Supplementary Material for Norm-ranging LSH for Maximum Inner Product Search

1 More examples of real datasets with long tails in 2-norm distribution

In this part, we provide 3 more examples of real datasets that have a long tail in their 2-norm distributions in Figure 1. Although we give example using the ImageNet dataset in the main text, we note that the ImageNet dataset is not an outlier and there are many real datasets with long tails in their 2-norm distributions.

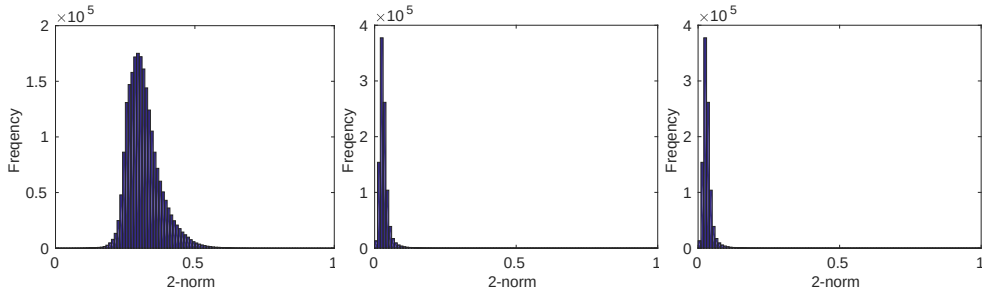

Figure 1: More datasets with long tails in their 2-norm distributions. From left to right, the datasets are glove2.2m, nuswide and msong. The maximum 2-norm is normalized to 1.

2 Adapting the proposed similarity metric to the inverted multi-index (IMI)

Our similarity metric $[I(q,b) - \frac{K}{2}]U_j$ can also be incorporated with IMI [Babenko and Lempitsky, 2012] to produce a generate-to-probe query processing scheme. Different from the query processing scheme introduced in the main text, a generate-to-probe scheme does not need to maintain the sorted structure. Moreover, a generate-to-probe scheme also does not need to sort all buckets according to their similarity metric before probing. Instead, the buckets are generated in an on-demand fashion. As only a small number of buckets are probed in most applications due to delay requirement, a generate-to-probe scheme is favorable as it can determine the buckets to probe with low complexity (without a full sorting), which helps in improving time-recall performance.

IMI is a very efficient algorithm to generate the buckets in ascending order of their distances to the query in product quantization based methods [Ge et al., 2013], which learn codebooks to quantize the dataset items. IMI first ranks the codewords in each codebook according to their distances to the query and then calculates the distance between a bucket and the query as the summation of the distances on two codebooks using $d(q,[c_1^i,c_2^j]) = d(q,c_1^i) + d(q,c_2^j)$, where the subscript is the index of codebook while the superscript is the distance ranking of a codeword in its codebook. IMI is based on the fact that $d(q,[c_1^{i+1},c_2^j]) \leq d(q,[c_1^{i+1},c_2^{j+1}])$ and $d(q,[c_1^i,c_2^{j+1}]) \leq d(q,[c_1^{i+1},c_2^{j+1}])$, and has the nice property that there are only $\mathcal{O}(\sqrt{t})$ buckets in its min-heap when t buckets are generated. For our similarity metric, we have:

$$\begin{aligned} [I(q,b) - \frac{K}{2}]U_j &\geq [I(q,b) - \frac{K}{2}]U_{j+1} \\ [I(q,b) - \frac{K}{2}]U_j &\geq [I(q,b) - 1 - \frac{K}{2}]U_j \end{aligned} \tag{1}$$

when $I(q,b) - \frac{K}{2} \geq 0$ and $U_j \geq U_{j+1}$. Note that this seems to be the opposite to the requirements of IMI, but it is not a problem as we want to generate the buckets in descending order of $[I(q,b) - \frac{K}{2}]U_j$.

Therefore, $I(q, b)$ and U_j can be treated as the two codebooks in IMI and sorted in descending order before querying, and IMI can decide the sub-dataset (according to U_j) and the bucket to probe (according to $I(q, b)$) with very low overhead. Note that when $I(q, b) < \frac{K}{2}$, U_j should be sorted in ascending order due to the flip of the sign.

3 The 2-norm distribution of the datasets used in the experiments

In the main paper, we motivate our RANGE-LSH with the long tail in the 2-norm distribution, e.g., that in the ImageNet dataset. However, the proof of Theorem 1 shows that RANGE-LSH is actually more general and can outperform SIMPLE-LSH as long as there are not too many sub-datasets having a maximum 2-norm equal to the global maximum in the entire dataset. In Figure 2, we show the 2-norm distributions of the item embeddings obtained via matrix factorization on the Netflix dataset and Yahoo! Music dataset along with the SIFT descriptors from the ImageNet dataset. The Netflix dataset and Yahoo! Music dataset do not have a long tail in 2-norm distribution and the median is close to the maximum. However, RANGE-LSH also significantly outperforms SIMPLE-LSH and L2-ALSH on these two datasets as reported in the main paper. Therefore, RANGE-LSH is robust to different 2-norm distributions and can handle a wider variety of real datasets.

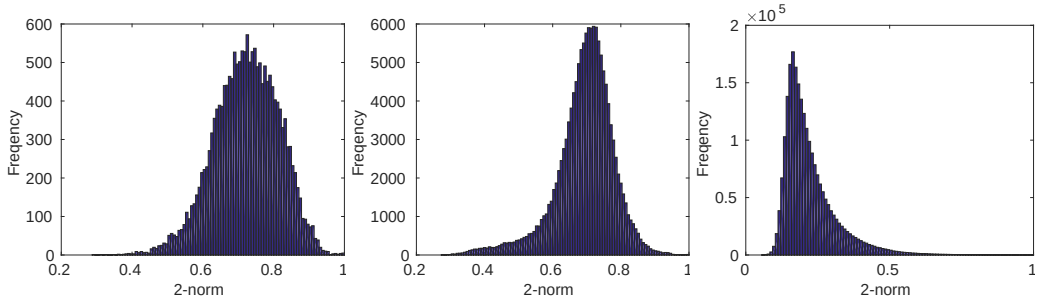

Figure 2: The 2-norm distribution of the datasets used in the experiments, the maximum 2-norm is normalized to 1. From left to right, the datasets are Netflix, Yahoo! Music and ImageNet.

4 More experimental results

In this part, we provide more experimental results on the Netflix, Yahoo! Music and ImageNet datasets under other configurations of k. Instead of time-recall curve, we report the item-recall curve as the two curves are very similar.

4.1 Top 1 MIPS

We report the performance of RANGE-LSH, SIMPLE-LSH and L2-ALSH for the top 1 MIPS in Figure 3. The results show that RANGE-LSH significantly outperforms SIMPLE-LSH and L2-ALSH for the top 1 MIPS.

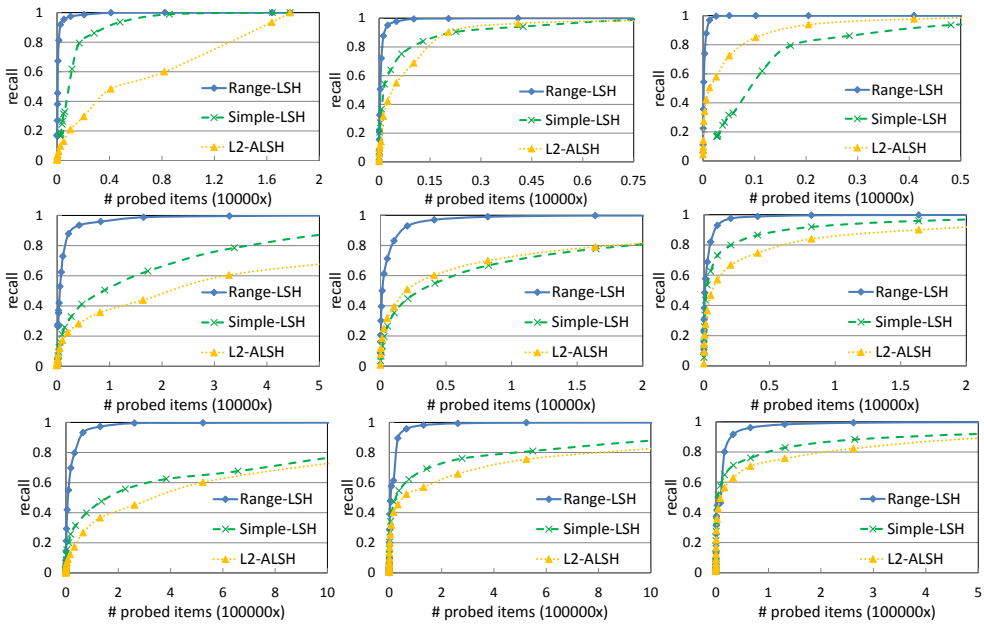

Figure 3: Recall versus the number of probed items (best viewed in colors) for the top 1 MIPS on Netflix (top row), Yahoo!Music (middle row), and ImageNet (bottom row). From left to right, the code lengths are 16, 32 and 64.

4.2 Top 10 MIPS

We report the performance of RANGE-LSH, SIMPLE-LSH and L2-ALSH for the top 10 MIPS in Figure 4. The results show that RANGE-LSH significantly outperforms SIMPLE-LSH and L2-ALSH for the top 10 MIPS.

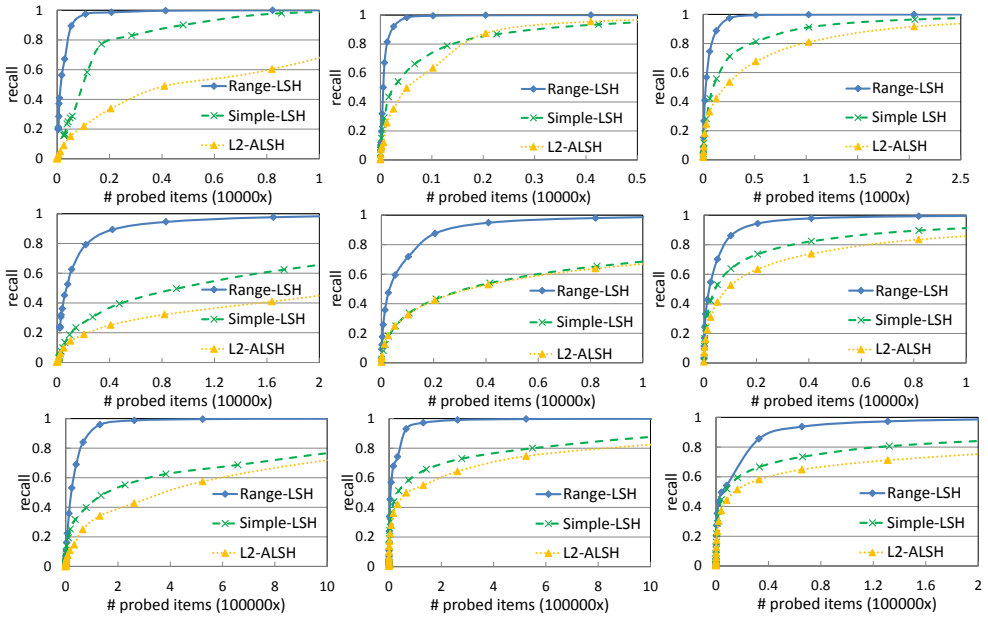

Figure 4: Recall versus the number of probed items (best viewed in colors) for the top 10 MIPS on Netflix (top row), Yahoo!Music (middle row), and ImageNet (bottom row). From left to right, the code lengths are 16, 32 and 64.

4.3 Top 20 MIPS

We report the performance of RANGE-LSH, SIMPLE-LSH and L2-ALSH for the top 20 MIPS in Figure 5. The results show that RANGE-LSH significantly outperforms SIMPLE-LSH and L2-ALSH for the top 20 MIPS.

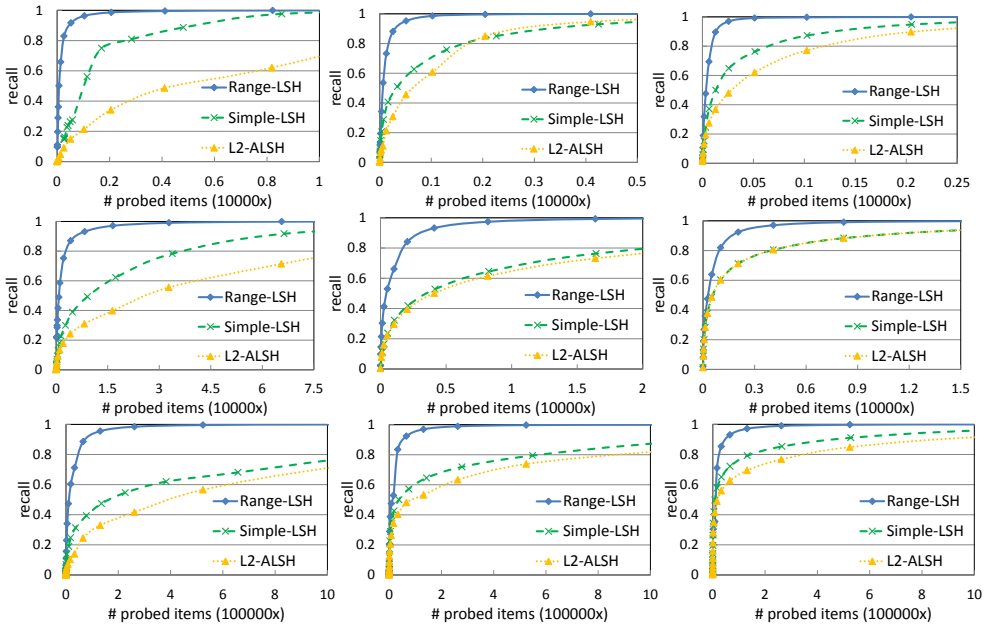

Figure 5: Recall versus the number of probed items (best viewed in colors) for the top 20 MIPS on Netflix (top row), Yahoo!Music (middle row), and ImageNet (bottom row). From left to right, the code lengths are 16, 32 and 64.

4.4 Top 50 MIPS

We report the performance of RANGE-LSH, SIMPLE-LSH and L2-ALSH for the top 50 MIPS in Figure 5. The results show that RANGE-LSH significantly outperforms SIMPLE-LSH and L2-ALSH for the top 50 MIPS.

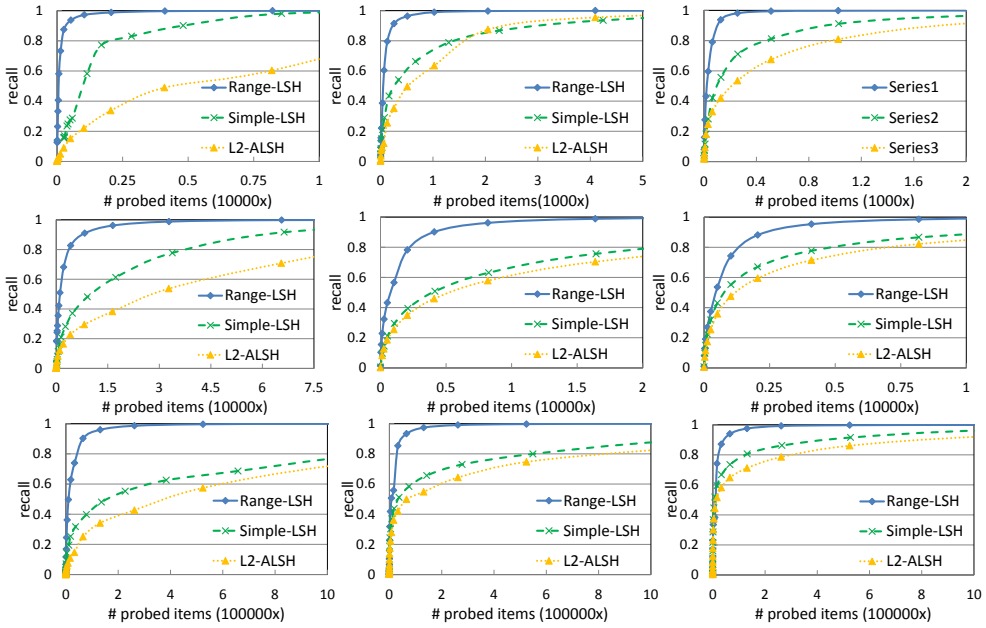

Figure 6: Recall versus the number of probed items (best viewed in colors) for the top 50 MIPS on Netflix (top row), Yahoo!Music (middle row), and ImageNet (bottom row). From left to right, the code lengths are 16, 32 and 64.

From the figures, we can conclude that the performance improvement of RANGE-LSH over SIMPLE-LSH and L2-ALSH is consistent over different configurations of k.

References

A. Babenko and V. S. Lempitsky. The inverted multi-index. In *CVPR*, pages 3069–3076, 2012.

T. Z. Ge, K. He, Q. Ke, and J. Sun. Optimized product quantization for approximate nearest neighbor search. In *ICCV*, pages 2946–2953, 2013.