8 Supplementary Materials

8.1 Proofs

8.1.1 Proof of Lemma 1

Proof. We use *law of iterated expectation* to prove the conclusion. We first compute

$$\begin{aligned} E[x_{jb}|t] &= \text{sgn}(w_t h_{jt}) P(j|t) \\ &= \text{sgn}(w_t h_{jt}) \frac{|h_{jt}|}{\sum_{j=1}^n |h_{jt}|} \\ &= \text{sgn}(w_t h_{jt}) \frac{|w_t h_{jt}|}{\sum_{j=1}^n |w_t h_{jt}|} \\ &= \frac{w_t h_{jt}}{\sum_{j=1}^n |w_t h_{jt}|}. \end{aligned}$$

This gives us:

$$\begin{aligned} E[x_{jb}] &= E[E[x_{jb}|t]] = E[\frac{w_t h_{jt}}{\sum_{j=1}^n |w_t h_{jt}|}] \\ &= \sum_{t=1}^d \frac{w_t h_{jt}}{\sum_{j=1}^n |w_t h_{jt}|} P(t|\mathbf{w}) \\ &= \sum_{t=1}^d \frac{w_t h_{jt}}{\sum_{j=1}^n |w_t h_{jt}|} \frac{\sum_{j=1}^n |w_t h_{jt}|}{\sum_{t=1}^d \sum_{j=1}^n |w_t h_{jt}|} \\ &= \frac{\sum_{t=1}^d w_t h_{jt}}{\sum_{t=1}^d \sum_{j=1}^n |w_t h_{jt}|} \\ &= \frac{\mathbf{w}^T \mathbf{h}_j}{\sum_{t=1}^d \sum_{j=1}^n |w_t h_{jt}|} = \frac{c_j}{S}. \end{aligned}$$

Therefore, $E[x_j] = E[\sum_{b=1}^B x_{jb}] = \frac{Bc_j}{S}$. □

8.1.2 Proof of Lemma 2

Proof. We still use *law of iterated expectation* to prove the conclusion and it is basically very similar to the proof of Lemma 1.

$$\begin{aligned} & E[(x_{jb} - x_{mb})^2|t] \\ = \; & [\text{sgn}(w_t h_{jt}) - 0]^2 P(j|t) + [0 - \text{sgn}(w_t h_{mt})]^2 P(m|t) \\ = \; & [\text{sgn}(w_t h_{jt})]^2 \frac{|h_{jt}|}{\sum_{j=1}^n |h_{jt}|} \\ & + [\text{sgn}(w_t h_{mt})]^2 \frac{|h_{mt}|}{\sum_{j=1}^n |h_{jt}|} \\ = \; & [\text{sgn}(w_t h_{jt})]^2 \frac{|w_t h_{jt}|}{\sum_{j=1}^n |w_t h_{jt}|} \\ & + [\text{sgn}(w_t h_{mt})]^2 \frac{|w_t h_{mt}|}{\sum_{j=1}^n |w_t h_{jt}|} \\ = \; & \frac{|w_t h_{jt}| + |w_t h_{mt}|}{\sum_{j=1}^n |w_t h_{jt}|}. \end{aligned}$$

This gives us:

$$\begin{aligned} E[(x_{jb} - x_{mb})^2] &= E\{E[(x_{jb} - x_{mb})^2|t]\} \\ &= \sum_{t=1}^d \frac{|w_t h_{jt}| + |w_t h_{mt}|}{\sum_{j=1}^n |w_t h_{jt}|} P(t|\mathbf{w}) \\ &= \frac{\sum_{t=1}^d (|w_t h_{jt}| + |w_t h_{mt}|)}{\sum_{t=1}^d \sum_{j=1}^n |w_t h_{jt}|} \\ &= \frac{a_j + a_m}{S}. \end{aligned}$$

Note that for each iteration b, $x_{jb} - x_{mb}$ is independently and identically distributed, so we have

$$\begin{aligned} \text{Var}(x_j - x_m) &= B\text{Var}(x_{jb} - x_{mb}) \\ &= B\left\{E[(x_{jb} - x_{mb})^2] - [E(x_{jb} - x_{mb})]^2\right\} \\ &= B\left[\frac{a_j + a_m}{S} - \frac{(c_j - c_m)^2}{S^2}\right]. \end{aligned}$$

□

8.1.3 Proof of Theorem 1

Proof. For some j, we know that x_{jb} and x_{mb} cannot be non-zero simultaneously. Therefore, all $(x_{jb} - x_{mb})$'s independently take values in $[-1, 1]$. We use Bennett's Inequality in (Bennett, 1962) to bound the probability of $P(x_j \geq x_m)$ for some $j \in \{c_j < c_m\}$.

$$\begin{aligned} P(x_j \geq x_m) &= P(x_j - x_m \geq 0) \\ &= P\left\{\sum_{b=1}^B (x_{jb} - x_{mb}) \geq 0\right\} \\ &= P(Y \geq y), \end{aligned} \tag{8}$$

where $Y_b = x_{jb} - x_{mb} - \frac{c_j - c_m}{S}$, $Y = \sum_{b=1}^B Y_b$ and $y = \frac{B(c_m - c_j)}{S}$. It is obvious that $Y_b \leq 1 - \frac{c_j - c_m}{S}$ almost surely. We denote $q = 1 - \frac{c_j - c_m}{S}$ and we compute a few quantities needed in Bennett's Inequality below:

$$\begin{aligned} & \Sigma_j^2 := BEY_b^2 = \text{Var}(x_j - x_m) = B\sigma_j^2, \\ & qy = B\frac{(S + c_m - c_j)(c_m - c_j)}{S^2} = BT_j, \\ & \frac{qy}{\Sigma_j^2} = \frac{T_j}{\sigma_j^2}, \\ & \frac{T_j}{\sigma_j^2} = \frac{(S + c_m - c_j)(c_m - c_j)}{S(a_j + a_m) - (c_m - c_j)^2}. \end{aligned} \tag{9}$$

From Bennett's Inequality, we can bound Equation (8):

$$\begin{aligned} & P(Y \geq y) \\ \leq \; & exp\left\{-\frac{\Sigma_j^2}{q^2}\left[(1 + \frac{qy}{\Sigma_j^2})\log(1 + \frac{qy}{\Sigma_j^2}) - \frac{qy}{\Sigma_j^2}\right]\right\} \\ = \; & exp\left\{-BM_j\sigma_j^2\left[(1 + \frac{T_j}{\sigma_j^2})\log(1 + \frac{T_j}{\sigma_j^2}) - \frac{T_j}{\sigma_j^2}\right]\right\} \end{aligned}$$

If we want $P(Y \geq y) \leq \frac{\delta}{n'}$, then it is equivalent to

$$BM_j\sigma_j^2\left[(1+\frac{T_j}{\sigma_j^2})\log(1+\frac{T_j}{\sigma_j^2})-\frac{T_j}{\sigma_j^2}\right] \geq \log\frac{n'}{\delta}.$$

Before proceeding to the result, we need to show that $(1+\frac{T_j}{\sigma_j^2})\log(1+\frac{T_j}{\sigma_j^2})-\frac{T_j}{\sigma_j^2} > 0$. Denote $v=\frac{T_j}{\sigma_j^2}$ and $g_1(v) = (1+v)\log(1+v)-v$. Then $g_1'(v) = 1+\log(1+v)-1 = \log(1+v) > 0$ when $v>0$. So $g_1(v) > g_1(0) = 0$ when $v>0$. It is not hard to see that $\frac{T_j}{\sigma_j^2} > 0$ from (9) for all $j \in \{c_j < c_m\}$. Therefore,

$$B \geq \frac{1}{M_j\sigma_j^2\left[(1+\frac{T_j}{\sigma_j^2})\log(1+\frac{T_j}{\sigma_j^2})-\frac{T_j}{\sigma_j^2}\right]}\log\frac{n'}{\delta}. \quad (10)$$

So far, we have proved that when B satisfies equation (6), $P(x_j \geq x_m) \leq \frac{\delta}{n'}$ for some $j \in \{j : c_j < c_m\}$. Since $\#\{j : c_j < c_m\} = n'$, we have when B satisfies equation (6) for all $j \in \{j : c_j < c_m\}$, the following holds:

$$\begin{aligned} & P(x_m > x_j, \forall j \in \{c_j < c_m\}) \\ = \;& 1 - P(x_m \leq x_j, \exists j \in \{c_j < c_m\}) \\ \geq \;& 1 - n'\frac{\delta}{n'} = 1-\delta. \end{aligned}$$

It is worthwhile to notice that $M_j \sim O(1)$, $T_j \sim O(\lambda)$ and $O(\lambda) \leq \sigma_j^2 \leq O(d\lambda)$ for all $j \in \{j : c_j < c_m\}$. Denote $g_2(\sigma_j^2, T_j) = (\sigma_j^2+T_j)\log(1+\frac{T_j}{\sigma_j^2}) - T_j$, then $\frac{\partial g_2}{\partial \sigma_j^2} = \log(1+\frac{T_j}{\sigma_j^2}) - \frac{T_j}{\sigma_j^2} < 0$. So $g_2(\sigma_j^2, T_j) \geq O\left(g_2(d\lambda, T_j)\right) = O(\frac{\lambda}{d})$. Therefore, the *r.h.s.* of Equation (10) is $\frac{1}{M_j g_2(\sigma_j^2,T_j)}\log\frac{n'}{\delta} \leq O(\frac{d}{\lambda}\log\frac{n'}{\delta})$. This implies $B \sim O(\frac{d}{\lambda}\log n')$ is sufficient.

□

8.1.4 Proof of Theorem 2

Proof. Take $m = \arg\max_{i=1,\ldots,n} c_i$. Since $w_t h_{jt} \geq 0, \forall j, t$, so $S = \sum_{i=1}^n c_i$.

Since $C = B$, not identifying maximum inner product is equivalent to index m not sampled within B samples. And we know that $P(\text{index } m \text{ sampled in a step}) = \frac{c_m}{S}$. Denote $A = \{\text{index m not sampled within B samples}\}$, then for any $\alpha \in (0, \frac{1}{2})$,

$$\begin{aligned} & P(\text{not identifying maximum inner product}) \\ = \;& E[\mathbb{1}(A)] = E[E[\mathbb{1}(A)|\frac{c_m}{S}]] \\ = \;& E[\left(1-\frac{c_m}{S}\right)^B] \leq E[exp\left(-B\frac{c_m}{S}\right)] \\ = \;& E[\exp\left(-\frac{Bc_m}{S}\right)\mathbb{1}(\frac{c_m}{S} > n^{-1}\alpha\log n)] \\ & +E[\exp\left(-\frac{Bc_m}{S}\right)\mathbb{1}(\frac{c_m}{S} \leq n^{-1}\alpha\log n)] \\ \leq \;& E[\exp\left(-Bn^{-1}\alpha\log n\right)] + P(\frac{c_m}{S} \leq n^{-1}\alpha\log n) \\ \leq \;& n^{-\rho} + P(\frac{c_m}{S} \leq n^{-1}\alpha\log n). \end{aligned} \quad (11)$$

For any $\alpha \in (0, \frac{1}{2})$ and $\forall \epsilon > 0$,

$$\begin{aligned} & P(\frac{c_m}{S} > n^{-1}\alpha\log n) = P(c_m > n^{-1}S\alpha\log n) \\ \geq \;& P(c_m > n^{-1}S\alpha\log n, n^{-1}S \leq \beta^{-1}+\epsilon) \\ = \;& P(c_m > n^{-1}S\alpha\log n | n^{-1}S \leq \beta^{-1}+\epsilon) \\ & P(n^{-1}S \leq \beta^{-1}+\epsilon) \\ \geq \;& P(c_m > (\beta^{-1}+\epsilon)\alpha\log n)P(n^{-1}S \leq \beta^{-1}+\epsilon) \end{aligned}$$

Since $c_j \overset{iid}{\sim} Exp(\beta)$, $E[S] = nE[c_j] = n\beta^{-1}$, $\text{Var}[S] = n\text{Var}[c_i] = \frac{n}{\beta^2}$. According to Chebyshev's Inequality, $\forall \epsilon > 0$,

$$\begin{aligned} & P(n^{-1}S \leq \beta^{-1}+\epsilon) = 1 - P(n^{-1}S - \beta^{-1} \geq \epsilon) \\ \geq \;& 1 - P(|n^{-1}S - \beta^{-1}| \geq \epsilon) \\ \geq \;& 1 - \frac{\text{Var}[n^{-1}S]}{\epsilon^2} = 1 - \frac{1}{n\beta^2\epsilon^2}. \end{aligned} \quad (12)$$

Since $c_j \overset{iid}{\sim} Exp(\beta)$, we have for $\alpha \in (0, \frac{1}{2})$,

$$\begin{aligned} & P(c_m > (\beta^{-1}+\epsilon)\alpha\log n) \\ = \;& 1 - P(c_m \leq (\beta^{-1}+\epsilon)\alpha\log n) \\ = \;& 1 - [1 - e^{-\beta(\beta^{-1}+\epsilon)\alpha\log n}]^n \\ = \;& 1 - (1 - \frac{1}{n^{(1+\beta\epsilon)\alpha}})^n \\ \sim \;& 1 - e^{-n^{1-(1+\beta\epsilon)\alpha}}. \end{aligned} \quad (13)$$

Since it holds for any $\epsilon > 0$, we can take $\epsilon = \beta^{-1}$, then from Equation (12, 13), we have

$$\begin{aligned} & P(\frac{c_m}{S} > n^{-1}\alpha\log n) \\ \geq \;& \left(1-\frac{1}{n\beta^2\epsilon^2}\right)\left(1-e^{-n^{1-(1+\beta\epsilon)\alpha}}\right) \\ = \;& (1-\frac{1}{n})(1-e^{-n^{1-2\alpha}}). \end{aligned} \quad (14)$$

From Equation (11), we know:

$$\begin{aligned} & P(\text{not identifying maximum inner product}) \\ \leq\ & n^{-\rho} + 1 - (1 - \frac{1}{n})(1 - e^{-n^{1-2\alpha}}) \\ \sim\ & O\left(\frac{1}{n^{\rho}}\right). \end{aligned}$$

□

8.1.5 Proof of Corollary 1

Proof. Assume $E(c_j) = \beta^{-1} > 0$ and $\text{Var}[c_j] = \gamma^{-2}$, then for any $\epsilon > 0$, the following equation still holds.

$$\begin{aligned} & P(n^{-1}S \leq \beta^{-1} + \epsilon) = 1 - P(n^{-1}S - \beta^{-1} \geq \epsilon) \\ \geq\ & 1 - P(|n^{-1}S - \beta^{-1}| \geq \epsilon) \\ \geq\ & 1 - \frac{\text{Var}[n^{-1}S]}{\epsilon^2} = 1 - \frac{1}{n\gamma^2\epsilon^2}. \end{aligned} \tag{15}$$

Take $\epsilon = \gamma^{-1}$ in the above equation, we have

$$P(n^{-1}S \leq \beta^{-1} + \epsilon) > 1 - \frac{1}{n}. \tag{16}$$

Since c_i is heavy-tailed, so for all $\mu > 0$, there exists $x_0 > 0$, such that for all $x \geq x_0$,

$$P(c_i \geq x) \geq e^{-\mu x}. \tag{17}$$

Take $\mu = \frac{1}{\beta^{-1}+\epsilon}$, when $(\beta^{-1} + \epsilon)\alpha \log n \geq x_0$, we have

$$\begin{aligned} & P(c_m > (\beta^{-1} + \epsilon)\alpha \log n) \\ =\ & 1 - P(c_m \leq (\beta^{-1} + \epsilon)\alpha \log n) \\ =\ & 1 - \left(P(c_i \leq (\beta^{-1} + \epsilon)\alpha \log n)\right)^n \\ \geq\ & 1 - \left(1 - e^{-\mu(\beta^{-1}+\epsilon)\alpha \log n}\right)^n \\ =\ & 1 - \left(1 - e^{-\alpha \log n}\right)^n \\ \sim\ & 1 - e^{-n^{1-\alpha}}. \end{aligned} \tag{18}$$

From Equation $(12, 16, 18)$, we know that

$$P(\frac{c_m}{S} > n^{-1}\alpha \log n) \geq (1 - \frac{1}{n})(1 - e^{-n^{1-\alpha}}).$$

From the proof of Theorem 2, we know the result of Corollary 1 holds. □

8.2 More comparison of experimental results

In order to get more informative comparisons, we summarize the results from Figure 3 in the following tables. Prec@1 and Prec@5 of MovieLens with $d = 50$ are shown in Table 1 and 2 respectively. From the results, we can see that Sampling-MIPS algorithm outperforms PCA-MIPS, LSH-MIPS and Diamond-MSIPS consistently on this dataset. Prec@1 and Prec@5 of Netflix with $d = 50$ are shown in Table 3 and 4 respectively. Note that the speedup of PCA-tree method depends on the depth of the PCA tree, which is corresponding to the number of candidates in each leaf node. A deeper PCA tree leads to a higher speedup with a tradeoff of a lower precision. The maximum depth in our experiment is too small to generate a point with a speedup greater than 5. This is the reason that in Table 3 and 4, there are no results shown for PCA-MIPS. But we can see from Figure 3 that with the current maximum depth chosen, the precision of PCA-MIPS is drastically reduced to almost 0 when speedup is less than 5. So with a bigger maximum depth, the result of PCA-MIPS will get even worst.

Table 1: Result of prec@1 for MovieLens $(d = 50)$

Speedup (prec@1)	5	10	20
Sampling-MIPS	99.95%	99.65%	65%
Diamond	68.35%	42.25%	25.85%
PCA	0.15%	0.00%	0.00%
LSH	4.75%	18.9%	0.05%

Table 2: Result of prec@5 for MovieLens $(d = 50)$

Speedup (prec@5)	5	10	20
Sampling-MIPS	87.38%	72%	13%
Diamond	11.65%	5.41%	1.52%
PCA	0.00%	0.00%	0.00%
LSH	9.73%	4.81%	0.00%

Table 3: Result of prec@1 for Netflix $(d = 50)$

Speedup (prec@1)	5	10	20
Sampling-MIPS	97.30%	77.15%	29.80%
Diamond	51.85%	31.4%	15.2%
PCA	NA	NA	NA
LSH	23%	1.05%	NA%

Table 4: Result of prec@5 for Netflix $(d = 50)$

Speedup (prec@5)	5	10	20
Sampling-MIPS	58.87%	28.7%	7.98%
Diamond	14.47%	7.26%	3.29%
PCA	NA	NA	NA
LSH	9.25%	0.29%	NA%

8.3 Extension to Maximum All-pair Dot-product (MAD) Search

8.3.1 Proposed algorithm for MAD Search

Our Sampling-MIPS algorithm can also be extended to solve the Maximum All-pair Dot-product (MAD) search problem. Instead of finding the maximum inner product for a specific query, MAD aims to find the maximum inner product over a set of m queries and n vectors in the database. More specifically, given two groups of d-dimensional vectors $\mathcal{W} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m\}$ and $\mathcal{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n\}$, the goal of MAD problem is to find the index pairs $(i, j) \in \{1, \ldots, m\} \times \{1, \ldots, n\}$ who have the maximum or top-K dot products. When $m = 1$, MAD problem reduces to the MIPS problem. MAD is also used in many recommender systems when we aim to find the best $(\text{user}, \text{item})$ among all the possible pairs.

We can use the same idea in Sampling-MIPS to solve the MAD problem. The only difference is the sampling procedure. In MIPS problem, we first sample t, then sample j conditioned on t. In MAD problem, we simply add one more step. We still sample t first, but then we sample i and j independently. We use a m-by-d matrix $W = [\mathbf{w}_1 \ldots \mathbf{w}_m]^T$ to denote the set of m queries and use w_{it} to denote the t-th feature of the i-th query. Define $P(t|W) \propto \sum_{i=1}^m |w_{it}| \sum_{j=1}^n |h_{jt}|$, $P(i|t) = \frac{|w_{it}|}{\sum_{i=1}^m |w_{it}|}$, and $P(j|t) = \frac{|h_{jt}|}{\sum_{j=1}^n |h_{jt}|}$. We also assume that $\sum_{i=1}^m |w_{it}| \neq 0$ and $\sum_{j=1}^n |h_{jt}| \neq 0$ for all t. So we have

$$\begin{aligned} P(i, j|W) &= \sum_{t=1}^d P(i, j, t|W) \\ &= \sum_{t=1}^d P(i|t)P(j|t)P(t|W) \\ &\propto \sum_{t=1}^d |w_{it}h_{jt}|. \end{aligned}$$

Here, the distribution for sampling (i, j) is very similar to the distribution for sampling j in MIPS problem.

Similarly, alias table for sampling j can be constructed in pre-processing phase. In query-dependent phase, we only need to construct alias table for sampling t and i. Details are shown in Algorithm 5. The time complexity of each step of is also shown in Algorithm 5. The total time complexity is $O(md + B + Cd)$, while the naive time complexity is $O(mnd)$. We expect this algorithm to perform as well as it does in MIPS problem. The theoretical guarantee of this sampling MAD algorithm can also be proved in a similar manner as Sampling-MIPS.

Algorithm 5 Sampling-MAD

Pre-processing:

$s_t \leftarrow \sum_{j=1}^n |h_{jt}|, \forall t = 1, \cdots, d$

$k_t \leftarrow \sum_{i=1}^m |w_{it}|, \forall t = 1, \cdots, d$

Use alias table to construct $P(j|t) \leftarrow \text{multinomial}([|h_{1t}|, \ldots, |h_{nt}|]), \forall t$

Candidate Screening:

Use alias table to construct $P(t|W) \leftarrow \text{multinomial}([|k_1 s_1|, \ldots, |k_d s_d|])$ $P(i|t) \leftarrow \text{multinomial}([|w_{1t}|, \ldots, |w_{mt}|]), \forall t$ $\cdots O(md)$

$\mathbf{x} = [x_{11} \ldots, x_{mn}] \leftarrow [0, \ldots, 0]$

Specify sample size B

for $b = 1, \ldots, B$ **do** $\cdots O(B)$

  Use alias method to sample $t \sim P(t|W)$

  Use alias method to sample $i \sim P(i|t)$

  Use alias method to sample $j \sim P(j|t)$

  $x_{ij} \leftarrow x_{ij} + \text{sgn}(w_{it}h_{jt})$

end for

Prediction Phase:

Find the biggest C elements in $\mathbf{x}$, i.e., $|\mathcal{C}(W)| = C$ and $\mathcal{C}(W) \leftarrow \{(i, j)|x_{ij} \geq x_{i'j'}, \forall (i', j') \notin \mathcal{C}(W)\}$ $\cdots O(B)$

for $(i, j) \in \mathcal{C}(W)$ **do** $\cdots O(Cd)$

  Compute inner product $\mathbf{w}_i^T \mathbf{h}_j$

end for

Output: indexes of the top-K elements of $\{\mathbf{w}_i^T \mathbf{h}_j | (i, j) \in C(W)\}$ $\cdots O(C)$

8.3.2 Mathematical analysis of Sampling-MAD

We also have the similar theoretical results for MAD problem. We omit the proofs since they are very similar to Lemma 1, 2, Theorem 1, 2 and Corollary 1 for the MIPS case.

Lemma 3. *Define $x_{ij,b} = sgn(w_{it}h_{jt})$ if index pair (i, j, t) is sampled in the b-th sampling step, $x_{ij,b} = 0$ otherwise. Note that $x_{ij} = \sum_{b=1}^B x_{ij,b}$. Define*

$$S = \sum_{t=1}^d \sum_{i=1}^m \sum_{j=1}^n |w_{it}h_{jt}|,$$

then we have

$$E[x_{ij,b}] = \frac{\mathbf{w}_i^T \mathbf{h}_j}{S}$$

and

$$E[x_{ij}] = \frac{B\mathbf{w}_i^T \mathbf{h}_j}{S}.$$

We can then show the ranking of x_{ij}'s will be correct when we have enough samples.

Lemma 4. *Define $A_{ij} = \sum_{t=1}^d |w_{it}h_{jt}|$, then $E[(x_{ij,b} - x_{i'j',b})^2] = \frac{A_{ij} + A_{i'j'}}{S}, \forall (i, j) \neq (i', j')$. Therefore, from*

Lemma 3, we have

$$\begin{aligned}&Var(x_{ij,b} - x_{i'j',b})\\ =&\frac{A_{ij} + A_{i'j'}}{S} - \frac{(\mathbf{w}_i^T\mathbf{h}_j - \mathbf{w}_{i'}^T\mathbf{h}_{j'})^2}{S^2}\end{aligned}$$

and $Var(x_{ij} - x_{i'j'}) = BVar(x_{ij,b} - x_{i'j',b})$.

Theorem 3. *For some index pair* (I, J)*, define for* $(i, j) \in \{(i, j) : \mathbf{w}_i^T\mathbf{h}_j < \mathbf{w}_I^T\mathbf{h}_J\}$*:*

$$\begin{aligned}N &= \#\{(i, j) : \mathbf{w}_i^T\mathbf{h}_j < \mathbf{w}_I^T\mathbf{h}_J\} \text{ (assume } N \neq 0),\\ \Lambda_{ij} &= \frac{\mathbf{w}_I^T\mathbf{h}_J - \mathbf{w}_i^T\mathbf{h}_j}{S},\\ \Lambda &= \min_{\{(i,j):\mathbf{w}_i^T\mathbf{h}_j < \mathbf{w}_I^T\mathbf{h}_J\}} \Lambda_{ij},\\ T_{ij} &= \frac{(S + \mathbf{w}_I^T\mathbf{h}_J - \mathbf{w}_i^T\mathbf{h}_j)(\mathbf{w}_I^T\mathbf{h}_J - \mathbf{w}_i^T\mathbf{h}_j)}{S^2},\\ M_{ij} &= \frac{S^2}{(S + \mathbf{w}_I^T\mathbf{h}_J - \mathbf{w}_i^T\mathbf{h}_j)^2},\\ \sigma_{ij}^2 &= Var(x_{ij,b} - x_{IJ,b}).\end{aligned}$$

If sample size B *satisfies the following equation for all* $(i, j) \neq (I, J)$*,*

$$B \geq \frac{\log\frac{N}{\delta}}{M_{ij}\sigma_{ij}^2\left[(1 + \frac{T_{ij}}{\sigma_{ij}^2})\log(1 + \frac{T_{ij}}{\sigma_{ij}^2}) - \frac{T_{ij}}{\sigma_{ij}^2}\right]}, \qquad (19)$$

implying that $B \sim O(\frac{1}{\Lambda}\log(N))$*, then with error probability* $\delta \in (0, 1)$*, we have*

$$P(x_{IJ} > x_{ij}, \forall(i, j) \in \{\mathbf{w}_i^T\mathbf{h}_j < \mathbf{w}_I^T\mathbf{h}_J\}) \geq 1 - \delta.$$

Similar to MIPS problem, when $\mathbf{w}_I^T\mathbf{h}_J$ has the maximum inner product, N in the theorem above can be replaced by mn. That means that sample size $B = O(\frac{1}{\Lambda}\log(mn))$ is sufficient for identifying the maximum inner product in MAD problem and $\Lambda < 1$ here.

Theorem 4. *Assume* $w_{it}h_{jt} \geq 0$*, for all* $(i, j, t) \in \{1, \ldots, m\} \times \{1, \ldots, n\} \times \{1, \ldots, d\}$ *and* $\mathbf{w}_i^T\mathbf{h}_j \overset{iid}{\sim} Exp(\beta)$*. In Sampling-MAD algorithm, if we take* $C = B$*, where* C *means we calculate inner products of indexes with top-*C *scores, then when* $B \geq \frac{\rho}{\alpha}mn$*, where* $\alpha \in (0, \frac{1}{2})$ *and* $\rho \in (0, \frac{\alpha d(n-1)}{(d+1)n})$*,*

$$P(\text{not identifying maximum inner product}) \leq O\left(\frac{1}{(mn)^\rho}\right).$$

Therefore, the overall time complexity of Sampling-MAD is $O(md + B + Bd) = O(md + \frac{\rho}{\alpha}(d + 1)mn) < O(mnd)$.

Corollary 2. *Assume* $w_{it}h_{jt} \geq 0$*, for all* $(i, j, t) \in \{1, \ldots, m\} \times \{1, \ldots, n\} \times \{1, \ldots, d\}$*. If* $\mathcal{F}$ *is a continuous, non-negative, heavy-tailed distribution,* $\mathbf{w}_i^T\mathbf{h}_j \overset{iid}{\sim} \mathcal{F}$ *and* $E[(\mathbf{w}_i^T\mathbf{h}_j)^2] < \infty$*, then when* $B \geq \frac{\rho}{\alpha}mn$*, where* $\alpha \in (0, 1)$ *and* $\rho \in (0, \frac{\alpha d(n-1)}{(d+1)n})$*, take* $C = B$ *in Sampling-MIPS algorithm, we have*

$$P(\text{not identifying maximum inner product}) \leq O\left(\frac{1}{(mn)^\rho}\right).$$