To Reviewer 1

We first restate Theorems 1 and 2 in our submitted paper.

Theorem 1 (FFD). *Given data $\mathbf{A} \in \mathbb{R}^{n\times d}$ and the sketching size $\ell \leq k = \min(m, d)$, let the small sketch $\mathbf{B} \in \mathbb{R}^{\ell\times d}$ be constructed by FFD. Then, with probability at least $1 - p\beta - (2p+1)\delta - \frac{2n}{e^k}$ we have*

$$\|\mathbf{A}^T\mathbf{A} - \mathbf{B}^T\mathbf{B}\|_2 \leq \widetilde{O}\Big(\frac{1}{\ell} + \Gamma(\ell, p, k)\Big)\|\mathbf{A}\|_F^2 \quad (1)$$

where $\Gamma(\ell, p, k) = \sqrt{\frac{k}{\ell p^2}} + \sqrt{\frac{1+\sqrt{k/\ell}}{p}}$ with $p = \frac{n}{m}$, and $\widetilde{O}(\cdot)$ hides logarithmic factors on (β, δ, k, d, m).

The running time of the algorithm is $\widetilde{O}(nl^2\frac{d}{m} + nd)$ and its space cost is $O(d\ell)$ before taking $m = \Theta(d)$.

Theorem 2 (FROSH). *Given data $\mathbf{A} \in \mathbb{R}^{n\times d}$ with its row mean vector $\boldsymbol{\mu} \in \mathbb{R}^{1\times d}$, let the sketching matrix $\mathbf{B}^{\ell\times d}$ be generated by FROSH in Algorithm 4. Then, with probability defined in Theorem 1 we have*

$$\begin{aligned}&\|(\mathbf{A}-\boldsymbol{\mu})^T(\mathbf{A}-\boldsymbol{\mu}) - \mathbf{B}^T\mathbf{B}\|_2 \\ &\qquad\leq \widetilde{O}\Big(\frac{1}{\ell} + \Gamma(\ell, p, k)\Big)\|\mathbf{A}-\boldsymbol{\mu}\|_F^2, \quad (2)\end{aligned}$$

where $(\mathbf{A}-\boldsymbol{\mu}) \in \mathbb{R}^{n\times d}$ means subtracting each row of $\mathbf{A}$ by $\boldsymbol{\mu}$, $\Gamma(\ell, p, k) = \sqrt{\frac{k}{\ell p^2}} + \sqrt{\frac{1+\sqrt{k/\ell}}{p}}$ with $p = \frac{n}{m}$, and the top r right singular vectors of $\mathbf{B}^{\ell\times d}$ are used for hashing projections $\mathbf{W}^T \in \mathbb{R}^{r\times d}$.

The algorithm requires $\widetilde{O}(n\ell^2 + nd + d\ell^2)$ running time with $O(d\ell)$ space cost after taking $m = O(d)$ in the FFD of FROSH.

Next, we explicitly show how the singular vectors $\mathbf{W}^T \in \mathbb{R}^{r\times d}$ can be approximated. We let $m = \Theta(d)$, and assume $n = \Omega(\ell^{3/2}d^{3/2})$ for simplicity, then the error bound of Eq. (2) in Theorem 2 becomes $\widetilde{O}(\frac{1}{\ell}\|(\mathbf{A}-\boldsymbol{\mu})\|_F^2)$. Based on it, we give Theorem 3.

Theorem 3. *Given data $\mathbf{A} \in \mathbb{R}^{n\times d}$ with its row mean vector $\boldsymbol{\mu} \in \mathbb{R}^{1\times d}$, let the sketching matrix $\mathbf{B}^{\ell\times d}$ be generated by FROSH in Algorithm 4. Let $m = \Theta(d)$, and assume $n = \Omega(\ell^{3/2}d^{3/2})$ for simplicity. Given $(\mathbf{A}-\boldsymbol{\mu}) \in \mathbb{R}^{n\times d}$ that means subtracting each row of $\mathbf{A}$ by $\boldsymbol{\mu}$, let $h = \|(\mathbf{A}-\boldsymbol{\mu})\|_F^2/\|(\mathbf{A}-\boldsymbol{\mu})\|_2^2$ and σ_i be the i-th largest singular value of $(\mathbf{A}-\boldsymbol{\mu})$. If the sketching size $\ell = \Omega(\frac{h\sigma_1^2}{\epsilon\sigma_{r+1}^2})$, then with probability defined in Theorem 1 we have*

$$\begin{aligned}&\|(\mathbf{A}-\boldsymbol{\mu}) - (\mathbf{A}-\boldsymbol{\mu})\mathbf{W_B}\mathbf{W_B}^T\|_2^2 \\ &\leq (1+\epsilon)\|(\mathbf{A}-\boldsymbol{\mu}) - (\mathbf{A}-\boldsymbol{\mu})\mathbf{W}\mathbf{W}^T\|_2^2, \quad (3)\end{aligned}$$

where $0 < \epsilon < 1$, $\mathbf{W_B}^T \in \mathbb{R}^{r\times d}$ contains the top r right singular vectors of $\mathbf{B}^{\ell\times d}$, and $\mathbf{W}^T \in \mathbb{R}^{r\times d}$ contains the top r right singular vectors of $(\mathbf{A}-\boldsymbol{\mu}) \in \mathbb{R}^{n\times d}$.

Remark. The bound on $\|(\mathbf{A}-\boldsymbol{\mu}) - (\mathbf{A}-\boldsymbol{\mu})\mathbf{W_B}\mathbf{W_B}^T\|_2^2$ shows the similarity between $\mathbf{W_B}\mathbf{W_B}^T$ and $\mathbf{W}\mathbf{W}^T$. If $\epsilon = 0$, we will have $\mathbf{W_B}\mathbf{W_B}^T = \mathbf{W}\mathbf{W}^T$. However, it cannot characterize the similarity between $\mathbf{W_B} \in \mathbb{R}^{d\times r}$ and $\mathbf{W} \in \mathbb{R}^{d\times r}$, because Eq. (3) of Theorem 3 may also indicate that $\mathbf{W_B}$ approximates $\mathbf{W}\boldsymbol{\Upsilon}$, where $\boldsymbol{\Upsilon} \in \mathbb{R}^{r\times r}$ is an arbitrary unitary matrix with $\boldsymbol{\Upsilon}\boldsymbol{\Upsilon}^T = \mathbf{I}_r$ and $\mathbf{I}_r$ being an identity matrix so that $\mathbf{W}\boldsymbol{\Upsilon}\boldsymbol{\Upsilon}^T\mathbf{W}^T = \mathbf{W}\mathbf{W}^T$. Fortunately, due to that $\boldsymbol{\Upsilon}\boldsymbol{\Upsilon}^T = \mathbf{I}_r$ (i.e., $\boldsymbol{\Upsilon} \in \mathbb{R}^{r\times r}$ is an orthogonal rotation), $\mathbf{W}\boldsymbol{\Upsilon}$ will still retain all information of $\mathbf{W}$ and even will empirically get better hashing accuracy, which has been mentioned in Remark 1 of our submitted paper. Therefore, Theorem 3 shows how $\mathbf{W_B}$ approximates $\mathbf{W}$ or $\mathbf{W}\boldsymbol{\Upsilon}$, which can be used to show the effectiveness of the related hashing algorithm.

We restate Remark 1 in our submitted paper: To address the problem that most of the information can be contained by only a small number of significant singular vectors in $\mathbf{W} \in \mathbb{R}^{d\times r}$, OSH [3] also empirically applies a random rotation $\boldsymbol{\Upsilon} \in \mathbb{R}^{r\times r}$ (the orthonormal bases of an $r \times r$ random Gaussian matrix) to all singular vectors $\mathbf{W} \in \mathbb{R}^{d\times r}$ returned by Algorithm 1 via $\mathbf{W}\boldsymbol{\Upsilon}$. This step resembles Iterative Quantization [2] but runs much more efficiently with streaming settings maintained and negligible computational cost incurred. Thus, following OSH, our method FROSH also applies $\mathbf{\Upsilon} \in \mathbb{R}^{r\times r}$ to the obtained top r right singular vectors of $\mathbf{B}^{\ell\times d}$.

Proof of Theorem 3. Due to that $\mathbf{W}_\mathbf{B}^T \in \mathbb{R}^{r\times d}$ contains the top r right singular vectors of $\mathbf{B}^{\ell\times d}$, we have $\mathbf{W}_\mathbf{B}\mathbf{W}_\mathbf{B}^T \in \mathbb{R}^{d\times d}$ as the projection matrix of $\mathbf{B}^{\ell\times d}$. With Lemma 4 in [1], we have

$$\begin{aligned}&\|(\mathbf{A}-\boldsymbol{\mu}) - (\mathbf{A}-\boldsymbol{\mu})\mathbf{W}_\mathbf{B}\mathbf{W}_\mathbf{B}^T\|_2^2 \\ &\le \sigma_{r+1}^2 + 2\|(\mathbf{A}-\boldsymbol{\mu})^T(\mathbf{A}-\boldsymbol{\mu}) - \mathbf{B}^T\mathbf{B}\|_2,\end{aligned} \tag{4}$$

where σ_i is the i-th largest singular value of $(\mathbf{A}-\boldsymbol{\mu})$.

For simplicity, when $m = \Theta(d)$ and $n = \Omega(\ell^{3/2}d^{3/2})$, the error bound of Eq. (2) in Theorem 2 will become $\widetilde{O}(\frac{1}{\ell}\|(\mathbf{A}-\boldsymbol{\mu})\|_F^2)$, which is then incorporated into Eq. (4) to get that

$$\begin{aligned}&\|(\mathbf{A}-\boldsymbol{\mu}) - (\mathbf{A}-\boldsymbol{\mu})\mathbf{W}_\mathbf{B}\mathbf{W}_\mathbf{B}^T\|_2^2 \\ &\le \sigma_{r+1}^2 + \widetilde{O}(\frac{1}{\ell}\|(\mathbf{A}-\boldsymbol{\mu})\|_F^2).\end{aligned} \tag{5}$$

Let $h = \|(\mathbf{A}-\boldsymbol{\mu})\|_F^2/\|(\mathbf{A}-\boldsymbol{\mu})\|_2^2$ be the numeric rank of $(\mathbf{A}-\boldsymbol{\mu}) \in \mathbb{R}^{n\times d}$, which could be much smaller than d for a low-rank matrix $(\mathbf{A}-\boldsymbol{\mu}) \in \mathbb{R}^{n\times d}$ with $d < n$. If $\ell = \Omega(\frac{h\sigma_1^2}{\epsilon\sigma_{r+1}^2})$, then from Eq. (5) we have

$$\begin{aligned}&\|(\mathbf{A}-\boldsymbol{\mu}) - (\mathbf{A}-\boldsymbol{\mu})\mathbf{W}_\mathbf{B}\mathbf{W}_\mathbf{B}^T\|_2^2 \le (1+\epsilon)\sigma_{r+1}^2 \\ &= (1+\epsilon)\|(\mathbf{A}-\boldsymbol{\mu}) - (\mathbf{A}-\boldsymbol{\mu})\mathbf{W}\mathbf{W}^T\|_2^2,\end{aligned} \tag{6}$$

where $\sigma_1^2 = \|(\mathbf{A}-\boldsymbol{\mu})\|_2^2$ and $\sigma_{r+1}^2 = \|(\mathbf{A}-\boldsymbol{\mu}) - (\mathbf{A}-\boldsymbol{\mu})\mathbf{W}\mathbf{W}^T\|_2^2$ according to the definition. $\square$

References

[1] P. Drineas and R. Kannan. Pass efficient algorithms for approximating large matrices. In *SODA*, volume 3, pages 223–232, 2003.

[2] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2916–2929, 2013.

[3] C. Leng, J. Wu, J. Cheng, X. Bai, and H. Lu. Online sketching hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2503–2511, 2015.