I - BB^\mathsf{T}))^+(\mathbf{L}A(I - BB^\mathsf{T})) - A(I - BB^\mathsf{T})\|_{1,2}\\ &\leq \|A(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T}\widetilde{X}(\mathbf{L}A(I - BB^\mathsf{T}))^+\mathbf{L}A(I - BB^\mathsf{T}) - A(I - BB^\mathsf{T})\|_{1,2}\\ &\leq 82 \cdot \text{SubApx}_{k,1}(A(I - BB^\mathsf{T})).\end{aligned}$$

The first inequality follows from the fact that for all x and y, $\|x^\mathsf{T}(\mathbf{L}A)^+(\mathbf{L}A) - x^\mathsf{T}\|_2 \leq \|y^\mathsf{T}(\mathbf{L}A)^+(\mathbf{L}A) - x^\mathsf{T}\|_2$.

By a union bound, with probability $\geq 1/2$, the matrix $\widehat{X}$ computed by Algorithm 1, which is an orthonormal basis for the rowspace of $\mathbf{L}A(I - BB^\top)$, satisfies

$$\|A(I - BB^\top)(I - \widehat{X}\widehat{X}^\top)\|_{1,2} \leq 82 \cdot \mathsf{SubApx}_{k,1}(A(I - BB^\top)).$$

Thus the matrix $\widehat{X}$ which has the minimum value over $\widetilde{O}(\log(1/\delta))$ trials satisfies with probability $\geq 1 - \delta$ that

$$\|A(I - BB^\top)(I - \widehat{X}\widehat{X}^\top)\|_{1,2} \leq O(1) \cdot \mathsf{SubApx}_{k,1}(A(I - BB^\top)).$$

The running time of Lewis weight sampling can be seen to be $O((\mathsf{nnz}(A) + k^2 d(c_1 + k)) \log(\log(n)))$ from (Cohen and Peng, 2015). Thus, the total running time is $\widetilde{O}((\mathsf{nnz}(A) + k^2 d(c_1 + k)) \log(1/\delta))$. □

B.4. Finding Best Solution Among Candidate Solutions

Algorithm 1 finds candidate solutions $\widehat{X}^{(1)}, \ldots, \widehat{X}^{(t)}$ for $t = O(\log(1/\delta))$ and returns the best candidate solution that minimizes the cost

$$\|A(I - BB^\top)(I - \widehat{X}\widehat{X}^\top)\|_{1,2}. \tag{5}$$

The proof of Theorem 4.1 shows that, for all $i = 1, \ldots, t$, with probability $\geq 3/5$, $\|A(I - BB^\top)(I - \widehat{X}^{(i)}(\widehat{X}^{(i)})^\top)\|_{1,2} \leq O(1) \cdot \mathsf{SubApx}_{k,1}(A(I - BB^\top))$. Therefore with probability $\geq 1 - \delta/2$

$$\min_i \|A(I - BB^\top)(I - \widehat{X}^{(i)}(\widehat{X}^{(i)})^\top)\|_{1,2} \leq O(1) \cdot \mathsf{SubApx}_{k,1}(A(I - BB^\top)) \tag{6}$$

i.e., with probability $\geq 1 - \delta$, there is a solution $\widehat{X}^{(i)}$ among the t potential solutions that has a cost at most $O(1) \cdot \mathsf{SubApx}_{k,1}(A(I - BB^\top))$. We first compute

$$\mathsf{apx}_i = \|A(I - BB^\top)(I - \widehat{X}^{(i)}(\widehat{X}^{(i)})^\top)\mathbf{G}\|_{1,2}$$

where $\mathbf{G}$ is a scaled Gaussian matrix with $O(\log(n/\delta))$ columns. Values of apx_j for all $j \in [t]$ can be computed in time $\widetilde{O}((\mathsf{nnz}(A) + (n + d)\mathsf{poly}(k/\varepsilon)) \cdot \log(1/\delta))$. We have using the union bound that, with probability $\geq 1 - \delta/2$, for all $j \in [n]$ and $i \in [t]$ that

$$\|A_{j*}(I - BB^\top)(I - \widehat{X}^{(i)}(\widehat{X}^{(i)})^\top)\mathbf{G}\|_2 = (1/2, 3/2)\|A_{j*}(I - BB^\top)(I - \widehat{X}^{(i)}(\widehat{X}^{(i)})^\top)\|_2. \tag{7}$$

Therefore with probability $\geq 1 - \delta/2$, for all $i \in [t]$,

$$\mathsf{apx}_i \in (1/2, 3/2)\|A(I - BB^\top)(I - \widehat{X}^{(i)}(\widehat{X}^{(i)})^\top)\|_{1,2}. \tag{8}$$

Let $\widetilde{i} = \arg\min_{i \in [t]} \mathsf{apx}_i$ and $i^* = \arg\min_{i \in [t]} \|A(I - BB^\top)(I - \widehat{X}^{(i)}(\widehat{X}^{(i)})^\top)\|_{1,2}$. By a union bound, with probability $\geq 1 - \delta$

$$\begin{aligned}
\|A(I - BB^\top)(I - \widehat{X}^{(\widetilde{i})}(\widehat{X}^{(\widetilde{i})})^\top)\|_{1,2} &\leq 2\mathsf{apx}_{\widetilde{i}} \\
&\leq 2\mathsf{apx}_{i^*} \\
&\leq 4\|A(I - BB^\top)(I - \widehat{X}^{(i^*)}(\widehat{X}^{(i^*)})^\top)\|_{1,2} \\
&\leq O(1) \cdot \mathsf{SubApx}_{k,1}(A(I - BB^\top)).
\end{aligned}$$

Thus, Algorithm 1, with probability $\geq 1 - \delta$, returns a subspace that has cost at most $O(\sqrt{k}) \cdot \mathsf{SubApx}_{k,1}(A(I - BB^\top))$ and has a running time of $\widetilde{O}((\mathsf{nnz}(A) + (n + d)\mathsf{poly}(k/\varepsilon)) \cdot \log(1/\delta))$.

B.5. Main Theorem for Constructing a $(1 + \varepsilon, k^{3.5}/\varepsilon^2)$ Bicriteria Solution

Theorem B.2 (Residual Sampling)**.** *Given matrix $A \in \mathbb{R}^{n \times d}$, matrices $B \in \mathbb{R}^{d \times c_1}$ and $\widehat{X} \in \mathbb{R}^{d \times c_2}$ with orthonormal columns such that $\|A(I - BB^\top)(I - \widehat{X}\widehat{X}^\top)\|_{1,2} \leq K \cdot \mathsf{SubApx}_{1,k}(A(I - BB^\top))$, Algorithm 2 returns a matrix U having $c = \widetilde{O}(c_2 + K \cdot k^3/\varepsilon^2 \cdot \log(1/\delta))$ orthonormal columns such that with probability $\geq 1 - \delta$*

$$\|A(I - BB^\top)(I - UU^\top)\|_{1,2} \leq (1 + \varepsilon)\mathsf{SubApx}_{1,k}(A(I - BB^\top)) \tag{9}$$

in time $\widetilde{O}(\mathsf{nnz}(A) + d \cdot \mathsf{poly}(k/\varepsilon))$. Moreover we also have that $U^\top B = 0$, i.e., the column spaces of U and B are orthogonal to each other.

Proof. As the matrix G is a Gaussian matrix with $t = O(\log(n/\delta))$ columns, we have that with probability $\geq 1-(\delta/2)$, for all $i \in [n]$,

$$\|M_{i*}\|_2 = \|A_{i*}(I - BB^\mathsf{T})(I - \widehat{X}\widehat{X}^\mathsf{T})G\|_2 = (1 \pm 1/10)\|A_{i*}(I - BB^\mathsf{T})(I - \widehat{X}\widehat{X}^\mathsf{T})G\|_2.$$

Therefore, the probabilities p_i computed by Algorithm 2 are such that

$$p_i = \frac{\|M_{i*}\|_2}{\|M\|_{1,2}} \geq \frac{(9/10)\|A_{i*}(I - BB^\mathsf{T})(I - \widehat{X}\widehat{X}^\mathsf{T})\|_2}{(11/10)\|A(I - BB^\mathsf{T})(I - \widehat{X}\widehat{X}^\mathsf{T})\|_{1,2}} \geq \frac{9}{11}\frac{\|A_{i*}(I - BB^\mathsf{T})(I - \widehat{X}\widehat{X}^\mathsf{T})\|_2}{\|A(I - BB^\mathsf{T})(I - \widehat{X}\widehat{X}^\mathsf{T})\|_{1,2}}.$$

Hence, by applying Lemma 4.2 to the matrix $A(I - BB^\mathsf{T})$, we obtain that with probability $\geq 1-\delta$, the matrix U returned by Algorithm 2 satisfies

$$\|A(I - BB^\mathsf{T})(I - UU^\mathsf{T})\|_{1,2} \leq (1+\varepsilon)\text{SubApx}_{1,k}(A(I - BB^\mathsf{T})).$$

The matrix M can be computed in time $O(\mathsf{nnz}(A)\log(n/\delta) + (c_1 + c_2)d\log(n/\delta))$. And $s = \widetilde{O}(K \cdot k^3/\varepsilon^2 \cdot \log(1/\delta))$ independent samples can be drawn from the distribution p in time $O(n+s)$. Finally, the orthonormal basis U can be computed in time $O(d(c+c_1)^2) = O(d\mathsf{poly}(k/\varepsilon))$. $\square$

C. Missing Proofs from Section 5

Lemma C.1. *With probability $\geq 2/3$, Algorithm 3 finds an $\widetilde{O}(k^3/\varepsilon^3)$-dimensional subspace S such that for all k-dimensional subspaces W,*

$$\|A(I - \mathbb{P}_S)\|_{1,2} - \|A(I - \mathbb{P}_{S+W})\|_{1,2} \leq 4\varepsilon \cdot \text{SubApx}_{k,1}(A).$$

Proof. Suppose that the loop in Algorithm 3 is run for all $t = 10/\varepsilon + 1$ iterations instead of stopping after i^* iterations. Let $\widehat{X}_i, U_i, B_i$ be the values of the matrices in the algorithm at the end of i iterations. Let $B_0 = []$ be the empty matrix. Condition on the event that all the calls to Algorithm 1 in the algorithm succeed. By a union bound over the failure event of each call to Algorithm 1, this event holds with probability $\geq 9/10$. Therefore, by Theorem 4.1, we obtain that

$$\begin{aligned}&\|A(I - \mathbb{P}_{B_{i-1}})(I - \mathbb{P}_{\widehat{X}_i})\|_{1,2}\\ &\leq \widetilde{O}(\sqrt{k}) \cdot \text{SubApx}_{k,1}(A(I - \mathbb{P}_{B_{i-1}}))\end{aligned}$$

for all $i \in [10/\varepsilon + 1]$ and also that $\widehat{X}_i$ has $\widetilde{O}(k)$ columns. Now we condition on the event that all the calls to Algorithm 2 succeed. By a union bound, this holds with probability $\geq 9/10$. Thus we have

$$\begin{aligned}&\|A(I - \mathbb{P}_{B_i})\|_{1,2} = \|A(I - \mathbb{P}_{B_{i-1}})(I - \mathbb{P}_{U_i})\|_{1,2}\\ &\leq (1+\varepsilon) \cdot \text{SubApx}_{k,1}(A(I - \mathbb{P}_{B_{i-1}}))\end{aligned}$$

for all iterations $i \in [10/\varepsilon + 1]$ and also that U_i has $\widetilde{O}(k^3/\varepsilon^2)$ columns which implies that B_i has $\widetilde{O}(ik^3/\varepsilon^2)$ columns. In particular, we have that $\|A(I - \mathbb{P}_{B_1})\|_{1,2} \leq (1+\varepsilon)\text{SubApx}_{k,1}(A)$. Therefore

$$\begin{aligned}&(1+\varepsilon)\text{SubApx}_{k,1}(A) - \|A(I - \mathbb{P}_{B_t})\|_{1,2}\\ &\geq \|A(I - \mathbb{P}_{B_1})\|_{1,2} - \|A(I - \mathbb{P}_{B_t})\|_{1,2}\\ &= \sum_{i=2}^{\mathsf{T}} \|A(I - \mathbb{P}_{B_{i-1}})\|_{1,2} - \|A(I - \mathbb{P}_{B_i})\|_{1,2} \geq 0.\end{aligned}$$

The last inequality follows from the fact that colspace$(B_i) \supseteq$ colspace(B_{i-1}). The summation in the above equation has $10/\varepsilon$ non-negative summands that all sum to at most $(1+\varepsilon)\text{SubApx}_{k,1}(A)$. Therefore, at least $9/\varepsilon$ summands have value $\leq \varepsilon(1+\varepsilon)\text{SubApx}_{k,1}(A)$. In particular, with probability $\geq 9/10$,

$$\|A(I - \mathbb{P}_{B_{i^*}})\|_{1,2} - \|A(I - \mathbb{P}_{B_{i^*+1}})\|_{1,2} \leq \varepsilon(1+\varepsilon)\text{SubApx}_{k,1}(A).$$

But we also have that

$$\begin{aligned}\|A(I-\mathbb{P}_{B_{i^*+1}})\|_{1,2} &= \|A(I-\mathbb{P}_{B_{i^*}})(I-\mathbb{P}_{U_{i^*}})\|_{1,2}\\ &\le (1+\varepsilon)\text{SubApx}_{k,1}(A(I-\mathbb{P}_{B_{i^*}}))\\ &\le (1+\varepsilon)\|A(I-\mathbb{P}_{B_{i^*}})(I-\mathbb{P}_W)\|_{1,2}\\ &= (1+\varepsilon)\|A(I-\mathbb{P}_{B_{i^*}+W})\|_{1,2}\end{aligned}$$

where W is any rank k matrix. The second inequality follows from the fact that $\text{SubApx}_{k,1}(A(I-\mathbb{P}_{B_{i^*}})) = \min_{\text{rank-}k\ W}\|A(I-\mathbb{P}_{B_{i^*}})(I-\mathbb{P}_W)\|_{1,2}$. Therefore, for any rank-k matrix W, we obtain that

$$\begin{aligned}&\|A(I-\mathbb{P}_{B_{i^*}})\|_{1,2} - \|A(I-\mathbb{P}_{B_{i^*}\cup W})\|_{1,2}\\ &\le \|A(I-\mathbb{P}_{B_{i^*}})\|_{1,2} - \frac{1}{1+\varepsilon}\|A(I-\mathbb{P}_{B_{i^*+1}})\|_{1,2}\\ &\le \|A(I-\mathbb{P}_{B_{i^*}})\|_{1,2} - (1-\varepsilon)\|A(I-\mathbb{P}_{B_{i^*+1}})\|_{1,2}\\ &\le (\|A(I-\mathbb{P}_{B_{i^*}})\|_{1,2} - \|A(I-\mathbb{P}_{B_{i^*+1}})\|_{1,2}) + \varepsilon\|A(I-\mathbb{P}_{B_{i^*+1}})\|_{1,2}\\ &\le 4\varepsilon\cdot\text{SubApx}_{k,1}(A).\end{aligned}$$

□

Theorem C.1. *Given a matrix $A\in\mathbb{R}^{n\times d}$, $k\in\mathbb{Z}$ and an accuracy parameter $\varepsilon>0$, Algorithm 4 returns a matrix B with $\widetilde{O}(k^3/\varepsilon^6)$ orthonormal columns and a matrix $Apx = [X\,v]$ such that for any k dimensional shape S, $\sum_i\sqrt{\text{dist}(BX_{i*}^{\mathsf{T}},S)^2+v_i^2} = (1\pm\varepsilon)\sum_i \text{dist}(A_i,S)$. The algorithm runs in time $O(\mathsf{nnz}(A)/\varepsilon^2 + (n+d)\mathsf{poly}(k/\varepsilon))$.*

Proof. From the above lemma, we have that the subspace B satisfies with probability $\ge 9/10$, that for any k dimensional subspace W,

$$\|A(I-\mathbb{P}_B)\|_{1,2} - \|A(I-\mathbb{P}_{B\cup W})\|_{1,2} \le \frac{\varepsilon^2}{80}\text{SubApx}_{k,1}(A). \tag{10}$$

From Theorem 2.10 of (Woodruff, 2014), we obtain that with probability $\ge 9/10$, for all $i\in[n]$, the matrix $\mathbf{S}_j$ found for $i\in[n]$ is such that $\mathbf{S}_j$ is a $\Theta(\varepsilon^2)$ subspace embedding for the matrix $[B\,A_{i*}^{\mathsf{T}}]$. Therefore, x_i is such that

$$\|Bx_i - A_{i*}^{\mathsf{T}}\|_2 \le (1+\Theta(\varepsilon^2))\|(I-BB^{\mathsf{T}})A_{i*}^{\mathsf{T}}\|_2.$$

and $v_i = (1\pm\Theta(\varepsilon^2))\|(I-BB^{\mathsf{T}})A_{i*}^{\mathsf{T}}\|_2$. Now the proof follows from Theorem 3.1. □

D. Missing Proofs from Section 6

D.1. Obtaining an $(O(1),\mathsf{poly}(k))$ Approximation

Theorem D.1. *Given $A\in\mathbb{R}^{n\times d}$, $B\in\mathbb{R}^{d\times c_1}$, $k\in\mathbb{Z}$ and δ, Algorithm 5 returns $\widehat{X}$ with $\widetilde{O}(k^{3.5})$ orthonormal columns that with probability $1-\delta$ satisfies*

$$\|A(I-BB^{\mathsf{T}})(I-\widehat{X}\widehat{X}^{\mathsf{T}})\|_{1,2} \le O(1)\cdot\text{SubApx}_{k,1}(A(I-BB^{\mathsf{T}})).$$

Given that the matrices $\mathbf{C}_1 A_{I_j}$ for all $j\in[b]$ and $\mathbf{W}A$ are precomputed for all $O(\log(1/\delta))$ trials, the algorithm can be implemented in time $\widetilde{O}(((nd/b)\cdot k^{3.5} + d\mathsf{poly}(k/\varepsilon))\log(1/\delta))$.

Proof. The proof is similar to proof of Theorem 4.1. That proof only makes use of the facts that

1. for any fixed matrix M, $\mathbb{E}[\|\mathbf{L}M\|_{1,2}] = \mathbb{E}[\|M\|_{1,2}]$,
2. with probability $\ge 9/10$, for all vectors x, $(1/2)\|Ax\|_1 \le \|\mathbf{L}Ax\|_1 \le (3/2)\|Ax\|_1$, and

to conclude with the statement in the theorem. We now show that the matrix $\mathbf{L}$ computed by Algorithm 5 satisfies all the above three properties.

Algorithm 5 POLYAPPROXDENSE

Input: $A \in \mathbb{R}^{n\times d}$, $B \in \mathbb{R}^{d\times c_1}$, $k \in \mathbb{Z}, \delta, b$
Output: $\widehat{X} \in \mathbb{R}^{d\times c_2}$
cols $\leftarrow O(k + 1/\delta^2)$
$\mathbf{S}^\mathsf{T} \leftarrow \mathcal{N}(0,1)^{d\times \text{cols}}$
$\mathbf{W} \leftarrow \ell_1$ embedding for $O(k)$ dimensions from (Wang and Woodruff, 2019)
$[Q, R] \leftarrow$ QR decomposition of $(\mathbf{W}A)(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T}$
$I_1, \ldots, I_b \leftarrow$ Equal size partition of $[n]$ into b parts
$\mathbf{C}_1 \leftarrow$ Cauchy matrix with $O(\log(n))$ rows
for $j = 1, \ldots, b$ **do**
  $M^{(j)} \leftarrow (\mathbf{C}_1 A_{I_j})(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T} R^{-1}$
  $\text{apx}_j \leftarrow \sum_{\text{col}\in\text{cols}(M^{(j)})} \text{median}(\text{abs}(M^{(j)}_{*\text{col}}))$
end for
$\mathbf{C} \leftarrow$ Cauchy matrix with $O(\log(n))$ columns
samples $\leftarrow \widetilde{O}(k^{3.5})$
$\mathbf{L} \leftarrow []$
for samples iterations **do**
  Sample $j \in [b]$ with probability proportional to apx_j
  $P^{(j)} \leftarrow A_{I_j}(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T} R^{-1}\mathbf{C}$
  For $i \in I_j$, $p^{(j)}_i \leftarrow \text{median}(\text{abs}(P^{(j)}_{i*}))$
  Sample $i \in I_j$ with probability proportional to $p^{(j)}_i$
  Append $\frac{1}{\frac{p^{(j)}_i}{\sum_{i\in I(j)} p^{(j)}_i}\frac{\text{apx}_j}{\sum_{j=1}^b \text{apx}_j}\cdot\text{samples}} e_i^\mathsf{T}$ to matrix $\mathbf{L}$
end for
$\widehat{X} \leftarrow$ Orthonormal Basis for rowspace$(\mathbf{L}A(I - BB^\mathsf{T}))$
Repeat the above O$(\log(1/\delta))$ times and return best $\widehat{X}$

Note that the random matrix $\mathbf{L}$ is constructed by sampling N rows, where each row is independently equal to $(1/Np_i)e_i^\mathsf{T}$ with probability p_i. Thus

$$\mathbb{E}[\|\mathbf{L}M\|_{1,2}] = \mathbb{E}[\sum_{i=1}^{N} \|\mathbf{L}_{i*}M\|_2] = N\mathbb{E}[\|\mathbf{L}_{1*}M\|_2] = N\sum_{j=1}^{n} \|(1/Np_j)e_j^\mathsf{T} M\|_2 p_j = \sum_{j=1}^{n} \|M_{j*}\|_2 = \|M\|_{1,2}. \quad (11)$$

We now prove property 2. From Theorem 1.3 of (Wang and Woodruff, 2019), we have that $\mathbf{W}$ has $O(k\log(k))$ rows and that with probability $\geq 99/100$, for all vectors x

$$\|A(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T} x\|_1 \leq \|\mathbf{W}A(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T} x\|_1 \leq O(k\log(k))\|A(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T} x\|.$$

Let $\ell_i = \|A_{i*}(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T} R^{-1}\|_1$ for $i \in [n]$. From Theorem 6.1, if the probability that the i^{th} row is sampled is $\geq (1/2)(\ell_i/\sum_{i'}\ell_{i'})$ for all $i \in [n]$, then the matrix $\mathbf{L}$ constructed is a $(1/2, 3/2)$ ℓ_1-subspace embedding with probability $\geq 99/100$. Now consider sampling a row of the matrix $\mathbf{L}$ in the algorithm. We have that the sampled row is in the direction of e_i with probability $(\text{apx}_{j(i)}/\sum_{j'\in[b]}\text{apx}_{j'}) \cdot (p^{j(i)}_i/\sum_{i'\in I_{j(i)}} p^{j(i)}_{i'})$. We use $j(i)$ to denote $j \in [b]$ such that $i \in I_j$. We show that this probability is at least $(1/2)(\ell_i/\sum_{i'}\ell_{i'})$. For $j \in [b]$,

$$\text{apx}_j = \sum_{\text{col}} \text{median}(\text{abs}(M^{(j)}_{*\text{col}})).$$

From Theorem 1 of (Indyk, 2006), we have with probability $\geq 1 - 1/100b$ that

$$\text{median}(\text{abs}(M^{(j)}_{*\text{col}})) = (1 \pm 1/6)\|A_{I_j}(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T} R^{-1}_{*\text{col}}\|_1.$$

Thus $\sum_{\text{col}} \text{median}(\text{abs}(M^{(j)}_{*\text{col}})) = (1 \pm 1/6) \sum_{\text{col}} \|A_{I_j}(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T} R^{-1}_{*\text{col}}\|_1 = (1 \pm 1/6) \sum_{i \in I_j} \|A_{i*}(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T} R^{-1}\|_1 = (1 \pm 1/6) \sum_{i \in I_j} \ell_i$. Therefore, by a union bound, with probability $\geq 99/100$, for all $j \in [b]$

$$\text{apx}_j = (1 \pm 1/6) \sum_{i \in I_j} \ell_i.$$

Again, from (Indyk, 2006), we obtain that with probability $\geq 99/100$, that for all $i \in [n]$

$$\text{median}(\text{abs}(A_{i*}(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T} R^{-1}\mathbf{C})) = (1 \pm 1/6)\|A_{i*}(I - BB^\mathsf{T})\mathbf{S}^\mathsf{T} R^{-1}\|_1 = (1 \pm 1/6)\ell_i.$$

Thus, with probability $\geq 99/100$, for all j and $i \in I_j$, we have $p^{(j)}_{(i)} = (1 \pm 1/6)\ell_i$. By a union bound, with probability $\geq 98/100$, the probability that an arbitrary row i is sampled in an iteration of the algorithm is

$$(\text{apx}_{j(i)} / \sum_{j' \in [b]} \text{apx}_{j'}) \cdot (p^{j(i)}_i / \sum_{i' \in I_{j(i)}} p^{j(i)}_{i'}) \geq \frac{5}{7} \frac{\sum_{i' \in I_j} \ell_{i'}}{\sum_{i' \in [n]} \ell_{i'}} \frac{5}{7} \frac{\ell_i}{\sum_{i' \in I_j} \ell_{i'}} \geq \frac{1}{2} \frac{\ell_i}{\sum_{i' \in [n]} \ell_{i'}}.$$

Thus by a union bound, $\mathbf{L}$ is a $(1/2, 3/2)$ subspace embedding. Now the proof and argument for the running time follow. $\square$

D.2. Obtaining a $(1 + \varepsilon, \mathsf{poly}(k/\varepsilon))$ Solution

Algorithm 6 EPSAPPROXDENSE

Input: $A \in \mathbb{R}^{n \times d}$, $B \in \mathbb{R}^{d \times c_1}$, $\widehat{X} \in \mathbb{R}^{d \times c_2}$, $k \in \mathbb{Z}, K, \varepsilon, \delta, b$
Output: $U \in \mathbb{R}^{d \times c}$
$t \leftarrow O(\log(n))$, $\mathbf{G} \leftarrow \mathcal{N}(0,1)^{d \times \text{cols}}$
$I_1, \dots, I_b \leftarrow$ Equal size partition of $[n]$ into b parts
$\mathbf{C}_1 \leftarrow$ Cauchy matrix with $O(\log(n))$ rows
for $j = 1, \dots, b$ **do**
  $M^{(j)} \leftarrow (\mathbf{C}_1 A_{I_j})(I - BB^\mathsf{T})(I - \widehat{X}\widehat{X}^\mathsf{T})(\mathbf{G}/t)\sqrt{\pi/2}$
  $\text{apx}_j \leftarrow \sum_{\text{col} \in \text{cols}(M^{(j)})} \text{median}(\text{abs}(M^{(j)}_{*\text{col}}))$
end for
samples $\leftarrow \widetilde{O}(K \cdot k^3/\varepsilon^2 \cdot \log(1/\delta))$, $\mathbf{S} \leftarrow \varnothing$
for samples iterations **do**
  Sample $j \in [b]$ with probability proportional to apx_j
  $P^{(j)} \leftarrow A_{I_j}(I - BB^\mathsf{T})(I - \widehat{X}\widehat{X}^\mathsf{T})\mathbf{G}$
  For $i \in I_j$, $p^{(j)}_i \leftarrow \|P^{(j)}_{i*}\|_2$
  Sample $i \in I_j$ with probability proportional to $p^{(j)}_i$
  $\mathbf{S} \leftarrow \mathbf{S} \cup i$
end for
$U \leftarrow \text{colspan}((I - BB^\mathsf{T})[\widehat{X}\ (A_\mathbf{S})^\mathsf{T}])$
Return U

Theorem D.2. *Given a matrix $A \in \mathbb{R}^{n \times d}$, orthonormal matrices $B \in \mathbb{R}^{n \times c_1}$ and $\widehat{X} \in \mathbb{R}^{n \times c_2}$ such that*

$$\|A(I - BB^\mathsf{T})(I - \widehat{X}\widehat{X}^\mathsf{T})\|_{1,2} \leq K \cdot \mathit{SubApx}_{1,k}(A(I - BB^\mathsf{T})),$$

and parameters k, ε, and δ, Algorithm 6 outputs a matrix U with $c = c_1 + \widetilde{O}(K \cdot k^3/\varepsilon^2 \cdot \log(1/\delta))$ orthonormal columns such that with probability $\geq 1 - \delta$,

$$\|A(I - BB^\mathsf{T})(I - UU^\mathsf{T})\|_{1,2} \leq (1 + \varepsilon)\mathit{SubApx}_{1,k}(A(I - BB^\mathsf{T})).$$

Given that $\mathbf{C}_1 A_{I_j}$ is precomputed for all $j \in [b]$, the algorithm runs in time $\widetilde{O}((nd/b) \cdot (K \cdot k^3/\varepsilon^2 \log(1/\delta)) + d\mathsf{poly}(k/\varepsilon))$.

Proof. We show that the probability that a row i is sampled in an iteration of the Algorithm is $\geq (1/12)\|A_{i*}(I-BB^{\mathsf{T}})(I-\widehat{X}\widehat{X}^{\mathsf{T}})\|_2/\|A(I-BB^{\mathsf{T}})(I-\widehat{X}\widehat{X}^{\mathsf{T}})\|_{1,2}$. Then the proof follows as in the proof of Theorem 4.2. First assume that apx_j for $j \in [b]$ computed by the algorithm satisfies

$$\text{apx}_j = (1/2, 2)\sum_{i \in I_j} \|A_{j*}(I-BB^{\mathsf{T}})(I-\widehat{X}\widehat{X}^{\mathsf{T}})\|_2.$$

Now the probability p_i with which a row i is sampled by the algorithm is given by

$$p_i = \frac{\text{apx}_{j(i)}}{\sum_{j\in[b]}\text{apx}_j} \cdot \frac{\|A_{i*}(I-BB^{\mathsf{T}})(I-\widehat{X}\widehat{X}^{\mathsf{T}})\mathbf{G}\|_2}{\sum_{i' \in I_{j(i)}}\|A_{i'*}(I-BB^{\mathsf{T}})(I-\widehat{X}\widehat{X}^{\mathsf{T}})\mathbf{G}\|_2}.$$

As $\mathbf{G}$ is a Gaussian matrix with $t = O(\log(n/\delta))$ columns, we have that with probability $\geq 1-\delta$ that for all $i' \in [n]$ $\|A_{i'*}(I-BB^{\mathsf{T}})(I-\widehat{X}\widehat{X}^{\mathsf{T}})\mathbf{G}\|_2 = (1 \pm 1/2)\|A_{i'*}(I-BB^{\mathsf{T}})(I-\widehat{X}\widehat{X}^{\mathsf{T}})\|_2 \cdot \sqrt{t}$. Therefore

$$\begin{aligned} p_i &= \frac{\text{apx}_{j(i)}}{\sum_{j\in[b]}\text{apx}_j} \cdot \frac{\|A_{i*}(I-BB^{\mathsf{T}})(I-\widehat{X}\widehat{X}^{\mathsf{T}})\mathbf{G}\|_2}{\sum_{i' \in I_{j(i)}}\|A_{i'*}(I-BB^{\mathsf{T}})(I-\widehat{X}\widehat{X}^{\mathsf{T}})\mathbf{G}\|_2} \\ &\geq \frac{1}{12}\frac{\|A_{i*}(I-BB^{\mathsf{T}})(I-\widehat{X}\widehat{X}^{\mathsf{T}})\|_2}{\|A(I-BB^{\mathsf{T}})(I-\widehat{X}\widehat{X}^{\mathsf{T}})\|_{1,2}}. \end{aligned}$$

We now prove our assumption which concludes the proof.

Let $x \in \mathbb{R}^d$ be an arbitrary vector. As $\mathbf{G}$ is a Gaussian matrix with $t = O(\log(n/\delta))$ columns, Lemma 5.3 of (Plan and Vershynin, 2013) states that

$$\Pr\left[\left|\frac{1}{t}\|x^{\mathsf{T}}G\|_1 - \sqrt{\frac{2}{\pi}}\|x\|_2\right| \geq \alpha\|x\|_2\right] \leq C\exp(-ct\alpha^2).$$

Picking an appropriate $\alpha = O(1)$, by a union bound, with probability $\geq 1-\delta/3$, we obtain

$$\begin{aligned} &\|A_{i*}(I-BB^{\mathsf{T}})(I-\widehat{X}\widehat{X}^{\mathsf{T}})(G/t)\sqrt{\pi/2}\|_1 \\ &= (4/5, 6/5)\|A_{i*}(I-BB^{\mathsf{T}})(I-\widehat{X}\widehat{X}^{\mathsf{T}})\|_2 \end{aligned}$$

for all $i \in [n]$. Now, if C is a Cauchy matrix with $O(\log(n/\delta))$ rows, then with probability $1-\delta/(3nb)$, we have that

$$\text{median}(\text{abs}(Cx)) = (1 \pm 1/5)\|x\|_1.$$

Therefore, by a union bound, we obtain that, with probability $\geq 1-\delta/3$, for all $j \in [b]$ and $i \in t$ that

$$\text{median}(\text{abs}(CA_{I_j*}(I-BB^{\mathsf{T}})(I-\widehat{X}\widehat{X}^{\mathsf{T}})\mathbf{G}_{*i})) = (1 \pm 1/5)\|A_{I_j*}(I-BB^{\mathsf{T}})(I-\widehat{X}\widehat{X}^{\mathsf{T}})\mathbf{G}_{*i}\|_1.$$

Therefore, with probability $\geq 1-2\delta/3$, for all $j \in [b]$,

$$\begin{aligned} \text{apx}_j &= \sum_i \text{median}(\text{abs}((M^{(j)})_{*i})) = \sum_{i=1}^{\mathsf{T}} \text{median}(\text{abs}(CA_{I_j*}(I-BB^{\mathsf{T}})(I-\widehat{X}\widehat{X}^{\mathsf{T}})(\mathbf{G}_{*i}/t)\sqrt{\pi/2})) \\ &= (1 \pm 1/5)\sum_{i=1}^{\mathsf{T}}\|A_{I_j*}(I-BB^{\mathsf{T}})(I-\widehat{X}\widehat{X}^{\mathsf{T}})(\mathbf{G}_{*i}/t)\sqrt{\pi/2}\|_1 \\ &= (1 \pm 1/5)\sum_{i \in I_j}\|A_{i*}(I-BB^{\mathsf{T}})(I-\widehat{X}\widehat{X}^{\mathsf{T}})(\mathbf{G}/t)\sqrt{\pi/2}\|_1 \\ &= (4/5, 6/5)(4/5, 6/5)\sum_{i \in I_j}\|A_{i*}(I-BB^{\mathsf{T}})(I-\widehat{X}\widehat{X}^{\mathsf{T}})\|_2 \\ &= (1/2, 2)\sum_{i \in I_j}\|A_{i*}(I-BB^{\mathsf{T}})(I-\widehat{X}\widehat{X}^{\mathsf{T}})\|_2. \end{aligned}$$

The only term in the running time that involves a factor nd is in computing the matrix $P^{(j)}$ for the chosen j. A total of $\widetilde{O}(K \cdot k^3/\varepsilon^2 \cdot \log(1/\delta))$ such $j \in [b]$ are sampled. Therefore, the total running time for computing the matrices $P^{(j)}$ for j sampled by the algorithm is equal to $(nd/b) \cdot \log(n) \cdot \widetilde{O}(K \cdot k^3/\varepsilon^2 \cdot \log(1/\delta)) + d \cdot \mathsf{poly}(k/\varepsilon)$. $\square$

Algorithm 7 DimensionReductionDense

Input: $A \in \mathbb{R}^{n \times d}$, $k, \varepsilon > 0$.
Output: $B \in \mathbb{R}^{d \times c}$ with orthonormal columns
$t \leftarrow 10/\varepsilon + 1$
$i^* \leftarrow$ uniformly random integer from $[10/\varepsilon + 1]$.
Initialize $B \leftarrow []$
$b \leftarrow k^{3.5}/\varepsilon^3$
$\delta = \Theta(\varepsilon)$
for $i = 1, \dots, i^*$ **do**
  $\widehat{X} \leftarrow$ PolyApproxDense(A, B, k, δ, b).
  $U \leftarrow$ EpsApproxDense$(A, B, \widehat{X}, k, \widetilde{O}(\sqrt{k}), \Theta(\varepsilon), \delta, b)$.
  $B \leftarrow [B \,|\, U]$.
end for
Return B.

D.3. Overall Algorithm

Lemma D.1. *Given matrix $A \in \mathbb{R}^{n\times d}$, $k \in \mathbb{Z}$ and $\varepsilon > 0$, Algorithm 7 returns a matrix B with $\widetilde{O}(k^{3.5}/\varepsilon^3)$ orthonormal columns such that, with probability $\geq 3/5$, for all k dimensional spaces W,*

$$\|A(I - \mathbb{P}_B)\|_{1,2} - \|A(I - \mathbb{P}_{B\cup W})\|_{1,2} \leq \varepsilon \cdot \mathit{SubApx}_{k,1}(A).$$

The Algorithm runs in time $\widetilde{O}(nd + (n+d)\mathsf{poly}(k/\varepsilon))$.

Proof. The proof of the lemma is similar to that of Lemma C.1. We now argue that all the pre-computed matrices required across all the iterations of the algorithm can be computed in time $\widetilde{O}(nd)$. The Cauchy matrix $\mathbf{C}_1$ used in Algorithm 5 has $O(\log(n))$ rows and the matrix $\mathbf{W}$ has $\widetilde{O}(k)$ rows. Note that we have

$$\begin{bmatrix} \mathbf{C}_1 A_{I_1} \\ \mathbf{C}_1 A_{I_2} \\ \\ \mathbf{C}_1 A_{I_b} \\ \mathbf{W} A \end{bmatrix} = \begin{bmatrix} \mathbf{C}_1 & & & \\ & \mathbf{C}_1 & & \\ & & & \\ & & & \mathbf{C}_1 \\ & & \mathbf{W} & \end{bmatrix} A.$$

Thus all the matrices required for Algorithm 5 can be computed by multiplying a $\mathsf{poly}(k/\varepsilon) \times n$ matrix with A. Similarly, we can compute all the matrices required for Algorithm 6 by computing the product of a $\mathsf{poly}(k/\varepsilon) \times n$ matrix with A. Thus, all the matrices required across all iterations of Algorithm 7 can be computed by multiplying a $\mathsf{poly}(k/\varepsilon) \times n$ matrix with A, which can be done in time $\widetilde{O}(nd)$ by the algorithm of Coppersmith (1982), assuming $n \gg \mathsf{poly}(k/\varepsilon)$. Now each iteration of the loop in Algorithm 7 takes $\widetilde{O}((nd/b)k^{3.5}/\varepsilon^2 + (n+d)\mathsf{poly}(k/\varepsilon))$ time. As there are $O(1/\varepsilon)$ iterations, the algorithm runs in time $\widetilde{O}((nd/b)k^{3.5}/\varepsilon^3 + (n+d)\mathsf{poly}(k/\varepsilon))$. Since the value of b is chosen to be $k^{3.5}/\varepsilon^3$, we obtain that the running time of the algorithm is $\widetilde{O}(nd + (n+d)\mathsf{poly}(k/\varepsilon))$, including the time to compute the required pre-computed matrices. □

E. Coreset Construction using Dimensionality Reduction

Algorithm 8 gives the general algorithm to construct a coreset for any objective involving the sum-of-distances metric. In this section, we discuss the coreset construction for two such problems: the k-median and k-subspace approximation problems.

For $(B, \text{Apx} = [X\ v])$ returned by Algorithm 4, we have the guarantee that, with probability $\geq 9/10$, for any k-dimensional shape S,

$$\sum_i \sqrt{\text{dist}(BX_{i*}^{\mathsf{T}}, S)^2 + v_i^2} = (1 \pm \varepsilon) \sum_i \text{dist}(A_{i*}, S).$$

Algorithm 8 CORESETCONSTRUCTION

Input: $A \in \mathbb{R}^{n\times d}$, k, ε
Output: Coreset
$(B, \text{Apx}) \leftarrow$ COMPLETEDIMREDUCE(A, k, ε)
Construct a coreset for the instance Apx $\begin{bmatrix} B^\top & 0 \\ 0 & 1 \end{bmatrix}$ and return

Given a set S, let S_{+1} denote the set $\{(s, 0) \mid s \in S\}$. Let $\text{diag}(B^\top, 1) = \begin{bmatrix} B^\top & 0 \\ 0 & 1 \end{bmatrix}$. Using this notation, we have that

$$\sum_i \text{dist}(\text{Apx}_{i*} \cdot \text{diag}(B^\top, 1), S_{+1}) = (1 \pm \varepsilon) \sum_i \text{dist}(A_{i*}, S).$$

Using the above relation, we give a coreset construction for the k-subspace approximation and k-median problems. These constructions are as in (Sohler and Woodruff, 2018). For any matrix M, let M_{+1} denote the matrix M with a new column of 0s appended at the end and let M_{-1} denote the matrix M with the last column deleted.

Theorem E.1 (Coreset for Subspace Approximation)**.** *There exists a sampling-and-scaling matrix T that samples and scales $\widetilde{O}(k^3/\varepsilon^8)$ rows of the matrix Apx such that, with probability $\geq 3/5$, for any projection matrix P of rank k that projects onto a subspace S of dimension at most k, we have*

$$\begin{aligned} \|((T \cdot \textit{Apx} \cdot \textit{diag}(B^\top, 1))_{-1}P)_{+1} - T \cdot \textit{Apx} \cdot \textit{diag}(B^\top, 1)\|_{1,2} &= (1 \pm O(\varepsilon))\|((\textit{Apx} \cdot \textit{diag}(B^\top, 1))_{-1}P)_{+1} - \textit{Apx} \cdot \textit{diag}(B^\top, 1)\|_{1,2} \\ &= (1 \pm O(\varepsilon)) \sum_i \text{dist}(A_i, S). \end{aligned}$$

This sampling matrix can be computed in time $O(n \cdot \mathsf{poly}(k/\varepsilon))$.

Proof. We first have $\|((\text{Apx} \cdot \text{diag}(B^\top, 1))_{-1}P)_{+1} - \text{Apx} \cdot \text{diag}(B^\top, 1)\|_{1,2} = \sum_i \|((\text{Apx}_{i*} \cdot \text{diag}(B^\top, 1))_{-1}P)_{+1} - \text{Apx}_{i*} \cdot \text{diag}(B^\top, 1)\|_2 = \sum_i \sqrt{\|(I - P)BX_{i*}^\top\|_2^2 + v_i^2} = \sum_i \sqrt{\text{dist}(BX_{i*}^\top, S)^2 + v_i^2} = (1 \pm \varepsilon) \sum_i \text{dist}(A_{i*}, S)$.

We now show $\|((T \cdot \text{Apx} \cdot \text{diag}(B^\top, 1))_{-1}P)_{+1} - T \cdot \text{Apx} \cdot \text{diag}(B^\top, 1)\|_{1,2} = (1 \pm O(\varepsilon))\|((\text{Apx} \cdot \text{diag}(B^\top, 1))_{-1}P)_{+1} - \text{Apx} \cdot \text{diag}(B^\top, 1)\|_{1,2}$ proving the claim. Let G be a Gaussian matrix with $\widetilde{O}(d/\varepsilon^2)$ columns. Then with probability $\geq 9/10$, for all $x \in \mathbb{R}^{d+1}$,

$$\|x^\top G\|_1 = (1 \pm \varepsilon)\|x\|_2.$$

See (Sohler and Woodruff, 2018) for references. Thus we have that with probability $\geq 9/10$, for all projection matrices P of rank at most k, we have

$$\|((\text{Apx} \cdot \text{diag}(B^\top, 1))_{-1}P)_{+1}G - \text{Apx} \cdot \text{diag}(B^\top, 1)G\|_{1,1} = (1 \pm \varepsilon)\|((\text{Apx} \cdot \text{diag}(B^\top, 1))_{-1}P)_{+1} - \text{Apx} \cdot \text{diag}(B^\top, 1)\|_{1,2}.$$

Note that for any P, the columns of the matrix $((\text{Apx} \cdot \text{diag}(B^\top, 1))_{-1}P)_{+1}G - \text{Apx} \cdot \text{diag}(B^\top, 1)G$ lie in the column space of the matrix Apx. Let T be a $(1 \pm \varepsilon)$ ℓ_1-subspace embedding constructed for the matrix Apx constructed using (Cohen and Peng, 2015). Therefore

$$\|T \cdot ((\text{Apx} \cdot \text{diag}(B^\top, 1))_{-1}P)_{+1}G - T \cdot \text{Apx} \cdot \text{diag}(B^\top, 1)G\|_{1,1} = (1 \pm \varepsilon)\|((\text{Apx} \cdot \text{diag}(B^\top, 1))_{-1}P)_{+1}G - \text{Apx} \cdot \text{diag}(B^\top, 1)G\|_{1,1}.$$

Again, using the fact that $\|x^\top G\|_1 = (1 \pm \varepsilon)\|x\|_2$ for all $d + 1$ dimensional vectors x, we obtain that

$$\begin{aligned} &\|T \cdot ((\text{Apx} \cdot \text{diag}(B^\top, 1))_{-1}P)_{+1} - T \cdot \text{Apx} \cdot \text{diag}(B^\top, 1)\|_{1,2} \\ &= (1 \pm \varepsilon)\|T \cdot ((\text{Apx} \cdot \text{diag}(B^\top, 1))_{-1}P)_{+1}G - T \cdot \text{Apx} \cdot \text{diag}(B^\top, 1)G\|_{1,1} \\ &= (1 \pm O(\varepsilon))\|((\text{Apx} \cdot \text{diag}(B^\top, 1))_{-1}P)_{+1}G - \text{Apx} \cdot \text{diag}(B^\top, 1)G\|_{1,1} \\ &= (1 \pm O(\varepsilon))\|((\text{Apx} \cdot \text{diag}(B^\top, 1))_{-1}P)_{+1} - \text{Apx} \cdot \text{diag}(B^\top, 1)\|_{1,2} \\ &= (1 \pm O(\varepsilon)) \sum_i \text{dist}(A_i, S). \end{aligned}$$

The matrix T is computed by Lewis Weight Sampling. As the matrix Apx has dimensions $n \times \widetilde{O}(k^3/\varepsilon^6)$, we see from (Cohen and Peng, 2015) that the matrix T can be computed in time $n \cdot \mathsf{poly}(k/\varepsilon)$. $\square$

Theorem E.2 (Coreset for k-median)**.** *There exists a subset $T \subseteq [n]$ with $|T| = \widetilde{O}(k^4/\varepsilon^8)$ and weights w_i for $i \in T$ such that, with probability $\geq 3/5$, for any set C of size k,*

$$\sum_{i \in T} w_i \mathrm{dist}(Apx_{i*} \cdot diag(B^\mathsf{T}, 1), C_{+1}) = (1 \pm \varepsilon) \sum_{i \in [n]} \mathrm{dist}(A_{i*}, C).$$

Recall that $C_{+1} = \big\{(c, 0) \mid c \in C\big\}$.

Proof. Let S denote the rowspan of the matrix $\mathrm{diag}(B^\mathsf{T}, 1)$. We have $\dim(S) = \widetilde{O}(k^3/\varepsilon^6)$. Let $\widehat{S}$ be the subspace S along with an orthogonal dimension. Thus $\widehat{S}$ is an $\widetilde{O}(k^3/\varepsilon^6)$ dimensional subspace of $\mathbb{R}^{d+1}$. Let $C = \{c_1, \ldots, c_k\}$ be an arbitrary set of k centers of $\mathbb{R}^{d+1}$. Now it is easy to see that we can find a set of k points $\widehat{C} = \{\widehat{c}_1, \ldots, \widehat{c}_k\} \subseteq \widehat{S}$ such that $\mathbb{P}_S c_i = \mathbb{P}_S \widehat{c}_i$ i.e., the projections of c_i and $\widehat{c}_i$ onto the subspace S are the same, and also that $\mathrm{dist}(c_i, \mathbb{P}_S(c_i)) = \mathrm{dist}(\widehat{c}_i, \mathbb{P}_S(\widehat{c}_i))$ and therefore, for any point $a \in S$, $\mathrm{dist}(a, C) = \mathrm{dist}(a, \widehat{C})$.

Now if $T \subseteq [n]$ and the weights w_i for $i \in T$ are such that

$$\sum_{i \in T} w_i \mathrm{dist}(\mathrm{Apx}_{i*} \cdot \mathrm{diag}(B^\mathsf{T}, 1), \widetilde{C}) = (1 \pm \varepsilon) \sum_{i=1}^{n} \mathrm{dist}(\mathrm{Apx}_{i*} \cdot \mathrm{diag}(B^\mathsf{T}, 1), \widetilde{C})$$

for all k-center sets $\widetilde{C} \subseteq \widehat{S}$, then for any k center set $C \subseteq \mathbb{R}^{d+1}$, we have

$$\begin{aligned}
\sum_{i \in T} w_i \mathrm{dist}(\mathrm{Apx}_{i*} \cdot \mathrm{diag}(B^\mathsf{T}, 1), C) &= \sum_{i \in T} w_i \mathrm{dist}(\mathrm{Apx}_{i*} \cdot \mathrm{diag}(B^\mathsf{T}, 1), \widehat{C}) \\
&= (1 \pm \varepsilon) \sum_{i=1}^{n} \mathrm{dist}(\mathrm{Apx}_{i*} \cdot \mathrm{diag}(B^\mathsf{T}, 1), \widehat{C}) \\
&= (1 \pm \varepsilon) \sum_{i=1}^{n} \mathrm{dist}(\mathrm{Apx}_{i*} \cdot \mathrm{diag}(B^\mathsf{T}, 1), C).
\end{aligned}$$

Thus, preserving the k-median distances with respect to the k center sets that lie in $\widehat{S}$, preserves the k-median distances to all the center sets in $\mathbb{R}^{d+1}$. Using the coreset construction of Feldman and Langberg (2011) on the matrix Apx, we can obtain a subset $T \subseteq [n]$ of size $\widetilde{O}(k^4/\varepsilon^8)$ along with weights w_i such that for any k-center set $C \subseteq \mathbb{R}^{d+1}$, we have

$$\sum_{i \in T} w_i \mathrm{dist}(\mathrm{Apx}_{i*} \cdot \mathrm{diag}(B^\mathsf{T}, 1), C) = (1 \pm \varepsilon) \sum_{i=1}^{n} \mathrm{dist}(\mathrm{Apx}_{i*} \cdot \mathrm{diag}(B^\mathsf{T}, 1), C).$$

As Apx is an $n \times \mathsf{poly}(k/\varepsilon)$-sized matrix, the algorithm of Feldman and Langberg (2011) can be run in time $n \cdot \mathsf{poly}(k/\varepsilon)$. Thus, the above subset T and weights w_i for $i \in T$ can be found in time $n\mathsf{poly}(k/\varepsilon)$. Now, for any k-center set $C \subseteq \mathbb{R}^d$, we have that

$$\begin{aligned}
\sum_{i=1}^{n} \mathrm{dist}(A_{i*}, C) &= (1 \pm \varepsilon) \sum_{i=1}^{n} \sqrt{\mathrm{dist}(BX_{i*}^\mathsf{T}, C) + v_i^2} \\
&= (1 \pm \varepsilon) \sum_{i=1}^{n} \mathrm{dist}(\mathrm{Apx}_{i*} \cdot \mathrm{diag}(B^\mathsf{T}, 1), C_{+1}) \\
&= (1 \pm \varepsilon) \sum_{i \in T} w_i \mathrm{dist}(\mathrm{Apx}_{i*} \cdot \mathrm{diag}(B^\mathsf{T}, 1), C_{+1}).
\end{aligned}$$

Therefore we obtain a coreset of size $\widetilde{O}(k^4/\varepsilon^8)$ in overall time $\widetilde{O}(\mathsf{nnz}(A)/\varepsilon^2 + (n + d)\mathsf{poly}(k/\varepsilon))$. $\square$

F. Near-Linear Time Coreset for k-Median

Let $A \in \mathbb{R}^{n \times d}$ be the dataset, where each row A_{i*} of A denotes a point in $\mathbb{R}^d$, for $i \in [n]$. We observe that the coreset construction of Huang and Vishnoi (2020) can be implemented in $\widetilde{O}(\mathsf{nnz}(A) + (n+d)\mathsf{poly}(k/\varepsilon))$ time. The authors only need to compute a constant factor approximation and assignment of each point to a center, which gives a constant factor approximation to the optimum. We show that we can compute such an assignment in time $O(\mathsf{nnz}(A) + (n+d)\mathsf{poly}(k/\varepsilon))$.

The usual k-median objective is the following

$$\min_{y_1,\ldots,y_k \in \mathbb{R}^d} \sum_{i=1}^{n} \min_j \|A_i^* - y_j\|_2.$$

We can restrict y_j to be a row of A_i^* and lose at most a factor of 2 as follows. Suppose $y_1^*, \ldots, y_k^*$ is the optimal solution. Let $\mathcal{C}^* = (\mathcal{C}_1^*, \mathcal{C}_2^*, \ldots, \mathcal{C}_n^*)$ be the partition of $[n]$ induced by the optimal solution $y_1^*, \ldots, y_k^*$, where $\mathcal{C}_j^*$ denotes all the indices i such that y_j^* is the closest center to A_{i*}. Therefore, the optimal cost for k-median is

$$\mathrm{OPT} = \sum_{j=1}^{k} \sum_{i \in \mathcal{C}_j^*} d(A_{i*}, y_j^*).$$

Let $A_{c(j)}$ be the point closest to y_j^*, i.e.,

$$\text{for all } i \in \mathcal{C}_j^*,\ d(A_{i*}, y_j^*) \geq d(A_{c(j)*}, y_j^*).$$

We claim that the k-median cost of the centers $A_{c(1)}, \ldots, A_{c(k)}$ is at most twice the optimum:

$$\sum_{j=1}^{k} \sum_{i \in \mathcal{C}_j^*} d(A_{i*}, A_{c(j)*}) \leq \sum_{j=1}^{k} \left(\sum_{i \in \mathcal{C}_j^*} d(A_{i*}, y_j^*) + d(A_{c(j)*}, y_j^*) \right) \leq \sum_{j=1}^{k} \sum_{i \in \mathcal{C}_j^*} 2d(A_{i*}, y_j^*) \leq 2\mathrm{OPT}.$$

Metric k-median In this version of k-median, we restrict to center sets C that are subsets of the data, i.e., we solve the optimization problem

$$\min_{y_1,\ldots,y_k \in A} \sum_{i=1}^{n} \min_j \|A_i^* - y_j\|_2.$$

Let $\mathrm{OPT}_{\text{metric}}$ denote the optimum objective value for metric k-median. From the above, we obtain that

$$\mathrm{OPT}_{\text{metric}} \leq 2\mathrm{OPT}.$$

Therefore, a c-approximate solution for metric k-median is at most a $2c$-approximate solution for Euclidean k-median. Let Π be a Johnson Lindenstrauss matrix embedding $\mathbb{R}^d$ into $\mathbb{R}^m$, where $m = O(\log(n))$, such that

$$\frac{1}{2} d(A_{i*}, A_{i'*}) \leq d(\Pi A_{i*}, \Pi A_{i'*}) \leq \frac{3}{2} d(A_{i*}, A_{i'*})$$

for all $i, i' \in [n]$. Now consider the metric k-median problem on the points $\Pi A_{1*}, \ldots, \Pi A_{n*}$. We can obtain an 11-approximate solution to the metric k-median problem in time $\widetilde{O}(nk + k^7)$ (see Theorem 6.2 of (Chen, 2009)). Let $A_{c^*(1)*}, \ldots, A_{c^*(k)*}$ be the optimal centers for the metric k-median problem on $A_{1*}, \ldots, A_{n*}$, and $\Pi A_{c'(1)}, \ldots, \Pi A_{c'(k)}$ be an 11-approximate solution to the metric k-median on $\Pi A_{1*}, \ldots, \Pi A_{n*}$. Let $\mathcal{C}' = (\mathcal{C}_1, \ldots, \mathcal{C}_k')$ be the partition of $[n]$ corresponding to this 11-approximate solution. Then the following shows that $A_{c'(1)}, \ldots, A_{c'(k)}$ is a good solution for the metric k-median problem on the original dataset:

$$\begin{aligned}\sum_{j=1}^{k}\sum_{i\in\mathcal{C}'_j} d(A_{i*},A_{c'(j)*}) &\le 2\sum_{j=1}^{k}\sum_{i\in\mathcal{C}'_j} d(\Pi A_{i*},\Pi A_{c'(j)*})\\ &\le 2\cdot 11\sum_{j=1}^{k}\sum_{i\in\mathcal{C}^*_j} d(\Pi A_{i*},\Pi A_{c^*(j)*})\\ &\le 2\cdot 11\cdot\frac{3}{2}\sum_{j=1}^{k}\sum_{i\in\mathcal{C}^*_j} d(A_{i*},A_{c^*(j)*})\\ &\le 33\text{OPT}_{\text{metric}} \le 66\text{OPT}.\end{aligned}$$

The time taken to compute $\Pi A_{1*},\ldots,\Pi A_{n*}$ is $O(\mathsf{nnz}(A)\log(n))$, and then we can compute the k centers and an assignment of points such that this is a 66-approximate solution in time $\widetilde{O}(nk+k^7)$. Using this assignment, we can implement the first stage of importance sampling in the algorithm of Huang and Vishnoi (2020) in time $\widetilde{O}(\mathsf{nnz}(A)+n\cdot\mathsf{poly}(k/\varepsilon))$. We note that the first stage of the algorithm of Huang and Vishnoi (2020) only needs a constant factor approximation of the distance of a point to its assigned centers, which can be computed as $d(\Pi A_{i*},\Pi A_{c'(j)*})$, in time $\widetilde{O}(\log(n))$, if the point i is assigned to cluster j. The second stage of their algorithm can be implemented in time $d\cdot\mathsf{poly}(k/\varepsilon)$. Thus, we can find a strong coreset for k-median in time

$$\widetilde{O}(\mathsf{nnz}(A)+(n+d)\cdot\mathsf{poly}(k/\varepsilon)).$$