Supplementary Material for StreamFP: Fingerprint-guided Data Selection for Efficient Stream Learning

Tongjun Shi[1,2], Shuhao Zhang[1], Binbin Chen[2], Bingsheng He[3]
[1] National Engineering Research Center for Big DataTechnology and System
Services Computing Technology and System Lab
Cluster and Grid Computing Lab
School of Computer Science and Technology
Huazhong University of Science and Technology, Wuhan, 430074, China
[2] Singapore University of Technology and Design [3] National University of Singapore

A APPENDIX

A.1 Gradient Correlations between different Tasks

As demonstrated in Fig. 1, the gradients of both fingerprints and the last layer demonstrate strong orthogonality across distinct tasks, as evidenced by near-zero correlation coefficients in off-diagonal elements. This gradient orthogonality indicates that fingerprint parameters evolve along task-specific trajectories during streaming learning. The gradient correlation analysis of the last layer further reveals that independently optimized fingerprints, leveraging model-specific knowledge, effectively minimize cross-task interference in the final layer outputs.

A.2 Comparison of Different Pretrained Models.

As *StreamFP* is based on the pretrained ViT, we also experiment with varying pretrained models based on the supervised (ImageNet-1K [4] and ImageNet-21K [3]) and the self-supervised (iBOT [5], DINO [1], and MoCo v3 [2]) datasets. Results are shown in Table 1. *StreamFP* consistently outperforms the *ER** method across different pretrained models in terms of accuracy and forgetting. Specifically, when using the ImageNet-1K pretrained model, *StreamFP* achieves the highest accuracy of 64.44%, compared to *ER**'s 59.99%. With ImageNet-21K, *StreamFP* even achieves the 0% forgetting that preserves all learned knowledge in the streaming setting. Notably, the performance differences highlight the impact of different pretraining strategies on stream learning outcomes, with *StreamFP* consistently providing significant improvements in accuracy, albeit with varying degrees of forgetting. These results underscore the robustness and versatility of *StreamFP* across different pretraining contexts.

A.3 Proof of Coreset Quality Guarantee

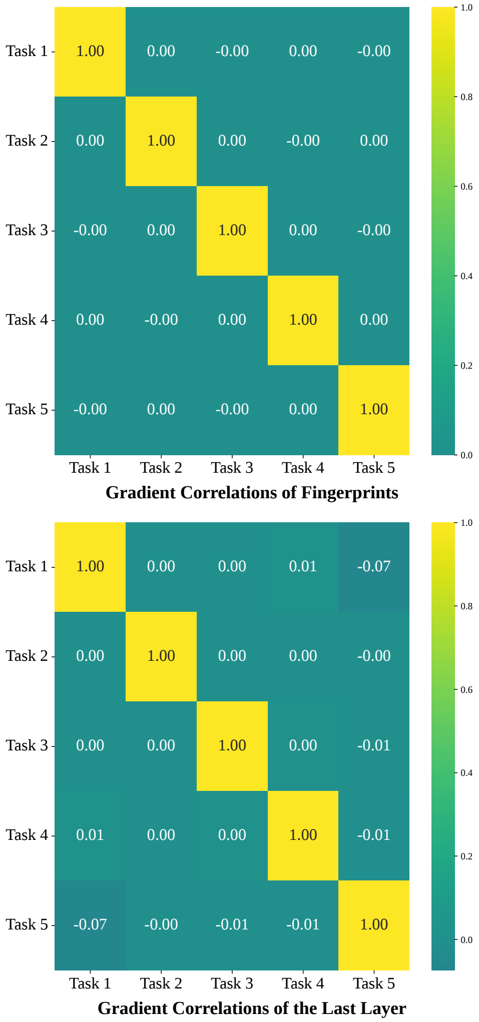
We prove that our fingerprint-based coreset selection method satisfies the standard definition of coreset with rigorous theoretical guarantees.

Theorem 1 (Coreset Quality Guarantee). *With probability at least $1-\delta$, the coreset C^t satisfies:*

$$(1-\varepsilon)cost(B^t) \le cost(C^t) \le (1+\varepsilon)cost(B^t),$$

where $cost(X) = \frac{1}{|X|}\sum_{x\in X} d(x)$, representing the average angular distance, $d(x) = \arccos(sim(x,P))$, and $\varepsilon = O(\sqrt{\log(1/\delta)/(\sigma b)})$.

Proof. Let us first define the notations:



Gradient Correlations of Fingerprints

Gradient Correlations of the Last Layer

Figure 1: Heatmap of gradient correlations for diverse tasks on Stream-51.

- B^t: original dataset with b points,
- C^t: selected coreset with $c = \sigma b$ points,
- $\mu = \text{sim}(x, P)$: cosine similarity between the embeddings of point x and fingerprints P,

Table 1: Comparison of ER* and *StreamFP* based on different pretrained models on Stream-51 with λ=6028.

Method	ImageNet-1K		ImageNet-21K		DINO-1K		iBOT-1K		iBOT-21K		MoCo-1K	
	Acc	Fgt	Acc	Fgt	Acc	Fgt	Acc	Fgt	Acc	Fgt	Acc	Fgt
ER*	59.99	3.70	55.68	2.09	44.83	5.43	39.07	2.50	2.12	9.71	2.00	10.03
StreamFP	**64.44**	**2.25**	**60.45**	**0.00**	**47.44**	**4.91**	**41.79**	**1.35**	**3.12**	**8.00**	**2.52**	**7.01**

- $\mu[1] \geq \mu[2] \geq \cdots \geq \mu[b]$: similarity values sorted in descending order,
- $k = \lfloor \frac{b}{2} \rfloor$: median position,
- $d(x) = \arccos(\mathrm{sim}(x, P))$: angular distance,
- $\mathrm{cost}(X) = \frac{1}{|X|}\sum_{x \in X} d(x)$: average angular distance.

We first establish four key lemmas:

Lemma 1 (Selection Interval Bounds). *For all $x \in C^t$:*

$$\mu[k + \lfloor \frac{c}{2} \rfloor] \leq sim(x, P) \leq \mu[k - \lfloor \frac{c}{2} \rfloor].$$

Lemma 2 (Single Point Change Bound). *For any single point change in C^t:*

$$|cost_{new}(C^t) - cost(C^t)| = \frac{1}{c}|d(x') - d(x)| \leq \frac{\pi}{c},$$

since $d(x) = \arccos(sim(x, P)) \in [0, \pi]$.

Lemma 3 (Similarity Difference Bound). *For any $x_1 \in C^t$, $x_2 \in B^t$:*

$$|sim(x_1, P) - sim(x_2, P)| \leq \max\{|\mu[1] - \mu[k + \lfloor \frac{c}{2} \rfloor]|, |\mu[b] - \mu[k - \lfloor \frac{c}{2} \rfloor]|\}.$$

Lemma 4 (Expected Value Approximation). *For coreset C^t selected from the middle region of sorted similarity sequence and original batch B^t:*

$$\begin{aligned} &|E[cost(C^t)] - cost(B^t)| \\ &\leq L \cdot \max\{|\mu[1] - \mu[k + \lfloor \frac{c}{2} \rfloor]|, |\mu[b] - \mu[k - \lfloor \frac{c}{2} \rfloor]|\} = M, \end{aligned}$$

where L is the Lipschitz constant of arccos *function and M is constant.*

Proof. Let us partition B^t based on similarity values:

$$\begin{aligned} B_l &= \{x \mid \mu[b] < \mathrm{sim}(x, P) < \mu[k + \lfloor \frac{c}{2} \rfloor]\}, \\ B_m &= \{x \mid \mu[k + \lfloor \frac{c}{2} \rfloor] \leq \mathrm{sim}(x, P) \leq \mu[k - \lfloor \frac{c}{2} \rfloor]\}, \\ B_h &= \{x \mid \mu[k - \lfloor \frac{c}{2} \rfloor] < \mathrm{sim}(x, P) < \mu[1]\}. \end{aligned}$$

Then, we have:

$$\mathrm{cost}(B^t) = \frac{|B_l|}{b} \cdot \mathrm{cost}(B_l) + \frac{|B_m|}{b} \cdot \mathrm{cost}(B_m) + \frac{|B_h|}{b} \cdot \mathrm{cost}(B_h).$$

Given Lemma 1, for any point $x \in B_m$:

$$\mu[k + \lfloor \frac{c}{2} \rfloor] \leq \mathrm{sim}(x, P) \leq \mu[k - \lfloor \frac{c}{2} \rfloor],$$

based on the monotonicity of arccos function, we have:

$$\arccos(\mu[k - \lfloor \frac{c}{2} \rfloor]) \leq d(x) \leq \arccos(\mu[k + \lfloor \frac{c}{2} \rfloor]).$$

Therefore, for any randomly selected point from B_m:

$$E[d(x)] \in [\arccos(\mu[k - \lfloor \frac{c}{2} \rfloor]), \arccos(\mu[k + \lfloor \frac{c}{2} \rfloor])].$$

By Lipschitz property of arccos function: for any $x_1 = \mathrm{sim}(x'_1, P)$ and $x_2 = \mathrm{sim}(x'_2, P)$ where $x_1, x_2 \in [-1, 1]$, there exists a Lipschitz constant L such that:

$$|\arccos(x_1) - \arccos(x_2)| \leq L|x_1 - x_2|.$$

This implies for the expected value:

$$\begin{aligned} &|E[\mathrm{cost}(B_m)] - \mathrm{cost}(B_m)| \\ &\leq \arccos(\mu[k + \lfloor \frac{c}{2} \rfloor]) - \arccos(\mu[k - \lfloor \frac{c}{2} \rfloor]) \\ &\leq L \cdot |\mu[k + \lfloor \frac{c}{2} \rfloor] - \mu[k - \lfloor \frac{c}{2} \rfloor]|, \end{aligned}$$

where L is the Lipschitz constant of arccos function. Similarly, for the differences between expected costs:

$$\begin{aligned} |E[\mathrm{cost}(B_m)] - \mathrm{cost}(B_l)| &\leq L|\mu[b] - \mu[k - \lfloor \frac{c}{2} \rfloor]|, \\ |E[\mathrm{cost}(B_m)] - \mathrm{cost}(B_h)| &\leq L|\mu[1] - \mu[k + \lfloor \frac{c}{2} \rfloor]|. \end{aligned}$$

Note that:

$$\begin{aligned} |\mu[k - \lfloor \frac{c}{2} \rfloor] - \mu[k + \lfloor \frac{c}{2} \rfloor]| &\leq |\mu[1] - \mu[k + \lfloor \frac{c}{2} \rfloor]|, \\ |\mu[k - \lfloor \frac{c}{2} \rfloor] - \mu[k + \lfloor \frac{c}{2} \rfloor]| &\leq |\mu[b] - \mu[k - \lfloor \frac{c}{2} \rfloor]|. \end{aligned}$$

By selection strategy:

$$E[\mathrm{cost}(C^t)] = E[\mathrm{cost}(B_m)],$$

we can get:

$$\begin{aligned}
&|E[\mathrm{cost}(C^t)] - \mathrm{cost}(B^t)| \\
&= |E[\mathrm{cost}(B_m)] - [\frac{|B_l|}{b} \cdot \mathrm{cost}(B_l) + \frac{|B_m|}{b} \cdot \mathrm{cost}(B_m) \\
&\quad + \frac{|B_h|}{b} \cdot \mathrm{cost}(B_h)]| \\
&= |\frac{|B_l|}{b} \cdot E[\mathrm{cost}(B_m)] - \frac{|B_l|}{b} \cdot \mathrm{cost}(B_l)| \\
&\quad + |\frac{|B_m|}{b} \cdot E[\mathrm{cost}(B_m)] - \frac{|B_m|}{b} \cdot \mathrm{cost}(B_m)| \\
&\quad + |\frac{|B_h|}{b} \cdot E[\mathrm{cost}(B_m)] - \frac{|B_h|}{b} \cdot \mathrm{cost}(B_h)| \\
&\le \frac{|B_l|}{b} \cdot L|\mu[b] - \mu[k - \lfloor \frac{c}{2} \rfloor]| \\
&\quad + \frac{|B_m|}{b} \cdot L|\mu[k - \lfloor \frac{c}{2} \rfloor] - \mu[k + \lfloor \frac{c}{2} \rfloor]| \\
&\quad + \frac{|B_h|}{b} \cdot L|\mu[1] - \mu[k + \lfloor \frac{c}{2} \rfloor]| \\
&\le \frac{|B_l|}{b} \cdot L \cdot \max\{|\mu[1] - \mu[k + \lfloor \frac{c}{2} \rfloor]|, |\mu[b] - \mu[k - \lfloor \frac{c}{2} \rfloor]|\} \\
&\quad + \frac{|B_m|}{b} \cdot L \cdot \max\{|\mu[1] - \mu[k + \lfloor \frac{c}{2} \rfloor]|, |\mu[b] - \mu[k - \lfloor \frac{c}{2} \rfloor]|\} \\
&\quad + \frac{|B_h|}{b} \cdot L \cdot \max\{|\mu[1] - \mu[k + \lfloor \frac{c}{2} \rfloor]|, |\mu[b] - \mu[k - \lfloor \frac{c}{2} \rfloor]|\} \\
&= L \cdot \max\{|\mu[1] - \mu[k + \lfloor \frac{c}{2} \rfloor]|, |\mu[b] - \mu[k - \lfloor \frac{c}{2} \rfloor]|\} \\
&\quad \cdot (\frac{|B_l|}{b} + \frac{|B_m|}{b} + \frac{|B_h|}{b}) \\
&= L \cdot \max\{|\mu[1] - \mu[k + \lfloor \frac{c}{2} \rfloor]|, |\mu[b] - \mu[k - \lfloor \frac{c}{2} \rfloor]|\} = M. \qquad \square
\end{aligned}$$

Now we proceed with the main proof:

Step 1. We have two bounds: by McDiarmid's inequality from Lemma 2 where $c = \sigma b$, for any $\xi > 0$:

$$\mathbb{P}(|\mathrm{cost}(C^t) - E[\mathrm{cost}(C^t)]| \ge \xi) \le 2\exp(-2c\xi^2/\pi^2).$$

Then, with probability at least $1 - 2\exp(-2c\xi^2/\pi^2)$:

$$|\mathrm{cost}(C^t) - E[\mathrm{cost}(C^t)]| \le \xi.$$

By Lemma 4:

$$|E[\mathrm{cost}(C^t)] - \mathrm{cost}(B^t)| \le M.$$

Step 2. Let $\xi + M = \varepsilon \cdot \mathrm{cost}(B^t)$, with probability at least $1 - 2\exp(-2c\xi^2/\pi^2) = 1 - 2\exp(-2c(\varepsilon \cdot \mathrm{cost}(B^t) - M)^2/\pi^2)$:

$$\begin{aligned}
&|\mathrm{cost}(C^t) - \mathrm{cost}(B^t)| \\
&= |\mathrm{cost}(C^t) - E[\mathrm{cost}(C^t)] + E[\mathrm{cost}(C^t)] - \mathrm{cost}(B^t)| \\
&\le |\mathrm{cost}(C^t) - E[\mathrm{cost}(C^t)]| + |E[\mathrm{cost}(C^t)] - \mathrm{cost}(B^t)| \\
&\le \xi + M \\
&= \varepsilon \cdot \mathrm{cost}(B^t).
\end{aligned}$$

Step 3. Setting this probability to be at least $1 - \delta$:

$$1 - 2\exp(-2c(\varepsilon \cdot \mathrm{cost}(B^t) - M)^2/\pi^2) = 1 - \delta.$$

Step 4. Solving for ε:

$$\varepsilon = M/\mathrm{cost}(B^t) + (\pi/\mathrm{cost}(B^t))\sqrt{-\ln(\delta/2)/(2c)}.$$

Since:

- $M = L \cdot \max\{|\mu[1] - \mu[k + \lfloor \frac{c}{2} \rfloor]|, |\mu[b] - \mu[k - \lfloor \frac{c}{2} \rfloor]|\}$ is constant,
- $0 < \mathrm{cost}(B^t) \le \pi$,
- $c = \sigma b$,
- $-\ln(\delta/2) = O(\log(1/\delta))$,

we have:

$$\varepsilon = O(\sqrt{\log(1/\delta)/(\sigma b)}).$$

Step 5. Therefore, with probability $\ge 1 - \delta$ and $\varepsilon = O(\sqrt{\log(1/\delta)/(\sigma b)})$:

$$\begin{aligned}
&|\mathrm{cost}(C^t) - \mathrm{cost}(B^t)| \le \varepsilon \cdot \mathrm{cost}(B^t) \iff \\
&-\varepsilon \cdot \mathrm{cost}(B^t) \le \mathrm{cost}(C^t) - \mathrm{cost}(B^t) \le \varepsilon \cdot \mathrm{cost}(B^t) \iff \\
&(1 - \varepsilon)\mathrm{cost}(B^t) \le \mathrm{cost}(C^t) \le (1 + \varepsilon)\mathrm{cost}(B^t). \qquad \square
\end{aligned}$$

A.4 Proof of Buffer Update Quality Guarantee

Theorem 2. *(Buffer Update Quality Guarantee) With probability at least $1 - \delta$, the distribution P_M obtained from the buffer satisfies:*

$$D(P_M, P_t) \le \varepsilon,$$

where $D(\cdot, \cdot)$ denotes the Maximum Mean Discrepancy (MMD) between distributions with kernel function $k(x, y) = sim(x, P)sim(y, P)$, P_M is the distribution of buffer data updated by StreamFP, P_t is the distribution of all seen data until time t, and $\varepsilon = O((m^3 \ln(1/\delta))^{1/4})$ with m being the buffer size.

Proof. Let us first define the notations:

- M^t: updated buffer with m points,
- P_M: distribution of buffer data,
- P_t: distribution of all seen data until time t,
- $k(x, y) = \mathrm{sim}(x, P)\mathrm{sim}(y, P)$: kernel function, where $\mathrm{sim}(\cdot)$ is the cosine similarity,
- H_m: m-th harmonic number: $\sum_{i=1}^{m} 1/i$,
- $w_i = 1 - \frac{\frac{1}{i}}{\sum_{j=1}^{m} 1/j} = 1 - \frac{1}{jH_m}$: based on the rank probability to get the weight for point i,
- $w_{ij} = (1 - \frac{1}{iH_m})(1 - \frac{1}{jH_m})$,
- $\mathbb{E}_{x,x' \sim P_M}[k(x, x')] = \sum_{i,j=1} w_{ij} k(x, x')$: since $x \sim P_M$ is sampled based on the rank probability,
- $\mathbb{E}_{y,y' \sim P_t}[k(y, y')] = \sum_{i,j=1} \frac{1}{n^2} k(y, y')$: since $x \sim P_M$ is sampled based on the unity probability,
- $\mathbb{E}_{x \sim P_M, y \sim P_t}[k(x, y)] = \sum_{i,j=1} w_i \frac{1}{n} k(x, y)$,
- $\mathrm{MMD}^2(P_M, P_t) = \mathbb{E}_{x,x' \sim P_M}[k(x, x')] + \mathbb{E}_{y,y' \sim P_t}[k(y, y')] - 2\mathbb{E}_{x \sim P_M, y \sim P_t}[k(x, y)]$.

Lemma 5 (MMD^2 Change Bound). *When changing the i-th sample in the buffer with m size from x_i to x_i', the change in MMD^2 satisfies:*

$$|\Delta MMD^2| \le 3 + 2m. \qquad (1)$$

Proof. We analyze the change in MMD^2 in 4 steps:

Step 1: This step is to analyze MMD^2 changes. Given the MMD^2:

$$\begin{aligned}\mathrm{MMD}^2(P_M, P_t) &= \mathbb{E}_{x_1,x_2\sim P_M}[k(x_1,x_2)] + \mathbb{E}_{y_1,y_2\sim P_t}[k(y_1,y_2)] \\ &\quad - 2\mathbb{E}_{x\sim P_M, y\sim P_t}[k(x,y)] \\ &= \sum_{i,j=1}^{m} w_{ij}k(x_i,x_j) + \mathbb{E}_{y_1,y_2\sim P_t}[k(y_1,y_2)] \\ &\quad - 2\sum_{i=1}^{m} w_i \mathbb{E}_{y\sim P_t}[k(x_i,y)].\end{aligned}$$

When x_i in P_M changes to x'_i, the change in MMD^2:

$$\begin{aligned}|\Delta\mathrm{MMD}^2| &\le |\sum_{i,j=1}^{m} w'_{ij}k'(x_i,x_j) - \sum_{i,j=1}^{m} w_{ij}k(x_i,x_j)| \\ &\quad + 2|\sum_{i=1}^{m} w'_i \mathbb{E}_{y\sim P_t}[k'(x_i,y)] - \sum_{i'=1}^{m} w_i \mathbb{E}_{y\sim P_t}[k(x_i,y)]| \\ &= \Delta(\text{first term}) + (\Delta\text{third term}),\end{aligned}$$

specifically:

$$\begin{aligned}(\text{first term})' &= w'_{ii}k(x'_i,x'_i) && (\text{self term}) \\ &\quad + \sum_{j\neq i}^{m} w'_i w_j k(x'_i,x_j) && (x'_i \text{ as the 1-st sample}) \\ &\quad + \sum_{j\neq i}^{m} w_j w'_i k(x_j,x'_i) && (x'_i \text{ as the 2-nd sample}) \\ &\quad + \sum_{p\neq i, q\neq i}^{m} w_{pq}k(x_p,x_q). && (\text{has no } x_i) \\ (\text{third term})' &= \sum_{i=1}^{m} w'_i \mathbb{E}_{y\sim P_t}[k(x'_i,y)].\end{aligned}$$

Therefore, the change in MMD^2 can be decomposed as:

$$\begin{aligned}|\Delta\mathrm{MMD}^2| &\le \Delta(\text{first term}) + \Delta(\text{third term}) \\ &= [w'_{ii}k(x'_i,x'_i) - w_{ii}k(x_i,x_i)] \\ &\quad + 2[w'_i \sum_{j\neq i}^{m} w_j k(x'_i,x_j) - w_i \sum_{j\neq i}^{m} w_j k(x_i,x_j)] \\ &\quad + 2[w'_i \mathbb{E}_{y\sim P_t}[k(x'_i,y)] - w_i \mathbb{E}_{y\sim P_t}[k(x_i,y)]] \\ &= \text{Self-term change} + 2(\text{Cross-term change}) \\ &\quad + \text{Expectation-term change}.\end{aligned}$$

Step 2: This step is to analyze weight changes. Since we replace the i-th sample, its rank would not change. Then:

$$|w'_i - w_i| = 0.$$

Step 3: Bound the change of each term.

1) Self-term change:

$$\begin{aligned}&|w'_{ii}k(x'_i,x'_i) - w_{ii}k(x_i,x_i)| \\ &= w_{ii}|k(x'_i,x'_i) - w_{ii}k(x_i,x_i)| \\ &= (1 - \frac{1}{iH_m})^2 |\mathrm{sim}^2(x'_i,P) - \mathrm{sim}^2(x_i,P)|.\end{aligned}$$

Since $\frac{1}{iH_m}$ is positive, $1 - \frac{1}{iH_m} < 1$, and $\mathrm{sim}^2(x'_i,P)$, $\mathrm{sim}^2(x_i,P)$ are in $[0,1]$, their absolute difference is at most 1:

$$|w'_{ii}k(x'_i,x'_i) - w_{ii}k(x_i,x_i)| \le 1.$$

2) Cross-term change:

$$\begin{aligned}&|w'_i \sum_{j\neq i}^{m} w_j k(x'_i,x_j) - w_i \sum_{j\neq i}^{m} w_j k(x_i,x_j)| \\ &= |w_i \sum_{j\neq i}^{m} w_j (k(x'_i,x_j) - k(x_i,x_j))| \\ &\le w_i \sum_{j\neq i}^{m} w_j \cdot 1 \\ &\le w_i \sum_{j=1}^{m} w_j.\end{aligned}$$

Since $\sum_{j=1}^{m} w_j = \sum_{j=1}^{m}(1 - \frac{1}{jH_m}) = m - \frac{1}{H_m}\sum_{j=1}^{m}\frac{1}{j} = m - 1$ and $w_i \le 1$, we have:

$$w_i \sum_{j=1}^{m} w_j \le m - 1.$$

3) Expectation term change:

$$\begin{aligned}&2|w'_i \mathbb{E}_{y\sim P_t}[k(x'_i,y)] - w_i \mathbb{E}_{y\sim P_t}[k(x_i,y)]| \\ &= 2w_i |\mathbb{E}_{y\sim P_t}[k(x'_i,y) - k(x_i,y)]|.\end{aligned}$$

To solve this, we have:

- $k(x'_i,y) = sim(x'_i,P)sim(y,P) \le 1,$
- $w_i \le 1,$
- $$\begin{aligned}&|E_{y\sim P_t}[k(x'_i,y) - k(x_i,y)]| \\ &= |E_{y\sim P_t}[sim(x'_i,P)sim(y,P) - sim(x_i,P)sim(y,P)]| \\ &= |E_{y\sim P_t}[(sim(x'_i,P) - sim(x_i,P))sim(y,P)]| \\ &= |sim(x'_i,P) - sim(x_i,P)||E_{y\sim P_t}[sim(y,P)]| \le 2.\end{aligned}$$

Therefore:

$$2w_i |\mathbb{E}_{y\sim P_t}[k(x'_i,y) - k(x_i,y)]| \le 4.$$

Step 4: Combine all the above bounds:

$$\begin{aligned}|\Delta\mathrm{MMD}^2| &= \text{Self-term change} + 2(\text{Cross-term change}) \\ &\quad + \text{Expectation-term change} \\ &\le 1 + 2(m-1) + 4 \\ &= 3 + 2m.\end{aligned}$$ □

Lemma 6 (MMD^2 Expectation Bound). *For $\mathbb{E}[MMD^2]$, it satisfies: $\mathbb{E}[MMD^2] \le A$, where $A = \frac{\pi^2}{6H_m^2} + 1 - \frac{2}{H_m^2}\sum_{i=1}^{m}\frac{H_i}{i}$.*

Proof. We analyze the expectation bound of MMD^2 in 2 steps:

Step 1: Analyze the MMD^2 expectation:

$$\begin{aligned}\mathbb{E}[\mathrm{MMD}^2] &= \mathbb{E}[\sum_{i,j=1}^{m} w_{ij}k(x_i,x_j)] + \mathbb{E}[\mathbb{E}_{y,y'\sim P_t}[k(y,y')]] \\ &\quad - 2\mathbb{E}[\sum_{i=1}^{m} w_i \mathbb{E}_{y\sim P_t}[k(x_i,y)]].\end{aligned}$$

1) First term expectation:

$$\begin{aligned}
&\mathbb{E}[\sum_{i=1}^{m}\sum_{j=1}^{m} w_i w_j k(x_i, x_j)] \\
&= \mathbb{E}[\sum_{i=1}^{m} w_i^2 k(x_i, x_i)] + \mathbb{E}[\sum_{i=1}^{m}\sum_{j=i+1}^{m} w_i w_j k(x_i, x_j)] \\
&\quad + \mathbb{E}[\sum_{j=1}^{m}\sum_{i=j+1}^{m} w_i w_j k(x_i, x_j)] \\
&= \mathbb{E}[\sum_{i=1}^{m} w_i^2 k(x_i, x_i)] + 2\mathbb{E}[\sum_{i=1}^{m}\sum_{j=i+1}^{m} w_i w_j k(x_i, x_j)].
\end{aligned}$$

We can get that:

$$\begin{aligned}
\mathbb{E}[\sum_{i=1}^{m} w_i^2 k(x_i, x_i)] &\le \mathbb{E}[\sum_{i=1}^{m}(\frac{1}{iH_m})^2] \quad \text{// } k(x_i, x_i) \le 1 \\
&= \mathbb{E}[\frac{1}{H_m^2}\sum_{i=1}^{m}\frac{1}{i^2}] \\
&\le \mathbb{E}[\frac{1}{H_m^2}\frac{\pi^2}{6}] \quad \text{// } \sum_{i=1}^{m}\frac{1}{i^2} \le \sum_{i=1}^{\infty}\frac{1}{i^2} = \frac{\pi^2}{6} \\
&= \frac{\pi^2}{6H_m^2}, \quad \text{// } H_m^2 \texttt{ is deterministic}
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}[\sum_{i=1}^{m}\sum_{j=i+1}^{m} w_i w_j k(x_i, x_j)] \\
&= \sum_{i=1}^{m}\sum_{j=i+1}^{m} w_i w_j \mathbb{E}[k(x_i, x_j)] \quad \texttt{//}w \texttt{ are deterministic} \\
&\le \sum_{i=1}^{m}\sum_{j=i+1}^{m} \frac{1}{iH_m}\cdot\frac{1}{jH_m}\cdot 1 \\
&= \frac{1}{H_m^2}\sum_{i=1}^{m}\frac{1}{i}\sum_{j=i+1}^{m}\frac{1}{j} \\
&= \frac{1}{H_m^2}\sum_{i=1}^{m}\frac{1}{i}(H_m - H_i) \\
&= \frac{1}{H_m^2}\left(\sum_{i=1}^{m}\frac{H_m}{i} - \sum_{i=1}^{m}\frac{H_i}{i}\right) \\
&= \frac{1}{H_m^2}(H_m \cdot H_m - \sum_{i=1}^{m}\frac{H_i}{i}) \\
&= 1 - \frac{1}{H_m^2}\sum_{i=1}^{m}\frac{H_i}{i}.
\end{aligned}$$

Therefore:

$$\begin{aligned}
\mathbb{E}[\sum_{i=1}^{m}\sum_{j=1}^{m} w_i w_j k(x_i, x_j)] &\le \frac{\pi^2}{6H_m^2} + 2(1 - \frac{1}{H_m^2}\sum_{i=1}^{m}\frac{H_i}{i}) \\
&= \frac{\pi^2}{6H_m^2} + 2 - \frac{2}{H_m^2}\sum_{i=1}^{m}\frac{H_i}{i}.
\end{aligned}$$

2) Second term expectation:

$$\mathbb{E}[\mathbb{E}_{y,y'\sim P_t}[k(y,y')]] = \mathbb{E}_{y,y'\sim P_t}[k(y,y')] \le 1.$$

3) Third term expectation:

$$\begin{aligned}
\mathbb{E}[\sum_{i=1}^{m} w_i \mathbb{E}_{y\sim P_t}[k(x_i, y)]] &\le \mathbb{E}[\sum_{i=1}^{m} w_i] \\
&= \sum_{i=1}^{m}\frac{1}{iH_m} \\
&= \frac{1}{H_m}\cdot H_m = 1.
\end{aligned}$$

Step 2: combine these expectation bounds:

$$\begin{aligned}
\mathbb{E}[\mathrm{MMD}^2] &\le (\frac{\pi^2}{6H_m^2} + 2 - \frac{2}{H_m^2}\sum_{i=1}^{m}\frac{H_i}{i}) + 1 - (2\times 1) \\
&= \frac{\pi^2}{6H_m^2} + 1 - \frac{2}{H_m^2}\sum_{i=1}^{m}\frac{H_i}{i}.
\end{aligned}$$

Let $A = \frac{\pi^2}{6H_m^2} + 1 - \frac{2}{H_m^2}\sum_{i=1}^{m}\frac{H_i}{i}$ and $\mathbb{E}[\mathrm{MMD}^2] \le A$. □

Now we proceed with the main proof:

Step 1: From Lemma 5 that $|\Delta\mathrm{MMD}^2| \le 3 + 2m$, we can apply McDiarmid's inequality:

$$\mathbb{P}(|\mathrm{MMD}^2 - \mathbb{E}[\mathrm{MMD}^2]| \ge \xi) \le 2exp(-2\xi^2/m(3+2m)^2),$$

where m is the buffer size. Then, with probability at least $1 - 2exp(-2\xi^2/m(3+2m)^2)$:

$$\begin{aligned}
&|\mathrm{MMD}^2 - \mathbb{E}[\mathrm{MMD}^2]| \le \xi \iff \\
&\mathbb{E}[\mathrm{MMD}^2] - \xi \le \mathrm{MMD}^2 \le \mathbb{E}[\mathrm{MMD}^2] + \xi.
\end{aligned}$$

Focus on upper bound, since MMD is non-negative:

$$\begin{aligned}
&\mathrm{MMD}^2 \le \mathbb{E}[\mathrm{MMD}^2] + \xi \implies \\
&\mathrm{MMD} \le \sqrt{\mathbb{E}[\mathrm{MMD}^2] + \xi} \le \sqrt{\mathbb{E}[\mathrm{MMD}^2]} + \sqrt{\xi} \le \sqrt{A} + \sqrt{\xi},
\end{aligned}$$

where the last inequality follows from $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$. Combining this with McDiarmid's inequality above, we have:

$$\begin{aligned}
&\mathbb{P}(|MMD^2 - E[MMD^2]| \le \xi) \ge 1 - 2exp(-2\xi^2/m(3+2m)^2) \iff \\
&\mathbb{P}(\mathrm{MMD} \le \sqrt{A} + \sqrt{\xi}) \ge 1 - 2exp(-2\xi^2/m(3+2m)^2).
\end{aligned}$$

Step 2: Let $\xi' = \sqrt{A} + \sqrt{\xi}$:

$$\begin{aligned}
\sqrt{\xi} &= \xi' - \sqrt{A}, \\
\xi &= (\xi' - \sqrt{A})^2.
\end{aligned}$$

Hence, we have:

$$\begin{aligned}
\mathbb{P}(\mathrm{MMD} \le \xi') &\ge 1 - 2exp(-2\xi^2/m(3+2m)^2) \\
&= 1 - 2exp(-2(\xi' - \sqrt{A})^4/m(3+2m)^2)
\end{aligned}$$

Step 3: Let $\delta = 2\exp(-2(\xi' - \sqrt{A})^4/m(3+2m)^2)$, then:

$$\mathbb{P}(MMD \le \varepsilon) \ge 1 - \delta$$

Step 4: Solve for ξ':

$$\delta = 2\exp(-2(\xi' - \sqrt{A})^4 / m(3+2m)^2)$$
$$\ln(\delta/2) = -2(\varepsilon - \sqrt{A})^4 / m(3+2m)^2$$
$$\varepsilon = \sqrt{A} + (-\frac{m(3+2m)^2}{2}\ln(\delta/2))^{1/4}$$

Step 5: Note that, since H_m is the harmonic series:

$$A = \frac{\pi^2}{6H_m^2} + 1 - 2\sum_{i=1}^{m}\frac{H_i}{iH_m^2} = O(1).$$

Therefore:

$$\sqrt{A} = O(1).$$

For the second term:

$$(-\frac{m(3+2m)^2}{2}\ln(\delta/2))^{1/4} = ((2m^3 + 6m^2 + 4.5m)\ln(2) - (2m^3 + 6m^2 + 4.5m)\ln(\delta))^{1/4} = O((m^3\ln(1/\delta))^{1/4}).$$

Step 6: Finally, renaming ξ' to ε, we obtain:

$$\mathbb{P}(MMD(P_M, P_t) \leq \varepsilon) \geq 1 - \delta,$$

where:

$$\varepsilon = O(1) + O((m^3\ln(1/\delta))^{1/4}) = O((m^3\ln(1/\delta))^{1/4}). \qquad \square$$

REFERENCES

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9650–9660.

[2] Xinlei Chen, Saining Xie, and Kaiming He. 2021. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9640–9649.

[3] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. 2021. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972* (2021).

[4] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115 (2015), 211–252.

[5] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. 2021. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832* (2021).