Locality-Sensitive Hashing for f-Divergences and Kreĭn Kernels: Mutual Information Loss and Beyond

Lin Chen[1,2], Hossein Esfandiari[1], Thomas Fu[1], and Vahab Mirrokni[1]
[1]Google Research, [2]Yale University

Introduction

In this paper, we first study LSH schemes for f-divergences[a] between two probability distributions. We first in **??** provide a simple reduction tool for designing LSH schemes for the family of f-divergence distance functions. This proposition is not hard to prove but might be of independent interest. Next we use this tool and provide LSH schemes for two examples of f-divergence distance functions, Jensen-Shannon divergence and triangular discrimination. Interestingly our result holds for a generalized version of Jensen-Shannon divergence. We apply this tool to design and analyze an LSH scheme for the generalized Jensen-Shannon (GJS) divergence through approximation by the squared Hellinger distance. We use a similar technique to provide an LSH for triangular discrimination. Our approximation is provably lower bounded by a factor 0.69 for the Jensen-Shannon divergence and is lower bounded by a factor 0.5 for triangular discrimination. The approximation result of the generalized Jensen-Shannon divergence by the squared Hellinger requires a more involved analysis and the lower and upper bounds depend on the weight parameter. This approximation result may be of independent interest for other machine learning tasks such as approximate information-theoretic clustering [2]. Our technique may be useful for designing LSH schemes for other f-divergences.
Next, we propose a general approach to designing an LSH for Kreĭn kernels. A Kreĭn kernel is a kernel function that can be expressed as the difference of two positive definite kernels. Our approach is built upon a reduction to the problem of maximum inner product search (MIPS) [8, 6, 9].
In contrast to our LSH schemes for f-divergence functions via approximation, our approach for Kreĭn kernels involves no approximation and is theoretically *lossless*. Contrary to [5], this approach is data-independent. We exemplify our approach by designing an LSH function specifically for mutual information loss. Mutual information loss is of our particular interest due to its several important applications such as model compression [1, 3], and compression in discrete memoryless channels [4, 7, 10].

[a]The formal definition of f-divergence is presented in ??.

References

[1] MohammadHossein Bateni, Lin Chen, Hossein Esfandiari, Thomas Fu, Vahab S Mirrokni, and Afshin Rostamizadeh. "Categorical Feature Compression via Submodular Optimization". In: *ICML* (2019).

[2] Kamalika Chaudhuri and Andrew McGregor. "Finding Metric Structure in Information Theoretic Clustering.". In: *COLT*. Vol. 8. 2008, p. 10.

[3] Inderjit S Dhillon, Subramanyam Mallela, and Rahul Kumar. "A divisive information-theoretic feature clustering algorithm for text classification". In: *JMLR* 3.Mar (2003), pp. 1265–1287.

[4] Assaf Kartowsky and Ido Tal. "Greedy-Merge Degrading has Optimal Power-Law". In: *IEEE Transactions on Information Theory* 65.2 (2018), pp. 917–934.

[5] Yadong Mu and Shuicheng Yan. "Non-Metric Locality-Sensitive Hashing.". In: *AAAI*. 2010, pp. 539–544.

[6] Behnam Neyshabur and Nathan Srebro. "On Symmetric and Asymmetric LSHs for Inner Product Search". In: *ICML*. 2015, pp. 1926–1934.

[7] Yuta Sakai and Ken-ichi Iwata. "Suboptimal quantizer design for outputs of discrete memoryless channels with a finite-input alphabet". In: *ISIT*. IEEE. 2014, pp. 120–124.

[8] Anshumali Shrivastava and Ping Li. "Asymmetric LSH (ALSH) for sublinear time maximum inner product search (MIPS)". In: *NeurIPS*. 2014, pp. 2321–2329.

[9] Xiao Yan, Jinfeng Li, Xinyan Dai, Hongzhi Chen, and James Cheng. "Norm-Ranging LSH for Maximum Inner Product Search". In: *NeurIPS*. 2018, pp. 2952–2961.

[10] Jiuyang Alan Zhang and Brian M Kurkoski. "Low-complexity quantization of discrete memoryless channels". In: *ISITA*. IEEE. 2016, pp. 448–452.