Improving Sign-Random-Projection via Count Sketch (Supplementary Material)

Punit Pankaj Dubey[1] **Bhisham Dev Verma**[1] **Rameshwar Pratap**[1] **Keegan Kang**[2]

[1]Indian Institute of Technology Mandi, H.P., India
[2]Bucknell University, Lewisburg, Pennsylvania, USA

1 EXTENDED EXPERIMENTAL RESULTS:

This section presents the extended experimental results for comparison among baseline and proposed methods (CSSRP and $\mathrm{CSSRP-L}$) using metrics: similarity search (defined in section 5.3 in the paper) and variance analysis via box plot (defined in section 5.4 in the paper). We summarized our findings for similarity search using recall in Figure 1, and for variance analysis via box plot in Figure 2.

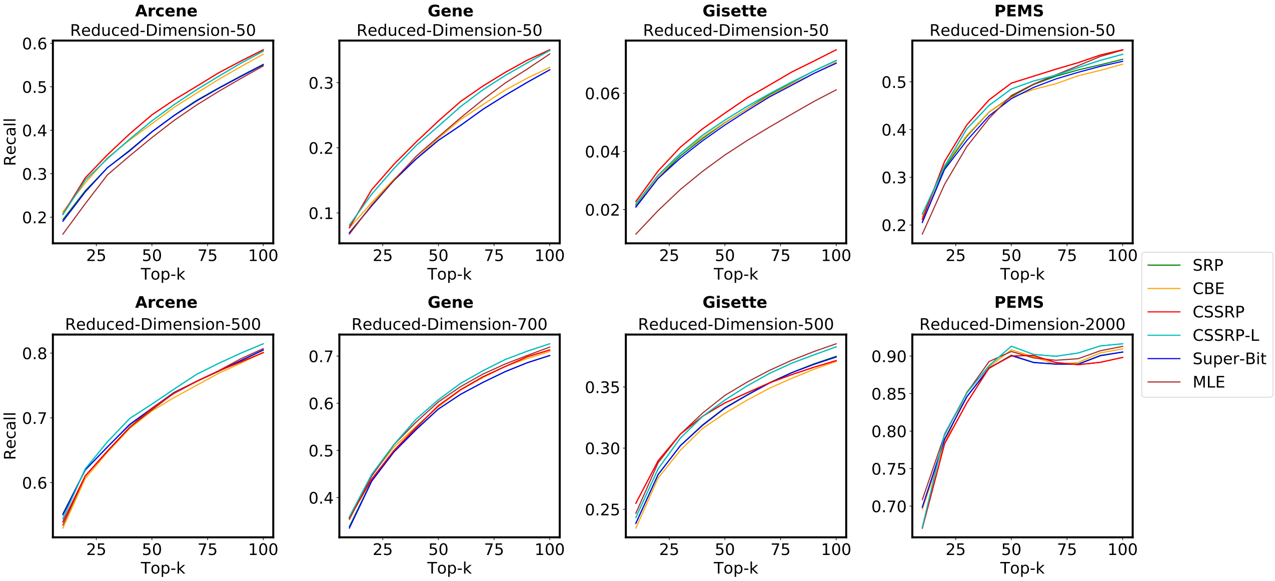

Figure 1: Comparison among the baselines on the task of top-k similarity search. A higher value of recall indicates a better performance.

2 MISSING PROOFS:

In this section, we present the missing proofs from the main paper. For convenience, we also restate them here.

Proof of Lemma 3:

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

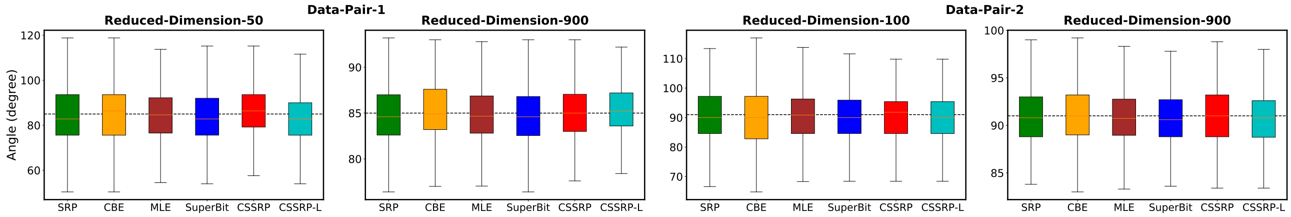

Figure 2: Comparison among baselines on the task of variance analysis via box plot. The sampled pairs are at angles 85° and 90°, respectively. The smaller interquartile range is an indicator of lower variance. The dotted line represents the actual angle in degree.

Lemma 3 (Adapted from Lemma 4 of **?**). *Let $\vec{r} = (r_1, \dots, r_j, \dots, r_D) \in \mathbb{R}^D$ s.t.*

$$r_j \sim \begin{cases} 1 & \textit{with probability } \frac{1}{2K} \\ 0 & \textit{with probability } \frac{K-1}{K} \\ -1 & \textit{with probability } \frac{1}{2K} \end{cases}$$

and $\vec{a} \in \mathbb{R}^D$. Denote $\alpha = \sum_{j=1}^{D} r_j a_j = \langle \vec{r}, \vec{a} \rangle$. Then if $D \to \infty$ and $K = o(D)$, we have $\alpha \overset{\mathcal{L}}{\Longrightarrow} \mathcal{N}\left(0, \frac{||\vec{a}||^2}{K}\right)$ with the rate of convergence

$$|F_\alpha(y) - \Phi(y)| \leq 0.8\sqrt{K}\frac{\sum_{i=1}^{D} |a_i|^3}{(\sum_{i=1}^{D} a_i^2)^{3/2}} = 0.8\sqrt{\frac{K}{D}}\frac{\mathbb{E}[|a_i|^3]}{(\mathbb{E}[a_i^2])^{3/2}} \to 0,$$

where $\overset{\mathcal{L}}{\Longrightarrow}$ denotes "convergence in distribution", $F_\alpha(y)$ is the empirical cumulative density function of α, and $\Phi(y)$ is the CDF of $\mathcal{N}\left(0, \frac{||\vec{a}||^2}{K}\right)$.

Proof. We know that

$$\alpha = \sum_{j=1}^{D} r_j a_j,$$

where

$$r_j \sim \begin{cases} 1 & \text{with probability } \frac{1}{2K}, \\ 0 & \text{with probability } \frac{K-1}{K}, \\ -1 & \text{with probability } \frac{1}{2K}. \end{cases}$$

Let

$$z_j = r_j a_j.$$

Then

$$\mathbb{E}[z_j] = \mathbb{E}[r_j a_j] = a_j \mathbb{E}[r_j] = 0.$$

$$\mathrm{Var}[z_j] = \mathbb{E}[(z_j - \mathbb{E}[z_j])^2] = \mathbb{E}[z_j^2] = \mathbb{E}[r_j^2 a_j^2] = \frac{a_j^2}{K}.$$

$$\mathbb{E}[|z_j|^{2+\delta}] = |a_j|^{2+\delta}\mathbb{E}[|r_j|^{2+\delta}] = |a_j|^{2+\delta}\left(1 \times \frac{1}{K} + 0 \times \frac{K-1}{K}\right) = \frac{|a_j|^{2+\delta}}{K}.$$

Let

$$S_D^2 = \sum_{j=1}^{D} \mathrm{Var}[z_j] = \frac{\sum_{j=1}^{D} a_j^2}{K}.$$

To prove that $\frac{\sum_{j=1}^{D} z_j}{S_D} \overset{\mathcal{L}}{\Longrightarrow} \mathcal{N}(0,1)$, we need to show that following Lindeberg condition is satisfied.

$$\frac{1}{S_D^2}\sum_{j=1}^{D}\mathbb{E}\left[z_j^2; |z_j| > \epsilon S_D\right] \to 0 \quad \text{for any } \epsilon > 0. \tag{1}$$

Now, we compute the LHS of the Equation (1):

$$\begin{aligned}
\frac{1}{S_D^2}\sum_{j=1}^{D}\mathbb{E}\left[z_j^2; |z_j| > \epsilon S_D\right] &\leq \frac{1}{S_D^2}\sum_{j=1}^{D}\mathbb{E}\left[\frac{|z_j|^{2+\delta}}{(\epsilon S_D)^\delta}\right]. \\
&= \frac{1}{\frac{\sum_{j=1}^{D} a_j^2}{K}} \cdot \frac{1}{\epsilon^\delta} \cdot \frac{\sum_{j=1}^{D} \frac{|a_j|^{2+\delta}}{K}}{\left(\frac{\sum_{j=1}^{D} a_j^2}{K}\right)^{\frac{\delta}{2}}}. \\
&= \frac{1}{\epsilon^\delta} \cdot \frac{\sum_{j=1}^{D} \frac{|a_j|^{2+\delta}}{K}}{\left(\frac{\sum_{j=1}^{D} a_j^2}{K}\right)^{\frac{2+\delta}{2}}}. \\
&= K^{\frac{\delta}{2}} \cdot \frac{1}{\epsilon^\delta} \cdot \frac{\sum_{j=1}^{D} \frac{|a_j|^{2+\delta}}{D}}{\left(\frac{\sum_{j=1}^{D} a_j^2}{D}\right)^{\frac{2+\delta}{2}}} \cdot \frac{1}{D^{\frac{\delta}{2}}}. \\
&= \left(\frac{K}{D}\right)^{\frac{\delta}{2}} \cdot \frac{1}{\epsilon^\delta} \cdot \frac{\mathbb{E}\left[|a_j|^{2+\delta}\right]}{\left(\mathbb{E}[a_j^2]\right)^{\frac{2+\delta}{2}}}. \\
&\to 0. \qquad (2)
\end{aligned}$$

Equation (2) holds as $K = o(D)$. Therefore, for $K = o(D)$, due to Lindeberg Central Limit theorem [**??**], we have

$$\begin{aligned}
&\frac{\sum_{j=1}^{D} z_j}{S_D} = \frac{\sum_{j=1}^{D} r_j a_j}{\sqrt{\frac{\sum_{j=1}^{D} a_j^2}{K}}} = \frac{\alpha}{\sqrt{||\vec{a}||^2/K}} \overset{\mathcal{L}}{\Longrightarrow} \mathcal{N}(0,1). \\
&\alpha \overset{\mathcal{L}}{\Longrightarrow} \mathcal{N}\left(0, \frac{||\vec{a}||^2}{K}\right). \qquad (3)
\end{aligned}$$

We remain to find the rate of convergence. For this we use Berry Esseen theorem [**??**]. Let us denote

$$\begin{aligned}
\rho_D &= \sum_{j=1}^{D}\mathbb{E}\left[|z_j|^3\right]. \\
&= \sum_{j=1}^{D}|a_j|^3\mathbb{E}\left[|r_j|^3\right]. \\
&= \frac{\sum_{j=1}^{D}|a_j|^3}{K}
\end{aligned}$$

Then, due to Berry Esseen theorem, we have

$$\begin{aligned}|F_\alpha(y)-\Phi(y)| &\leq 0.8\frac{\rho_D}{S_D^3}.\\ &= 0.8\frac{\frac{\sum_{j=1}^D |a_j|^3}{K}}{\left(\frac{\sum_{j=1}^D a_j^2}{K}\right)^{\frac{3}{2}}}.\\ &= 0.8\times\sqrt{\frac{K}{D}}\frac{\sum_{j=1}^D |a_j|^3/D}{\left(\sum_{j=1}^D a_j^2/D\right)^{\frac{3}{2}}}.\\ &= 0.8\times\sqrt{\frac{K}{D}}\frac{\mathbb{E}[|a_j|^3]}{\left(\mathbb{E}[a_j^2]\right)^{\frac{3}{2}}}.\\ &\to 0 \qquad \text{as} \qquad D\to\infty. \end{aligned} \tag{4}$$

Equation (4) holds for $K=o(D)$. Equation (3) and (4) completes a proof of the Lemma 3. □

Proof of Theorem 6:

Theorem 6. *Let $\vec{a},\vec{b}\in\mathbb{R}^D$, and $h(\vec{a})$, $h(\vec{b})$ be their K-dimensional binary vector obtained via our proposal (Definition 2 define in the paper). If $K=o(D)$, then as $D\to\infty$ we have the following*

$$\begin{aligned}&\mathrm{Var}\left[\frac{\pi}{K}||h(\vec{a})-h(\vec{b})||_1\right]\\ &=\frac{\pi^2}{K^2}\left(\frac{K\theta_{(\vec{a},\vec{b})}}{\pi}+K(K-1)\frac{\theta_{(\vec{a},\vec{b})}}{\pi}\times\eta\right)-\theta^2_{(\vec{a},\vec{b})}.\end{aligned}$$

where, $k_1\neq k_2$, $k_1,k_2\in[K]$, and
$\eta=\Pr\left[\left(h^{(k_2)}(\vec{a})\neq h^{(k_2)}(\vec{b})\right)\mid\left(h^{(k_1)}(\vec{a})\neq h^{(k_1)}(\vec{b})\right)\right]$.

Proof. We know that

$$\begin{aligned}\mathrm{Var}\left[\frac{\pi}{K}||h(\vec{a})-h(\vec{b})||_1\right] &= \frac{\pi^2}{K^2}\mathrm{Var}\left[||h(\vec{a})-h(\vec{b})||_1\right].\\ &=\frac{\pi^2}{K^2}\mathrm{Var}\left[\sum_{k=1}^K Y_k\right].\\ &\text{where } Y_k := \mathbb{1}_{\{h^{(k)}(\vec{a})\neq h^{(k)}(\vec{b})\}}.\end{aligned} \tag{5}$$

We focus on the term

$$\begin{aligned}\mathrm{Var}\left[\sum_{k=1}^K Y_k\right] &= \mathbb{E}\left[\left(\sum_{k=1}^K Y_k\right)^2\right]-\left(\mathbb{E}\left[\sum_{k=1}^K Y_k\right]\right)^2.\\ &=\mathbb{E}\left[\sum_{k=1}^K Y_k^2+\sum_{k_1\neq k_2,k_1,k_2\in[K]} Y_{k_1}Y_{k_2}\right]-\left(\mathbb{E}\left[\sum_{k=1}^K Y_k\right]\right)^2.\\ &=\sum_{k=1}^K\mathbb{E}\left[Y_k\right]+\sum_{k_1\neq k_2,k_1,k_2\in[K]}\mathbb{E}\left[Y_{k_1}Y_{k_2}\right]-\left(\mathbb{E}\left[\sum_{k=1}^K Y_k\right]\right)^2.\end{aligned} \tag{6}$$

We compute the value of each term one-by-one as follows:

$$\begin{aligned}\sum_{k=1}^{K}\mathbb{E}\left[Y_k\right] &= \sum_{k=1}^{K}\Pr\left[h^{(k)}(\vec{a})\neq h^{(k)}(\vec{b})\right].\\ &= \sum_{k=1}^{K}\frac{\theta_{(\vec{a},\vec{b})}}{\pi}\\ &= \frac{K\theta_{(\vec{a},\vec{b})}}{\pi}. \qquad (7)\end{aligned}$$

Now, we compute the following

$$\begin{aligned}&\sum_{k_1\neq k_2, k_1,k_2\in[K]}\mathbb{E}\left[Y_{k_1}Y_{k_2}\right] = \sum_{k_1\neq k_2}\Pr\left[\left(h^{(k_2)}(\vec{a})\neq h^{(k_2)}(\vec{b})\right)\cap\left(h^{(k_1)}(\vec{a})\neq h^{(k_1)}(\vec{b})\right)\right].\\ &= \sum_{k_1\neq k_2, k_1,k_2\in[K]}\Pr\left[\left(h^{(k_2)}(\vec{a})\neq h^{(k_2)}(\vec{b})\right)\mid\left(h^{(k_1)}(\vec{a})\neq h^{(k_1)}(\vec{b})\right)\right]\times\Pr\left[\left(h^{(k_1)}(\vec{a})\neq h^{(k_1)}(\vec{b})\right)\right].\\ &= \sum_{k_1\neq k_2, k_1,k_2\in[K]}\frac{\theta_{(\vec{a},\vec{b})}}{\pi}\cdot\Pr\left[\left(h^{(k_2)}(\vec{a})\neq h^{(k_2)}(\vec{b})\right)\mid\left(h^{(k_1)}(\vec{a})\neq h^{(k_1)}(\vec{b})\right)\right]. \qquad (8)\end{aligned}$$

From linearity of expectation and Equation (7), we have

$$\left(\mathbb{E}\left[\sum_{k=1}^{K}Y_k\right]\right)^2 = \left(\frac{K\theta_{(\vec{a},\vec{b})}}{\pi}\right)^2.$$

We denote $\Pr\left[\left(h^{(k_2)}(\vec{a})\neq h^{(k_2)}(\vec{b})\right)\mid\left(h^{(k_1)}(\vec{a})\neq h^{(k_1)}(\vec{b})\right)\right] := \eta$ in Equation (8), and $\eta\in\left[0,\frac{\theta}{\pi}\right]$. Therefore

$$\mathrm{Var}\left[\sum_{k=1}^{K}Y_k\right] = \frac{K\theta_{(\vec{a},\vec{b})}}{\pi} + K(K-1)\frac{\theta_{(\vec{a},\vec{b})}}{\pi}\times\eta - \left(\frac{K\theta_{(\vec{a},\vec{b})}}{\pi}\right)^2$$

Hence, the variance of our estimate is

$$\begin{aligned}\mathrm{Var}\left[\frac{\pi}{K}||h(\vec{a})-h(\vec{b})||_1\right] &= \frac{\pi^2}{K^2}\left(\frac{K\theta_{(\vec{a},\vec{b})}}{\pi} + K(K-1)\frac{\theta_{(\vec{a},\vec{b})}}{\pi}\times\eta - \left(\frac{K\theta_{(\vec{a},\vec{b})}}{\pi}\right)^2\right).\\ &= \frac{\pi^2}{K^2}\left(\frac{K\theta_{(\vec{a},\vec{b})}}{\pi} + K(K-1)\frac{\theta_{(\vec{a},\vec{b})}}{\pi}\times\eta\right) - \theta^2_{(\vec{a},\vec{b})}. \qquad (9)\end{aligned}$$

Equation (9) completes a proof of the theorem. □

Proof of Theorem 9:

In order to prove Theorem 9, we require the following lemma which is similar to Lemma 3. We first complete its proof and then conclude with the proof of Theorem 9.

Lemma 12. *[Adapted from Lemma 4 of* **?***] Let $\vec{r}' = (r'_1,\ldots,r'_j,\ldots,r'_D)\in\mathbb{R}^D$ s.t.*

$$r'_j\sim\begin{cases}1 & \textit{with probability } \frac{l}{2K}\\ 0 & \textit{with probability } \frac{K-l}{K}\\ -1 & \textit{with probability } \frac{l}{2K}\end{cases}$$

and $\vec{a}\in\mathbb{R}^D$. Denote $\alpha' = \sum_{j=1}^{D}r'_j a_j = \langle\vec{r}',\vec{a}\rangle$. Then if $D\to\infty$ and $K = o(lD)$, we have $\alpha'\overset{\mathcal{L}}{\Longrightarrow}\mathcal{N}\left(0,\frac{||\vec{a}||^2}{K}\right)$ with the rate of convergence

$$|F_{\alpha'}(y)-\Phi(y)|\leq 0.8\sqrt{\frac{K}{l}}\frac{\sum_{i=1}^{D}|a_i|^3}{(\sum_{i=1}^{D}a_i^2)^{3/2}} = 0.8\sqrt{\frac{K}{lD}}\frac{\mathbb{E}[|a_i|]^3}{(\mathbb{E}[a_i^2])^{3/2}}\to 0,$$

where $\overset{\mathcal{L}}{\Longrightarrow}$ denotes "convergence in distribution", $F_{\alpha'}(y)$ is the empirical cumulative density function of α', and $\Phi(y)$ is the CDF of $\mathcal{N}\left(0, \frac{||\vec{a}||^2}{K}\right)$.

Proof. We know that

$$\alpha' = \sum_{j=1}^{D} r'_j a_j,$$

where

$$r'_j \sim \begin{cases} \;\;1 & \text{with probability } \frac{l}{2K}, \\ \;\;0 & \text{with probability } \frac{K-l}{K}, \\ -1 & \text{with probability } \frac{l}{2K}. \end{cases}$$

Let

$$z'_j = r'_j a_j.$$

Then

$$\begin{aligned} \mathbb{E}[z'_j] &= \mathbb{E}[r'_j a_j] = a_j \mathbb{E}[r'_j] = 0. \\ \mathrm{Var}[z'_j] &= \mathbb{E}[(z'_j - \mathbb{E}[z'_j])^2] = \mathbb{E}[z'^2_j] = \mathbb{E}[r'^2_j a^2_j] = \frac{l}{K} a^2_j. \\ \mathbb{E}[|z'_j|^{2+\delta}] &= |a_j|^{2+\delta} \mathbb{E}[|r'_j|^{2+\delta}] = |a_j|^{2+\delta} \left(1 \times \frac{l}{K} + 0 \times \frac{K-l}{K}\right) = \frac{l}{K} |a_j|^{2+\delta}. \end{aligned}$$

Let

$$S'^2_D = \sum_{j=1}^{D} \mathrm{Var}[z_j] = \frac{l}{K} \sum_{j=1}^{D} a^2_j.$$

To prove that $\frac{\sum_{j=1}^{D} z'_j}{S'_D} \overset{\mathcal{L}}{\Longrightarrow} \mathcal{N}(0,1)$, we need to show that following Lindeberg condition [**??**] is satisfied.

$$\frac{1}{S'^2_D} \sum_{j=1}^{D} \mathbb{E}\left[z'^2_j; |z'_j| > \epsilon S'_D\right] \to 0 \quad \text{for any } \epsilon > 0. \tag{10}$$

Now, we compute the LHS of the Equation (10):

$$\begin{aligned} \frac{1}{S'^2_D} \sum_{j=1}^{D} \mathbb{E}\left[z'^2_j; |z'_j| > \epsilon S'_D\right] &\le \frac{1}{S'^2_D} \sum_{j=1}^{D} \mathbb{E}\left[\frac{|z'_j|^{2+\delta}}{(\epsilon S'_D)^{\delta}}\right] \\ &= \frac{1}{\frac{l}{K}\sum_{j=1}^{D} a^2_j} \cdot \frac{1}{\epsilon^{\delta}} \cdot \frac{\frac{l}{K}\sum_{j=1}^{D} |a_j|^{2+\delta}}{\left(\frac{l}{K}\sum_{j=1}^{D} a^2_j\right)^{\frac{\delta}{2}}}. \\ &= \frac{1}{\epsilon^{\delta}} \cdot \frac{\frac{l}{K}\sum_{j=1}^{D} |a_j|^{2+\delta}}{\left(\frac{l}{K}\sum_{j=1}^{D} a^2_j\right)^{\frac{2+\delta}{2}}}. \\ &= \left(\frac{K}{l}\right)^{\frac{\delta}{2}} \cdot \frac{1}{\epsilon^{\delta}} \cdot \frac{\sum_{j=1}^{D} \frac{|a_j|^{2+\delta}}{D}}{\left(\frac{\sum_{j=1}^{D} a^2_j}{D}\right)^{\frac{2+\delta}{2}}} \cdot \frac{1}{D^{\frac{\delta}{2}}}. \\ &= \left(\frac{K}{lD}\right)^{\frac{\delta}{2}} \cdot \frac{1}{\epsilon^{\delta}} \cdot \frac{\mathbb{E}\left[|a_j|^{2+\delta}\right]}{\left(\mathbb{E}[a^2_j]\right)^{\frac{2+\delta}{2}}}. \\ &\to 0. \end{aligned} \tag{11}$$

Equation (11) holds when $K = o(lD)$. Therefore, for $K = o(lD)$, due to Lindeberg Central Limit theorem [??], we have

$$\frac{\sum_{j=1}^{D} z'_j}{S'_D} = \frac{\sum_{j=1}^{D} r'_j a_j}{\sqrt{\frac{\sum_{j=1}^{D} a_j^2}{K}}} = \frac{\alpha'}{\sqrt{||\vec{a}||^2/K}} \overset{\mathcal{L}}{\Rightarrow} \mathcal{N}(0,1).$$

$$\alpha' \overset{\mathcal{L}}{\Rightarrow} \mathcal{N}\left(0, \frac{||\vec{a}||^2}{K}\right). \tag{12}$$

We remain to find the rate of convergence. For this we use Berry Esseen theorem [??]. Let us denote

$$\begin{aligned}\rho'_D &= \sum_{j=1}^{D} \mathbb{E}\left[|z'_j|^3\right].\\ &= \sum_{j=1}^{D} |a_j|^3 \mathbb{E}\left[|r'_j|^3\right].\\ &= \frac{l}{K}\sum_{j=1}^{D} |a_j|^3.\end{aligned}$$

From Berry Esseen theorem [??], we have

$$\begin{aligned}&|F_{\alpha'}(y) - \Phi(y)| \leq 0.8\frac{\rho'_D}{S'^3_D}.\\ &= 0.8\frac{\frac{l}{K}\sum_{j=1}^{D}|a_j|^3}{\left(\frac{l}{K}\sum_{j=1}^{D}a_j^2\right)^{\frac{3}{2}}}.\\ &= 0.8 \times \sqrt{\frac{K}{lD}}\frac{\sum_{j=1}^{D}|a_j|^3/D}{\left(\sum_{j=1}^{D}a_j^2/D\right)^{\frac{3}{2}}}.\\ &= 0.8 \times \sqrt{\frac{K}{lD}}\frac{\mathbb{E}[|a_j|^3]}{\left(\mathbb{E}[a_j^2]\right)^{\frac{3}{2}}}.\\ &\to 0 \qquad \text{as} \quad D \to \infty.\end{aligned} \tag{13}$$

Equation (13) holds for $K = o(lD)$. Equation (12) and (13) completes a proof of the Lemma 12. □

We now complete a proof of Theorem 9.

Theorem 9. *Let $\vec{a}, \vec{b} \in \mathbb{R}^D$, and $h'(\vec{a})$, $h'(\vec{b})$ be their K-dimensional binary vector obtained via our improved estimator proposal (stated in Definition 8 in the paper). If $K = o(lD)$, then as $D \to \infty$ we have the following*

$$\mathbb{E}\left[\frac{\pi}{K}||h'(\vec{a}) - h'(\vec{b})||_1\right] = \theta_{(\vec{a},\vec{b})}.$$

Proof. Let R' be a $K \times D$ projection matrix (defined in Definition 8 in the paper) such that each column of R' has exactly l non-zero entries. These l positions are sampled uniformly at random and each of them takes value $\{\pm 1\}$ with probability $1/2$

$$R' = \begin{bmatrix} \vec{r'}_1 \\ \vdots \\ \vec{r'}_k \\ \vdots \\ \vec{r'}_K \end{bmatrix}_{K \times D}. \tag{14}$$

We first consider each row $\vec{r}'_k, 1 \leq k \leq K$ of the random matrix in Equation (14). The goal is to find the distribution of each $\vec{r}'_k$, and hence compute

$$\mathbb{E}\left[\sum_{k=1}^{K} |h'^{(k)}(\vec{a}) - h'^{(k)}(\vec{b})|\right] = \sum_{k=1}^{K} \mathbb{E}\left[|h'^{(k)}(\vec{a}) - h'^{(k)}(\vec{b})|\right].$$

Suppose we denote $Z'_k := |h'^{(k)}(\vec{a}) - h'^{(k)}(\vec{b})|$. While each Z'_k are not independent due to our construction of R', let us briefly consider how each $\vec{r}'_k$ is distributed.

When $k = 1$, we have that each entry in $\vec{r}'_1$ comes from a sparse Bernoulli distribution with

$$r'_{1j} \sim \begin{cases} 1 & \text{with probability } \frac{l}{2K} \\ 0 & \text{with probability } \frac{K-l}{K} \\ -1 & \text{with probability } \frac{l}{2K}. \end{cases} \tag{15}$$

where $\mathbb{E}[r'_{1j}] = 0$, with $\text{Var}[r'_{1j}] = \frac{l}{K}$. Here, we note that each entry in $\vec{r}'_1$ is i.i.d.

We can also compute the moment generating function of each r'_{1j} and get

$$\mathbb{E}\left[e^{sr'_{1j}}\right] = \frac{K-l}{K} + \frac{l \cdot (\exp\{s\} + \exp\{-s\})}{2K}. \tag{16}$$

Now let us consider the case $k = 2$, and compute the moment generating function for each r'_{2j}. By using the Law of Total Expectation, we have that

$$\mathbb{E}\left[e^{sr'_{2j}}\right] = \mathbb{E}\left[e^{sr'_{2j}} \mid r'_{1j} = 0\right] \mathbb{P}\left[r'_{1j} = 0\right] + \mathbb{E}\left[e^{sr'_{2j}} \mid r'_{1j} = 1\right] \mathbb{P}\left[r'_{1j} = 1\right] + \mathbb{E}\left[e^{sr'_{2j}} \mid r'_{1j} = -1\right] \mathbb{P}\left[r'_{1j} = -1\right]. \tag{17}$$

$$= \left(\frac{l(\exp\{s\} + \exp\{-s\})}{2(K-1)} + \frac{K-l-1}{K-1}\right)\frac{K-l}{K} + \left(\frac{(l-1)(\exp\{s\} + \exp\{-s\})}{2(K-1)} + \frac{K-l}{K-1}\right)\frac{l}{2K}$$
$$+ \left(\frac{(l-1)(\exp\{s\} + \exp\{-s\})}{2(K-1)} + \frac{K-l}{K-1}\right)\frac{l}{2K}. \tag{18}$$

$$= l(K-1) \times \frac{(\exp\{s\} + \exp\{-s\})}{2K(K-1)} + \frac{(K-l)(K-1)}{(K-1)K}. \tag{19}$$

$$= \frac{l \cdot (\exp\{s\} + \exp\{-s\})}{2K} + \frac{(K-l)}{K}. \tag{20}$$

which is the same moment generating function as the Sparse Bernoulli distribution mentioned in Equation (15).

Now consider $\vec{r}_k, 2 < k \leq K - l$, and suppose we have seen l' non-zero entries so far. Denote $\lambda = \exp\{s\} + \exp\{-s\}$.

For ease of notation, if we have seen l' non-zero entries so far, then we have $\max(l - l', 0)$ non-zero entries to choose, out of the remaining $K - k + 1$ terms, and the probability of drawing a non-zero for our k^{th} entry has to be given by $\frac{1}{2}p^{(l-l')}_{(K-k+1)} = \frac{\max(l-l',0)}{(K-k+1)}$.

Then we can write

$$\mathbb{E}\left[e^{sr'_{kj}}\right] = \sum_{i=0}^{l'} \left(\mathbb{E}\left[e^{sr'_{kj}} \mid i \text{ non-zeroes for } r'_{k'j}, k' < k\right] \times \mathbb{P}\left[i \text{ non-zeroes for } r'_{k'j}, k' < k\right]\right).$$
$$= \sum_{i=0}^{l'} \left[\left(\frac{1}{2}\lambda p^{(l-i)}_{(K-k+1)} + \left(1 - p^{(l-i)}_{(K-k+1)}\right)\right) \times \mathbb{P}\left[i \text{ non-zeroes for } r'_{k'j}, k' < k\right]\right].$$

Consider each term in the above summation. For the right term, we have

$$\begin{aligned}
&\sum_{i=1}^{l'}(1 - p_{(K-k+1)}^{(l-i)})\mathbb{P}\left[i \text{ non-zeroes for } r'_{k'j}, k' < k\right] \\
&= \sum_{i=1}^{l'} p_{(K-k+1)}^{(K-l-(k-1-i))}\mathbb{P}\left[k-1-i \text{ zeroes for } r'_{k'j}, k' < k\right] \\
&= p_K^{K-l} \\
&= \frac{K-l}{K}.
\end{aligned} \tag{21}$$

The left term is more straightforward, and we have

$$\frac{1}{2}\sum_{i=0}^{l'} p_{(K-k+1)}^{(l-i)}\mathbb{P}\left[i \text{ non-zeroes for } r_{k'j}, k' < k\right] = \frac{1}{2}\frac{l}{K}\lambda.$$

Adding both terms together, we get the MGF of the Sparse Bernoulli distribution mentioned in Equation (15). Using Lemma 12, we can show that $\alpha'_k = \langle \vec{r'}_k, \vec{a} \rangle$ and $\beta'_k = \langle \vec{r'}_k, \vec{b} \rangle$ converge in distribution to $\mathcal{N}\left(0, \frac{||\vec{a}||^2}{K}\right)$ as D grows large. This fact and Lemma 4 (defined in paper), concludes a proof of the theorem. □

3 OTHER DETAILS:

GUARANTEE OF COUNT-SKETCH ALGORITHM:

The following theorem states that for a pair of real-valued vectors their sketches obtained via COUNT-SKETCH closely approximate the original pairwise inner product.

Theorem 13 (??). *Given vectors $\vec{a} = (a_1, \ldots a_D)$, $\vec{b} = (b_1, \ldots b_D)$ get compressed into vectors $\vec{\alpha} = (\alpha_1, ..\alpha_k, ..\alpha_K)$ and $\vec{\beta} = (\beta_1, ..\beta_k, ..\beta_K)$, respectively, using the* COUNT-SKETCH *algorithm, where $k \in [K]$. Then, we have the following*

$$\mathbb{E}[\vec{\alpha}] = \mathbb{E}[\vec{\beta}] = \vec{0}. \tag{22}$$

$$\mathbb{E}[\langle \vec{\alpha}, \vec{\beta} \rangle] = \langle \vec{a}, \vec{b} \rangle. \tag{23}$$

$$\mathrm{Var}[\langle \vec{\alpha}, \vec{\beta} \rangle] = \frac{1}{K}\sum_{i \neq j, i,j=1}^{D}\left(a_i^2 b_j^2 + a_i b_i a_j b_j\right). \tag{24}$$

References

Moses Charikar, Kevin C. Chen, and Martin Farach-Colton. Finding frequent items in data streams. *Theor. Comput. Sci.*, 312 (1):3–15, 2004. doi: 10.1016/S0304-3975(03)00400-6. URL https://doi.org/10.1016/S0304-3975(03)00400-6.

William Feller. *An introduction to probability theory and its applications. Vol. II.* Second edition. John Wiley & Sons Inc., New York, 1971.

F. Gotze. On the Rate of Convergence in the Multivariate CLT. *The Annals of Probability*, 19(2):724 – 739, 1991. doi: 10.1214/aop/1176990448. URL https://doi.org/10.1214/aop/1176990448.

Ping Li, Trevor Hastie, and Kenneth Ward Church. Very sparse random projections. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006* **?**, pages 287–296. doi: 10.1145/1150402.1150436. URL https://doi.org/10.1145/1150402.1150436.

I. S. Shiganov. Refinement of the upper bound of the constant in the central limit theorem. *Journal of Soviet Mathematics*, 35:2545–2550, 1986.

Kilian Q. Weinberger, Anirban Dasgupta, John Langford, Alexander J. Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, pages 1113–1120, 2009. doi: 10.1145/1553374.1553516. URL `http://doi.acm.org/10.1145/1553374.1553516`.