Hashing: Construction of Hash Families With Smaller Description Length

Mingkun Ni, Yuanhao Zhang

{m8ni, y2384zha}@uwaterloo.ca

University of Waterloo

Waterloo, ON, Canada

August 8, 2020

Abstract

TO BE REMOVED: The following is a placeholder for abstract. After reading through some related research papers, we decide to look into one of them, titled *Ball and bins: smaller hash families and faster evaluation*. This paper introduces two new constructions that we could guarantee $O(\log n/\log\log n)$ maximum load when throw n balls into n with either a smaller description length or a faster calculation time. It is well-known that, with high probability, a $O(\log n/\log\log n)$-wise independent hash family would guarantee max load of $O(\log n/\log\log n)$. Such $O(\log n/\log\log n)$-wise independent function can be described by $O(\log^2 n/\log\log n)$ bits, which already yields a dramatic improvement over a truly random function. This paper aims to find an even smaller description length or a faster calculation time of the function while maintaining the guarantee of $O(\log n/\log\log n)$ max load. The special part of this research, specifically about the first construction, is that it constructs an innovative structure of a multi-layer random graph. With such construction of multi-layer graph, each layer is considered as a different hash process with different input and output sizes, which is the number of bins in each layer. Thus, we want to look deep in the first construction of this research and elaborate more details.

1. Introduction

A traditional analysis of randomized algorithm to map m balls into n bins independently and uniformly guarantees that each bin contains at most $O(\log n/\log\log n)$ balls with high probability, as known as the maximum load of the balls and bins problem. For a truly random hash function $h(x) : M \rightarrow N$, it would take $O(mlogn)$ space to store it. The traditional analysis with the use of truly random hash functions is impractical in various real-world applications because of the space to store the hash functions. Hence, a weaker notion of randomness, k-wise independence, is introduced to solve this issue. It is specifically well-studied in the case of mapping n balls into n bins that any $O(\log n/\log\log n)$-wise independent hash families can guarantee the maximum load of $O(\log n/\log\log n)$ with high probability.

This paper will continue to study the problems of mapping n balls into n bins with a construction of hash functions that require a smaller description length given the inspiration from the paper, titled *Ball and bins: smaller hash families and faster evaluation*. By using a $O(\log n/\log\log n)$-wise independent hash families, the hash functions can be described by $O(\log^2 n/\log\log n)$ bits, which itself yields a dramatic improvement over the description length of a truly random functions. We would like to provide an explicit family of hash functions to guarantee the same maximum load of $O(\log n/\log\log n)$ with high probability, and each hash function can be strictly described by $o(\log^2 n/\log\log n)$ bits. We provide an overview of the construction below.

1.1 Construction COPIED AND PASTED:

Our construction is based on concatenating the outputs of $O(loglogn)$ functions which are gradually more independent: each function f in our construction is described using d functions

$$f(x) = h_1(x) \circ ... \circ h_d(x),$$

where we view the output of each h_i as a binary string, and a $\circ$ denotes the concatenation operator on binary strings. The first function h_1 is only $O(1)$-wise independent, and the level of independence gradually increases to $O(logn/loglogn)$-wise independence for the last function h_d. As we increase the level of independence, we decrease the output length of the functions from $\Omega(logn)$ bits for h_1 to $O(loglogn)$ bits for h_d. We instantiate these $O(loglogn)$ functions using ϵ-biased distributions. The trade-off between the level of independence and the output length implies that each of these functions can be described using only $O(logn)$ bits and evaluated in time $O(logn)$.

1.2 Contribution:

We note that the above construction is from the paper, titled *Ball and bins: smaller hash families and faster evaluation*. This paper was written to fully understand the construction presented in Section 1.1. This construction was the study results of L. Elisa Celis, Omer Reingold, Gil Segev, and Udi Wieder. This paper will present a full and complete tour guide in understanding that this construction indeed guaranteed the same maximum load with a smaller description length. In the original paper, proofs of many lemmas and theorems are neglected. We have expanded many of lemmas and theorems given in the original paper with elaboration and proofs.

1.2 Outline:
In Section 3, we will introduce a few pieces of terminology, definitions, lemmas, and theorems that we will be using in latter sections. Section 4 is the essential part of this research paper. It will contain the formal introduction of our construction and formal proof of why this construction works. It will first give a formal description of the construction, and we will be analyzing the construction in which we will be essentially explaining the interpretation of such construction. Finally, we will prove step-by-step that the construction guarantees the description length of hashing functions. Finally, Section 5 will introduce some extensional use of this construction. It will also analyze this construction with correspondence to the trade-off it makes to obtain the smaller description length of the function.

References

[1] L. E. Celis, O. Reingold, G. Segev and U. Wieder, "Balls and Bins: Smaller Hash Families and Faster Evaluation," 2011 IEEE 52nd Annual Symposium on Foundations of Computer Science, Palm Springs, CA, 2011, pp. 599-608, doi: 10.1109/FOCS.2011.49.

[2] A. Pagh and R. Pagh. Uniform hashing in constant time and optimal space. SIAM Journal on Computing, 38(1):85–96, 2008.

[3] J. Naor and M. Naor. Small-bias probability spaces: Efficient constructions and applications. SIAM Journal on Computing, 22(4):838–856, 1993.

[4] M. Dietzfelbinger and F. Meyer auf der Heide. A new universal class of hash functions and dynamic hashing in real time. In Proceedings of the 17th International Colloquium on Automata, Languages and Programming, pages 6–19, 1990.

[5] N. Alon, M. Dietzfelbinger, P. B. Miltersen, E. Petrank, and G. Tardos. Linear hash functions. Journal of the ACM, 46(5):667–683, 1999.

[6] N. Alon, O. Goldreich, J. Hastad, and R. Peralta. Simple construction of almost kwise independent random variables. Random Structures and Algorithms, 3(3):289–304, 1992.

[7] O. Goldreich and A. Wigderson. Tiny families of functions with random properties: A quality-size trade-off for hashing. Random Structures and Algorithms, 11(4):315–343, 1997.