THEORY OF COMPUTING, Volume 9 (30), 2013, pp. 897–945
www.theoryofcomputing.org

Why Simple Hash Functions Work: Exploiting the Entropy in a Data Stream

Kai-Min Chung* Michael Mitzenmacher† Salil Vadhan‡

Received September 28, 2012; Revised December 17, 2013; Published December 31, 2013

Abstract: Hashing is fundamental to many algorithms and data structures widely used in practice. For the theoretical analysis of hashing, there have been two main approaches. First, one can assume that the hash function is truly random, mapping each data item independently and uniformly to the range. This idealized model is unrealistic because a truly random hash function requires an exponential number of bits (in the length of a data item) to describe. Alternatively, one can provide rigorous bounds on performance when explicit families of hash functions are used, such as 2-universal or $O(1)$-wise independent families. For such families, performance guarantees are often noticeably weaker than for ideal hashing.

In practice, however, it is commonly observed that simple hash functions, including 2-universal hash functions, perform as predicted by the idealized analysis for truly random

The paper is a merger of the following two conference papers: *Why Simple Hash Functions Work: Exploiting the Entropy in a Data Stream* by M. M. and S. V., Proc. 19th Ann. ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 746-755, 2008, and *Tight Bounds for Hashing Block Sources* by K-M. C., Proc. 11th International Workshop, APPROX 2008, and 12th International Workshop, RANDOM 2008, on Approximation, Randomization and Combinatorial Optimization: Algorithms and Techniques, pages 357 - 370, 2008.

*Work done when visiting U.C. Berkeley, supported by US-Israel BSF grant 2006060 and NSF grant CNS-0430336.

†Supported in part by NSF grants CCF-0915922 and IIS-0964473.

‡Work done in part while visiting U.C. Berkeley. Supported by ONR grant N00014-04-1-0478, NSF grant CCF-0133096, US-Israel BSF grant 2002246, a Guggenheim Fellowship, and the Miller Institute for Basic Research in Science.

ACM Classification: F.2

AMS Classification: 68W20, 68Q25, 68W40

Key words and phrases: algorithms, hashing, extractors, derandomization, average case, pairwise independence, Bloom filters, linear probing, balanced allocations

DOI: 10.4086/toc.2013.v009a030

hash functions. In this paper, we try to explain this phenomenon. We demonstrate that the strong performance of universal hash functions in practice can arise naturally from a combination of the randomness of the hash function and the data. Specifically, following the large body of literature on random sources and randomness extraction, we model the data as coming from a "block source," whereby each new data item has some "entropy" given the previous ones. As long as the Rényi entropy per data item is sufficiently large, it turns out that the performance when choosing a hash function from a 2-universal family is essentially the same as for a truly random hash function. We describe results for several sample applications, including linear probing, chained hashing, balanced allocations, and Bloom filters.

Towards developing our results, we prove tight bounds for hashing block sources, determining the entropy required per block for the distribution of hashed values to be close to uniformly distributed.

1 Introduction

Hashing is at the core of many fundamental algorithms and data structures, including all varieties of hash tables [20], Bloom filters and their many variants [7], summary algorithms for data streams [21], and many others. Traditionally, applications of hashing are analyzed as if the hash function is a truly random function (a. k. a. "random oracle") mapping each data item independently and uniformly to the range of the hash function. However, this idealized model is unrealistic, because a truly random function mapping $\{0,1\}^n$ to $\{0,1\}^m$ requires an exponential (in n) number of bits to describe.

For this reason, a line of theoretical work, starting with the seminal paper of Carter and Wegman [8] on universal hashing, has sought to provide rigorous bounds on performance when explicit families of hash functions are used, e. g., ones whose description and computational complexity are polynomial in n and m. While many beautiful results of this type have been obtained, they are not always as strong as we would like. In some cases, the types of hash functions analyzed can be implemented very efficiently (e. g., universal or $O(1)$-wise independent hash functions), but the performance guarantees are noticeably weaker than for ideal hashing. (A recent motivating example is the analysis of linear probing under 5-wise independence [25], discussed more below.) In other cases, the performance guarantees are (essentially) optimal, but the hash functions are more complex and expensive (e. g., with a super-linear time or space requirement). For example, if at most T items are going to be hashed, then a T-wise independent hash function will have precisely the same behavior as an ideal hash function. But a T-wise independent hash function mapping to $\{0,1\}^m$ requires at least $T \cdot m$ bits to represent, which is often too large. For some applications, it has been shown that less independence, such as $O(\log T)$-wise independence, suffices, e. g., [36, 26], but such functions are still substantially less efficient than 2-universal hash functions. A series of works [38, 24, 11] have improved the time complexity of (almost) T-wise independence to a *constant* number of word operations, but the space complexity necessarily remains at least $T \cdot m$.

In practice, however, the performance of standard universal hashing seems to match what is predicted for ideal hashing. This phenomenon was experimentally observed long ago in the setting of Bloom filters [31]; other reported examples include [6, 10, 26, 30, 32]. Thus, it does not seem truly necessary to use the more complex hash functions for which this kind of performance can be proven. We view this as a significant gap between the theory and practice of hashing.

In this paper, we aim to bridge this gap. Specifically, we suggest that it is due to the use of worst-case analysis. Indeed, in some cases, it can be proven that there exist sequences of data items for which universal hashing does not provide optimal performance. But these bad sequences may be pathological cases that are unlikely to arise in practice. That is, the strong performance of universal hash functions in practice may arise from a *combination* of the randomness of the hash function and the randomness of the data.

Of course, doing an average-case analysis, whereby each data item is independently and uniformly distributed in $\{0,1\}^n$, is also very unrealistic (not to mention that it trivializes many applications). Here we propose that an intermediate model, previously studied in the literature on randomness extraction [9], may be an appropriate data model for hashing applications. Under the assumption that the data fits this model, we show that relatively weak hash functions achieve essentially the same performance as ideal hash functions.

Our model We model the data as coming from a random source in which the data items can be far from uniform and have arbitrary correlations, provided that each (new) data item is sufficiently unpredictable given the previous items. This is formalized by Chor and Goldreich's notion of a *block source* [9],[1] where we require that the i-th item (block) X_i has at least some k bits of "entropy" conditioned on the previous items (blocks) $X_1,\dots,X_{i-1}$. There are various choices for the entropy measure that can be used here; Chor and Goldreich use *min-entropy*, but most of our results hold even for the less stringent measure of *Rényi entropy*.

We believe that a block source is a plausible model for many real-life data sources, provided the entropy k required per-block is not too large. However, in some settings, the data may have structure that violates the block-source property, in which case our results will not apply. Indeed, recent experimental and theoretical results [40, 27] have identified some natural classes of data sets (e. g., where the items are densely packed in an interval) where existing universal hash families perform poorly (e. g., when used in linear probing, as described below).

Our work is very much in the same spirit as previous works that have examined intermediate models between worst-case and average-case analysis of algorithms for other kinds of problems. Examples include the semi-random graph model of Blum and Spencer [5], and the smoothed analysis of Spielman and Teng [39]. Interestingly, Blum and Spencer's semi-random graph models are based on Santha and Vazirani's model of semi-random sources [35], which in turn were the precursor to the Chor–Goldreich model of block sources [9]. Chor and Goldreich suggest using block sources as an input model for communication complexity, but surprisingly it seems that no one has considered them as an input model for hashing applications.

Our results Our first observation is that standard results in the literature on randomness extractors already imply that universal hashing performs nearly as well ideal hashing, provided the data items have enough entropy [3, 17, 9, 44]. Specifically, if we have T data items coming from a block source

[1] Chor and Goldreich called these *probability-bounded sources,* but the term *block source* has become more common in the literature.

$(X_1,\ldots,X_T)$ where each data item has Rényi entropy at least $m+2\log(T/\varepsilon)$ and H is a random 2-universal hash function mapping to $\{0,1\}^m$, then $(H(X_1),\ldots,H(X_T))$ has statistical difference at most ε from T uniform and independent elements of $\{0,1\}^m$. Thus, any event that would occur with some probability p under ideal hashing now occurs with probability $p\pm\varepsilon$. This allows us to automatically translate existing results for ideal hashing into results for universal hashing in our model.

In our remaining results, we focus on reducing the amount of entropy required from the data items. Assuming our hash function has a description size $o(mT)$, then we must have at least $(1-o(1))m$ bits of entropy per item for the hashing to "behave like" ideal hashing (because the entropy of $(H(X_1),\ldots,H(X_T))$ is at most the sum of the entropies of H and the X_i's). The standard analysis mentioned above requires an additional $2\log(T/\varepsilon)$ bits of entropy per block. In the randomness extraction literature, the additional entropy required is typically not significant because $\log(T/\varepsilon)$ is much smaller than m. However, it can be significant in our applications. For example, a typical setting is hashing $T=\Theta(M)$ items into $2^m=M$ bins. Here $m+2\log(T/\varepsilon)\geq 3m-O(1)$ and thus the standard analysis requires 3 times more entropy than the lower bound of $(1-o(1))m$. (The bounds obtained for the specific applications mentioned below are even larger, sometimes due to the need for a subconstant $\varepsilon=o(1)$ and sometimes due to the fact that several independent hash values are needed for each item.)

We use a variety of general techniques to reduce the entropy required. These include switching from statistical difference (equivalently, ℓ_1 distance) to Rényi entropy (equivalently, ℓ_2 distance or collision probability) and/or Hellinger distance (corresponding to $\ell_{1/2}$ distance under appropriate normalization) throughout the analysis and decoupling the probability that a hash function is "good" from the uniformity of the hashed values $h(X_i)$. In particular, we reduce the required entropy, for $(H(X_1),\ldots,H(X_T))$ to be ε-close to uniform in statistical distance, from $m+2\log(T/\varepsilon)$ to $m+\log T+2\log(1/\varepsilon)$, which we show is tight. We can reduce the entropy required even further for some applications by measuring the quality of the output differently (not using statistical distance) or by using 4-wise independent hash functions (which also have very fast implementations [40]).

Applications We illustrate our approach with several specific applications. Here we informally summarize the results; definitions and discussions appear in Sections 3 and 4. In the following discussion, T is the number of data to be hashed, M is the size of the hash table, and we focus on the typical setting where $T=O(M)$.

The classic analysis of Knuth [20] gives a tight bound for the insertion time in a hash table with *linear probing* in terms of the "load" of the table (the number of items divided by the size of the table), under the assumption that an idealized, truly random hash function is used. Resolving a longstanding open problem, Pagh, Pagh, and Ružić [25] recently showed that while pairwise independence does not suffice to bound the insertion time in terms of the load alone (for worst-case data), such a bound is possible with 5-wise independent hashing. However, their bound for 5-wise independent hash functions is polynomially larger than the bound for ideal hashing. We show that 2-universal hashing actually achieves the same asymptotic performance as ideal hashing, provided that the data comes from a block source with roughly $3\log M$ bits of (Rényi) entropy per item, where M is the size of the hash table.

With standard *chained hashing*, when T items are hashed into T buckets by a single truly random hash function, the maximum load is known to be $(1+o(1))\cdot(\log T/\log\log T)$ with high probability [15, 28]. In contrast, Alon et al. [1] show that for a natural family of 2-universal hash functions, it is possible for

Type of Hash Family	Required Entropy
Linear Probing	
2-universal hashing	$3\log T$
4-wise independence	$2\log T$
Chained Hashing	
2-universal hashing	$2\log T$
Balanced Allocations with d Choices	
2-universal hashing	$(d+1)\log T$
Bloom Filters	
2-universal hashing	$3\log T$

Table 1: Each entry denotes the (Rényi) entropy required per item to ensure that the performance of the given application is "close" to the performance when using truly random hash functions. In all cases, the bounds omit additive terms that depend on how close a performance is desired, and we restrict to the (standard) case that the size of the hash table is linear in the number of items being hashed. That is, $m = \log T + O(1)$.

an adversary to choose a set of T items so that the maximum load is always $\Omega(T^{1/2})$. Our results in turn show that 2-universal hashing achieves the same performance as ideal hashing asymptotically, provided that the data comes from a block source with roughly $2\log T$ bits of (Rényi) entropy per item.

With the *balanced allocation* paradigm [2], it is known that when T items are hashed to T buckets, with each item being sequentially placed in the least loaded of d choices (e. g., $d = 2$), the maximum load is $\log\log T/\log d + O(1)$ with high probability. We show that the same result holds when the hash function is chosen from a 2-universal hash family, provided the data items come from a block source with roughly $(d+1)\log T$ bits of entropy per data item.

Bloom filters [4] are data structures for approximately storing sets in which membership tests can result in false positives with some bounded probability. We begin by showing that there is a constant gap in the false positive probability for worst-case data when $O(1)$-wise independent hash functions are used instead of truly random hash functions. On the other hand, we show that if the data comes from a block source with roughly $3\log M$ bits of (Rényi) entropy per item, where M is the size of the Bloom filter, then the false positive probability with 2-universal hashing asymptotically matches that of ideal hashing.

A summary of required (Rényi) entropy per item for the above applications can be found in Table 1.

2 Preliminaries

Notation $[N]$ denotes the set $\{1,\ldots,N\}$. All logs are base 2. For a random variable X and an event E, $X|_E$ denotes X conditioned on E. The *support* of X is $\mathrm{supp}(X) = \{x : \Pr[X = x] > 0\}$. For a real-valued function f, $\mathrm{E}[f(X)] \triangleq \sum_x \Pr[X = x] \cdot f(x)$ denotes the expectation of $f(X)$, which is also denoted by $\mathrm{E}_{x\leftarrow X}[f(x)]$. For a finite set S, U_S denotes a random variable uniformly distributed on S.

Hashing Let $\mathcal{H}$ be a family (multiset) of hash functions $h : [N] \to [M]$ and let H be uniformly distributed over $\mathcal{H}$. We use $h \leftarrow H$ to denote that h is sampled according to the distribution H. We say that $\mathcal{H}$ is a *truly random family* if $\mathcal{H}$ is the set all functions mapping $[N]$ to $[M]$, i. e., the N random variables $\{H(x)\}_{x\in[N]}$ are independent and uniformly distributed over $[M]$. For $s \in \mathbb{N}$, $\mathcal{H}$ is *s-wise independent* (a. k. a. *strongly s-universal* [42]) if for every sequence of distinct elements $x_1, \dots, x_s \in [N]$, the random variables $H(x_1), \dots, H(x_s)$ are independent and uniformly distributed over $[M]$. $\mathcal{H}$ is *s-universal* if for every sequence of distinct elements $x_1, \dots, x_s \in [N]$, we have $\Pr[H(x_1) = \cdots = H(x_s)] \le 1/M^s$. The description size of $H \in \mathcal{H}$ is the number of bits to describe H, which is simply $\log|\mathcal{H}|$. For a hash family $\mathcal{H}$ mapping $[N] \to [M]$ and $k \in \mathbb{N}$, we define $\mathcal{H}^k$ to be the family mapping $[N] \to [M]^k$ consisting of the functions of the form $h(x) = (h_1(x), \dots, h_k(x))$, where each $h_i \in \mathcal{H}$. Observe that if $\mathcal{H}$ is s-wise independent (resp., s-universal), then so is $\mathcal{H}^k$. However, description size and computation time for functions in $\mathcal{H}^k$ are k times larger than for $\mathcal{H}$.

3 Hashing worst-case data

In this section, we describe the four main hashing applications we study in this paper—linear probing, chained hashing, balanced allocations, and Bloom filters—and describe mostly known results about what can be achieved for worst-case data.

3.1 Linear probing

A hash table using linear probing stores a sequence $\bar{x} = (x_1, \dots, x_T)$ of data items from $[N]$ using M memory locations. Given a hash function $h : [N] \to [M]$, we place the data items $x_1, \dots, x_T$ sequentially as follows. The data item x_i first attempts placement at $h(x_i)$, and if this location is already filled, locations $(h(x_i) + 1) \bmod M$, $(h(x_i) + 2) \bmod M$, and so on are tried until an empty location is found. The ratio $\alpha = T/M$ is referred to as the *load* of the table. The efficiency of linear probing is measured by the insertion time for a new data item. (Other measures, such as the average time to search for items already in the table, are also often studied, and our results can be generalized to these measures as well.)

Definition 3.1. Given $h : [N] \to [M]$, a set $\bar{x} = \{x_1, \dots, x_{T-1}\}$ of data items from $[N]$ stored via linear probing using h, and an extra data item $y \notin \bar{x}$, we define the *insertion time* $\mathrm{Time}_{\mathrm{LP}}(h, \bar{x}, y)$ to be the value of j such that y is placed at location $h(y) + (j - 1) \bmod M$.

It is well known that with ideal hashing (i. e., hashing using truly random hash functions), the expected insertion time can be bounded quite tightly as a function of the load [20].

Theorem 3.2 ([20]). *Let H be a truly random hash function mapping $[N]$ to $[M]$. For every sequence $\bar{x} \in [N]^{T-1}$ and $y \notin \bar{x}$, we have*

$$\mathrm{E}[\mathrm{Time}_{\mathrm{LP}}(H, \bar{x}, y)] \le 1/(1-\alpha)^2 ,$$

where $\alpha = T/M$ is the load.

Resolving a longstanding open problem, Pagh, Pagh, and Ružić [25] recently showed that the expected lookup time could be bounded in terms of α using only $O(1)$-wise independence. Specifically, with 5-wise independence, the expected time for an insertion is $O\big(1/(1-\alpha)^{2.5}\big)$ for any sequence $\bar{x}$. On the other hand, in [25] it is also shown that there are examples of sequences $\bar{x}$ and pairwise independent hash families such that the expected time for a lookup is logarithmic in T (even though the load α is independent of T). In contrast, our work demonstrates that pairwise independent hash functions yield expected lookup times that are asymptotically the same as under the idealized analysis, assuming there is sufficient entropy in the data items themselves.

3.2 Chained hashing

A hash table using *chained hashing* stores a set $\bar{x} = \{x_1, \dots, x_T\} \in [N]^T$ in an array of M buckets. Let h be a hash function mapping $[N]$ to $[M]$. We place each item x_i in the bucket $h(x_i)$. The *load* of a bucket when the process terminates is the number of items in it.

Definition 3.3. Given $h : [N] \to [M]$ and a sequence $\bar{x} = \{x_1, \dots, x_T\}$ of data items from $[N]$ stored via chained hashing using h, we define the *maximum load* $\mathrm{MaxLoad}_{\mathrm{CH}}(\bar{x}, h)$ to be the maximum load among the buckets after all data items have been placed.

Gonnet [15] proved that when $M = T$, the expected maximum load is $\log T / \log\log T$ asymptotically. This bound also holds with high probability, as noted in [28]. More precisely, we have:

Theorem 3.4 ([15])**.** *Let H be a truly random hash function mapping $[N]$ to $[T]$. For every sequence $\bar{x} \in [N]^T$ of distinct data items, we have*

$$\mathrm{E}\left[\mathrm{MaxLoad}_{\mathrm{CH}}(\bar{x}, H)\right] = (1 + o(1)) \cdot \frac{\log T}{\log\log T}$$

and there is a function $g(T) = o(1)$ such that

$$\Pr\left[\mathrm{MaxLoad}_{\mathrm{CH}}(\bar{x}, H) \le (1 + g(T)) \cdot \frac{\log T}{\log\log T}\right] = 1 - o(1)\,,$$

where the $o(1)$ terms tend to zero as $T \to \infty$.

The calculation underlying this theorem requires that the hash function be $\Omega(\log T / \log\log T)$-wise independent. Indeed, Alon et al. [1] demonstrate that this result does not hold in general for 2-universal hash functions. For example, they show that when the hash function is chosen from the (2-universal) family of linear transformations $F^2 \to F$ for a finite field F whose size $T = |F|$ is a square, it is possible for an adversary to choose a set of T items so that the maximum load is always at least $\sqrt{T}$.

3.3 Balanced allocations

A hash table using the *balanced allocation paradigm* [2] with $d \in \mathbb{N}$ choices stores a sequence $\bar{x} = (x_1, \dots, x_T) \in [N]^T$ in an array of M buckets. Let h be a hash function mapping $[N]$ to $[M]^d \cong [M^d]$, where we view the components of $h(x)$ as $(h_1(x), \dots, h_d(x))$. We place the items sequentially by putting x_i in the least loaded of the d buckets $h_1(x_i), \dots, h_d(x_i)$ at time i (breaking ties arbitrarily), where the *load* of a bucket at time i is the number of items from $x_1, \dots, x_{i-1}$ placed in it.

Definition 3.5. Given $h : [N] \to [M]^d$, a sequence $\bar{x} = (x_1, \ldots, x_T)$ of data items from $[N]$ stored via the balanced allocation paradigm (with d choices) using h, we define the *maximum load* $\mathrm{MaxLoad}_{\mathrm{BA}}(\bar{x}, h)$ to be the maximum load among the buckets at time $T+1$.

In the case that the number T of items is the same as the number M of buckets and we do balanced allocation with $d = 1$ choice (i. e., chained hashing), it is proved [28] that the maximum load is $\Theta(\log T / \log\log T)$ with high probability. Remarkably, when the number of choices d is two or larger, the maximum load drops to be double-logarithmic.

Theorem 3.6 ([2, 41])**.** *For every $d \geq 2$ and $\gamma > 0$, there is a constant c such the following holds. Let H be a truly random hash function mapping $[N]$ to $[T]^d$. For every sequence $\bar{x} \in [N]^T$ of distinct data items, we have*

$$\Pr\left[\mathrm{MaxLoad}_{\mathrm{BA}}(\bar{x}, H) > \frac{\log\log T}{\log d} + c\right] \leq \frac{1}{T^\gamma}.$$

There are other variations on this scheme, including the asymmetric version due to Vöcking [41] and cuckoo hashing [26]; we choose to study the original setting for simplicity.

The asymmetric scheme has been recently studied under explicit functions [43], similar to those of [11]. At this point, we know of no non-trivial upper or lower bounds for the balanced allocation paradigm using families of hash functions with constant independence, although performance has been tested empirically [6]. Such bounds have been a long-standing open question in this area.

3.4 Bloom filters

A *Bloom filter* [4] represents a set $\bar{x} = \{x_1, \ldots, x_T\} \subset [N]$ using an array of M bits and ℓ hash functions. For our purposes, it will be somewhat easier to work with a *segmented Bloom filter*, where the M bits are partitioned into ℓ disjoint subarrays of size M/ℓ, with one subarray for each hash function. We assume that M/ℓ is an integer. (This splitting does not substantially change the results from the standard approach of having all hash functions map into a single array of size M.) As in the previous section, we denote the components of a hash function $h : [N] \to [M/\ell]^\ell \cong [(M/\ell)^\ell]$, as providing ℓ hash values $h(x) = (h_1(x), \ldots, h_\ell(x)) \in [M/\ell]^\ell$ in the natural way. The Bloom filter is initialized by setting all bits to 0, and then setting the $h_i(x_j)$'th bit to be 1 in the i'th subarray for all $i \in [\ell]$ and $j \in [T]$. Given a data item y, one tests for membership in $\bar{x}$ by accepting if the $h_i(y)$'th bit is 1 in the i'th subarray for all $i \in [\ell]$, and rejecting otherwise. Clearly, if $y \in \bar{x}$, then the algorithm will always accept. However, the algorithm may err if $y \notin \bar{x}$.

Definition 3.7. Given $h : [N] \to [M/\ell]^\ell$ (where ℓ divides M), a set $\bar{x} = \{x_1, \ldots, x_T\}$ of data items from $[N]$ stored in an ℓ-segment Bloom filter using h, and an additional data item $y \in [N]$, we define the *false positive predicate* $\mathrm{FalsePos}_{\mathrm{BF}}(h, \bar{x}, y)$ to be 1 if $y \notin \bar{x}$ and the membership test accepts (i. e., if $y \notin \bar{x}$ yet

$$h_i(y) \in h_i(\bar{x}) \stackrel{\text{def}}{=} \{h_i(x_j) : j = 1, \ldots, T\}$$

for all $i = 1, \ldots, \ell$).

For truly random families of hash functions, it is easy to compute the false positive probability.

Theorem 3.8 ([4]). *Let H be a truly random hash function mapping $[N]$ to $[M/\ell]^\ell$ (where ℓ divides M). For every set $\bar{x} \in [N]^T$ of data items and every $y \notin \bar{x}$, we have*

$$\Pr[\mathrm{FalsePos}_{\mathrm{BF}}(H,\bar{x},y)=1] = \left(1-\left(1-\frac{\ell}{M}\right)^T\right)^\ell \approx \left(1-e^{-\ell T/M}\right)^\ell .$$

In the typical case that $M=\Theta(T)$, the asymptotically optimal number of hash functions is $\ell = (M/T)\cdot \ln 2$, and the false positive probability is approximately $2^{-\ell}$.

We now turn to the worst-case performance of Bloom filters under $O(1)$-wise independence. It is folklore that 2-universal hash functions can be used with a constant-factor loss in space efficiency. Indeed, a union bound shows that $\Pr[h_i(y) \in h_i(\bar{x})]$ is at most $T\cdot(\ell/M)$, compared to $1-(1-\ell/M)^T$ in the case of truly random hash functions. This can be generalized to s-wise independent families using the following inclusion-exclusion formula.

Lemma 3.9. *Let $H:[N]\to[M/\ell]$ be a hash function chosen at random from a family $\mathcal{H}$ (where $\ell \mid M$). For every sequence $\bar{x}\in[N]^T$, every $y\notin\bar{x}$, and every even $s\le T$, we have*

$$\begin{aligned}\Pr[H(y)\in H(\bar{x})] &= \sum_{j=1}^{T}(-1)^{j+1}\sum_{U\subseteq T,|U|=j}\Pr[\forall k\in U: H(y)=H(x_k)]\\ &\le \sum_{j=1}^{s-1}(-1)^{j+1}\sum_{U\subseteq T,|U|=j}\Pr[\forall x_k\in U: H(y)=H(x_k)]\,.\end{aligned}$$

If $\mathcal{H}$ is an s-universal hash family, then the first $s-1$ terms of the outer sum above are the same as for a truly random function (namely $(-1)^{j+1}\cdot\binom{T}{j}(\ell/M)^j$). This gives the following bound.

Proposition 3.10. *Let s be an even constant. Let $\mathcal{H}$ be an s-universal family mapping $[N]$ to $[M/\ell]$ (where ℓ divides M), and let $H=(H_1,\dots,H_\ell)$ be a random hash function from $\mathcal{H}^\ell$. For every sequence $\bar{x}\in[N]^T$ of $T\le M/\ell$ data items and every $y\notin\bar{x}$, we have*

$$\Pr[\mathrm{FalsePos}_{\mathrm{BF}}(H,\bar{x},y)=1] \le \left(1-\left(1-\frac{\ell}{M}\right)^T+\left(\frac{\ell T}{M}\right)^s\right)^\ell .$$

Proof. By Lemma 3.9, for each $i=1,\dots,\ell$, we have:

$$\begin{aligned}\Pr[H_i(y)\in H_i(\bar{x})] &\le -\sum_{j=1}^{s-1}(-1)^j\sum_{U\subseteq T,|U|=j}\Pr[H_i(y)=H_i(x_k)\ \forall k\in U]\\ &= -\sum_{j=1}^{s-1}(-1)^j\cdot\binom{T}{j}\left(\frac{\ell}{M}\right)^j \qquad \text{(by s-universality)}\\ &= -\left[\left(1-\frac{\ell}{M}\right)^T-1-\sum_{j=s}^{T}(-1)^j\cdot\binom{T}{j}\left(\frac{\ell}{M}\right)^j\right]\\ &\le 1-\left(1-\frac{\ell}{M}\right)^T+\left(\frac{\ell T}{M}\right)^s,\end{aligned}$$

where the last inequality follows by observing that the sum is alternating and thus bounded by

$$\binom{T}{s}(\ell/M)^2 \leq (\ell T/M)^s .$$

Thus,

$$\Pr[\mathrm{FalsePos}_{\mathrm{BF}}(H,\bar{x},y)=1] = \Pr[H_i(y) \in H_i(\bar{x})\ \forall i] \leq \left(1-\left(1-\frac{\ell}{M}\right)^T + \left(\frac{\ell T}{M}\right)^s\right)^\ell . \qquad \square$$

Notice that in the common case that $\ell = \Theta(1)$ and $\ell T \leq M/2$, so that the false positive probability is constant, the above bound differs from the one for ideal hashing by an amount that shrinks rapidly with s. However, when s is constant, the difference remains an additive constant. Another way of interpreting this is that to obtain a given guarantee on the false positive probability using $O(1)$-wise independence instead of ideal hashing, one must pay a constant factor in the space for the Bloom filter. The following proposition shows that no better bound can be proved based solely on $O(1)$-wise independence.

Proposition 3.11. *Let s be an even constant. For all $N,M,\ell,T \in \mathbb{N}$ such that M/ℓ is a prime power and $T < \min\{M/\ell,N\}$, there exists an $(s+1)$-wise independent family of hash functions $\mathcal{H}$ mapping $[N]$ to $[M/\ell]$ a sequence $\bar{x} \in [N]^T$ of data items, and a $y \in [N]\setminus\bar{x}$, such that if $H=(H_1,\ldots,H_\ell)$ is a random hash function from $\mathcal{H}^\ell$, we have*

$$\Pr[\mathrm{FalsePos}_{\mathrm{BF}}(H,\bar{x},y)=1] \geq \left(1-\left(1-\frac{\ell}{M}\right)^T + \Omega\left(\left(\frac{\ell T}{M}\right)^s\right)\right)^\ell .$$

Proof. Let $q = M/\ell$, and let $\mathbb{F}$ be the finite field of size q. Associate the elements of $[M/\ell]$ with elements of $\mathbb{F}$, and similarly for the first M/ℓ elements of $[N]$. Let $\mathcal{H}_1$ consist of all polynomials $f:\mathbb{F}\to\mathbb{F}$ of degree at most s; this is an $(s+1)$-wise independent family. Let $\mathcal{H}_2$ consist of any $(s+1)$-wise independent family mapping $[N]\setminus\mathbb{F}$ to $\mathbb{F}$. For a function $f\in\mathcal{H}_1$ and $g\in\mathcal{H}_2$, define $h = f\cup g : [N]\to[M/\ell]$ by $h(x) = f(x)$ if $x\in\mathbb{F}$ and $h(x)=g(x)$ if $x\notin\mathbb{F}$. Let $\mathcal{H}$ be the family of all such functions $f\cup g$. We let $\bar{x}$ be an arbitrary sequence of T distinct elements of $\mathbb{F}$, and y any other element of $\mathbb{F}$.

Again we compute the false positive probability using Lemma 3.9. The first s terms can be computed exactly as before, using $(s+1)$-wise independence. For the terms beyond s, we observe that when $|U|\geq s$, it is the case that $h_i(y) = h_i(x_k)$ for all $k\in U$ if and only if $h_i = f\cup g$ for a *constant* polynomial f. The reason is that no nonconstant polynomial of degree at most s can take on the same value more than s times. The probability that a random polynomial of degree at most s is a constant polynomial is $1/q^s = (\ell/M)^s$.

$$\begin{aligned}
\Pr[H_i(y) \in H_i(\bar{x})] &= \sum_{j=1}^{T} (-1)^{j+1} \sum_{U \subseteq T, |U|=j} \Pr[\forall k \in U : H_i(y) = H_i(x_k)] \\
&= \left[\sum_{j=1}^{s-1} (-1)^{j+1} \cdot \binom{T}{j} \left(\frac{\ell}{M}\right)^j \right] + \left[\sum_{j=s}^{T} (-1)^{j+1} \cdot \binom{T}{j} \left(\frac{\ell}{M}\right)^s \right] \\
&= \left[1 - \left(1 - \frac{\ell}{M}\right)^T + \sum_{j=s}^{T} (-1)^j \cdot \binom{T}{j} \left(\frac{\ell}{M}\right)^j \right] + \left[\sum_{j=0}^{s-1} (-1)^j \cdot \binom{T}{j} \left(\frac{\ell}{M}\right)^s \right] \\
&\geq \left[1 - \left(1 - \frac{\ell}{M}\right)^T + \Omega\left(\frac{\ell T}{M}\right)^s \right] - O\left(T^{s-1} \cdot \left(\frac{\ell}{M}\right)^s\right) \\
&= 1 - \left(1 - \frac{\ell}{M}\right)^T + \Omega\left(\left(\frac{\ell T}{M}\right)^s\right).
\end{aligned}$$

Again, to bound the false positive probability, we simply raise the above to the ℓ-th power. □

4 Hashing block sources

4.1 Block sources

We view our data items as being random variables distributed over a finite set of size N, which we identify with $[N]$. We use the following quantities to measure the amount of randomness in a data item. For a random variable X, the *max probability* of X is $\mathrm{mp}(X) = \max_x \Pr[X = x]$. The *collision probability* of X is $\mathrm{cp}(X) = \sum_x \Pr[X = x]^2$. Measuring these quantities is equivalent to measuring the *min-entropy*

$$\mathrm{H}_\infty(X) = \min_x \log(1/\Pr[X = x]) = \log(1/\mathrm{mp}(X))$$

and the *Rényi entropy*

$$\mathrm{H}_2(X) = \log(1/\Pr[X = X']) = \log(1/\mathrm{cp}(X)),$$

where X' is an i. i. d. copy of X. If X is supported on a set of size K, then $\mathrm{mp}(X) \geq \mathrm{cp}(X) \geq 1/K$ (i. e., $\mathrm{H}_\infty(X) \leq \mathrm{H}_2(X) \leq \log K$), with equality iff X is uniform on its support. It also holds that $\mathrm{mp}(X) \leq \mathrm{cp}(X)^{1/2}$ (i. e., $\mathrm{H}_\infty(X) \geq \mathrm{H}_2(X)/2$), so min-entropy and Rényi entropy are always within a factor of 2 of each other.

We model a sequence of data items as a sequence $(X_1, \ldots, X_T)$ of correlated random variables where each item is guaranteed to have some entropy even conditioned on the previous items.

Definition 4.1 (Block Sources)**.** A sequence of random variables $(X_1, \ldots, X_T)$ taking values in $[N]^T$ is a *block source with collision probability p per block* (respectively, *max probability p per block*) if for every $i \in [T]$ and every $(x_1, \ldots, x_{i-1}) \in \mathrm{supp}(X_1, \ldots, X_{i-1})$, we have $\mathrm{cp}(X_i|_{X_1=x_1,\ldots,X_{i-1}=x_{i-1}}) \leq p$ (respectively, $\mathrm{mp}(X_i|_{X_1=x_1,\ldots,X_{i-1}=x_{i-1}}) \leq p$).

When *max probability* is used as the measure of entropy, then this is precisely the model of sources suggested in the randomness extractor literature by Chor and Goldreich [9]. We will mainly use the *collision probability* formulation as the entropy measure, since it makes our results more general.

Definition 4.2. $(X_1,\ldots,X_T)$ is a *black K-source* if it is a block source with collision probability $p=1/K$ per block.

The following simple fact relates the collision probability of a joint distribution with its marginal.

Lemma 4.3. *Let (X,Y) be a joint distribution. We have* $\mathrm{cp}(Y)\le|\mathrm{supp}(X)|\cdot\mathrm{cp}(X,Y)$.

Proof. It follows by an application of the Cauchy-Schwarz inequality.

$$\begin{aligned}|\mathrm{supp}(X)|\cdot\mathrm{cp}(X,Y) &= |\mathrm{supp}(X)|\cdot\sum_{x,y}\Pr[X=x\wedge Y=y]^2\\ &=\left(\sum_{x\in\mathrm{supp}(X)}1^2\right)\cdot\left(\sum_y\Pr[Y=y]^2\cdot\sum_{x\in\mathrm{supp}(X)}\Pr[X=x\mid Y=y]^2\right)\\ &=\sum_y\Pr[Y=y]^2\cdot\left(\sum_{x\in\mathrm{supp}(X)}\Pr[X=x\mid Y=y]^2\right)\cdot\left(\sum_{x\in\mathrm{supp}(X)}1^2\right)\\ &\ge\sum_y\Pr[Y=y]^2\cdot\left(\sum_x\Pr[X=x\mid Y=y]\right)^2\\ &=\mathrm{cp}(Y)\,.\end{aligned}$$

□

Let (X,Y) be jointly distributed random variables. We can define the conditional collision probability of X conditioning on Y as follows.

Definition 4.4. The *conditional collision probability* of X conditioning on Y is

$$\mathrm{cp}(X\mid Y)=\mathop{\mathrm{E}}_{y\leftarrow Y}[\mathrm{cp}(X|_{Y=y})]\,.$$

The *conditional Rényi entropy* is $\mathrm{H}_2(X\mid Y)=\log(1/\mathrm{cp}(X\mid Y))$.

We note that in general, the chain rule (i. e., $H(X,Y)=H(X)+H(Y\mid X)$) does not hold for Rényi entropy; that is, it is not true in general that $\mathrm{cp}(X,Y)=\mathrm{cp}(X)\cdot\mathrm{cp}(Y\mid X)$. But this fact is true when Y is uniformly distributed.

Lemma 4.5. *Let (X,Y) be jointly distributed random variables such that X is uniformly distributed. We have*

$$\mathrm{cp}(X,Y)=\mathrm{cp}(X)\cdot\mathrm{cp}(Y\mid X)\,.$$

Proof. Let (X',Y') be an i. i. d. copy of (X,Y). We have

$$\mathrm{cp}(X,Y)=\Pr[X=X'\wedge Y=Y']=\Pr[X=X']\cdot\Pr[Y=Y'\mid X=X']\,.$$

The first term is $\mathrm{cp}(X)$ by definition. For the second term, note that when X is uniformly distributed, the distribution of X remains uniform after conditioning on $X=X'$. Thus,

$$\Pr[Y=Y'\mid X=X']=\mathop{\mathrm{E}}_{x\leftarrow X}[\Pr[Y=Y'\mid X=X'=x]]=\mathop{\mathrm{E}}_{x\leftarrow X}[\mathrm{cp}(Y|_{X=x})]=\mathrm{cp}(Y\mid X)\,.$$

□

On the other hand, the following lemma says that as in the case of Shannon entropy, conditioning can only decrease the entropy.

Lemma 4.6. *Let* (X,Y,Z) *be jointly distributed random variables. We have*

$$\mathrm{cp}(X) \le \mathrm{cp}(X \mid Y) \le \mathrm{cp}(X \mid Y,Z)\,.$$

Proof. For the first inequality, we have

$$\begin{aligned}
\mathrm{cp}(X) &= \sum_x \Pr[X=x]^2 \\
&= \sum_{y,y'} \Pr[Y=y]\cdot\Pr[Y=y']\cdot\left(\sum_x \Pr[X=x \mid Y=y]\cdot\Pr[X=x \mid Y=y']\right) \\
&\le \sum_{y,y'} \Pr[Y=y]\cdot\Pr[Y=y']\cdot\left(\sum_x \Pr[X=x \mid Y=y]^2\right)^{1/2}\cdot\left(\sum_x \Pr[X=x \mid Y=y']^2\right)^{1/2} \\
&= \mathop{\mathrm{E}}_{y\leftarrow Y}\left[\mathrm{cp}(X|_{Y=y})^{1/2}\right]^2 \\
&\le \mathrm{cp}(X \mid Y)\,.
\end{aligned}$$

For the second inequality, observe that for every y in the support of Y, we have

$$\mathrm{cp}(X|_{Y=y}) \le \mathrm{cp}((X|_{Y=y}) \mid (Z|_{Y=y}))$$

from the first inequality. It follows that

$$\begin{aligned}
\mathrm{cp}(X \mid Y) &= \mathop{\mathrm{E}}_{y\leftarrow Y}[\mathrm{cp}(X|_{Y=y})] \\
&\le \mathop{\mathrm{E}}_{y\leftarrow Y}[\mathrm{cp}((X|_{Y=y}) \mid (Z|_{Y=y}))] \\
&= \mathop{\mathrm{E}}_{y\leftarrow Y}[\mathop{\mathrm{E}}_{z\leftarrow (Z|_{Y=y})}[\mathrm{cp}(X|_{Y=y,Z=z})] \\
&= \mathrm{cp}(X \mid Y,Z)\,.
\end{aligned}$$

□

4.2 Extracting randomness

A *randomness extractor* [23] can be viewed as a family of hash functions with the property that for any random variable X with enough entropy, if we pick a random hash function h from the family, then $h(X)$ is "close" to being uniformly distributed on the range of the hash function. Randomness extractors are a central object in the theory of pseudorandomness and have many applications in theoretical computer science. Thus there is a large body of work on the construction of randomness extractors. (See the surveys [22, 37].) A major emphasis in this line of work is constructing extractors where it takes extremely few (e. g., a logarithmic number of) random bits to choose a hash function from the family. This parameter is less crucial for us, so instead our emphasis is on using simple and very efficient hash functions (e. g., universal hash functions) and minimizing the amount of entropy needed from the source X. To do this, we will measure the quality of a hash family in ways that are tailored to our application, and thus we do not necessarily work with the standard definitions of extractors.

In requiring that the hashed value $h(X)$ be "close" to uniform, the standard definition of an extractor uses the most natural measure of "closeness." Specifically, for random variables X and Y, taking values in $[N]$, their *statistical difference* is defined as

$$\Delta(X,Y) = \max_{S\subseteq[N]} \big|\Pr[X\in S] - \Pr[Y\in S]\big| .$$

X and Y are called ε-*close* (resp., ε-far) if $\Delta(X,Y)\le\varepsilon$ (resp., $\Delta(X,Y)\ge\varepsilon$).

The classic Leftover Hash Lemma shows that universal hash functions are randomness extractors with respect to statistical difference.

Lemma 4.7 (The Leftover Hash Lemma [3, 17])**.** *Let $H:[N]\to[M]$ be a random hash function from a 2-universal family $\mathcal{H}$. For every random variable X taking values in $[N]$ with $\mathrm{cp}(X)\le 1/K$, we have*

$$\mathrm{cp}(H(X)\mid H)\le 1/M+1/K \quad \textit{and} \quad \mathrm{cp}(H,H(X))\le (1/|\mathcal{H}|)\cdot(1/M+1/K),$$

and thus $(H,H(X))$ is $(1/2)\cdot\sqrt{M/K}$-close to $(H,U_{[M]})$.

Notice that the above lemma says that the *joint* distribution of $(H,H(X))$ is ε-close to uniform (for $\varepsilon=(1/2)\cdot\sqrt{M/K}$); a family of hash functions achieving this property is referred to as a "strong" randomness extractor. Up to some loss in the parameter ε (which we will later want to save), this strong extraction property is equivalent to saying that with high probability over $h\leftarrow H$, the random variable $h(X)$ is close to uniform. The above formulation of the Leftover Hash Lemma, passing through collision probability, is attributed to Rackoff [18].

To prove the lemma, let (H',X') be an i. i. d. copy of (H,X). We have

$$\begin{aligned}\mathrm{cp}(H(X)\mid H) &= \mathop{\mathrm{E}}_{h\leftarrow H}[\mathrm{cp}(h(X))] = \Pr[H(X)=H(X')]\\ &\le \Pr[X=X'] + \Pr[H(X)=H(X')\mid X\ne X'] \le \frac{1}{K}+\frac{1}{M}.\end{aligned}$$

Also, since H is uniformly distributed, by Lemma 4.5,

$$\mathrm{cp}(H,H(X)) = \mathrm{cp}(H)\cdot\mathrm{cp}(H(X)\mid H)\le \frac{1}{|\mathcal{H}|}\cdot\left(\frac{1}{M}+\frac{1}{K}\right).$$

It then relies on the fact that if the collision probability of a random variable is close to that of the uniform distribution, then the random variable is close to uniform in statistical difference. This fact is captured (in a more general form) by the following lemma.

Lemma 4.8. *If X takes values in $[M]$ and $\mathrm{cp}(X)\le 1/M+1/K$, then:*

(a) For every function $f:[M]\to\mathbb{R}$,

$$|\mathrm{E}[f(X)]-\mu|\le\sigma\cdot\sqrt{M/K},$$

where μ is the expectation of $f(U_{[M]})$ and σ is its standard deviation. In particular, if f takes values in the interval $[a,b]$, then

$$|\mathrm{E}[f(X)]-\mu| \leq \sqrt{(\mu-a)\cdot(b-\mu)}\cdot\sqrt{M/K}\,.$$

(b) X is $(1/2)\cdot\sqrt{M/K}$-close to $U_{[M]}$.

Proof. By the premise of the lemma,

$$\begin{aligned}
|\mathrm{E}[f(X)]-\mu| &= \left|\sum_{x\in[M]}(f(x)-\mu)\cdot(\Pr[X=x]-1/M)\right| \\
&\leq \sqrt{\sum_{x\in[M]}(f(x)-\mu)^2}\cdot\sqrt{\sum_{x\in[M]}(\Pr[X=x]-1/M)^2} \qquad \text{(Cauchy-Schwarz)} \\
&= \sqrt{M\cdot\mathrm{Var}[f(U_{[M]})]}\cdot\sqrt{\sum_{x\in[M]}(\Pr[X=x]^2-2\Pr[X=x]/M+1/M^2)} \\
&= \sqrt{M}\cdot\sigma\cdot\sqrt{\mathrm{cp}(X)-2/M+1/M} \\
&\leq \sigma\cdot\sqrt{M/K}\,.
\end{aligned}$$

The "in particular" follows from the fact that $\sigma[Y]\leq\sqrt{(\mu-a)\cdot(b-\mu)}$ for every random variable Y taking values in $[a,b]$ and having expectation μ. (Proof: $\sigma[Y]^2=\mathrm{E}[(Y-a)^2]-(\mu-a)^2\leq(b-a)\cdot(\mu-a)-(\mu-a)^2=(\mu-a)\cdot(b-\mu)$.)

Item (b) follows by noting that the statistical difference between X and $U_{[M]}$ is the maximum of $|\mathrm{E}[f(X)]-\mathrm{E}[f(U_{[M]})]|$ over Boolean functions f, which by Item (a) is at most $\sqrt{\mu(f)\cdot(1-\mu(f))}\cdot\sqrt{M/K}\leq(1/2)\cdot\sqrt{M/K}$. □

While the bound on statistical difference given by Lemma 4.8 (b) is simpler to state, Lemma 4.8 (a) often provides substantially stronger bounds. To see this, suppose there is a bad event S of vanishing density, i. e., $|S|=o(M)$, and we would like to say that $\Pr[X\in S]=o(1)$. Using Lemma 4.8 (b), we would need $K=\omega(M)$, i. e., $\mathrm{cp}(X)=(1+o(1))/M$. But applying Lemma 4.8 (a) with f equal to the characteristic function of S, we get the desired conclusion assuming only $K=O(M)$, i. e., $\mathrm{cp}(X)=O(1/M)$. Variations of Lemma 4.8 (a) can be obtained by using Hölder's inequality instead of Cauchy-Schwarz in the proof; these variations provide bounds in terms of Rényi entropy of different orders (and different moments of $f(U_{[M]})$).

The classic approach to extracting randomness from block sources is to simply apply a (strong) randomness extractor, like the one in Lemma 4.7, to each block of the source, and uses a union bound over blocks. The bound on the distance from the uniform distribution grows linearly with the number of blocks.

Theorem 4.9 ([9, 44]). *Let $H:[N]\to[M]$ be a random hash function from a 2-universal family $\mathcal{H}$. For for every block source $(X_1,\ldots,X_T)$ with collision probability $1/K$ per block, the random variable $(H,H(X_1),\ldots,H(X_T))$ is $(T/2)\cdot\sqrt{M/K}$-close to $(H,U_{[M]^T})$.*

Thus, if we have enough entropy per block, universal hash functions behave just like ideal hash functions. How much entropy do we need? To achieve an error ε with the above theorem, we need $K \geq MT^2/(4\varepsilon^2)$. In the next section, we will explore how to improve the quadratic dependence on ε and T.

4.3 Optimized block-source extraction

In this section, we present several optimized variants of Theorem 4.9. Working with statistical distance, we shave a factor of $\sqrt{T}$ from Theorem 4.9, which translates to a factor of T saving in the needed K for the distribution of hashed value to be ε-close to uniform. Recall that a block K-source $(X_1,\dots,X_T)$ is simply a block source with collision probability $1/K$ per block.

Theorem 4.10. *Let $H : [N] \to [M]$ be a random hash function from a 2-universal family $\mathcal{H}$. For for every block K-source $(X_1,\dots,X_T)$, the random variable $(H,H(X_1),\dots,H(X_T))$ is $\sqrt{MT/K}$-close to $(H,U_{[M]^T})$.*

Recall that Theorem 4.9 is proven by passing to statistical distance first, and then measuring the growth of distance using statistical distance, which incurs a linear loss in the number of blocks T. As the linear loss in statistical distance is tight in the worst case, we instead measure the growth of distance using *Hellinger distance* (cf. [14]), and only pass to statistical distance in the end.

In addition to working with the stringent notion of statistical distance, it turns out that for several applications, it suffices to ensure that the hashed value $(H(X_1),\dots,H(X_T))$ has (or is statistically close to having) sufficiently small collision probability, say, within an $O(1)$ factor of that of the uniform distribution. We prove theorems of this form with smaller required entropy from the block source, where Theorem 4.11 uses only 2-universal hash functions and Theorem 4.12 achieves better bounds using 4-wise independent hash functions.

Theorem 4.11. *Let $H : [N] \to [M]$ be a random hash function from a 2-universal family $\mathcal{H}$. For every block K-source $(X_1,\dots,X_T)$ and every $\varepsilon > 0$, the random variable $(H,\overline{Y}) = (H,H(X_1),\dots,H(X_T))$ is ε-close to a distribution $(H,\overline{Z})$ with collision probability*

$$\mathrm{cp}(H,\overline{Z}) \leq \frac{1}{|\mathcal{H}|\cdot M^T}\left(1+\frac{M}{K\varepsilon}\right)^T .$$

In particular, if $K \geq MT/\varepsilon$, then $(H,\overline{Z})$ has collision probability at most $(1+2MT/(\varepsilon K))/(|\mathcal{H}|\cdot M^T)$.

Theorem 4.12. *Let $H : [N] \to [M]$ be a random hash function from a 4-wise independent family $\mathcal{H}$. For every block K-source $(X_1,\dots,X_T)$ and every $\varepsilon > 0$, the random variable $(H,\overline{Y}) = (H,H(X_1),\dots,H(X_T))$ is ε-close to a distribution $(H,\overline{Z})$ with collision probability*

$$\mathrm{cp}(H,\overline{Z}) \leq \frac{1}{|\mathcal{H}|\cdot M^T}\left(1+\frac{M}{K}+\sqrt{\frac{2M}{K^2\varepsilon}}\right)^T .$$

In particular, if $K \geq MT+\sqrt{2MT^2/\varepsilon}$, then $(H,\overline{Z})$ has collision probability at most $(1+\gamma)/(|\mathcal{H}|\cdot M^T)$, for $\gamma = 2\cdot(MT+\sqrt{2MT^2/\varepsilon})/K$.

Note that by Lemma 4.3, the conclusions of Theorems 4.11 and 4.12 imply that the collision probability $\mathrm{cp}(\overline{Z})$ is at most $(1+2MT/(\varepsilon K))/M^T$ and $(1+\gamma)/M^T$, for $\gamma = 2\cdot(MT+\sqrt{2MT^2/\varepsilon})/K$, respectively.

We shall prove the above three theorems in the following subsections. As the proof of Theorem 4.10 is more involved, we prove Theorems 4.11 and 4.12 first in Sections 4.3.1 and 4.3.2, and then prove Theorem 4.10 in Section 4.3.3. We further provide several lower bounds in Section 4.4 showing that the above theorems are optimal in several aspects.

4.3.1 Small collision probability using 2-universal hash functions

Let $H : [N] \to [M]$ be a random hash function from a 2-universal family $\mathcal{H}$. We first study the conditions under which $(H,\overline{Y}) = (H, H(X_1), \dots, H(X_T))$ is ε-close to having collision probability $O(1/(|\mathcal{H}|\cdot M^T))$. This requirement is less stringent than $(H,\overline{Y})$ being ε-close to uniform in statistical distance, and so requires less bits of entropy.

The starting point of our analysis is the Leftover Hash Lemma stated in Lemma 4.7 above, which asserts that if $\mathrm{cp}(X) \le 1/K$, then $\mathrm{cp}(H(X)\mid H) \le 1/M + 1/K$. Using the Leftover Hash Lemma, we show that for every hashed block Y_i, the conditional collision probability $\mathrm{cp}(Y_i \mid H, Y_{<i})$ is at most $1/M+1/K$.

Lemma 4.13. *Let $H : [N] \to [M]$ be a random hash function from a 2-universal family $\mathcal{H}$. Let $\overline{X} = (X_1,\dots,X_T)$ be a block K-source over $[N]^T$. Let $(H,\overline{Y}) = (H, H(X_1), \dots, H(X_T))$. Then $\mathrm{cp}(H) = 1/|\mathcal{H}|$ and for every $i \in [T]$, $\mathrm{cp}(Y_i \mid H, Y_{<i}) \le 1/M + 1/K$.*

Proof. $\mathrm{cp}(H) = 1/|\mathcal{H}|$ is trivial since H is uniformly chosen from $\mathcal{H}$. Fix $i \in [T]$. By the definition of block K-source, for every $x_{<i}$ in the support of $X_{<i}$, $\mathrm{cp}(X_i|_{X_{<i}=x_{<i}}) \le 1/K$. By the Leftover Hash Lemma, we have

$$\mathrm{cp}\big((Y_i|_{X_{<i}=x_{<i}})|(H|_{X_{<i}=x_{<i}})\big) \le 1/M + 1/K$$

for every $x_{<i}$. It follows that $\mathrm{cp}(Y_i \mid H, X_{<i}) \le 1/M+1/K$. Now, noting that the value of $Y_{<i}$ is determined by that of $H, X_{<i}$, we can think of $(Y_i \mid H, X_{<i})$ as Y_i first conditioning on $(H, Y_{<i})$, and then further conditioning on $X_{<i}$. By Lemma 4.6, we have

$$\mathrm{cp}(Y_i \mid H, Y_{<i}) \le \mathrm{cp}(Y_i \mid H, Y_{<i}, X_{<i}) = \mathrm{cp}(Y_i \mid H, X_{<i}) \le 1/M + 1/K\,,$$

as desired. □

Lemma 4.13 implies that $(1/T)\cdot\sum_i \mathrm{cp}(Y_i \mid H, Y_{<i}) \le 1/M+1/K$, which by definition can be rewrite as

$$\mathop{\mathrm{E}}_{(h,\overline{y})\leftarrow(H,\overline{Y})}\left[\frac{1}{T}\sum_{i=1}^{T}\mathrm{cp}(Y_i|_{(H,Y_{<i})=(h,y_{<i})})\right] \le \frac{1}{M}+\frac{1}{K}.$$

Noting that the collision probability is at least $1/M$, Markov's inequality implies that with probability at least $1-\varepsilon$ over $(h,\overline{y}) \leftarrow (H,\overline{Y})$,

$$\frac{1}{T}\sum_{i=1}^{T}\mathrm{cp}(Y_i|_{(H,Y_{<i})=(h,y_{<i})}) \le \frac{1}{M}+\frac{1}{K\varepsilon} = \frac{1}{M}\cdot\left(1+\frac{M}{K\varepsilon}\right). \tag{4.1}$$

We proceed in the following two steps to finish the proof of Theorem 4.11.

1. First, we show how to fix the ε-fraction of *bad* $(h,\overline{y})$'s. Namely, we modify at most ε-fraction of the distribution $(H,\overline{Y})$ to obtain a distribution $(H,\overline{Z}) = (H,Z_1,\ldots,Z_T)$ such that for every $(h,\overline{z}) \leftarrow (H,\overline{Z})$,
$$\frac{1}{T}\sum_{i=1}^{T} \mathrm{cp}(Z_i|_{(H,Z_{<i})=(h,z_{<i})}) \leq \frac{1}{M}\cdot\left(1+\frac{M}{K\varepsilon}\right).$$

2. Then we show that the above condition is sufficient to imply that
$$\mathrm{cp}(H,\overline{Z}) \leq (1/|\mathcal{H}|\cdot M^T)\cdot(1+(M/K\varepsilon))^T.$$

We use Lemmas 4.14 and 4.15 below to formalize the above two steps.

Lemma 4.14. *Let $(H,\overline{Y}) = (H,Y_1,\ldots,Y_T)$ be jointly distributed random variables over $\mathcal{H}\times[M]^T$ such that with probability at least $1-\varepsilon$ over $(h,\overline{y}) \leftarrow (H,\overline{Y})$, the average conditional collision probability satisfies*
$$\frac{1}{T}\cdot\sum_{i=1}^{T} \mathrm{cp}(Y_i|_{(H,Y_{<i})=(h,y_{<i})}) \leq \frac{1}{M}+\alpha.$$
Then there exists a distribution $(H,\overline{Z}) = (H,Z_1,\ldots,Z_T)$ such that $(H,\overline{Z})$ is ε-close to $(H,\overline{Y})$, and for every $(h,\overline{z}) \in \mathrm{supp}(H,\overline{Z})$, we have
$$\frac{1}{T}\cdot\sum_{i=1}^{T} \mathrm{cp}(Z_i|_{(H,Z_{<i})=(h,z_{<i})}) \leq \frac{1}{M}+\alpha.$$
Furthermore, the marginal distribution of H is unchanged.

Proof. We define the distribution $(H,\overline{Z})$ as follows.

- Sample $(h,\overline{y}) \leftarrow (H,\overline{Y})$.
- If $(1/T)\cdot\sum_{i=1}^{T}\mathrm{cp}(Y_i|_{(H,Y_{<i})=(h,y_{<i})}) \leq 1/M+\alpha$, then output $(h,\overline{y})$.
- Otherwise, let $j\in[T]$ be the least index such that
$$\frac{1}{T}\sum_{i=1}^{j}\left(\mathrm{cp}(Y_i|_{(H,Y_{<i})=(h,y_{<i})})-\frac{1}{M}\right) \leq \alpha \text{ and } \frac{1}{T}\sum_{i=1}^{j+1}\left(\mathrm{cp}(Y_i|_{(H,Y_{<i})=(h,y_{<i})})-\frac{1}{M}\right) > \alpha.$$
- Choose $w_{j+1},\ldots,w_T \leftarrow U_{[M]}$, and output $(h,y_1,\ldots,y_j,w_{j+1},\ldots,w_T)$.

It is easy to check that (i) $(H,\overline{Z})$ is well-defined, (ii) $(H,\overline{Y})$ is ε-close to $(H,\overline{Z})$, (iii) for every $(h,\overline{z}) \in (H,\overline{Z})$,
$$\frac{1}{T}\cdot\sum_{i=1}^{T} \mathrm{cp}(Z_i|_{(H,Z_{<i})=(h,z_{<i})}) \leq \frac{1}{M}+\alpha,$$
and (iv) the marginal distribution of H is unchanged. □

Lemma 4.15. *Let $\overline{X} = (X_1, \ldots, X_T)$ be a sequence of random variables such that for every $\overline{x} \in \operatorname{supp}(\overline{X})$,*

$$\frac{1}{T} \sum_{i=1}^{T} \operatorname{cp}(X_i|_{X_{<i}=x_{<i}}) \le \alpha .$$

Then the overall collision probability satisfies $\operatorname{cp}(\overline{X}) \le \alpha^T$.

Proof. By the Arithmetic Mean-Geometric Mean inequality, the inequality in the premise implies

$$\prod_{i=1}^{T} \operatorname{cp}(X_i|_{X_{<i}=x_{<i}}) \le \alpha^T .$$

Therefore, it suffices to prove

$$\operatorname{cp}(\overline{X}) \le \max_{\overline{x} \in \operatorname{supp}(\overline{X})} \prod_{i=1}^{T} \operatorname{cp}(X_i|_{X_{<i}=x_{<i}}) .$$

We prove the inequality by induction on T. The base case $T = 1$ is trivial. Suppose the inequality is true for $T - 1$. We have

$$\begin{aligned}
\operatorname{cp}(\overline{X}) &= \sum_{x_1} \Pr[X_1 = x_1]^2 \cdot \operatorname{cp}(X_2, \ldots, X_T|_{X_1=x_1}) \\
&\le \left(\sum_{x_1} \Pr[X_1 = x_1]^2 \right) \cdot \max_{x_1} \operatorname{cp}(X_2, \ldots, X_T|_{X_1=x_1}) \\
&\le \operatorname{cp}(X_1) \cdot \max_{x_1} \left(\max_{x_2, \ldots, x_T} \prod_{i=2}^{T} \operatorname{cp}(X_i|_{X_{<i}=x_{<i}}) \right) \\
&= \max_{\overline{x}} \prod_{i=1}^{T} \operatorname{cp}(X_i|_{X_{<i}=x_{<i}}) ,
\end{aligned}$$

as desired. □

We now finish the proof of Theorem 4.11. By Lemma 4.14, $(H, \overline{Y})$is ε-close to a distribution $(H, \overline{Z}) = (H, Z_1, \ldots, Z_T)$ such that for every $(h, \overline{z}) \leftarrow (H, \overline{Z})$,

$$\frac{1}{T} \sum_{i=1}^{T} \operatorname{cp}(Z_i|_{(H, Z_{<i})=(h, z_{<i})}) \le \frac{1}{M} \cdot \left(1 + \frac{M}{K\varepsilon} \right) .$$

Applying Lemma 4.15 on $(\overline{Z}|_{H=h})$ for every $h \in \operatorname{supp}(\mathcal{H})$, we have

$$\operatorname{cp}(H, \overline{Z}) = \frac{1}{|\mathcal{H}|} \cdot \operatorname*{E}_{h \leftarrow H} \left[\operatorname{cp}(\overline{Z}|_{H=h}) \right] \le \frac{1}{|\mathcal{H}| \cdot M^T} \cdot \left(1 + \frac{M}{K\varepsilon} \right)^T .$$

4.3.2 Small collision probability using 4-wise independent hash functions

Here we improve the bound in the previous section using 4-wise independent hash functions. The improvement comes from the fact that when we use 4-wise independent hash functions, we have a concentration result on the conditional collision probability for each block, via the following lemma.

Lemma 4.16. *Let $H : [N] \to [M]$ be a random hash function from a 4-wise independent family $\mathcal{H}$, and X a random variable over $[N]$ with $\mathrm{cp}(X) \le 1/K$. Then we have*

$$\operatorname*{Var}_{h \leftarrow H}[\mathrm{cp}(h(X))] \le \frac{2}{MK^2}.$$

Proof. Let $f(h) = \mathrm{cp}(h(X))$. We compute the variance $\mathrm{Var}[f(H)] = \mathrm{E}[f(H)^2] - \mathrm{E}[f(H)]^2$. Let X', Y, Y' be i. i. d. copies of X. We first recall

$$\begin{aligned}
\mathrm{E}[f(H)] &= \Pr[H(X) = H(X')] \\
&= \Pr[X = X'] + \Pr[X \neq X'] \cdot \Pr[H(X) = H(X') \mid X \neq X'] \\
&= \mathrm{cp}(X) + (1 - \mathrm{cp}(X)) \cdot \frac{1}{M},
\end{aligned}$$

and note that

$$\mathrm{E}[f(H)^2] = \operatorname*{E}_{h \leftarrow H}[\Pr[h(X) = h(X')]^2] = \Pr[H(X) = H(X') \wedge H(Y) = H(Y')].$$

The lemma then follows by the following calculation.

$$\begin{aligned}
&\Pr[H(X) = H(X') \wedge H(Y) = H(Y')] \\
&\le \Pr[X = X' \wedge Y = Y'] \\
&\quad + \Pr[(X = X' \wedge Y \neq Y') \vee (X \neq X' \wedge Y = Y')] \\
&\qquad \cdot \Pr[H(X) = H(X') \wedge H(Y) = H(Y') \mid (X = X' \wedge Y \neq Y') \vee (X \neq X' \wedge Y = Y')] \\
&\quad + \Pr[X \neq X' \wedge Y \neq Y' \wedge \{X, X'\} \neq \{Y, Y'\}] \\
&\qquad \cdot \Pr[H(X) = H(X') \wedge H(Y) = H(Y') \mid X \neq X' \wedge Y \neq Y' \wedge \{X, X'\} \neq \{Y, Y'\}] \\
&\quad + \Pr[\{X, X'\} = \{Y, Y'\} \wedge X \neq X'] \\
&\qquad \cdot \Pr[H(X) = H(X') \wedge H(Y) = H(Y') \mid \{X, X'\} = \{Y, Y'\} \wedge X \neq X'] \\
&\le \mathrm{cp}(X)^2 + 2\mathrm{cp}(X)(1 - \mathrm{cp}(X)) \cdot \frac{1}{M} + (1 - \mathrm{cp}(X))^2 \cdot \frac{1}{M^2} + 2\mathrm{cp}(X)^2 \cdot \frac{1}{M} \\
&\le \mathrm{E}[f(H)]^2 + \frac{2\mathrm{cp}(X)^2}{M}.
\end{aligned}$$

Thus, $\mathrm{Var}[f(H)] \le 2/(MK^2)$. □

We can then replace the application of Markov's inequality in the proof of Theorem 4.11 by Chebychev's inequality to get a stronger result. Formally, we prove the following lemma, which suffices to prove Theorem 4.12.

Lemma 4.17. *Let $H : [N] \to [M]$ be a random hash function from a 4-wise independent family $\mathcal{H}$. Let $\overline{X} = (X_1, \dots, X_T)$ be a block K-source over $[N]^T$. Let $(H, \overline{Y}) = (H, H(X_1), \dots, H(X_T))$. Then with probability at least $1 - \varepsilon$ over $(h, \overline{y}) \leftarrow (H, \overline{Y})$,*

$$\frac{1}{T}\sum_{i=1}^{T} \mathrm{cp}(Y_i|_{(H,Y_{<i})=(h,y_{<i})}) \le \frac{1}{M} \cdot \left(1 + \frac{M}{K} + \sqrt{\frac{2M}{K^2\varepsilon}}\right).$$

Theorem 4.12 follows immediately by composing Lemmas 4.17, 4.14, and 4.15 in the same way as the proof of Theorem 4.11.

Proof of Lemma 4.17. Recall that we have

$$\mathop{\mathrm{E}}_{(h,\overline{y})\leftarrow(H,\overline{Y})}\left[\frac{1}{T}\sum_{i=1}^{T} \mathrm{cp}(Y_i|_{(H,Y_{<i})=(h,y_{<i})})\right] \le \frac{1}{M} + \frac{1}{K}.$$

Hence, our goal is to upper bound the probability of the value

$$\frac{1}{T}\sum_{i=1}^{T} \mathrm{cp}(Y_i|_{(H,Y_{<i})=(h,y_{<i})})$$

deviating from its mean by $\sqrt{2/MK^2\varepsilon}$. Our strategy is to bound the variance of a properly defined random variable, and then apply Chebychev's inequality. By Lemma 4.16, for every $i \in [T]$, we have

$$\mathop{\mathrm{Var}}_{h\leftarrow H}\left[\mathrm{cp}(Y_i|_{(H,X_{<i})=(h,x_{<i})})\right] \le \frac{2}{MK^2}, \qquad \forall x_{<i} \in \mathrm{supp}(X_{<i}). \tag{4.2}$$

Fix $i \in [T]$, let us try to bound the variance of the i-th block. There are two issues to take care of. First, the variance we have is conditioning on $X_{<i}$ instead of $Y_{<i}$. Second, even when conditioning on $X_{<i}$, it is possible that the variance is too large:

$$\mathop{\mathrm{Var}}_{(h,\overline{x})\leftarrow(H,\overline{X})}\left[\mathrm{cp}(Y_i|_{(H,X_{<i})=(h,x_{<i})})\right] = \Omega\left(\frac{1}{K^2}\right) \gg \frac{2}{MK^2}.$$

The reason is that conditioning on different $X_{<i} = x_{<i}$, the collision probability of $(Y_i|_{X_{<i}=x_{<i}})$ may have different expectations over $h \leftarrow \mathcal{H}$. Thus, we have to subtract the mean first. Let us define

$$f(h, x_{<i}) = \mathrm{cp}(Y_i|_{(H,X_{<i})=(h,x_{<i})}) - \mathop{\mathrm{E}}_{h\leftarrow H}\left[\mathrm{cp}(Y_i|_{(H,X_{<i})=(h,x_{<i})})\right].$$

Now, for every $x_{<i} \in \mathrm{supp}(X_{<i})$, $f(H, x_{<i})$ has mean 0, and variance $\le 2/MK^2$. It follows that

$$\mathop{\mathrm{Var}}_{(h,\overline{x})\leftarrow(H,\overline{X})}\left[f(h, x_{<i})\right] \le \frac{2}{MK^2}.$$

We now deal with the issue of conditioning on $X_{<i}$ versus $Y_{<i}$. Let us define

$$g(h, y_{<i}) = \mathop{\mathrm{E}}_{x_{<i}\leftarrow(X_{<i}|_{(H,Y_{<i})=(h,y_{<i})})}\left[f(h, x_{<i})\right].$$

We claim that

$$\mathrm{cp}(Y_i|_{(H,Y_{<i})=(h,y_{<i})}) \le \frac{1}{M} + \frac{1}{K} + g(h,y_{<i}) .$$

Indeed, by Lemma 4.6 and the definition of f and g,

$$\begin{aligned}
\mathrm{cp}(Y_i|_{(H,Y_{<i})=(h,y_{<i})}) &\le \mathrm{cp}\big((Y_i|_{(H,Y_{<i})=(h,y_{<i})}) \mid (X_i|_{(H,Y_{<i})=(h,y_{<i})})\big) \\
&= \mathop{\mathrm{E}}_{x_{<i} \leftarrow (X_{<i}|_{(H,Y_{<i})=(h,y_{<i})})} \left[\mathrm{cp}(Y_i|_{(H,X_{<i})=(h,x_{<i})})\right] \\
&= \mathop{\mathrm{E}}_{x_{<i} \leftarrow (X_{<i}|_{(H,Y_{<i})=(h,y_{<i})})} \left[f(h,x_{<i}) + \mathop{\mathrm{E}}_{h \leftarrow H} \left[\mathrm{cp}(Y_i|_{(H,X_{<i})=(h,x_{<i})})\right] \right] \\
&\le g(h,y_{<i}) + \frac{1}{M} + \frac{1}{K} .
\end{aligned}$$

Also note that $g(H,Y_{<i})$ has mean 0 and small variance:

$$\mathop{\mathrm{E}}_{(h,y_{<i}) \leftarrow (H,y_{<i})} [g(h,y_{<i})] = \mathop{\mathrm{E}}_{(h,\overline{x}) \leftarrow (H,\overline{X})} [f(h,x_{<i})] = 0 ,$$

$$\mathop{\mathrm{Var}}_{(h,y_{<i}) \leftarrow (H,y_{<i})} [g(h,y_{<i})] \le \mathop{\mathrm{Var}}_{(h,\overline{x}) \leftarrow (H,\overline{X})} [f(h,x_{<i})] \le \frac{2}{MK^2} .$$

The above argument holds for every block $i \in [T]$. Taking average over blocks, we get

$$\mathop{\mathrm{E}}_{(h,\overline{y}) \leftarrow (H,\overline{Y})} \left[\frac{1}{T} \sum_{i=1}^{T} g(h,y_{<i}) \right] = 0 ,$$

$$\mathop{\mathrm{Var}}_{(h,\overline{y}) \leftarrow (H,\overline{Y})} \left[\frac{1}{T} \sum_{i=1}^{T} g(h,y_{<i}) \right] \le \frac{2}{MK^2} ,$$

and

$$\frac{1}{T} \sum_{i=1}^{T} \mathrm{cp}(Y_i|_{(H,Y_{<i})=(h,y_{<i})}) \le \frac{1}{M} + \frac{1}{K} + \left(\frac{1}{T} \sum_{i=1}^{T} g(h,y_{<i}) \right) .$$

Finally, we can apply Chebychev's inequality to random variable $(1/T) \cdot \sum_i g(H,Y_{<i})$ to get the desired result: with probability $1-\varepsilon$ over $(h,\overline{y}) \leftarrow (H,\overline{Y})$,

$$\frac{1}{T} \sum_{i=1}^{T} \mathrm{cp}(Y_i|_{(H,Y_{<i})=(h,y_{<i})}) \le \frac{1}{M} \cdot \left(1 + \frac{M}{K} + \sqrt{\frac{2M}{K^2 \varepsilon}} \right) . \qquad \square$$

4.3.3 Statistical distance to uniform distribution

Let $H : [N] \to [M]$ be a random hash function form a 2-universal family $\mathcal{H}$. Let $\overline{X} = (X_1, \ldots, X_T)$ be a block K-source over $[N]^T$. In this subsection, we study the statistical distance between the distribution of hashed sequence $(H,\overline{Y}) = (H, H(X_1), \ldots, H(X_T))$ and the uniform distribution $(H, U_{[M]^T})$.

As mentioned, the classic result in Theorem 4.9 showed that $(H,\overline{Y})$ is $(T/2)\cdot\sqrt{M/K}$-close to $(H,U_{[M]^T})$. The result was proven by passing to statistical distance first, and then measuring the growth of statistical distance using a hybrid argument, which incurs a linear loss in the number of blocks T. Since without further information, the hybrid argument is tight, to avoid linear loss in T, we have to measure the increase of distance over blocks in another way, and pass to statistical distance only in the end. It turns out that the *Hellinger distance* (cf. [14]) is a good measure for our purposes:

Definition 4.18 (Hellinger distance)**.** Let X and Y be two random variables over $[M]$. The *Hellinger distance* between X and Y is

$$d(X,Y) \stackrel{\text{def}}{=} \left(\frac{1}{2}\sum_i \left(\sqrt{\Pr[X=i]}-\sqrt{\Pr[Y=i]}\right)\right)^{1/2} = \sqrt{1-\sum_i \sqrt{\Pr[X=i]\cdot\Pr[Y=i]}}\,.$$

Like statistical distance, Hellinger distance is a distance measure for distributions, and it takes value in $[0,1]$. The following standard lemma says that the two distance measures are closely related. We remark that the lemma is tight in both directions even if Y is the uniform distribution.

Lemma 4.19 (cf. [14])**.** *Let X and Y be two random variables over $[M]$. We have*

$$d(X,Y)^2 \le \Delta(X,Y) \le \sqrt{2}\cdot d(X,Y)\,.$$

In particular, the lemma allows us to upper-bound the statistical distance by upper-bounding the Hellinger distance. Since our goal is to bound the distance to uniform, it is convenient to work with the following definition.

Definition 4.20 (Bhattacharyya Coefficient to Uniform)**.** Let X be a random variable over $[M]$. The *Bhattacharyya coefficient* of X to uniform $U_{[M]}$ is

$$C(X) \stackrel{\text{def}}{=} \frac{1}{M}\sum_i \sqrt{M\cdot\Pr[X=i]} = 1-d(X,U_{[M]})^2\,.$$

(In general, the *Bhattacharyya coefficient* of random variables X and Y is defined to be $1-d(X,Y)^2$.)

Note that $C(X,Y)=C(X)\cdot C(Y)$ when X and Y are independent random variables, so the Bhattacharyya coefficient is well-behaved with respect to products (unlike statistical distance). By Lemma 4.19, if the Bhattacharyya coefficient $C(X)$ is close to 1, then X is close to uniform in statistical distance. Recall that collision probability behaves similarly. If the collision probability $\mathrm{cp}(X)$ is close to $1/M$, then X is close to uniform. In fact, by the following normalization, we can view the collision probability as the 2-norm of X, and the Bhattacharyya coefficient as the $1/2$-norm of X.

Let $f(i)=M\cdot\Pr[X=i]$ for $i\in[M]$. In terms of $f(\cdot)$, the collision probability is $\mathrm{cp}(X)=(1/M^2)\cdot\sum_i f(i)^2$, and Lemma 4.8 says that if the "2-norm" $M\cdot\mathrm{cp}(X)=\mathrm{E}_i[f(i)^2]\le 1+\varepsilon$ where the expectation is over uniform $i\in[M]$, then $\Delta(X,U)\le\sqrt{\varepsilon}$. Similarly, Lemma 4.19 says that if the "1/2-norm" $C(X)=\mathrm{E}_i[\sqrt{f(i)}]\ge 1-\varepsilon$, then $\Delta(X,U)\le\sqrt{\varepsilon}$.

We now discuss our approach to prove Theorem 4.10. We want to show that $(H,\overline{Y})$ is close to uniform. All we know is that the conditional collision probability $\mathrm{cp}(Y_i\mid H,Y_{<i})$ is close to $1/M$ for every block. If

all blocks are independent, then the overall collision probability $\mathrm{cp}(H,\overline{Y})$ is small, and so $(H,\overline{Y})$ is close to uniform. However, this is not true without independence, since 2-norm tends to over-weight heavy elements. In contrast, the $1/2$-norm does not suffer this problem. Therefore, our approach is to show that small conditional collision probability implies large Bhattacharyya coefficient. Formally, we have the following lemma.

Lemma 4.21. *Let $\overline{X} = (X_1,\ldots,X_T)$ be jointly distributed random variables over $[M_1]\times\cdots\times[M_T]$ such that $\mathrm{cp}(X_i \mid X_{<i}) \le \alpha_i/M_i$ for every $i \in [T]$. Then the Bhattacharyya coefficient satisfies*

$$C(\overline{X}) \ge \sqrt{\frac{1}{\alpha_1 \ldots \alpha_T}}\,.$$

With this lemma, the proof of Theorem 4.10 is immediate.

Proof of Theorem 4.10. By Lemma 4.13, $\mathrm{cp}(H) = 1/|\mathcal{H}|$, and $\mathrm{cp}(Y_i \mid H, Y_{<i}) \le (1+M/K)/M$ for every $i \in [T]$. By Lemma 4.21, the Bhattacharyya coefficient satisfies $C(H,\overline{Y}) \ge (1+M/K)^{-T/2} \ge 1 - MT/2K$ (recall that $K \ge MT/\varepsilon^2$). It follows by Lemma 4.19 that

$$\Delta((H,\overline{Y}),(H,U_{[M]^T})) \le \sqrt{2}\cdot d((H,\overline{Y}),(H,U_{[M]^T})) = \sqrt{2}\cdot\sqrt{1-C(H,\overline{Y})} \le \sqrt{MT/K} \le \varepsilon\,. \qquad \square$$

We proceed to prove Lemma 4.21. The main idea is to use Hölder's inequality to relate two different norms. We recall Hölder's inequality.

Lemma 4.22 (Hölder's inequality [13])**.**

- *Let F,G be two non-negative functions from $[M]$ to $\mathbb{R}$, and $p,q > 0$ satisfying $1/p + 1/q = 1$. Let x be a uniformly random index over $[M]$. We have*

$$\mathop{\mathrm{E}}_x[F(x)\cdot G(x)] \le \mathop{\mathrm{E}}_x[F(x)^p]^{1/p} \cdot \mathop{\mathrm{E}}_x[G(x)^q]^{1/q}\,.$$

- *In general, let $F_1,\ldots,F_n$ be non-negative functions from $[M]$ to $\mathbb{R}$, and $p_1,\ldots p_n > 0$ satisfying $1/p_1 + \ldots 1/p_n = 1$. We have*

$$\mathop{\mathrm{E}}_x[F_1(x)\cdots F_n(x)] \le \mathop{\mathrm{E}}_x[F_1(x)^{p_1}]^{1/p_1} \cdots \mathop{\mathrm{E}}_x[F_n(x)^{p_n}]^{1/p_n}\,.$$

Towards proving Lemma 4.21, we first prove the following lemma that relates the collision probability and the Bhattacharyya coefficient of a random variable, which may be of independent interest. The lemma is in fact a special case of Lemma 4.21 with $T = 1$.

Lemma 4.23. *Let X be a random variable over $[M]$ with $\mathrm{cp}(X) \le \alpha/M$. Then the Bhattacharyya coefficient of X satisfies $C(X) \ge \sqrt{1/\alpha}$. That is, the Hellinger distance satisfies*

$$d(X,U_{[M]}) \le \sqrt{1-(1/(M\cdot\mathrm{cp}(X)))^{1/2}}\,.$$

Proof. We use Hölder's inequality to relate the two notions. To do so, we express them using the normalization we mentioned before. Let $f(x) = M \cdot \Pr[X = x]$ for every $x \in [M]$. In terms of $f(\cdot)$, we want to show that $\mathrm{E}_x[f(x)^2] \leq \alpha$ implies $\mathrm{E}_x[\sqrt{f(x)}] \geq \sqrt{1/\alpha}$. Note that $\mathrm{E}_x[f(x)] = 1$. We now apply Hölder's inequality with $F = f^{2/3}$, $G = f^{1/3}$, $p = 3$, and $q = 3/2$. We have

$$\mathop{\mathrm{E}}_x[f(x)] \leq \mathop{\mathrm{E}}_x[f(x)^2]^{1/3} \cdot \mathop{\mathrm{E}}_x[f(x)^{1/2}]^{2/3},$$

which implies

$$C(X) = \mathop{\mathrm{E}}_x[\sqrt{f(x)}] \geq \mathop{\mathrm{E}}_x[f(x)]^{3/2} / \mathop{\mathrm{E}}_x[f(x)^2]^{1/2} \geq \sqrt{1/\alpha}. \qquad \square$$

Proof of Lemma 4.21. We prove it by induction on T. The base case $T = 1$ is exactly Lemma 4.23 above. Suppose the lemma is true for $T - 1$, we show that it is true for T. To apply the induction hypothesis, we consider the conditional random variables $(X_2, \ldots, X_T|_{X_1=x})$ for every $x \in [M_1]$. For every $x \in [M_1]$ and $j = 2, \ldots, T$, we define

$$g_j(x) = M_j \cdot \mathrm{cp}((X_j|_{X_1=x}) \mid (X_2, \ldots, X_{j-1}|_{X_1=x}))$$

to be the "normalized" conditional collision probability. By the induction hypothesis, we have

$$C(X_2, \ldots, X_T|_{X_1=x}) \geq \sqrt{1/g_2(x) \cdots g_T(x)}$$

for every $x \in [M_1]$. Now, let $f(x) = M_1 \cdot \Pr[X_1 = x]$, and note that by definition,

$$C(\overline{X}) = \mathop{\mathrm{E}}_{x \leftarrow X_1}\left[\sqrt{f(x)} \cdot C(X_2, \ldots, X_T|_{X_1=x})\right].$$

It follows that

$$C(\overline{X}) = \mathop{\mathrm{E}}_{x \leftarrow X_1}\left[\sqrt{f(x)} \cdot C(X_2, \ldots, X_T|_{X_1=x})\right] \geq \mathop{\mathrm{E}}_{x \leftarrow X_1}\left[\sqrt{f(x)/g_2(x) \cdots g_T(x)}\right].$$

We use Hölder's inequality twice to show that

$$\mathop{\mathrm{E}}_x\left[\sqrt{f(x)/g_2(x) \cdots g_T(x)}\right] \geq \sqrt{1/\alpha_1 \cdots \alpha_T}.$$

Let us first summarize the constraints we have. By definition, we have $\mathrm{E}_x[f(x)^2] \leq \alpha_1$. Fix $j \in \{2, \ldots, T\}$. Note that

$$\mathrm{cp}(X_j \mid X_{<j}) = \mathop{\mathrm{E}}_{x \leftarrow X_1}\left[\mathrm{cp}((X_j|_{X_1=x}) \mid (X_2, \ldots, X_{j-1}|_{X_1=x}))\right] = \mathop{\mathrm{E}}_{x \leftarrow X_1}[g_j(x)/M_j] = \mathop{\mathrm{E}}_{x \leftarrow U_{[M_1]}}[f(x)g_j(x)]/M_j.$$

It follows that $\mathrm{E}_x[f(x)g_j(x)] \leq \alpha_j$ for $j = 2, \ldots, T$. Now, we apply the second version of Hölder's inequality with $F_1 = (f/g_2 \cdots g_T)^{1/2}$, $F_j = (fg_j)^{1/(T+1)}$ for $j = 2, \ldots, T$, $p_1 = 2/(T+1)$, and $p_j = 1/(T+1)$ for $j = 2, \ldots, T$, which gives

$$\mathop{\mathrm{E}}_x\left[f(x)^{T/(T+1)}\right] \leq \mathop{\mathrm{E}}_x\left[\sqrt{f(x)/g_2(x) \cdots g_T(x)}\right]^{2/(T+1)} \cdot \mathop{\mathrm{E}}_x[f(x)g_2(x)]^{1/(T+1)} \cdots \mathop{\mathrm{E}}_x[f(x)g_T(x)]^{1/(T+1)},$$

so

$$\mathop{\mathrm{E}}_x\left[\sqrt{f(x)/g_2(x)\cdots g_T(x)}\right] \geq \mathop{\mathrm{E}}_x\left[f(x)^{T/(T+1)}\right]^{(T+1)/2} \cdot \prod_{j=2}^{T} \mathop{\mathrm{E}}_x[f(x)g_j(x)]^{-1/2}$$
$$\geq \mathop{\mathrm{E}}_x\left[f(x)^{T/(T+1)}\right]^{(T+1)/2} \cdot \sqrt{1/\alpha_2\cdots\alpha_T}\,.$$

It remains to lower bound the first term by $\sqrt{1/\alpha_1}$. We apply Hölder again with $F = f^{2/(T+2)}$, $G = f^{T/(T+2)}$, $p = T+2$, and $q = (T+2)/(T+1)$, which gives

$$\mathop{\mathrm{E}}_x[f(x)] \leq \mathop{\mathrm{E}}_x\left[f(x)^2\right]^{1/(T+2)} \cdot \mathop{\mathrm{E}}_x\left[f(x)^{T/(T+1)}\right]^{(T+1)/(T+2)},$$

so

$$\mathop{\mathrm{E}}_x\left[f(x)^{T/(T+1)}\right]^{(T+1)/2} \geq \mathop{\mathrm{E}}_x[f(x)]^{(T+2)/2} / \mathop{\mathrm{E}}_x\left[f(x)^2\right]^{1/2} \geq \sqrt{1/\alpha_1}\,.$$

Combining the inequalities, we have $C(\overline{X}) \geq \sqrt{1/\alpha_1\cdots\alpha_T}$. □

4.4 Lower bounds

In this section, we provide lower bounds on the entropy needed for the data items. We show that if K is not large enough, then for every hash family $\mathcal{H}$, there exists a block K-source $\overline{X} = (X_1,\ldots,X_T)$ such that the hashed sequence $\overline{Y} = (H(X_1),\ldots,H(X_T))$ does not satisfy the desired closeness requirements to uniform (possibly in conjunction with the hash function H).

4.4.1 Lower bound for statistical distance to uniform distribution

Let us first consider the requirement for the joint distribution of $(H,\overline{Y})$ being ε-close to uniform. When there is only one block, this is exactly the requirement for a "strong extractor." The lower bound in the extractor literature, due to Radhakrishnan and Ta-Shma [29] shows that $K \geq \Omega(M/\varepsilon^2)$ is necessary, which is tight up to a constant factor. Our goal is to show that when hashing T blocks, the value of K required for each block increases by a factor of T. Intuitively, each block will produce some error (i. e., the hashed value is not close to uniform), and the overall error will accumulate over the blocks, so we need to inject more randomness per block to reduce the error. Indeed, we use this intuition to show that $K \geq \Omega(MT/\varepsilon^2)$ is necessary for the hashed sequence to be ε-close to uniform, matching the upper bound in Theorem 4.10. Note that the lower bound holds even for a truly random hash family. Formally, we prove the following theorem.

Theorem 4.24. *Let N, M, and T be positive integers and $\varepsilon \in (0,\varepsilon_0)$ a real number such that $N \geq MT/\varepsilon^2$, where $\varepsilon_0 > 0$ is a small absolute constant. Let $H : [N] \to [M]$ be a random hash function from an hash family $\mathcal{H}$. Then there exists an integer $K = \Omega(MT/\varepsilon^2)$, and a block K-source $\overline{X} = (X_1,\ldots,X_T)$ such that $(H,\overline{Y}) = (H,H(X_1),\ldots,H(X_T))$ is ε-far from uniform $(H,U_{[M]^T})$ in statistical distance.*

To prove the theorem, we need to find such an $\overline{X}$ for every hash family $\mathcal{H}$. Following the intuition, we find an X that incurs a certain error on a single block, and take $\overline{X} = (X_1, \ldots, X_T)$ to be T i. i. d. copies of X. More precisely, we first find a K-source X such that for $\Omega(1)$-fraction of hash functions $h \in \mathcal{H}$, $h(X)$ is $\Omega(\varepsilon/\sqrt{T})$-far from uniform. This step is the same as the lower bound proof for extractors [29], which uses the probabilistic method. We pick X to be a random *flat* K-source, i. e., a uniform distribution over a random set of size K, and show that X satisfies the desired property with nonzero probability. The next step is to measure how the error accumulates over independent blocks. Note that for a fixed hash function h, the hashed sequence $(h(X_1), \ldots, h(X_T))$ consists of T i. i. d. copies of $h(X)$. Reyzin [33] has shown that the statistical distance increases by a factor of $\sqrt{T}$ when we have T independent copies for small T. However, Reyzin's result only shows an increase up to distance $O(\delta^{1/3})$, where δ is the statistical distance of the original random variables. We improve Reyzin's result to show that the $\Omega(\sqrt{T})$ growth continues until the distance reaches some absolute constant. We then use it to show that the joint distribution $(H, \overline{Y})$ is far from uniform.

The following lemma corresponds to the first step.

Lemma 4.25. *Let N and M be positive integers and $\varepsilon \in (0, 1/4), \delta \in (0, 1)$ real numbers such that $N \geq M/\varepsilon^2$. Let $H : [N] \to [M]$ be a random hash function from an hash family $\mathcal{H}$. Then there exists an integer $K = \Omega(\delta^2 M/\varepsilon^2)$, and a flat K-source X over $[N]$, such that with probability at least $1 - \delta$ over $h \leftarrow H$, $h(X)$ is ε-far from uniform.*

Proof. Let $K = \lfloor \min\{\alpha \cdot M/\varepsilon^2, N/2\} \rfloor$ for some α to be determined later. Let X be a random flat K-source over $[N]$. That is, $X = U_S$ where $S \subset [N]$ is a uniformly random size K subset of $[N]$. We claim that for every hash function $h : [N] \to [M]$,

$$\Pr_S[\, h(U_S) \text{ is } \varepsilon\text{-far from uniform} \,] \geq 1 - c \cdot \sqrt{\alpha} \tag{4.3}$$

for some absolute constant c. Let us assume (4.3), and prove the lemma first. Since the claim holds for every hash function h,

$$\Pr_{h \leftarrow H, S}[\, h(U_S) \text{ is } \varepsilon\text{-far from uniform} \,] \geq 1 - c \cdot \sqrt{\alpha}\,.$$

Thus, there exists a flat K-source U_S such that

$$\Pr_{h \leftarrow H}[\, h(U_S) \text{ is } \varepsilon\text{-far from uniform} \,] \geq 1 - c \cdot \sqrt{\alpha}\,.$$

The lemma follows by setting $\alpha = \min\{\delta^2/c^2, 1/32\}$. We proceed to prove (4.3). It suffices to show that for every $y \in [M]$, with probability at least $1 - c' \cdot \sqrt{\alpha}$ over random U_S, the deviation of $\Pr[h(U_S) = y]$ from $1/M$ is at least $4\varepsilon/M$, where c' is another absolute constant. That is,

$$\Pr_S\left[\left|\Pr[h(U_S) = y] - \frac{1}{M}\right| \geq \frac{4\varepsilon}{M}\right] \geq 1 - c' \cdot \sqrt{\alpha}\,. \tag{4.4}$$

Again, let us see why (4.4) is sufficient to prove (4.3) first. Let us call $y \in [M]$ is *bad* for S if

$$\left|\Pr[h(U_S) = y] - \frac{1}{M}\right| \geq \frac{4\varepsilon}{M}\,.$$

Since inequality (4.4) holds for every $y \in [M]$, we have

$$\Pr_{S,y}[y \text{ is bad for } S] \geq 1 - c' \cdot \sqrt{\alpha}\,,$$

where y is uniformly random over $[M]$. It follows that

$$\Pr_{S}[\text{at least } 1/2\text{-fraction of } y \text{ are bad for } S] \geq 1 - 2c' \cdot \sqrt{\alpha}\,.$$

Observe that if at least $1/2$-fraction of y are bad for S, then $\Delta(h(X), U_{[M]}) \geq \varepsilon$. Inequality (4.3) follows by setting $c = 2c'$.

It remains to prove (4.4). Let $T = h^{-1}(y)$. We have $\Pr_S[h(U_S) = y] = |S \cap T|/|S|$. Thus, recall that $K \leq \alpha M/\varepsilon^2$, (4.4) follows from inequality

$$\Pr_{S}\left[\left|\,|S \cap T| - \frac{K}{M}\right| < \frac{4K\varepsilon}{M}\right] \leq c' \cdot \sqrt{\frac{K\varepsilon^2}{M}}\,,$$

which follows by the claim below by setting $L = K/M$, and $\beta = 4\varepsilon\sqrt{K/M}$. (Working out the parameters, we have $c' = 4c''$, $\varepsilon < 1/4$ implies $\beta < \sqrt{L}$, and $\alpha \leq 1/32$ implies $\beta < 1$.)

Claim 4.26. *Let $N, K > 1$ be positive integers such that $N > 2K$, and $L \in [0, K/2]$, $\beta \in (0, \min\{1, \sqrt{L}\})$ real numbers. Let $S \subset [N]$ be a random subset of size K, and $T \subset [N]$ be a fixed subset of arbitrary size. We have*

$$\Pr_{S}\left[\big|\,|S \cap T| - L\big| \leq \beta\sqrt{L}\right] \leq c'' \cdot \beta\,,$$

for some absolute constant c''.

Intuitively, the probability in the claim is maximized when the set T has size NL/K so that $L = \mathrm{E}_S[|S \cap T|]$, and the claim follows by observing that in this case, the distribution has deviation $\Theta(\sqrt{L})$, and each possible outcome has probability $O(\sqrt{1/L})$. The formal proof of the claim is in Appendix A and is proved by expressing the probability in terms of binomial coefficients, and estimating them using Stirling formula. □

The next step is to measure the increase of statistical distance over independent random variables.

Lemma 4.27. *Let X and Y be random variables over $[M]$ such that $\Delta(X, Y) \geq \varepsilon$. Let $\overline{X} = (X_1, \ldots, X_T)$ be T i. i. d. copies of X, and let $\overline{Y} = (Y_1, \ldots, Y_T)$ be T i. i. d. copies of Y. We have*

$$\Delta(\overline{X}, \overline{Y}) \geq \min\{\varepsilon_0, c\sqrt{T} \cdot \varepsilon\}\,,$$

where ε_0, c are absolute constants.

Proof. We prove the lemma by the following two claims. The first claim reduces the lemma to the special case that X is a Bernoulli random variable with bias $\Omega(\varepsilon)$, and Y is a uniform coin. The second claim proves the special case.

For our first claim, we make use of the notion of a randomized function. Recall that with a randomized function, the output $f(x)$ for an input x is a random variable that may take on different values each time $f(x)$ is evaluated.

Claim 4.28. *Let X and Y be random variables over $[M]$ such that $\Delta(X,Y)=\varepsilon$. Then there exists a randomized function $f:[M]\to\{0,1\}$ such that $f(Y)=U_{\{0,1\}}$, and $\Delta(f(X),f(Y))\geq\varepsilon/2$.*

Proof. By the definition, there exists a set $T\subset[M]$ such that

$$\left|\Pr[X\in T]-\Pr[Y\in T]\right|=\varepsilon\,.$$

Without loss of generality, we can assume that $\Pr[Y\in T]=p\leq 1/2$ (because we can take the complement of T). Let $g:[M]\to\{0,1\}$ be the indicator function of T, so we have $\Pr_Y[g(Y)=1]=p$. For every $x\in[M]$, we define $f(x)=g(x)\vee b$, where b is a biased coin with $\Pr[b=0]=1/(2(1-p))$. The claim follows by observing that

$$\Pr[f(Y)=0]=\Pr[g(Y)=0\wedge b=0]=(1-p)\cdot 1/(2(1-p))=1/2\,,$$

and

$$\Delta(f(X),f(Y))\geq\Delta(X,Y)\cdot\Pr[b=0]\geq\varepsilon/2\,.$$

□

Claim 4.29. *Let X be a Bernoulli random variable over $\{0,1\}$ such that $\Pr[X=0]=1/2-\varepsilon$. Let $\overline{X}=(X_1,\dots,X_T)$ be T independent copies of X. Then*

$$\Delta(\overline{X},U_{\{0,1\}^T})\geq\min\{\varepsilon_0,c\sqrt{T}\varepsilon\}\,,$$

where ε_0,c are absolute constants independent of ε and T.

Proof. For $\bar{x}\in\{0,1\}^T$, let the *weight* $\mathrm{wt}(\bar{x})$ of $\bar{x}$ to be the number of 1's in $\bar{x}$. Let

$$S=\left\{x\in\{0,1\}^T:\mathrm{wt}(x)\leq\frac{T}{2}-\sqrt{T}\right\}$$

be the subset of $\{0,1\}^T$ with small weight. (This choice of S is the main source of improvement in our proof compared to that of Reyzin [33], who instead considers the set of all x with weight at most $T/2$.) For every $\bar{x}\in S$, we have

$$\Pr[\overline{X}=\bar{x}]\leq\frac{1}{2^T}\cdot(1-\varepsilon)^{T/2+\sqrt{T}}\cdot(1+\varepsilon)^{T/2-\sqrt{T}}\leq\left(1-\min\left\{\frac{\sqrt{T}\cdot\varepsilon}{2},\frac{1}{2}\right\}\right)\cdot\Pr[U_{\{0,1\}^T}=\bar{x}]\,.$$

By standard results on large deviation, we have

$$\Pr[U_{\{0,1\}^T}\in S]\geq\Omega(1)\,.$$

Combining the above two inequalities, we get

$$\begin{aligned}\Delta(\overline{X},U_{\{0,1\}^T})&\geq\Pr[U_{\{0,1\}^T}\in S]-\Pr[\overline{X}\in S]\\&\geq\left(1-\left(1-\min\left\{\frac{\sqrt{T}\cdot\varepsilon}{2},\frac{1}{2}\right\}\right)\right)\cdot\Pr[U_{\{0,1\}^T}\in S]\\&\geq\min\left\{\frac{\sqrt{T}\cdot\varepsilon}{2},\frac{1}{2}\right\}\cdot\Omega(1)\\&=\min\{c\sqrt{T}\varepsilon,\varepsilon_0\}\end{aligned}$$

for some absolute constants c,ε_0, which completes the proof.

□

Note that applying the same randomized function f on two random variables X and Y cannot increase the statistical distance. I. e., $\Delta(f(X), f(Y)) \leq \Delta(X,Y)$. The lemma follows immediately by the above two claims:

$$\begin{aligned}\Delta(\overline{X},\overline{Y}) &\geq \Delta(((f_1(X_1),\ldots,f_T(X_T)),((f_1(Y_1),\ldots,f_T(Y_T)) \\ &\geq \min\{\varepsilon_0, c\sqrt{T}\varepsilon\}\end{aligned}$$

where $f_1,\ldots,f_T$ are independent copies of the randomized function defined in Claim 4.28, and ε_0, c are absolute constants from Claim 4.29. □

Proof of Theorem 4.24. The absolute constant ε_0 in the theorem is a half of the ε_0 in Lemma 4.27. By Lemma 4.25 there is a flat K-source such that for $1/2$-fraction of hash functions $h \in \mathcal{H}$, $h(X)$ is $(2\varepsilon/c\sqrt{T})$-far from uniform, for $K = \Omega((1/2)^2 M/(2\varepsilon/c\sqrt{T})^2) = \Omega(MT/\varepsilon^2)$. We set $\overline{X} = (X_1,\ldots,X_T)$ to be T independent copies of X. Consider a hash function h such that $h(X)$ is $(2\varepsilon/c\sqrt{T})$-far from uniform. By Lemma 4.27, $(h(X_1),\ldots,h(X_T))$ is 2ε-far from uniform. Note that this holds for $1/2$-fraction of hash functions h. It follows that

$$\Delta((H,\overline{Y}),(H,U_{[M]})) = \mathop{\mathrm{E}}_{h\leftarrow H}\left[\Delta((h(X_1),\ldots,h(X_T),U_{[M]^T})\right] \geq \frac{1}{2}\cdot 2\varepsilon = \varepsilon\,. \qquad \square$$

4.4.2 Lower bound for small collision probability

In this subsection, we prove lower bounds on the entropy needed per item to ensure that the sequence of hashed values is close to having small collision probability. Since this requirement is less stringent than being close to uniform, less entropy is needed from the source. The interesting setting in applications is to require the hashed sequence $(H,\overline{Y}) = (H,H(X_1),\ldots,H(X_T))$ to be ε-close to having collision probability $O(1/(|\mathcal{H}|\cdot M^T))$. Recall that in this setting, instead of requiring $K \geq MT/\varepsilon^2$, $K \geq \Omega(MT/\varepsilon)$ is sufficient for 2-universal hash functions (Theorem 4.11), and $K \geq \Omega(MT + T\sqrt{M/\varepsilon})$ is sufficient for 4-wise independent hash functions (Theorem 4.12). The main improvement from 2-universal to 4-wise independent hashing is the better dependency on ε. Indeed, it can be shown that if we use truly random hash functions, we can reduce the dependency on ε to $\log(1/\varepsilon)$. Since we are now proving lower bounds for arbitrary hash families, we focus on the dependency on M and T. Specifically, our goal is to show that $K = \Omega(MT)$ is necessary. More precisely, we show that when $K \ll MT$, it is possible for the hashed sequence $(H,\overline{Y})$ to be .99-far from any distribution that has collision probability less than $100/(|\mathcal{H}|\cdot M^T)$.

We use the same strategy as in the previous subsection to prove this lower bound. Fixing a hash family $\mathcal{H}$, we take T independent copies $(X_1,\ldots,X_T)$ of the worst-case X found in Lemma 4.25, and show that $(H,H(X_1),\ldots,H(X_T))$ is far from having small collision probability. The new ingredient is to show that when we have T independent copies, and $K \ll MT$, then $(h(X_1),\ldots,h(X_T))$ is very far from uniform (say, 0.99-far) for many $h \in \mathcal{H}$. We then argue that in this case, we can not reduce the collision probability of $(h(X_1),\ldots,h(X_T))$ by changing a small fraction of distribution, which implies the overall distribution $(H,\overline{Y})$ is far from any distribution $(H',\overline{Z})$ with small collision probability. Formally, we prove the following theorem.

Theorem 4.30. *Let N, M, and T be positive integers such that $N \geq MT$. Let $\delta \in (0,1)$ and $\alpha > 1$ be real numbers such that $\alpha < \delta^3 \cdot e^{T/32}/128$. Let $H : [N] \to [M]$ be a random hash function from a hash family $\mathcal{H}$. There exists an integer $K = \Omega(\delta^2 MT/\log(\alpha/\delta))$, and a block K-source $\overline{X} = (X_1, \ldots, X_T)$ such that $(H, \overline{Y}) = (H, H(X_1), \ldots, H(X_T))$ is $(1-\delta)$-far from any distribution $(H', \overline{Z})$ with $\mathrm{cp}(H', \overline{Z}) \leq \alpha/(|\mathcal{H}| \cdot M^T)$.*

Think of α and δ as constants. Then the theorem says that $K = \Omega(MT)$ is necessary for the hashed sequence $(H, H(X_1), \ldots, H(X_T))$ to be close to having small collision probability, matching the upper bound in Theorem 4.11. In the previous proof, we used Lemma 4.27 to measure the increase of distance over blocks. However, the lemma can only measure the progress up to some small constant. It is known that if the number of copies T is larger then $\Omega(1/\varepsilon^2)$, where ε is the statistical distance of original copy, then the statistical distance goes to 1 exponentially fast. Formally, we use the following lemma.

Lemma 4.31 ([34])**.** *Let X and Y be random variables over $[M]$ such that $\Delta(X,Y) \geq \varepsilon$. Let $\overline{X} = (X_1, \ldots, X_T)$ be T i. i. d. copies of X, and let $\overline{Y} = (Y_1, \ldots, Y_T)$ be T i. i. d. copies of Y. We have*

$$\Delta(\overline{X}, \overline{Y}) \geq 1 - e^{-T\varepsilon^2/2}.$$

We remark that Lemmas 4.27 and 4.31 are incomparable. In the parameter range of Lemma 4.27, Lemma 4.31 only gives $\Delta(\overline{X}, \overline{Y}) \geq \Omega(T\varepsilon^2)$ instead of $\Omega(\sqrt{T}\varepsilon)$. To argue that the overall distribution is far from having small collision probability, we introduce the following notion of nonuniformity.

Definition 4.32. Let X be a random variable over $[M]$ with probability mass function p. X is (δ, β)-*nonuniform* if for every function $q : [M] \to \mathbb{R}$ such that $0 \leq q(x) \leq p(x)$ for all $x \in [M]$, and $\sum_x q(x) \geq \delta$, the function satisfies

$$\sum_{x \in [M]} q(x)^2 > \beta/M.$$

Intuitively, a distribution X over $[M]$ is (δ, β)-nonuniform means that even if we remove $(1-\delta)$-fraction of probability mass from X, the "collision probability" remains greater than β/M. In particular, X is $(1-\delta)$-far from any random variable Y with $\mathrm{cp}(Y) \leq \beta/M$.

Lemma 4.33. *Let X be a random variable over $[M]$. If X is $(1-\eta)$-far from uniform, then X is $(2\sqrt{\beta \cdot \eta}, \beta)$-nonuniform for every $\beta \geq 1$.*

Proof. Let p be the probability mass function of X, and $q : [M] \to \mathbb{R}$ be a function such that $0 \leq q(x) \leq p(x)$ for every $x \in [M]$, and $\sum_x q(x) \geq 2\sqrt{\beta \cdot \eta}$. Our goal is to show that $\sum_x q(x)^2 > \beta/M$. Let $T = \{x \in [M] : p(x) \geq 1/M\}$. Note that

$$\Delta(X, U_{[M]}) = \Pr[X \in T] - \Pr[U_{[M]} \in T] \geq 1 - \eta.$$

This implies $\Pr[X \in T] \geq 1 - \eta$, and $\mu(T) = \Pr[U_{[M]} \in T] \leq \eta$. Now,

$$\sum_{x \in T} q(x) \geq 2\sqrt{\beta \cdot \eta} - \Pr[X \notin T] \geq 2\sqrt{\beta \cdot \eta} - \eta > \sqrt{\beta \cdot \eta},$$

and $\mu(T) \leq \eta$ implies

$$\sum_{x \in [M]} q(x)^2 \geq \sum_{x \in T} q(x)^2 \geq \frac{(\sum_{x \in T} q(x))^2}{|T|} > \frac{\beta}{M}. \qquad \square$$

We are ready to prove Theorem 4.30.

Proof of Theorem 4.30. By Lemma 4.25 with $\varepsilon = \sqrt{2\ln(128\alpha/\delta^3)/T} < 1/4$, there is a flat K-source X such that for $(1-\delta/4)$-fraction of hash function $h \in \mathcal{H}$, $h(X)$ is ε-far from uniform, for $K = \Omega((\delta/4)^2 M/\varepsilon^2) = \Omega(\delta^2 MT/\log(\alpha/\delta))$. We set $\overline{X} = (X_1,\ldots,X_T)$ to be T independent copies of X. Consider a hash function h such that $h(X)$ is ε-far from uniform. By Lemma 4.31, $(h(X_1),\ldots,h(X_T))$ is $(1-\eta)$-far from uniform, for $\eta = e^{-\varepsilon^2 T/2} = \delta^3/128\alpha$. By Lemma 4.33, $(h(X_1),\ldots,h(X_T))$ is $(\delta/4, 2\alpha/\delta)$-nonuniform for $(1-\delta/4)$-fraction of hash functions h. By the first statement of Lemma 4.34 below, this implies that $(H,\overline{Y})$ is $(1-\delta)$-far from any distribution $(H',\overline{Z})$ with collision probability $\alpha/(|\mathcal{H}|\cdot M^T)$. □

Lemma 4.34. *Let (H,Y) be a joint distribution over $\mathcal{H}\times[M]$ such that the marginal distribution H is uniform over $\mathcal{H}$. Let $\varepsilon,\delta,\alpha$ be positive real numbers.*

1. *If $Y|_{H=h}$ is $(\delta/4, 2\alpha/\delta)$-nonuniform for at least $(1-\delta/4)$-fraction of $h\in\mathcal{H}$, then (H,Y) is $(1-\delta)$-far from any distribution (H',Z) with $\mathrm{cp}(H',Z)\le\alpha/(|\mathcal{H}|\cdot M)$.*
2. *If $Y|_{H=h}$ is $(0.1, 2\alpha/\varepsilon)$-nonuniform for at least 2ε-fraction of $h\in\mathcal{H}$, then (H,Y) is ε-far from any distribution (H',Z) with $\mathrm{cp}(H',Z)\le\alpha/(|\mathcal{H}|\cdot M)$.*

Proof. We introduce the following notations first. For every $h\in\mathcal{H}$, we define $q_h : [M]\to\mathbb{R}$ by

$$q_h(y) = \min\{\Pr[(H,Y)=(h,y)], \Pr[(H',Z)=(h,y)]\}$$

for every $y\in[M]$. We also define $f:\mathcal{H}\to\mathbb{R}$ by

$$f(h) = \sum_{y\in[M]} q_h(y) \le \Pr[H=h] = \frac{1}{|\mathcal{H}|}.$$

For the first statement, let (H',Z) be a random variable over $\mathcal{H}\times[M]$ that is $(1-\delta)$-close to (H,Y). We need to show that $\mathrm{cp}(H',Z) > \alpha/(|\mathcal{H}|\cdot M)$. Note that $\sum_h f(h) = 1-\Delta((H,Y),(H',Z)) \ge \delta$. So there are at least a $(3\delta/4)$-fraction of hash functions h with $f(h)\ge(\delta/4)/|\mathcal{H}|$. At least a $(3\delta/4)-(\delta/4) = \delta/2$-fraction of h satisfy both $f(h)\ge(\delta/4)/|\mathcal{H}|$ and $Y|_{H=h}$ is $(\delta/4, 2\alpha/\delta)$-nonuniform. By the definition of nonuniformity, for each such h, we have

$$\sum_{y\in[M]^T} (|\mathcal{H}|\cdot q_h(y))^2 > \frac{2\alpha}{\delta\cdot M}.$$

Therefore,

$$\mathrm{cp}(H',Z) \ge \sum_{h,y} q_h(y)^2 > \left(\frac{\delta}{2}\cdot|\mathcal{H}|\right)\cdot\frac{2\alpha}{\delta\cdot|\mathcal{H}|^2 M} = \frac{\alpha}{|\mathcal{H}|\cdot M}.$$

Similarly, for the second statement, let (H',Z) be a random variable over $\mathcal{H}\times[M]$ that is ε-close to (H,Y). We need to show that $\mathrm{cp}(H',Z) > \alpha/(|\mathcal{H}|\cdot M)$. Note that $\sum_h f(h) = 1-\Delta((H,Y),(H',Z)) \ge 1-\varepsilon$. So there are at least a $1-\varepsilon/0.9$-fraction of h with $f(h)\ge 0.1/|\mathcal{H}|$. At least a $2\varepsilon-\varepsilon/0.9 > \varepsilon/2$-fraction of hash functions satisfy both $f(h) \geq 0.1/|\mathcal{H}|$ and $Y|_{H=h}$ is $(0.1, 2\alpha/\varepsilon)$-nonuniform. By Lemma 4.33, for each such h, we have

$$\sum_{y \in [M]} (|\mathcal{H}| \cdot q_h(y))^2 > \frac{2\alpha}{\varepsilon \cdot M}.$$

Therefore,

$$\mathrm{cp}(H', Z) \geq \sum_{h,y} q_h(y)^2 > \left(\frac{\varepsilon}{2} \cdot |\mathcal{H}|\right) \cdot \frac{2\alpha}{\varepsilon \cdot |\mathcal{H}|^2 M} = \frac{\alpha}{|\mathcal{H}| \cdot M}. \qquad \square$$

4.4.3 Lower bounds for the distribution of hashed values only

We can extend our lower bounds to the distribution of hashed sequence $\overline{Y} = (H(X_1), \ldots, H(X_T))$ along (without H) for both closeness requirements, at the price of losing the dependency on ε and incurring some dependency on the size of the hash family. Let $2^d = |\mathcal{H}|$ be the size of the hash family. The dependency on d is necessary. Intuitively, the hashed sequence $\overline{Y}$ contains at most $T \cdot m$ bits of entropy, and the input $(H, X_1, \ldots, X_T)$ contains at least $d + T \cdot k$ bits of entropy. When d is large enough, it is possible that all the randomness of hashed sequence comes from the randomness of the hash family. Indeed, if H is T-wise independent (which is possible with $d \simeq T \cdot m$), then $(H(X_1), \ldots, H(X_T))$ is uniform when $X_1, \ldots, X_T$ are all distinct. Therefore,

$$\Delta((H(X_1), \ldots, H(X_T)), U_{[M]^T}) \leq \Pr[\text{not all } X_1, \ldots, X_T \text{ are distinct}].$$

Thus, $K = \Omega(T^2)$ (independent of M) suffices to make the hashed value close to uniform.

Theorem 4.35. *Let N, M, T be positive integers, and d a positive real number such that $N \geq MT/d$. Let $\delta \in (0,1)$, $\alpha > 1$ be real numbers such that $\alpha \cdot 2^d < \delta^3 \cdot e^{T/32}/128$. Let $H : [N] \to [M]$ be a random hash function from an hash family $\mathcal{H}$ of size at most 2^d. There exists an integer $K = \Omega(\delta^2 MT/d \cdot \log(\alpha/\delta))$, and a block K-source $\overline{X} = (X_1, \ldots, X_T)$ such that $\overline{Y} = (H(X_1), \ldots, H(X_T))$ is $(1-\delta)$-far from any distribution $\overline{Z} = (Z_1, \ldots, Z_T)$ with $\mathrm{cp}(\overline{Z}) \leq \alpha/M^T$. In particular, $\overline{Y}$ is $(1-\delta)$-far from uniform.*

Think of α and δ as constants. Then the theorem says that when the hash function contains $d \leq T/(32 \ln 2) - O(1)$ bits of randomness, $K = \Omega(MT/d)$ is necessary for the hashed sequence to be close to uniform. For example, in some typical hash applications, $N = \mathrm{poly}(M)$ and the hash function is 2-universal or $O(1)$-wise independent. In this case, $d = O(\log M)$ and we need $K = \Omega(MT/\log M)$. (Recall that our upper bound in Theorem 4.11 says that $K = O(MT)$ suffices.)

Proof. We will deduce the theorem from Theorem 4.30. Replacing the parameter α by $\alpha \cdot 2^d$ in Theorem 4.30, we know that there exists an integer $K = \Omega(\delta^2 MT/d \cdot \log(\alpha/\delta))$ and a block K-source $\overline{X} = (X_1, \ldots, X_T)$ such that $(H, \overline{Y}) = (H, H(X_1), \ldots, H(X_T))$ is $(1-\delta)$-far from any distribution $(H', \overline{Z})$ with $\mathrm{cp}(H', \overline{Z}) \leq \alpha \cdot 2^d/(2^d \cdot M^T) = \alpha/M^T$. Now, suppose we are given a random variable $\overline{Z}$ on $[M]^T$ with $\Delta(\overline{Y}, \overline{Z}) \leq 1 - \delta$. Then we can define an H' such that $\Delta((H, \overline{Y}), (H', \overline{Z})) = \Delta(\overline{Y}, \overline{Z})$ (Indeed, define the conditional distribution $H'|_{\overline{Z}=\overline{z}}$ to equal $H|_{\overline{Y}=\overline{z}}$ for every $\overline{z} \in [M]^T$.) Then we have

$$\mathrm{cp}(\overline{Z}) \geq \mathrm{cp}(H', \overline{Z}) > \frac{\alpha}{M^T}. \qquad \square$$

One limitation of the above lower bound is that it only works when $d \leq T/(32\ln 2) - O(1)$. For example, the lower bound cannot be applied when the hash function is T-wise independent. Although $d = \Omega(T)$ may not be interesting in practice, for the sake of completeness, we provide another simple lower bound to cover this parameter region.

Theorem 4.36. *Let N,M,T be positive integers, and $\delta \in (0,1)$, $\alpha > 1$, $d > 0$ real numbers. Let $H : [N] \to [M]$ be a random hash function from an hash family $\mathcal{H}$ of size at most 2^d. Suppose $K \leq N$ be an integer such that $K \leq (\delta^2/4\alpha \cdot 2^d)^{1/T} \cdot M$. Then there exists a block K-source $\overline{X} = (X_1,\dots,X_T)$ such that $\overline{Y} = (H(X_1),\dots,H(X_T))$ is $(1-\delta)$-far from any distribution $\overline{Z} = (Z_1,\dots,Z_T)$ with $\mathrm{cp}(\overline{Z}) \leq \alpha/M^T$. In particular, $\overline{Y}$ is $(1-\delta)$-far from uniform.*

Again, think of α and δ as constants. The theorem says that $K = \Omega(M/2^{d/T})$ is necessary for the hashed sequence to be close to uniform. In particular, when $d = \Theta(T)$, $K = \Omega(M)$ is necessary. Theorem 4.35 gives the same conclusion, but only works for $d \leq T/(32\ln 2) - O(1)$. On the other hand, when $d = o(T)$, Theorem 4.35 gives better lower bound $K = \Omega(MT/d)$.

Proof. Let X be any flat K-source, i. e., a uniform distribution over a set of size K. We simply take $\overline{X} = (X_1,\dots,X_T)$ to be T independent copies of X. Note that $\overline{Y}$ has support at most as large as $(H,\overline{X})$. Thus,

$$|\mathrm{supp}(\overline{Y})| \leq |\mathrm{supp}(H,\overline{X})| = 2^d \cdot K^T \leq \frac{\delta^2}{4\alpha} \cdot M^T .$$

Therefore, $\overline{Y}$ is $(1-\delta^2/4\alpha)$-far from uniform. By Lemma 4.33, $\overline{Y}$ is $(1-\delta)$-far from any distribution $\overline{Z} = (Z_1,\dots,Z_T)$ with $\mathrm{cp}(\overline{Z}) \leq \alpha/M^T$. □

4.4.4 Lower bound for 2-universal hash functions

In this subsection, we show Theorem 4.11 is almost tight in the following sense. We show that there exists $K = \Omega(MT/\varepsilon \cdot \log(1/\varepsilon))$, a 2-universal hash family $\mathcal{H}$, and a block K-source $\overline{X}$ such that $(H,\overline{Y})$ is ε-far from having collision probability $100/(|\mathcal{H}| \cdot M^T)$. The improvement over Theorem 4.30 is the almost tight dependency on ε. Recall that Theorem 4.11 says that for 2-universal hash family, $K = O(MT/\varepsilon)$ suffices. The upper and lower bound differs by a factor of $\log(1/\varepsilon)$. In particular, our result for 4-wise independent hash functions (Theorem 4.12) cannot be achieved with 2-universal hash functions. The lower bound can further be extended to the distribution of hashed sequence $\overline{Y} = (H(X_1),\dots,H(X_T))$ as in the previous subsection. Furthermore, since the 2-universal hash family we use has small size, we only pay a factor of $O(\log M)$ in the lower bound on K. Formally we prove the following results.

Theorem 4.37. *For every prime power M, real numbers $\varepsilon \in (0,1/4)$ and $\alpha \geq 1$, the following holds. For all integers t and N such that $\varepsilon \cdot M^{t-1} \geq 1$ and $N \geq 6\varepsilon M^{2t}$, and for $T = \lceil \varepsilon^2 M^{2t-1} \log(\alpha/\varepsilon) \rceil$,[2] there exists an integer $K = \Omega(MT/\varepsilon \cdot \log(\alpha/\varepsilon))$, and a 2-universal hash family $\mathcal{H}$ from $[N]$ to $[M]$, and a block K-source $\overline{X} = (X_1,\dots,X_T)$ such that $(H,\overline{Y}) = (H,H(X_1),\dots,H(X_T))$ is ε-far from any distribution $(H',\overline{Z})$ with $\mathrm{cp}(H',\overline{Z}) \leq \alpha/(|\mathcal{H}| \cdot M^T)$.*

[2] For technical reasons, our lower bound proof does not work for every sufficiently large T. However, note that the density of T such that the lower bound holds is $1/M^2$.

Theorem 4.38. *For every prime power M, real numbers $\varepsilon \in (0,1/4)$ and $\alpha \geq 1$, the following holds. For all integers t and N such that $\varepsilon \cdot M^{t-1} \geq 1$ and $N \geq 6\varepsilon M^{2t}$, and for $T = \lceil \varepsilon^2 M^{2t-1} \log(\alpha M/\varepsilon) \rceil$, there exists an integer $K = \Omega(MT/\varepsilon \cdot \log(\alpha M/\varepsilon))$, and a 2-universal hash family $\mathcal{H}$ from $[N]$ to $[M]$, and a block K-source $\overline{X} = (X_1, \dots, X_T)$ such that $\overline{Y} = (H(X_1), \dots, H(X_T))$ is ε-far from any distribution $\overline{Z}$ with $\mathrm{cp}(\overline{Z}) \leq \alpha/M^T$.*

Basically, the idea is to show that the Markov inequality applied in the proof of Theorem 4.11 (see inequality (4.1)) is tight for a single block. More precisely, we show that there exists a 2-universal hash family $\mathcal{H}$, and a K-source X such that with probability ε over $h \leftarrow H$, $\mathrm{cp}(h(X)) \geq 1/M + \Omega(1/K\varepsilon)$. Intuitively, if we take $T = \Theta(K\varepsilon \cdot \log(\alpha/\varepsilon)/M)$ independent copies of such X, then the collision probability will satisfy $\mathrm{cp}(h(X_1), \dots, h(X_T)) \geq (1 + \Omega(M/K\varepsilon))^T/M^T \geq \alpha/(\varepsilon M^T)$, and so the overall collision probability is $\mathrm{cp}(H, \overline{Y}) \geq \alpha/(|\mathcal{H}| \cdot M^T)$. Formally, we analyze our construction below using Hellinger distance, and show that the collision probability remains high even after modifying a $\Theta(\varepsilon)$-fraction of distribution.

Proof of Theorem 4.37. Fix a prime power M, and $\varepsilon > 0$, we identify $[M]$ with the finite field $\mathbb{F}$ of size M. Let t be an integer parameter such that $M^{t-1} > 1/\varepsilon$. Recall that the set $\mathcal{H}_0$ of linear functions $\{h_{\vec{a}} : \mathbb{F}^t \to \mathbb{F}\}_{\vec{a} \in \mathbb{F}^t}$ where $h_{\vec{a}}(\vec{x}) = \sum_i a_i x_i$ is 2-universal. Note that picking a random hash function $h \leftarrow \mathcal{H}_0$ is equivalent to picking a random vector $\vec{a} \leftarrow \mathbb{F}^t$. Two special properties of $\mathcal{H}_0$ are (i) when $\vec{a} = \vec{0}$, the whole domain $\mathbb{F}^t$ is sent to $0 \in \mathbb{F}$, and (ii) the size of hash family $|\mathcal{H}_0|$ is the same as the size of the domain, namely $|\mathbb{F}^t|$. We will use $\mathcal{H}_0$ as a building block in our construction.

We proceed to construct the hash family $\mathcal{H}$. We partition the domain $[N]$ into several sub-domains, and apply different hash function to each sub-domain. Let s be an integer parameter to be determined later. We require $N \geq s \cdot M^t$, and partition $[N]$ into $D_0, D_1, \dots, D_s$, where each of $D_1, \dots, D_s$ has size M^t and is identified with $\mathbb{F}^t$, and D_0 is the remaining part of $[N]$. In our construction, the data $\overline{X}$ will never come from D_0. Thus, w. l. o. g., we can assume D_0 is empty. For every $i = 1, \dots, s$, we use a linear hash function $h_{\vec{a}_i} \in \mathcal{H}_0$ to send D_i to $\mathbb{F}$. Thus, a hash function $h \in \mathcal{H}$ consists of s linear hash function $(h_{\vec{a}_1}, \dots, h_{\vec{a}_s})$, and can be described by s vectors $\vec{a}_1, \dots, \vec{a}_s \in \mathbb{F}^t$. Note that to make $\mathcal{H}$ 2-universal, it suffices to pick $\vec{a}_1, \dots, \vec{a}_s$ pairwise independently. Specifically, we identify $\mathbb{F}^t$ with the finite field $\hat{\mathbb{F}}$ of size M^t, and pick $(\vec{a}_1, \dots, \vec{a}_s)$ by picking $a, b \in \hat{\mathbb{F}}$, and output $(a + \alpha_1 \cdot b, a + \alpha_2 \cdot b, \dots, a + \alpha_s \cdot b)$ for some distinct constants $\alpha_1, \dots, \alpha_s \in \hat{\mathbb{F}}$. Formally, we define the hash family to be

$$\mathcal{H} = \{h^{a,b} : [N] \to [M]\}_{a,b \in \hat{\mathbb{F}}}, \text{ where } h^{a,b} = (h_{a+\alpha_1 b}, \dots, h_{a+\alpha_s b}) \stackrel{\text{def}}{=} (h_1^{a,b}, \dots, h_s^{a,b}).$$

It is easy to verify that $\mathcal{H}$ is indeed 2-universal, and $|\mathcal{H}| = M^{2t}$.

We next define a single block K-source X that makes the Markov inequality (4.1) tight. We simply take X to be a uniform distribution over $D_1 \cup \dots \cup D_s$, and so $K = s \cdot M^t$. Consider a hash function $h^{a,b} \in \mathcal{H}$. If all $h_i^{a,b}$ are non-zero and distinct, then $h^{a,b}(X)$ is the uniform distribution. If exactly one $h_i^{a,b} = 0$, then $h^{a,b}$ sends $M^t + (s-1)M^{t-1}$ elements in $[N]$ to 0, and $(s-1)M^{t-1}$ elements to each nonzero

$y \in \mathbb{F}$. Let us call such $h^{a,b}$ *bad* hash functions. Thus, if $h^{a,b}$ is bad, then

$$\begin{aligned}\mathrm{cp}(h^{a,b}(X)) &= \left(\frac{M^t + (s-1)M^{t-1}}{K}\right)^2 + (M-1)\cdot\left(\frac{(s-1)M^{t-1}}{K}\right)^2 \\ &= \frac{1}{M} + \frac{M-1}{s^2 M} \geq \frac{1}{M} + \frac{1}{2s^2}\,.\end{aligned}$$

Note that $h^{a,b}$ is bad with probability

$$\Pr[\text{exactly one } h_i^{a,b} = 0] = \Pr[b \neq 0 \wedge \exists i\ (a + \alpha_i b = 0)] = \left(1 - \frac{1}{M^t}\right)\cdot\frac{s}{M^t} \geq \frac{s}{2M^t}\,.$$

We set $s = \lceil 4\varepsilon M^t \rceil \leq M^t$. It follows that with probability at least 2ε over $h \leftarrow \mathcal{H}$, the collision probability satisfies $\mathrm{cp}(h(X)) \geq 1/M + 1/(4K\varepsilon)$, as we intuitively desired. However, instead of working with collision probability directly, we need to use Bhattacharyya coefficient to measure the growth of distance to uniform (see Definition 4.18.) The following claim upper bounds the Bhattacharyya coefficient of $h(X)$ for bad hash functions h. The proof of the claim is deferred to the end of this section.

Claim 4.39. *Suppose h is a bad hash function defined as above, then the Bhattacharyya coefficient of $h(X)$ satisfies $C(h(X)) \leq 1 - M/(64K\varepsilon)$.*

Finally, for every integer $T \in [\varepsilon^2 M^{2t-1}\log(\alpha/\varepsilon), c_0 \cdot \varepsilon^2 M^{2t-1}\log(\alpha/\varepsilon)]$, we can write $T = c\cdot(64K\varepsilon/M)\cdot\ln(800\alpha/\varepsilon)$ for some constant $c < c_0$. Let $\overline{X} = (X_1,\ldots,X_T)$ be T independent copies of X. We now show that $K,\mathcal{H},\overline{X}$ satisfy the conclusion of the theorem. That is, $K = \Omega(MT/(\varepsilon\log(\alpha/\varepsilon)))$ (as follows from the definition of T) and $(H,\overline{Y}) = (H,H(X_1),\ldots,H(X_T))$ is ε-far from any distribution $(H',\overline{Z})$ with $\mathrm{cp}(H',\overline{Z}) \leq \alpha/(|\mathcal{H}|\cdot M^T)$.

Consider the distribution $(h(X_1),\ldots,h(X_T))$ for a bad hash function $h \in \mathcal{H}$. From the above claim, the Bhattacharyya coefficient satisfies

$$C(h(X_1),\ldots,h(X_T)) = C(h(X))^T \leq (1 - M/64K\varepsilon)^T \leq e^{MT/64K\varepsilon} \leq \frac{800\alpha}{\varepsilon}\,.$$

By Lemma 4.19 and the definition of Bhattacharyya coefficient, we have

$$\Delta((h(X_1),\ldots,h(X_T)), U_{[M]^T}) \geq 1 - C(h(X_1),\ldots,h(X_T)) \geq 1 - \frac{800\alpha}{\varepsilon}\,.$$

By Lemma 4.33, $(h(X_1),\ldots,h(X_T))$ is $(0.1, 2\alpha/\varepsilon)$-nonuniform for at least 2ε-fraction of bad hash functions h. By the second statement of Lemma 4.34, this implies $(H,\overline{Y})$ is ε-far from any distribution $(H',\overline{Z})$ with $\mathrm{cp}(H',\overline{Z}) \leq \alpha/(|\mathcal{H}|\cdot M^T)$.

In sum, given M,ε,α,t that satisfies the premise of the theorem, we set $K = \lceil 4\varepsilon M^t \rceil \cdot M^t$, and proved that for every $N \geq K$, and $T = \Theta((K\varepsilon/M)\cdot\ln(\alpha/\varepsilon))$, the conclusion of the theorem is true. It remains to prove Claim 4.39.

Proof of Claim 4.39. Let $p(x) = M\cdot\Pr[h(X) = x]$ for every $x \in \mathbb{F}$. For a bad hash function h, we have $p(0) = (1 + (M-1)/s)$, and $p(x) = (1 - 1/s)$ for every $x \neq 0$. We will upper bound $C(h(X)) =$

$(1/M) \cdot \sum_x \sqrt{p(x)}$ using Taylor series. Recall that for $z \geq 0$, there exists some $z', z'' \in [0, z]$ such that

$$\sqrt{1+z} = 1 + \frac{z}{2} + \frac{z^2}{2} \cdot \left(-\frac{1}{4(1+z')^{3/2}} \right) \leq 1 + \frac{z}{2} - \frac{z^2}{8(1+z)^{3/2}}$$

and

$$\sqrt{1-z} = 1 - z \cdot \frac{1}{2\sqrt{1-z''}} \leq 1 - \frac{z}{2}.$$

We have

$$\begin{aligned} C(h(X)) &= \frac{1}{M} \sum_x \sqrt{p(x)} \\ &\leq \frac{1}{M} \left(1 + \frac{M-1}{2s} - \frac{(M-1)^2}{8s^2 \cdot (1+(M-1)/s)^{3/2}} + (M-1) \cdot \left(1 - \frac{1}{2s} \right) \right) \\ &= 1 - \frac{(M-1)^2}{8Ms^2(1+(M-1)/s)^{3/2}}. \end{aligned}$$

Recall that $M \geq 2$, $s = \varepsilon M^t \geq M$, and $s^2 = K\varepsilon$, we have

$$C(h(X)) \leq 1 - \frac{M^2}{64K\varepsilon}. \qquad \square$$

This concludes the proof of Theorem 4.37. $\square$

Recall that $|\mathcal{H}| = M^{2t}$. Theorem 4.38 follows from Theorem 4.37 by exactly the same argument as in the proof of Theorem 4.35.

5 Applications

5.1 Linear probing

We consider data items come as a block K-source $(X_1, \ldots, X_{T-1}, X_T)$ where the item $Y = X_T$ to be inserted is the last block. An immediate application of Theorem 4.10, using just a 2-universal hash family, gives that if $K \geq MT/\varepsilon^2$, the resulting distribution of the element hashes is ε-close to uniform. The effect of the ε statistical difference on the expected insertion time is at most εT, because the maximum insertion time is T. That is, if we let E_U be the expected time for an insertion when using a truly random hash function, and E_P be the expected time for an insertion using pairwise independent hash functions, we have

$$E_P \leq E_U + \varepsilon T.$$

A natural choice is $\varepsilon = o(1/T)$, so that the εT term is $o(1)$, giving that K needs to be $\omega(MT^3) = \omega(M^4)$ in the standard case where $T = \alpha M$ for a constant $\alpha \in (0,1)$ (which we assume henceforth). An alternative interpretation is that with probability $1 - \varepsilon$, our hash table behaves exactly as though a truly random hash function was used. In some applications, constant ε may be sufficient, in which case $K = O(M^2)$ suffices.

Better results can be obtained by applying Lemma 4.8, in conjunction with Theorem 4.11 or Theorem 4.12. In particular, for linear probing, the standard deviation σ of the insertion time is known (see, e. g., [16, p.52]) and is $O(1/(1-\alpha)^2)$. With a 2-universal family, as long as $K \geq MT/\varepsilon$, from Theorem 4.11 the resulting hash values are ε-close to a block source with collision probability at most $(1+2MT/(\varepsilon K))/M^T$. Using this, we apply Lemma 4.8 to bound the expected insertion time as

$$E_P \leq E_U + \varepsilon T + \sigma\sqrt{\frac{2MT}{\varepsilon K}}\,.$$

Choosing $\varepsilon = o(1/T)$ gives that E_P and E_U are the same up to lower order terms when K is $\omega(M^3)$. Theorem 4.12 gives a further improvement; for $K \geq MT + \sqrt{2MT^2/\varepsilon}$, we have

$$E_P \leq E_U + \varepsilon T + \sigma\sqrt{\frac{2MT + 2\sqrt{2MT^2/\varepsilon}}{K}}\,.$$

Choosing $\varepsilon = o(1/T)$ now allows for K to be only $\omega(M^2)$.

In other words, the Rényi entropy needs only to be $2\log M + \omega(1)$ bits when using 4-wise independent hash functions, and $3\log M + \omega(1)$ for 2-universal hash functions. These numbers seem quite reasonable for practical situations. We formalize the result for the case of 2-universal hash functions as follows:

Theorem 5.1. *Let H be chosen at random from a 2-universal hash family $\mathcal{H}$ mapping $[N]$ to $[M]$. For every block K-source $(\overline{X}, Y)$ taking values in $[N]^T$ with $K \geq MT/\varepsilon$, we have*

$$\mathrm{E}[\mathrm{Time}_{\mathrm{LP}}(H, \overline{X}, Y)] \leq 1/(1-\alpha)^2 + \varepsilon T + \sigma\sqrt{\frac{2MT}{\varepsilon K}}\,.$$

Here $\alpha = T/M$ is the load and $\sigma = O(1/(1-\alpha)^2)$ is the standard deviation in the insertion time in the case of truly random hash functions.

5.2 Chained hashing

We can follow essentially the same line of argument as in the previous section. Recall here T elements are hashed into a table of size $M = T$. Theorem 4.10 again implies that using just a 2-universal hash family, if $K \geq MT/\varepsilon^2 = T^2/\varepsilon^2$, the resulting distribution of the element hashes is ε-close to uniform. In this case, if we let E_U be the expected maximum load when using a truly random hash function, and E_P be the expected maximum load using a 2-universal hash function, we again have

$$E_P \leq E_U + \varepsilon T\,,$$

and similarly having K be $\omega(T^4)$ suffices.

Similarly, extending the argument for Theorem 5.1, we deduce that if $K \geq T^2/\varepsilon$, then

$$\mathrm{E}[\mathrm{MaxLoad}_{\mathrm{CH}}(\overline{X}, H)] \leq (1+o(1)) \cdot \frac{\log T}{\log\log T} + \varepsilon T + \sigma\sqrt{\frac{2T^2}{\varepsilon K}}\,,$$

where the $o(1)$ term goes to zero as $T \to \infty$ and σ is the standard deviation in the maximum load in the case of a truly random hash function.

However, here we get a cleaner "high-probability" result by using Theorem 4.11:

Theorem 5.2. *Let H be chosen at random from a 2-universal hash family $\mathcal{H}$ mapping $[N]$ to $[T]$. For every block K-source $\overline{X}$ taking values in $[N]^T$ with $K = \omega(T^2)$, we have*

$$\Pr\left[\mathrm{MaxLoad}_{\mathrm{CH}}(\overline{X}, H) \leq \frac{\log T}{\log\log T}(1+o(1))\right] = 1 - o(1)\,,$$

where the $o(1)$ terms tend to zero as $T \to \infty$.

Proof. Set $M = T$. Note that the value of $\mathrm{MaxLoad}_{\mathrm{CH}}(\bar{x}, h)$ can be determined from the hashed sequence $(h(x_1), \ldots, h(x_T)) \in [M]^T$ alone, and does not otherwise depend on the data sequence $\bar{x}$ or the hash function h. Thus for a function $\lambda : \mathbb{N} \to \mathbb{N}$, we can let $S \subseteq [M]^T$ be the set of all sequences of hashed values that produce an allocation with a max load greater than $\lambda(T)$. By Theorem 3.4, we can take $\lambda(T) = (1+o(1)) \cdot (\log T)/(\log\log T)$ so that we have:

$$\Pr[U_{[M]^T} \in S] = \Pr[\mathrm{MaxLoad}_{\mathrm{CH}}(\bar{x}, I) > \lambda(T)] = o(1)\,,$$

where I is a truly random hash function mapping $[N]$ to $[M] = [T]$ and $\bar{x}$ is an arbitrary sequence of distinct data items.

We are interested in the quantity:

$$\Pr[\mathrm{MaxLoad}_{\mathrm{CH}}(\overline{X}, H) > \lambda(T)] = \Pr[(H(X_1), \ldots, H(X_T)) \in S]\,,$$

where H is a random hash function from a 2-universal family. Given $K = \omega(T^2)$, we set $\varepsilon = MT/K = o(1)$. By Theorem 4.11, $(H(X_1), \ldots, H(X_T))$ is ε-close to a random variable Z with collision probability at most $(1 + 2MT/(\varepsilon K))/M^T = 3/M^T$ per block. Thus, applying Lemma 4.8 with f as the characteristic function of S and $\mu = \mathrm{E}[f(U_{[M]^T})] \leq o(1)$, we have

$$\begin{aligned}\Pr[(H(X_1), \ldots, H(X_T)) \in S] &\leq \Pr[Z \in S] + \varepsilon \\ &\leq \mu + \sqrt{\mu \cdot (1-\mu)} \cdot \sqrt{3} + \varepsilon \\ &= o(1)\,. \end{aligned}$$

□

5.3 Balanced allocations

By combining the known analysis for ideal hashing (Theorem 3.6), our optimized bounds for block-source extraction (Theorems 4.11 and 4.12), and the effect of collision probability on expectations (Lemma 4.8), we obtain:

Theorem 5.3. *For every $d \geq 2$ and $\gamma > 0$, there is a constant c such the following holds. Let H be chosen at random from a 2-universal hash family $\mathcal{H}$ mapping $[N]$ to $[T]^d$. For every block K-source $\overline{X}$ taking values in $[N]^T$ with $K \geq 2T^{d+1+\gamma}$, we have*

$$\Pr\left[\mathrm{MaxLoad}_{\mathrm{BA}}(\overline{X}, H) > \frac{\log\log T}{\log d} + c\right] \leq \frac{1}{T^\gamma}\,.$$

Proof. Set $M = T^d$. Note that the value of $\mathrm{MaxLoad}_{\mathrm{BA}}(\bar{x}, h)$ can be determined from the hashed sequence $(h(x_1), \ldots, h(x_T)) \in M^T$ alone, and does not otherwise depend on the data sequence $\bar{x}$ or the hash function h. Thus we can let $S \subseteq M^T$ be the set of all sequences of hashed values that produce an allocation with a max load greater than $(\log\log T)/(\log d) + c$. By Theorem 3.6, we can choose the constant c such that

$$\Pr\left[U_{[M]^T} \in S\right] = \Pr\left[\mathrm{MaxLoad}_{\mathrm{BA}}(\bar{x}, I) > \frac{\log\log T}{\log d} + c\right] \leq \frac{1}{T^{3\gamma}},$$

where I is a truly random hash function mapping $[N]$ to $[M] = [T]^d$ and $\bar{x}$ is an arbitrary sequence of distinct data items.

We are interested in the quantity

$$\Pr\left[\mathrm{MaxLoad}_{\mathrm{BA}}(\overline{X}, H) > \frac{\log\log T}{\log d} + c\right] = \Pr\left[(H(X_1), \ldots, H(X_T)) \in S\right].$$

Set $\varepsilon = 1/2T^\gamma$ and $K = 2T^{d+1+\gamma} = MT/\varepsilon$. By Theorem 4.11, $(H(X_1), \ldots, H(X_T))$ is ε-close to a random variable Z with collision probability at most $(1 + 2MT/(\varepsilon K))/M^T = 3/M^T$ per block. Thus, applying Lemma 4.8 with f as the characteristic function of S and $\mu = \mathrm{E}[f(U_{[M]^T})] \leq 1/T^{3\gamma}$, we have

$$\begin{aligned}
\Pr[(H(X_1), \ldots, H(X_T)) \in S] &\leq \Pr[Z \in S] + \varepsilon \\
&\leq \mu + \sqrt{\mu \cdot (1-\mu)} \cdot \sqrt{3} + \varepsilon \\
&\leq \frac{1}{T^{3\gamma}} + \sqrt{\frac{3}{T^{3\gamma}}} + \frac{1}{2T^\gamma} \\
&\leq \frac{1}{T^\gamma},
\end{aligned}$$

for sufficiently large T. (Small values of T can be handled by increasing the constant c in the theorem.) □

Theorem 5.4. *For every $d \geq 2$ and $\gamma > 0$, there is a constant c such the following holds. Let H be chosen at random from a 4-wise independent hash family $\mathcal{H}$ mapping $[N]$ to $[T]^d$. For every block K-source $\overline{X}$ taking values in $[N]^T$ with $K \geq (T^{d+1} + 2T^{(d+2+\gamma)/2})$, we have*

$$\Pr\left[\mathrm{MaxLoad}_{\mathrm{BA}}(\overline{X}, H) > \frac{\log\log T}{\log d} + c\right] \leq \frac{1}{T^\gamma}.$$

Proof. The proof is identical to that of Theorem 5.3, except we use Theorem 4.12 instead of Theorem 4.11 and set $K = T^{d+1} + T^{(d+2+\gamma)/2} = MT + \sqrt{2MT^2/\varepsilon}$. □

5.4 Bloom filters

We consider the following setting: our block source takes on values in $[N]^{T+1}$, producing a collection $(x_1, \ldots, x_T, y) = (\bar{x}, y)$, where $\bar{x}$ constitutes the set represented by the filter, and y represents an additional data item that will not be equal to any data items of $\bar{x}$ (with high probability).

We first take the advantage of the following result by [19], which reduces the number of required hash function from ℓ to 2.

Theorem 5.5 ([19]). *Let $H = (H_1, H_2)$ be a truly random hash function mapping $[N]$ to $[M/\ell]^2$, where M/ℓ is a prime integer. Define $H' = (H'_1, \ldots, H'_\ell) : [N] \to [M/\ell]^\ell$ by*

$$H'_i(w) = H_1(w) + (i-1)H_2(w) \bmod M/\ell\,.$$

Then for every sequence $\overline{x} \in [N]^T$ of T data items and every $y \notin \overline{x}$, we have

$$\Pr[\mathrm{FalsePos}_{\mathrm{BF}}(H', \overline{x}, y) = 1] \le \left(1 - \left(1 - \frac{\ell}{M}\right)^T\right)^\ell + O(1/M)\,.$$

The restriction to prime integers M/ℓ is not strictly necessary in general; for more complete statements of when 2 truly random hash functions suffice, see [19].

If we allow the false positive probability to increase by some $\varepsilon > 0$ over truly random hash functions, we can use Theorem 4.10 to immediately obtain the following parallel to Theorem 5.5:[3]

Theorem 5.6. *Let $H = (H_1, H_2)$ be chosen at random from a 2-universal hash family $\mathcal{H}$ mapping $[N]$ to $[M/\ell]^2$, where M/ℓ is a prime integer. Define $H' = (H'_1, \ldots, H'_\ell) : [N] \to [M/\ell]^\ell$ by*

$$H'_i(w) = H_1(w) + (i-1)H_2(w) \bmod M/\ell\,.$$

For every $\varepsilon > 1/M$ and every block K-source $(\overline{X}, Y)$ taking values in $[N]^T \times [N] \cong [N]^{T+1}$ with $K \ge M^2T/\varepsilon^2\ell^2$, we have

$$\Pr[\mathrm{FalsePos}_{\mathrm{BF}}(H', \overline{X}, Y) = 1] \le \left(1 - \left(1 - \frac{\ell}{M}\right)^T\right)^\ell + O(\varepsilon)\,.$$

If we set $\varepsilon = o(1)$, then we obtain the same asymptotic false positive probabilities as with truly random hash functions. When $T = \Theta(M)$, the Rényi entropy per block needs only to be $3\log M + \omega(1)$ bits for 2-universal hash functions.

6 Alternative approaches

The results we have described in Section 5 rely on very general arguments, referring to the collision probability of the entire sequence of hashed data values. We suggest, however, that it may prove useful in the future to view the results of hashing block sources in this paper as a collection of tools that can be applied in various ways to specific applications. For example, here we present a variant of Theorems 4.11 and 4.12, asserting that the hashed values are close to a block source with bounded collision probability *per block*, which may yield improved results in some cases.

Theorem 6.1. *Let $H : [N] \to [M]$ be a random hash function from a 2-universal family $\mathcal{H}$. For every block K-source $(X_1, \ldots, X_T)$ and every $\varepsilon > 0$, the random variable $Y = (H(X_1), \ldots, H(X_T))$ is ε-close to a block source Z with collision probability $1/M + T/(\varepsilon K)$ per block.*

[3] We note that the approach of using Lemma 4.8 along with Theorems 4.11 and 4.12 in the previous sections does not yield improvement here, since in the typical case of Bloom filters, the false positive probability is a constant, instead of $o(1)$.

Theorem 6.2. *Let $H : [N] \to [M]$ be a random hash function from a 4-wise independent family $\mathcal{H}$. For every block K-source $(X_1, \ldots, X_T)$ and for every $\varepsilon > 0$, the random variable $Y = (H(X_1), \ldots, H(X_T))$ is ε-close to a block source Z with collision probability $1/M + 1/K + \sqrt{2T/(\varepsilon M)} \cdot 1/K$ per block.*

Theorem 6.1 and 6.2 can be proved in a similar way to the proof of Theorem 4.11 and 4.12, where instead of applying Markov/Chebychev's inequality to the whole sequence once, here we apply the inequalities to each block to achieve the stronger conclusion.

We sketch an example of how these results can be applied to more specific arguments for an application. In the standard layered induction argument for balanced allocations [2], the following key step is used. Suppose that there are at most $\beta_i T$ buckets with load at least i throughout the process. Then (using truly random hash functions) the probability that a data item with d choices lands in a bin with i or more balls already present is bounded above by $(\beta_i)^d$. When using 2-universal hash functions, we can bound this probability, but with slightly weaker results. The choices for a data item correspond to the hash of one of the blocks from the input block source. Let S be the set of size at most $(\beta_i)^d$ possible hash values for the item's choices that would place the item in a bin with i or more balls. We can bound the probability that the item hashes to a value in S by bounding the collision probability per block (via Theorem 6.1) and applying Lemma 4.8 with f equal to the characteristic function of S. We have applied this technique to generalize the standard layered induction proof of [2] to this setting. This approach turns out to require slightly less entropy from the source for 2-universal hash functions than Theorem 5.3, but the loss incurred in applying Lemma 4.8 means that the analysis only works for $d \geq 3$ choices and the maximum load changes by a constant factor (although the $O(\log \log n)$ behavior is still apparent). We omit the details.

7 Conclusion

We have started to build a link between previous work on randomness extraction and the practical performance of simple hash functions, specifically 2-universal hash functions. In the future, we hope that there will be a collaboration between theory and systems researchers aimed at fully understanding the behavior of hashing in practice. Indeed, while our view of data as coming from a block source is a natural initial suggestion, theory–systems interaction could lead to more refined and realistic models for real-life data (and in particular, provide estimates for the amount of entropy in the data). A complementary direction is to show that hash functions used in practice (such as those based on cryptographic functions, which may not even be 2-universal) behave similarly to truly random hash functions for these data models. Some results in this direction can be found in [12].

Acknowledgments

We thank Adam Kirsch for his careful reading of and helpful comments on the paper, Wei-Chun Kao for helpful discussions, and David Zuckerman for telling us about Hellinger distance. We also thank the anonymous reviewers for helpful corrections and suggestions.

A Technical lemma on binomial coefficients

Claim A.1 (Claim 4.26, restated)**.** *Let $N,K > 1$ be integers such that $N > 2K$, and $L \in [0,K/2]$, $\beta \in (0,\min\{1,\sqrt{L}\})$ real numbers. Let $S \subset [N]$ be a random subset of size K, and $T \subset [N]$ be a fixed subset of $[N]$ of arbitrary size. We have*

$$\Pr_S\left[||S \cap T| - L| \le \beta\sqrt{L}\right] \le O(\beta)\,.$$

Proof. By an abuse of notation, we use T to denote the size of set T. The probability can be expressed as a sum of binomial coefficients as follows.

$$\Pr_S\left[||S \cap T| - L| \le \beta\sqrt{L}\right] = \sum_{R=\lceil L-\beta\sqrt{L}\rceil}^{\lfloor L+\beta\sqrt{L}\rfloor} \frac{\binom{T}{R}\binom{N-T}{K-R}}{\binom{N}{K}}\,.$$

Note that there are at most $\lfloor 2\beta\sqrt{L}\rfloor + 1$ terms, it suffices to show that for every $R \in \left[L-\beta\sqrt{L}, L+\beta\sqrt{L}\right]$,

$$f(T) \stackrel{\text{def}}{=} \frac{\binom{T}{R}\binom{N-T}{K-R}}{\binom{N}{K}} \le O\left(\sqrt{\frac{1}{L}}\right)\,.$$

We use the following bound on binomial coefficients, which can be derived from Stirling's formula.

Claim A.2. *For integers $0 < i < a$, $0 < j < b$, we have*

$$\frac{\binom{a}{i}\binom{b}{j}}{\binom{a+b}{i+j}} \le O\left(\sqrt{\frac{a \cdot b \cdot (i+j) \cdot (a+b-i-j)}{i \cdot (a-i) \cdot j \cdot (b-j) \cdot (a+b)}}\right)\,.$$

Note that $L \in [0,K/2]$ implies $K - R = \Omega(K)$. When $2R \le T \le N - 2K + 2R$, we have

$$\begin{aligned} f(T) &= \frac{\binom{T}{R}\binom{N-T}{K-R}}{\binom{N}{K}} \\ &= O\left(\sqrt{\frac{T(N-T)K(N-K)}{R(T-R)(K-R)(N-T-K+R)N}}\right) \\ &= O\left(\sqrt{\frac{1}{R} \cdot \frac{K}{K-R} \cdot \frac{N-K}{N} \cdot \frac{T(N-T)}{(T-R)(N-T-K+R)}}\right) \\ &= O\left(\sqrt{\frac{1}{R}}\right) = O\left(\sqrt{\frac{1}{L}}\right), \end{aligned}$$

as desired. Note that when $N > 2K$, such T exists. Finally, observe that $\beta^2 < L$ implies $R \ge 1$, and

$$\frac{f(T)}{f(T+1)} = \frac{(T-R+1)(N-T)}{(T+1)(N-T-K+R)}\,.$$

It follows that $f(T)$ is increasing when $T \le 2R$, and $f(T)$ is decreasing when $T \ge N-2K+2R$. Therefore, $f(T) \le f(2R) = O(\sqrt{1/L})$ for $T \le 2R$, and $f(T) \le f(N-2K+2R) = O(\sqrt{1/L})$ for $T \ge N-2K+2R$, which complete the proof. ☐

References

[1] NOGA ALON, MARTIN DIETZFELBINGER, PETER BRO MILTERSEN, EREZ PETRANK, AND GÁBOR TARDOS: Linear hash functions. *J. ACM*, 46(5):667–683, 1999. Preliminary version in STOC'97. [doi:10.1145/324133.324179] 900, 903

[2] YOSSI AZAR, ANDREI Z. BRODER, ANNA R. KARLIN, AND ELI UPFAL: Balanced allocations. *SIAM J. Comput.*, 29(1):180–200, 1999. Preliminary version in STOC'94. [doi:10.1137/S0097539795288490] 901, 903, 904, 938

[3] CHARLES H. BENNETT, GILLES BRASSARD, AND JEAN-MARC ROBERT: Privacy amplification by public discussion. *SIAM J. Comput.*, 17(2):210–229, 1988. Preliminary version in CRYPTO'85. [doi:10.1137/0217014] 899, 910

[4] BURTON H. BLOOM: Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426, 1970. [doi:10.1145/362686.362692] 901, 904, 905

[5] AVRIM BLUM AND JOEL SPENCER: Coloring random and semi-random k-colorable graphs. *J. Algorithms*, 19(2):204–234, 1995. [doi:10.1006/jagm.1995.1034] 899

[6] ANDREI Z. BRODER AND MICHAEL MITZENMACHER: Using multiple hash functions to improve IP lookups. In *Proc. 20th Ann. Joint Conf. IEEE Computer and Communications Societies (INFOCOM'01)*, pp. 1454–1463. IEEE Comp. Soc. Press, 2001. [doi:10.1109/INFCOM.2001.916641] 898, 904

[7] ANDREI Z. BRODER AND MICHAEL MITZENMACHER: Network applications of Bloom filters: A survey. *Internet Mathematics*, 1(4):485–509, 2004. [doi:10.1080/15427951.2004.10129096] 898

[8] J. LAWRENCE CARTER AND MARK N. WEGMAN: Universal classes of hash functions. *J. Comput. System Sci.*, 18(2):143–154, 1979. Preliminary version in STOC'77. [doi:10.1016/0022-0000(79)90044-8] 898

[9] BENNY CHOR AND ODED GOLDREICH: Unbiased bits from sources of weak randomness and probabilistic communication complexity. *SIAM J. Comput.*, 17(2):230–261, 1988. Preliminary version in FOCS'85. [doi:10.1137/0217015] 899, 908, 911

[10] SARANG DHARMAPURIKAR, PRAVEEN KRISHNAMURTHY, TODD S. SPROULL, AND JOHN W. LOCKWOOD: Deep packet inspection using parallel Bloom filters. *IEEE Micro*, 24(1):52–61, 2004. Preliminary version in HotI'03. [doi:10.1109/MM.2004.1268997] 898

[11] MARTIN DIETZFELBINGER AND PHILIPP WOELFEL: Almost random graphs with simple hash functions. In *Proc. 35th STOC*, pp. 629–638. ACM Press, 2003. [doi:10.1145/780542.780634] 898, 904

[12] YEVGENIY DODIS, ROSARIO GENNARO, JOHAN HÅSTAD, HUGO KRAWCZYK, AND TAL RABIN: Randomness extraction and key derivation using the CBC, Cascade and HMAC modes.

In *Proc. 24th. Ann. Internat. Cryptology Conference (CRYPTO'04)*, pp. 494–510. Springer, 2004. [doi:10.1007/978-3-540-28628-8_30] 938

[13] RICHARD DURRETT: *Probability: Theory and Examples. Third Edition.* Duxbury, 2004. 920

[14] ALISON L. GIBBS AND FRANCIS EDWARD SU: On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002. See also at arXiv. [doi:10.1111/j.1751-5823.2002.tb00178.x] 912, 919

[15] GASTON H. GONNET: Expected length of the longest probe sequence in hash code searching. *J. ACM*, 28(2):289–304, 1981. [doi:10.1145/322248.322254] 900, 903

[16] GASTON H. GONNET AND RICARDO BAEZA-YATES: *Handbook of algorithms and data structures: in Pascal and C.* Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1991. [ACM:103324] 934

[17] JOHAN HÅSTAD, RUSSELL IMPAGLIAZZO, LEONID A. LEVIN, AND MICHAEL LUBY: A pseudorandom generator from any one-way function. *SIAM J. Comput.*, 28(4):1364–1396, 1999. Preliminary versions in STOC'89 and STOC'90. [doi:10.1137/S0097539793244708] 899, 910

[18] RUSSELL IMPAGLIAZZO AND DAVID ZUCKERMAN: How to recycle random bits. In *Proc. 30th FOCS*, pp. 248–253. IEEE Comp. Soc. Press, 1989. [doi:10.1109/SFCS.1989.63486] 910

[19] ADAM KIRSCH AND MICHAEL MITZENMACHER: Less hashing, same performance: Building a better Bloom filter. *Random Structures & Algorithms*, 33(2):187–218, 2008. Preliminary version in ESA'06. [doi:10.1002/rsa.20208] 936, 937

[20] DONALD E. KNUTH: *The Art of Computer Programming, Volume 3: Sorting and Searching*. Addison Wesley Longman Publishing Co., Inc. Redwood City, CA, USA, 1998. [ACM:280635] 898, 900, 902

[21] S. MUTHU MUTHUKRISHNAN: Data Streams: Algorithms and Applications. *Foundations and Trends in Theoretical Computer Science*, 1(2):117–236, 2005. Preliminary version in SODA'03. [doi:10.1561/0400000002] 898

[22] NOAM NISAN AND AMNON TA-SHMA: Extracting randomness: A survey and new constructions. *J. Comput. System Sci.*, 58(1):148–173, 1999. Preliminary versions in CCC'96 and STOC'96. [doi:10.1006/jcss.1997.1546] 909

[23] NOAM NISAN AND DAVID ZUCKERMAN: Randomness is linear in space. *J. Comput. System Sci.*, 52(1):43–52, 1996. [doi:10.1006/jcss.1996.0004] 909

[24] ANNA PAGH AND RASMUS PAGH: Uniform hashing in constant time and optimal space. *SIAM J. Comput.*, 38(1):85–96, 2008. Preliminary version in STOC'03. [doi:10.1137/060658400] 898

[25] ANNA PAGH, RASMUS PAGH, AND MILAN RUŽIĆ: Linear probing with constant independence. *SIAM J. Comput.*, 39(3):1107–1120, 2009. Preliminary version in STOC'07. [doi:10.1137/070702278] 898, 900, 903

[26] RASMUS PAGH AND FLEMMING FRICHE RODLER: Cuckoo hashing. *J. Algorithms*, 51(2):122–144, 2004. Preliminary version in ESA'01. [doi:10.1016/j.jalgor.2003.12.002] 898, 904

[27] MIHAI PĂTRAŞCU AND MIKKEL THORUP: On the k-independence required by linear probing and minwise independence. In *Proc. 37th Internat. Colloq. on Automata, Languages and Programming (ICALP'10)*, pp. 715–726, 2010. Extended version on arXiv. [doi:10.1007/978-3-642-14165-2_60] 899

[28] MARTIN RAAB AND ANGELIKA STEGER: "Balls into bins"—a simple and tight analysis. In *Proc. 2nd Internat. Workshop on Randomization and Computation (RANDOM'98)*, pp. 159–170. Springer, 1998. [doi:10.1007/3-540-49543-6_13] 900, 903, 904

[29] JAIKUMAR RADHAKRISHNAN AND AMNON TA-SHMA: Bounds for dispersers, extractors, and depth-two superconcentrators. *SIAM J. Discrete Math.*, 13(1):2–24, 2000. Preliminary version in FOCS'97. [doi:10.1137/S0895480197329508] 922, 923

[30] M. V. RAMAKRISHNA: Hashing practice: analysis of hashing and universal hashing. In *Proc. 1988 ACM SIGMOD Internat. Conf. on Management of Data (SIGMOD'88)*, pp. 191–199, New York, NY, USA, 1988. ACM Press. [doi:10.1145/50202.50223] 898

[31] M. V. RAMAKRISHNA: Practical performance of Bloom filters and parallel free-text searching. *Commun. ACM*, 32(10):1237–1239, 1989. [doi:10.1145/67933.67941] 898

[32] M. V. RAMAKRISHNA, E. FU, AND E. BAHCEKAPILI: Efficient hardware hashing functions for high performance computers. *IEEE Trans. Comput.*, 46(12):1378–1381, 1997. [doi:10.1109/12.641938] 898

[33] LEONID REYZIN: A note on the statistical difference of small direct products. Technical Report BUCS-TR-2004-032, Boston University Computer Science Department, 2004. Boston University. 923, 925

[34] AMIT SAHAI AND SALIL VADHAN: Manipulating statistical difference. In PANOS PARDALOS, SANGUTHEVAR RAJASEKARAN, AND JOSÉ ROLIM, editors, *Proc. DIMACS Workshop on Randomization Methods in Algorithm Design (DIMACS'97)*, volume 43 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pp. 251–270. Amer. Math. Soc., 1997. 927

[35] MIKLOS SANTHA AND UMESH V. VAZIRANI: Generating quasi-random sequences from semi-random sources. *J. Comput. System Sci.*, 33(1):75–87, 1986. Preliminary version in FOCS'84. [doi:10.1016/0022-0000(86)90044-9] 899

[36] JEANETTE P. SCHMIDT AND ALAN SIEGEL: The analysis of closed hashing under limited randomness (extended abstract). In *Proc. 22nd STOC*, pp. 224–234. ACM Press, 1990. [doi:10.1145/100216.100245] 898

[37] RONEN SHALTIEL: Recent developments in explicit constructions of extractors. In *Current and Trends in Theoretical Computer Science: The Challenge of the New Century. Volume 1: Algorithms and Complexity*, pp. 189–228. World Scientific, 2002. Available at World Scientific. 909

[38] ALAN SIEGEL: On universal classes of extremely random constant-time hash functions. *SIAM J. Comput.*, 33(3):505–543, 2004. [doi:10.1137/S0097539701386216] 898

[39] DANIEL A. SPIELMAN AND SHANG-HUA TENG: Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *J. ACM*, 51(3):385–463, 2004. Preliminary version in STOC'01. See also at arXiv. [doi:10.1145/990308.990310] 899

[40] MIKKEL THORUP AND YIN ZHANG: Tabulation-based 5-independent hashing with applications to linear probing and second moment estimation. *SIAM J. Comput.*, 41(2):293–331, 2012. Preliminary work in ALENEX'10 and SODA'04. [doi:10.1137/100800774] 899, 900

[41] BERTHOLD VÖCKING: How asymmetry helps load balancing. *J. ACM*, 50(4):568–589, 2003. Preliminary version in FOCS'99. [doi:10.1145/792538.792546] 904

[42] MARK N. WEGMAN AND J. LAWRENCE CARTER: New hash functions and their use in authentication and set equality. *J. Comput. System Sci.*, 22(3):265–279, 1981. Preliminary version in FOCS'79. [doi:10.1016/0022-0000(81)90033-7] 902

[43] PHILIPP WOELFEL: Asymmetric balanced allocation with simple hash functions. In *Proc. 17th Ann. ACM-SIAM Symp. on Discrete Algorithms (SODA'06)*, pp. 424–433. ACM Press, 2006. [ACM:1109557.1109605] 904

[44] DAVID ZUCKERMAN: Simulating BPP using a general weak random source. *Algorithmica*, 16(4/5):367–391, 1996. Preliminary version in FOCS'91. [doi:10.1007/BF01940870] 899, 911

AUTHORS

Kai-Min Chung
Assistant Research Fellow
Academia Sinica, Taiwan
kmchung@iis.sinica.edu.tw
http://www.iis.sinica.edu.tw/~kmchung/

Michael Mitzenmacher
Professor
Harvard University, Cambridge, MA
michaelm@eecs.harvard.edu
http://www.eecs.harvard.edu/~michaelm/

Salil Vadhan
Vicky Joseph Professor
Harvard University, Cambridge, MA
salil@seas.harvard.edu
http://people.seas.harvard.edu/~salil/

ABOUT THE AUTHORS

KAI-MIN CHUNG received a bachelor's degree from National Taiwan University in 2003, and a Ph. D. from Harvard University in 2011. His advisor was Salil Vadhan. After his Ph. D., he was a postdoctoral researcher at Cornell University for three years and supported by a Simons postdoctoral fellowship in 2010-2012. He is currently an assistant research fellow at Academia Sinica in Taiwan. His research interests include cryptography, complexity theory, and pseudorandomness. His work on parallel repetition for interactive arguments received a Best Student Paper award from the Theory of Cryptography Conference in 2010.

MICHAEL MITZENMACHER is a Professor of Computer Science in the School of Engineering and Applied Sciences at Harvard University; from 2010 to 2013, he also served as Area Dean of Computer Science. He graduated summa cum laude with a B. A. in mathematics and computer science from Harvard in 1991. After studying mathematics for a year in Cambridge, England, on the Churchill Scholarship, he obtained his Ph. D. in computer science at U.C. Berkeley in 1996 under the supervision of Alistair Sinclair. He then worked at Digital Systems Research Center until joining the Harvard faculty in 1999. His work on low-density parity-check codes shared the 2002 IEEE Information Theory Society Best Paper Award and received the 2009 ACM SIGCOMM Test of Time Award. His textbook with Eli Upfal on randomized algorithms and probabilistic techniques in computer science was published in 2005 by Cambridge University Press.

SALIL VADHAN is the Vicky Joseph Professor of Computer Science and Applied Mathematics at Harvard University. He received an A. B. summa cum laude in mathematics and computer science from Harvard University in 1995, a Certificate of Advanced Study with distinction in mathematics from Cambridge University in 1996, and a Ph. D. in applied mathematics from MIT in 1999 (under the supervision of Shafi Goldwasser). Vadhan was an NSF mathematical sciences postdoctoral fellow at MIT and the Institute for Advanced Study before joining the Harvard faculty in 2001. He has held visiting positions at the Radcliffe Institute for Advanced Study at Harvard University, the Miller Institute for Basic Research in Science at U.C. Berkeley, Microsoft Research Silicon Valley, and Stanford University, and was director of the Harvard Center for Research on Computation and Society from 2008-11.

Vadhan's research is in computational complexity and cryptography, with specific interests including zero-knowledge proofs, pseudorandomness, and differential privacy. His Ph. D. thesis on statistical zero-knowledge proofs received the ACM Doctoral Dissertation Award 2000 and his work on expander graphs with Omer Reingold and Avi Wigderson received a Gödel Prize in 2009. He has also received a Sloan Fellowship and a Guggenheim Fellowship, and was named a Simons Investigator in 2013. He currently leads a large, multidisciplinary project on "Privacy Tools for Sharing Research Data" at Harvard University.