Tight Bounds for Hashing Block Sources*

Kai-Min Chung** and Salil Vadhan***

School of Engineering & Applied Sciences
Harvard University
Cambridge, MA
{kmchung,salil}@eecs.harvard.edu

Abstract. It is known that if a 2-universal hash function H is applied to elements of a *block source* $(X_1, \ldots, X_T)$, where each item X_i has enough min-entropy conditioned on the previous items, then the output distribution $(H, H(X_1), \ldots, H(X_T))$ will be "close" to the uniform distribution. We provide improved bounds on how much min-entropy per item is required for this to hold, both when we ask that the output be close to uniform in statistical distance and when we only ask that it be statistically close to a distribution with small collision probability. In both cases, we reduce the dependence of the min-entropy on the number T of items from $2 \log T$ in previous work to $\log T$, which we show to be optimal. This leads to corresponding improvements to the recent results of Mitzenmacher and Vadhan (SODA '08) on the analysis of hashing-based algorithms and data structures when the data items come from a block source.

1 Introduction

A *block source* is a sequence of items $\mathbf{X} = (X_1, \ldots, X_T)$ in which each item has at least some k bits of "entropy" conditioned on the previous ones [CG88]. Previous works [CG88, Zuc96, MV08] have analyzed what happens when one applies a 2-universal hash function to each item in such a sequence, establishing results of the following form:

> **Block-Source Hashing Theorems (informal):** *If $(X_1, \ldots, X_T)$ is a block source with k bits of "entropy" per item and H is a random hash function from a 2-universal family mapping to $m \ll k$ bits, then $(H(X_1), \ldots, H(X_T))$ is "close" to the uniform distribution.*

In this paper, we prove new results of this form, achieving improved (in some cases, optimal) bounds on how much entropy k per item is needed to ensure that

* A full version of this paper can be found on [CV08].

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

*** Work done when visiting U.C. Berkeley, supported by the Miller Institute for Basic Research in Science, a Guggenheim Fellowship, US-Israel BSF grant 2006060, and ONR grant N00014-04-1-0478.

the output is close to uniform, as a function of the other parameters (the output length m of the hash functions, the number T of items, and the "distance" from the uniform distribution). But first we discuss the two applications that have motivated the study of Block-Source Hashing Theorems.

1.1 Applications of Block-Source Hashing

Randomness Extractors. A *randomness extractor* is an algorithm that extracts almost-uniform bits from a source of biased and correlated bits, using a short *seed* of truly random bits as a catalyst [NZ96]. Extractors have many applications in theoretical computer science and have played a central role in the theory of pseudorandomness. (See the surveys [NT99, Sha04, Vad07].) Block-source Hashing Theorems immediately yield methods for extracting randomness from block sources, where the seed is used to specify a universal hash function. The gain over hashing the entire T-tuple at once is that the blocks may be much shorter than the entire sequence, and thus a much shorter seed is required to specify the universal hash function. Moreover, many subsequent constructions of extractors for general sources (without the block structure) work by first converting the source into a block source and performing block-source hashing.

Analysis of Hashing-Based Algorithms. The idea of hashing has been widely applied in designing algorithms and data structures, including hash tables [Knu98], Bloom filters [BM03], summary algorithms for data streams [Mut03], etc. Given a stream of data items $(x_1, \ldots, x_T)$, we first hash the items into $(H(x_1), \ldots, H(x_T))$, and carry out a computation using the hashed values. In the literature, the analysis of a hashing algorithm is typically a worst-case analysis on the input data items, and the best results are often obtained by unrealistically modelling the hash function as a truly random function mapping the items to uniform and independent m-bit strings. On the other hand, for realistic, efficiently computable hash functions (eg., 2-universal or $O(1)$-wise independent hash functions), the provable performance is sometimes significantly worse. However, such gaps seem to not show up in practice, and even standard 2-universal hash functions empirically seem to match the performance of truly random hash functions. To explain this phenomenon, Mitzenmacher and Vadhan [MV08] have suggested that the discrepancy is due to worst-case analysis, and propose to instead model the input items as coming from a block source. Then Block-Source Hashing Theorems imply that the performance of universal hash functions is close to that of truly random hash functions, provided that each item has enough bits of entropy.

1.2 How Much Entropy is Required?

A natural question about Block-Source Hashing Theorems is: how large does the "entropy" k per item need to be to ensure a certain amount of "closeness" to uniform (where both the entropy and closeness can be measured in various ways). This also has practical significance for the latter motivation regarding hashing-based algorithms, as it corresponds to the amount of entropy we need

Setting	Previous Results	Our Results
2-universal hashing ε-close to uniform	$m + 2\log T + 2\log(1/\varepsilon)$ [CG88, ILL89, Zuc96]	$m + \log T + 2\log(1/\varepsilon)$
2-universal hashing ε-close to small cp.	$m + 2\log T + \log(1/\varepsilon)$ [MV08]	$m + \log T + \log(1/\varepsilon)$
4-wise indep. hashing ε-close to small cp.	$\max\{m + \log T,$ $1/2(m + 3\log T + \log 1/\varepsilon)\}$ [MV08]	$\max\{m + \log T,$ $1/2(m + 2\log T + \log(1/\varepsilon)\}$

Table 1. Our Results: Each entry denotes the min-entropy (actually, Renyi entropy) required per item when hashing a block source of T items to m-bit strings to ensure that the output has statistical distance at most ε from uniform (or from having collision probability within a constant factor of uniform). Additive constants are omitted for readability.

to assume in data items. In [MV08], they provide bounds on the entropy required for two measures of closeness, and use these as basic tools to bound the required entropy in various applications. The requirement is usually some small constant multiple of $\log T$, where T is the number of items in the source, which can be on the borderline between a reasonable and unreasonable assumption about real-life data. Therefore, it is interesting to pin down the optimal answers to these questions. In what follows, we first summarize the previous results, and then discuss our improved analysis and corresponding lower bounds.

A standard way to measure the distance of the output from the uniform distribution is by *statistical distance*.[1] In the randomness extractor literature, classic results [CG88, ILL89, Zuc96] show that using 2-universal hash functions, $k = m + 2\log(T/\varepsilon) + O(1)$ bits of min-entropy (or even Renyi entropy)[2] per item is sufficient for the output distribution to be ε-close to uniform in statistical distance. Sometimes a less stringent closeness requirement is sufficient, where we only require that the output distribution is ε-close to a distribution having "small" *collision probability*[3]. A result of [MV08] shows that $k = m + 2\log T + \log(1/\varepsilon) + O(1)$ suffices to achieve this requirement. Using 4-wise independent hash functions, [MV08] further reduce the required entropy to $k = \max\{m + \log T, 1/2(m + 3\log T + \log(1/\varepsilon))\} + O(1)$.

Our Results. We reduce the entropy required in the previous results, as summarized in Table 1. Roughly speaking, we save an additive $\log T$ bits of min-entropy (or Renyi entropy) for all cases. We show that using universal hash functions,

[1] The *statistical distance* of two random variables X and Y is $\Delta(X, Y) = \max_T |\Pr[X \in T] - \Pr[Y \in T]|$, where T ranges over all possible events.

[2] The *min-entropy* of a random variable X is $\mathrm{H}_\infty(X) = \min_x \log(1/\Pr[X = x])$. All of the results mentioned actually hold for the less stringent measure of *Renyi entropy* $\mathrm{H}_2(X) = \log(1/\mathrm{E}_{x \leftarrow X}[\Pr[X = x]])$.

[3] The *collision probability* of a random variable X is $\sum_x \Pr[X = x]^2$. By "small collision probability," we mean that the collision probability is within a constant factor of the collision probability of uniform distribution.

$k = m + \log T + 2\log 1/\varepsilon + O(1)$ bits per item is sufficient for the output to be ε-close to uniform, and $k = m + \log(T/\varepsilon) + O(1)$ is enough for the output to be ε-close to having small collision probability. Using 4-wise independent hash functions, the entropy k further reduces to $\max\{m + \log T, 1/2(m + 2\log T + \log 1/\varepsilon)\} + O(1)$. The results hold even if we consider the joint distribution $(H, H(X_1), \ldots, H(X_T))$ (corresponding to "strong extractors" in the literature on randomness extractors). Substituting our improved bounds in the analysis of hashing-based algorithms from [MV08], we obtain similar reductions in the min-entropy required for every application with 2-universal hashing. With 4-wise independent hashing, we obtain a slight improvement for Linear Probing, and for the other applications, we show that the previous bounds can already be achieved with 2-universal hashing. The results are summarized in Table 2.

Although the $\log T$ improvement seems small, we remark that it could be significant for practical settings of parameter. For example, suppose we want to hash 64 thousand internet traffic flows, so $\log T \approx 16$. Each flow is specified by the 32-bit IP addresses and 16-bit port numbers for the source and destination plus the 8-bit transport protocol, for a total of 104 bits. There is a noticeable difference between assuming that each flow contains $3\log T \approx 48$ vs. $4\log T \approx 64$ bits of entropy as they are only 104 bits long, and are very structured.

We also prove corresponding lower bounds showing that our upper bounds are almost tight. Specifically, we show that when the data items have not enough entropy, then the joint distribution $(H, H(X_1), \ldots, H(X_T))$ can be "far" from uniform. More precisely, we show that if $k = m + \log T + 2\log 1/\varepsilon - O(1)$, then there exists a block source $(X_1, \ldots, X_T)$ with k bits of min-entropy per item such that the distribution $(H, H(X_1), \ldots, H(X_T))$ is ε-far from uniform in statistical distance (for H coming from any hash family). This matches our upper bound up to an additive constant. Similarly, we show that if $k = m + \log T - O(1)$, then there exists a block source $(X_1, \ldots, X_T)$ with k bits of min-entropy per item such that the distribution $(H, H(X_1), \ldots, H(X_T))$ is 0.99-far from having small collision probability (for H coming from any hash family). This matches our upper bound up to an additive constant in case the statistical distance parameter ε is constant; we also exhibit a specific 2-universal family for which the $\log(1/\varepsilon)$ in our upper bound is nearly tight — it cannot be reduced below $\log(1/\varepsilon) - \log\log(1/\varepsilon)$. Finally, we also extend all of our lower bounds to the case that we only consider distribution of hashed values $(H(X_1), \ldots, H(X_T))$, rather than their joint distribution with Y. For this case, the lower bounds are necessarily reduced by a term that depends on the size of the hash family. (For standard constructions of universal hash functions, this amounts to $\log n$ bits of entropy, where n is the bit-length of an individual item.)

Techniques. At a high level, all of the previous analyses for hashing block sources were loose due to summing error probabilities over the T blocks. Our improvements come from avoiding this linear blow-up by choosing more refined measures of error. For example, when we want the output to have small statistical distance from uniform, the classic Leftover Hash Lemma [ILL89] says that min-entropy $k = m + 2\log(1/\varepsilon_0)$ suffices for a single hashed block to be ε_0-close to uniform,

Type of Hash Family	Previous Results [MV08]	Our Results
Linear Probing		
2-universal hashing	$4 \log T$	$3 \log T$
4-wise independence	$2.5 \log T$	$2 \log T$
Balanced Allocations with d Choices		
2-universal hashing	$(d+2) \log T$	$(d+1) \log T$
4-wise independence	$(d+1) \log T$	—
Bloom Filters		
2-universal hashing	$4 \log T$	$3 \log T$
4-wise independence	$3 \log T$	—

Table 2. Applications: Each entry denotes the min-entropy (actually, Renyi 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, $m = \log T + O(1)$.

and then a "hybrid argument" implies that the joint distribution of T hashed blocks is $T\varepsilon_0$-close to uniform [Zuc96]. Setting $\varepsilon_0 = \varepsilon/T$, this leads to a min-entropy requirement of $k = m + 2\log(1/\varepsilon) + 2\log T$ per block. We obtain a better bound, reducing $2 \log T$ to $\log T$, by using *Hellinger distance* to analyze the error accumulation over blocks, and only passing to statistical distance at the end.

For the case where we only want the output to be close to having small collision probability, the previous analysis of [MV08] worked by first showing that the expected collision probability of each hashed block $h(X_i)$ is "small" even conditioned on previous blocks, then using Markov's Inequality to deduce that each hashed block has small collision probability except with some probability ε_0, and finally doing a union bound to deduce that all hashed blocks have small collision probability except with probability $T\varepsilon_0$. We avoid the union bound by working with more refined notions of "conditional collision probability," which enable us to apply Markov's Inequality on the entire sequence rather than on each block individually.

The starting point for our negative results is the tight lower bound for randomness extractors due to Radhakrishnan and Ta-Shma [RT00]. Their methods show that if the min-entropy parameter k is not large enough, then for any hash family, there exists a (single-block) source X such that $h(X)$ is "far" from uniform (in statistical distance) for "many" hash functions h. We then take our block source $(X_1, \ldots, X_T)$ to consist of T iid copies of X, and argue that the statistical distance from uniform grows sufficiently fast with the number T of copies taken. For example, we show that if two distributions have statistical distance ε, then their T-fold products have statistical distance $\Omega(\min\{1, \sqrt{T} \cdot \varepsilon\})$, strengthening a previous bound of Reyzin [Rey04], who proved a bound of $\Omega(\min\{\varepsilon^{1/3}, \sqrt{T} \cdot \varepsilon\})$. Due to space constraints, we skip the precise statements and proofs of our negative results. Please refer to the full version of this paper [CV08] for details.

2 Preliminaries

Notations. All logs are based 2. We use the convention that $N = 2^n$, $K = 2^k$, and $M = 2^m$. We think of a data item X as a random variable over $[N] = \{1, \ldots, N\}$, which can be viewed as the set of n-bit strings. A hash function $h : [N] \to [M]$ hashes an item to a m-bit string. A *hash function family* $\mathcal{H}$ is a multiset of hash functions, and H will usually denote a uniformly random hash function drawn from $\mathcal{H}$. $U_{[M]}$ denotes the uniform distribution over $[M]$. Let $\mathbf{X} = (X_1, \ldots, X_T)$ be a sequence of data items. We use $X_{<i}$ to denote the first $i-1$ items $(X_1, \ldots, X_{i-1})$. We refer to X_i as an item or a block interchangeably. Our goal is to study the distribution of hashed sequence $(H, \mathbf{Y}) = (H, Y_1, \ldots, Y_T) \stackrel{\text{def}}{=} (H, H(X_1), \ldots, H(X_T))$.

Hash Families. The *truly random hash family* $\mathcal{H}$ is the set of all functions from $[N]$ to $[M]$. A hash family $\mathcal{H}$ is *s-universal* if for every sequence of distinct elements $x_1, \ldots, x_s \in [N]$, $\Pr_H[H(x_1) = \cdots = H(x_s)] \leq 1/M^s$. $\mathcal{H}$ is *s-wise* independent if for every sequence of distinct elements $x_1, \ldots, x_s \in [N]$, $H(x_1), \ldots, H(x_s)$ are independent and uniform random variables over $[M]$.

Block Sources and Collision Probability. For a random variable X, the collision probability of X is $\mathrm{cp}(X) = \Pr[X = X'] = \sum_x \Pr[X = x]^2$, where X' is an independent copy of X. The *Renyi entropy* $\mathrm{H}_2(X) = \log(1/\mathrm{cp}(X))$ can be viewed as a measure of the amount of randomness in X (In the randomness extractor literature, the entropy is measured by *min-entropy* $H_\infty(X) = \min_{x \in \mathrm{supp}(X)} \log(1/\Pr[X = x])$, but using the less stringent measure Renyi entropy makes our results stronger since $H_2(X) \geq H_\infty(X)$.) For an event E, $(X|_E)$ is the random variable defined by conditioning X on E.

Definition 2.1 (Block Sources). *A sequence of random variables $(X_1, \ldots, X_T)$ over $[N]^T$ is a* block K-source *if for every $i \in [T]$, and every $x_{<i}$ in the support of $X_{<i}$, we have $\mathrm{cp}(X_i | X_{<i} = x_{<i}) \leq 1/K$. That is, each item X_i has at least $k = \log K$ bits of Renyi entropy even after conditioning on the previous items.*

Let $\mathbf{X} = (X_1, \ldots, X_T)$ be a sequence of random variables over $[M]^T$. We are interested in bounding the overall collision probability $\mathrm{cp}(\mathbf{X})$ by the collision probability of each blocks. Suppose all X_i's are independent, then $\mathrm{cp}(\mathbf{X}) = \prod_{i=1}^T \mathrm{cp}(X_i)$. The following lemma generalizes Lemma 4.2 in [MV08], which says that if for every $\mathbf{x} \in \mathbf{X}$, the average collision probability of every block X_i conditioning on $X_{<i} = x_{<i}$ is small, then the overall collision probability $\mathrm{cp}(\mathbf{X})$ is also small. In particular, if $\mathbf{X}$ is a block K-source, then $\mathrm{cp}(\mathbf{X}) \leq 1/K^T$.

Lemma 2.2. *Let $\mathbf{X} = (X_1, \ldots, X_T)$ be a sequence of random variables such that for every $\mathbf{x} \in \mathrm{supp}(\mathbf{X})$,*

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

Then the overall collision probability satisfies $\mathrm{cp}(\mathbf{X}) \leq \alpha^T$.

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

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

Therefore, it suffices to prove

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

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

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

as desired.

Statistical Distance. The statistical distance is a standard way to measure the distance of two distributions. Let X and Y be two random variables. The *statistical distance* of X and Y is $\Delta(X,Y) = \max_T |\Pr[X \in T] - \Pr[Y \in T]| = (1/2) \cdot \sum_x |\Pr[X = x] - \Pr[Y = x]|$, where T ranges over all possible events. When $\Delta(X,Y) \le \varepsilon$, we say that X is *ε-close* to Y. Similarly, if $\Delta(X,Y) \ge \varepsilon$, then X is *ε-far* from Y. The following standard lemma says that if X has small collision probability, then X is close to uniform in statistical distance.

Lemma 2.3. *Let X be a random variable over $[M]$ such that $\mathrm{cp}(X) \le (1+\varepsilon)/M$. Then $\Delta(X, U_{[M]}) \le \sqrt{\varepsilon}$.*

Conditional Collision Probability. Let (X,Y) be jointly distributed random variables. We can define the conditional Renyi entropy of X conditioning on Y as follows.

Definition 2.4. *The* conditional collision probability *of X conditioning on Y is* $\mathrm{cp}(X|Y) = \mathrm{E}_{y\leftarrow Y}[\mathrm{cp}(X|_{Y=y})]$. *The* conditional Renyi entropy *is* $\mathrm{H}_2(X|Y) = \log 1/\mathrm{cp}(X|Y)$.

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

Lemma 2.5. *Let (X, Y, Z) be jointly distributed random variables. We have* $\mathrm{cp}(X) \le \mathrm{cp}(X|Y) \le \mathrm{cp}(X|Y, Z)$.

Proof. For the first inequality, we have

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

For the second inequality, observe that for every y in the support of Y, we have $\mathrm{cp}(X|_{Y=y}) \le \mathrm{cp}((X|_{Y=y})|(Z|_{Y=y}))$ from the first inequality. It follows that

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

3 Positive Results: How Much Entropy is Sufficient?

In this section, we present our positive results, showing that the distribution of hashed sequence $(H, \mathbf{Y}) = (H, H(X_1), \ldots, H(X_T))$ is close to uniform when H is a random hash function from a 2-universal hash family, and $\mathbf{X} = (X_1, \ldots, X_T)$ has sufficient entropy per block. The new contribution is that we will not need $K = 2^k$ to be as large as in previous works, and so save the required randomness in the block source $\mathbf{X} = (X_1, \ldots, X_T)$.

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 $(H, \mathbf{Y}) = (H, H(X_1), \ldots, H(X_T))$ is ε-close to having collision probability $O(1/(|\mathcal{H}| \cdot M^T))$. This requirement is less stringent than $(H, \mathbf{Y})$ being ε-close to uniform in statistical distance, and so requires less bits of entropy. Mitzenmacher and Vadhan [MV08] show that this

guarantee suffices for some hashing applications. They show that $K \geq MT^2/\varepsilon$ is enough to satisfy the requirement. We save a factor of T, and show that in fact, $K \geq MT/\varepsilon$, is sufficient. (Taking logs yields the first entry in Table 1, i.e. it suffices to have Renyi entropy $k = m + \log T + \log(1/\varepsilon)$ per block.) Formally, we prove the following theorem.

Theorem 3.1. *Let $H : [N] \to [M]$ be a random hash function from a 2-universal family $\mathcal{H}$. Let $\mathbf{X} = (X_1, \ldots, X_T)$ be a block K-source over $[N]^T$. For every $\varepsilon > 0$, the hashed sequence $(H, \mathbf{Y}) =$ $(H, H(X_1), \ldots, H(X_T))$ is ε-close to a distribution $(H, \mathbf{Z}) = (H, Z_1, \ldots, Z_T)$ such that*

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

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

To analyze the distribution of the hashed sequence $(H, \mathbf{Y})$, the starting point is the following version of the Leftover Hash Lemma [BBR85, ILL89], which says that when we hash a random variable X with enough entropy using a 2-universal hash function H, the conditional collision probability of $H(X)$ conditioning on H is small.

Lemma 3.2 (The Leftover Hash Lemma). *Let $H : [N] \to [M]$ be a random hash function from a 2-universal family $\mathcal{H}$. Let X be a random variable over $[N]$ with $\mathrm{cp}(X) \leq 1/K$. We have $\mathrm{cp}(H(X)|H) \leq 1/M + 1/K$.*

We now sketch how the hashed block source $\mathbf{Y} = (Y_1, \ldots, Y_T)$ $= (H(X_1), \ldots, H(X_T))$ is analyzed in [MV08], and how we improve the analysis. The following natural approach is taken in [MV08]. Since the data $\mathbf{X}$ is a block K-source, the Leftover Hash Lemma tells us that for every block $i \in [T]$, if we condition on the previous blocks $X_{<i} = x_{<i}$, then the hashed value $(Y_i|_{X_{<i}=x_{<i}})$ has small conditional collision probability, i.e. $\mathrm{cp}((Y_i|_{X_{<i}=x_{<i}})|H) \leq 1/M + 1/K$. This is equivalent to saying that the average collision probability of $(Y_i|_{X_{<i}=x_{<i}})$ over the choice of the hash function H is small, i.e.,

$$\mathop{\mathrm{E}}_{h \leftarrow H}[\mathrm{cp}(h(X_i)|_{X_{<i}=x_{<i}})] = \mathrm{cp}((Y_i|_{X_{<i}=x_{<i}})|H) \leq \frac{1}{M} + \frac{1}{K}.$$

We can then use a Markov argument to say that for every block, with probability at least $1 - \varepsilon/T$ over $h \leftarrow H$, the collision probability is at most $1/M + T/(K\varepsilon)$. We can then take a union bound to say that for every $\mathbf{x} \in \mathrm{supp}(\mathbf{X})$, at least $(1-\varepsilon)$-fraction of hash functions h are good in the sense that $\mathrm{cp}(h(X_i)|_{X_{<i}=x_{<i}})$ is small for all blocks $i = 1, \ldots, T$. [MV08] shows that if this condition is true for every $(h, \mathbf{x}) \in \mathrm{supp}(H, \mathbf{X})$, then $\mathbf{Y}$ is a block $(1/M + T/(K\varepsilon))$-source, and thus the overall collision probability is at most $(1 + MT/K\varepsilon)^T/M^T$. [MV08] also shows how to modify an ε-fraction of the distribution to fix the bad hash functions, and thus complete the analysis.

The problem of the above analysis is that taking a Markov argument for each block, and then taking a union bound incurs a loss of factor T. To avoid this, we want to apply Markov argument only once to the whole sequence. For example, a natural thing to try is to sum over blocks to get

$$\mathop{\mathrm{E}}_{h \leftarrow H}\left[\frac{1}{T}\sum_{i=1}^{T}\mathrm{cp}(h(X_i)|_{X_{<i}=x_{<i}})\right] = \frac{1}{T}\sum_{i=1}^{T}\mathrm{cp}((Y_i|_{X_{<i}=x_{<i}})|H) \le \frac{1}{M}+\frac{1}{K},$$

and use a Markov argument to deduce that for every $\mathbf{x} \in \mathrm{supp}(\mathbf{X})$, with probability $1-\varepsilon$ over $h \leftarrow H$, the average collision probability per block satisfies

$$\frac{1}{T}\cdot\sum_{i=1}^{T}\mathrm{cp}(h(X_i)|_{X_{<i}=x_{<i}}) \le \frac{1}{M}+\frac{1}{K\varepsilon}.$$

We need to bound the collision probability of $\mathbf{Y}$ using this information. We may try to apply Lemma 2.2, but it requires the information on $(1/T)\sum_i \mathrm{cp}(Y_i|_{Y_{<i}=y_{<i}})$ instead of $(1/T)\sum_i \mathrm{cp}(h(X_i)|_{X_{<i}=x_{<i}})$. That is, Lemma 2.2 requires us to condition on previous *hashed values* $Y_{<i}$, whereas the above argument refers to conditioning on the un-hashed values $X_{<i}$. The difficulty with directly reasoning about the former is that conditioned on the hashed values $Y_{<i}$, the hash function H may no longer be uniform (as it is correlated with $Y_{<i}$) and thus the Leftover Hash Lemma no longer applies.

To get around with the issues, we work with the averaged form of conditional collision probability $\mathrm{cp}(Y_i|H,Y_{<i})$, as from Definition 2.4. Our key observation is that now we can apply Lemma 2.5 to deduce that for every block $i \in [T]$, the conditional collision probability satisfies $\mathrm{cp}(Y_i|H,Y_{<i}) \le \mathrm{cp}(Y_i|H,X_{<i}) \le 1/M+1/K$. Then, by a Markov argument, it follows that with probability $1-\varepsilon$ over $(h,\mathbf{y}) \leftarrow (H,\mathbf{Y})$, the average collision probability satisfies

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

We can then modify an ε-fraction of distribution, and apply Lemma 2.2 to complete the analysis.

The following lemma formalizes our claim about that the conditional collision probability of every block of $(H,\mathbf{Y})$ is small.

Lemma 3.3. *Let $H:[N]\to[M]$ be a random hash function from a 2-universal family $\mathcal{H}$. Let $\mathbf{X}=(X_1,\ldots,X_T)$ be a block K-source over $[N]^T$. Let $(H,\mathbf{Y}) = (H,H(X_1),\ldots,H(X_T))$. Then $\mathrm{cp}(H)=1/|\mathcal{H}|$ and for every $i\in[T]$, $\mathrm{cp}(Y_i|H,Y_{<i}) \le 1/M+1/K$.*

Proof. $\mathrm{cp}(H)=1/|\mathcal{H}|$ is trivial since H is the uniform distribution. 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 $\mathrm{cp}((Y_i|_{X_{<i}=x_{<i}})|(H|_{X_{<i}=x_{<i}})) \le 1/M+1/K$ for every $x_{<i}$. It follows that

$\mathrm{cp}(Y_i|H, X_{<i}) \le 1/M + 1/K$. Now, we can think of $(Y_i|H, X_{<i})$ as Y_i first conditioning on $(H, Y_{<i})$, and then further conditioning on $X_{<i}$. By Lemma 2.5, we have

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

as desired.

The remaining part of the proof follows the above sketch closely. Details can be found in the full version of this paper[CV08].

3.2 Small Collision Probability Using 4-wise Independent Hash Functions

As discussed in [MV08], using 4-wise independent hash functions $H : [N] \to [M]$ from $\mathcal{H}$, we can further reduce the required randomness in the data $\mathbf{X} = (X_1, \ldots, X_T)$. [MV08] shows that in this case, $K \ge MT + \sqrt{2MT^3/\varepsilon}$ is enough for the hashed sequence $(H, \mathbf{Y})$ to be ε-close to having collision probability $O(1/|\mathcal{H}| \cdot M^T)$. As discussed in the previous subsection, by avoiding using union bounds, we show that $K \ge MT + \sqrt{2MT^2/\varepsilon}$ suffices. (Taking logs yields the second entry in Table 1, i.e. it suffices to have Renyi entropy $k = \max\{m + \log T, (1/2) \cdot (m + 2\log T + \log(1/\varepsilon))\} + O(1)$ per block.) Formally, we prove the following theorem.

Theorem 3.4. *Let $H : [N] \to [M]$ be a random hash function from a 4-wise independent family $\mathcal{H}$. Let $\mathbf{X} = (X_1, \ldots, X_T)$ be a block K-source over $[N]^T$. For every $\varepsilon > 0$, the hashed sequence $(H, \mathbf{Y}) = (H, H(X_1), \ldots, H(X_T))$ is ε-close to a distribution $(H, \mathbf{Z}) = (H, Z_1, \ldots, Z_T)$ such that*

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

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

The improvement of Theorem 3.4 over Theorem 3.1 comes from that when we use 4-wise independent hash families, we have a concentration result on the conditional collision probability for each block . For the proof of the theorem, please refer to [CV08].

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 $\mathbf{X} = (X_1, \ldots, X_T)$ be a block K-source over $[N]^T$. In this subsection, we study the statistical distance between the distribution of hashed sequence $(H, \mathbf{Y}) = (H, H(X_1), \ldots, H(X_T))$ and the uniform distribution $(H, U_{[M]^T})$. Classic results

of [CG88, ILL89, Zuc96] show that if $K \geq MT^2/\varepsilon^2$, then $(H, \mathbf{Y})$ is ε-close to uniform. The proof idea is as follows. The Leftover Hash Lemma together with Lemma 2.3 tells us that the joint distribution of hash function and a hashed value $(H, Y_i) = (H, H(X_i))$ is $\sqrt{M/K}$-close to uniform $U_{[M]}$ even conditioning on the previous blocks $X_{<i}$. One can then use a hybrid argument to show that the distance grows linearly with the number of blocks, so $(H, \mathbf{Y})$ is $T \cdot \sqrt{M/K}$-close to uniform. Taking $K \geq MT^2/\varepsilon^2$ completes the analysis.

We save a factor of T, and show that in fact, $K = MT/\varepsilon^2$ is sufficient. (Taking logs yields the third entry in Table 1, i.e. it suffices to have Renyi entropy $k = m + \log T + 2\log(1/\varepsilon)$ per block.) Formally, we prove the following theorem.

Theorem 3.5. *Let $H : [N] \to [M]$ be a random hash function from a 2-universal family $\mathcal{H}$. Let $\mathbf{X} = (X_1, \ldots, X_T)$ be a block K-source over $[N]^T$. For every $\varepsilon > 0$ such that $K > MT/\varepsilon^2$, the hashed sequence $(H, \mathbf{Y}) = (H, H(X_1), \ldots, H(X_T))$ is ε-close to uniform $(H, U_{[M]^T})$.*

Recall that the previous analysis goes by passing to statistical distance first, and then measuring the growth of distance using statistical distance. This incurs a quadratic dependency of K on T. Since without further information, the hybrid argument is tight, to save a factor of 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., [GS02]) is a good measure for our purposes:

Definition 3.6 (Hellinger distance). *Let X and Y be two random variables over $[M]$. The* Hellinger distance *between X and Y is*

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

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 3.7 (cf., [GS02]). *Let X and Y be two random variables over $[M]$. We have*

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

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 introduce the following definition.

Definition 3.8 (Hellinger Closeness to Uniform). *Let X be a random variable over $[M]$. The* Hellinger closeness *of X to uniform $U_{[M]}$ is*

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

Note that $C(X,Y) = C(X) \cdot C(Y)$ when X and Y are independent random variables, so the Hellinger closeness is well-behaved with respect to products (unlike statistical distance). By Lemma 3.7, if the Hellinger closeness $C(X)$ 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 Hellinger closeness 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 2.3 says that if the "2-norm" $M \cdot \mathrm{cp}(X) = \mathrm{E}_i[f(i)^2] \leq 1 + \varepsilon$ where the expectation is over uniform $i \in [M]$, then $\Delta(X,U) \leq \sqrt{\varepsilon}$,. Similarly, Lemma 3.7 says that if the "1/2-norm" $C(X) = \mathrm{E}_i[\sqrt{f(i)}] \geq 1 - \varepsilon$, then $\Delta(X,U) \leq \sqrt{\varepsilon}$.

We now discuss our approach to prove Theorem 3.5. We want to show that $(H, \mathbf{Y})$ is close to uniform. All we know is that the conditional collision probability $\mathrm{cp}(Y_i|H, Y_{<i})$ is close to $1/M$ for every block. If all blocks are independent, then the overall collision probability $\mathrm{cp}(H, \mathbf{Y})$ is small, and so $(H, \mathbf{Y})$ 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 Hellinger closeness. Formally, we have the following lemma. The main idea is to use Hölder's inequality to relate two different norms.

Lemma 3.9. *Let $\mathbf{X} = (X_1, \ldots, X_T)$ be jointly distributed random variables over $[M_1] \times \cdots \times [M_T]$ such that $\mathrm{cp}(X_i|X_{<i}) \leq \alpha_i/M_i$ for every $i \in [T]$. Then the Hellinger closeness satisfies*

$$C(\mathbf{X}) \geq \sqrt{\frac{1}{\alpha_1 \ldots \alpha_T}}.$$

The proof of this lemma can be found in the full version of this paper[CV08]. With this lemma, the proof of Theorem 3.5 is immediate.

Proof of Theorem 3.5: By Lemma 3.3, $\mathrm{cp}(H) = 1/|\mathcal{H}|$, and $\mathrm{cp}(Y_i|H, Y_{<i}) \leq (1 + M/K)/M$ for every $i \in [T]$. By Lemma 3.9, the Hellinger closeness satisfies $C(H, \mathbf{Y}) \geq (1 + M/K)^{-T/2} \geq 1 - MT/2K$ (recall that $K \geq MT/\varepsilon^2$). It follows by Lemma 3.7 that

$$\begin{aligned}\Delta((H, \mathbf{Y}), (H, U_{[M]^T})) &\leq \sqrt{2} \cdot d((H, \mathbf{Y}), (H, U_{[M]^T})) \\ &= \sqrt{2} \cdot \sqrt{1 - C(H, \mathbf{Y})} \leq \sqrt{MT/K} \leq \varepsilon.\end{aligned}$$

■

Acknowledgments

We thank Wei-Chun Kao for helpful discussions in the early stages of this work, David Zuckerman for telling us about Hellinger distance, and Michael Mitzenmacher for suggesting parameter settings useful in practice.

References

[BBR85] Charles H. Bennett, Gilles Brassard, and Jean-Marc Robert. How to reduce your enemy's information (extended abstract). In Hugh C. Williams, editor, *Advances in Cryptology—CRYPTO '85*, volume 218 of *Lecture Notes in Computer Science*, pages 468–476. Springer-Verlag, 1986, 18–22 August 1985.

[BM03] Andrei Z. Broder and Michael Mitzenmacher. Survey: Network applications of bloom filters: A survey. *Internet Mathematics*, 1(4), 2003.

[CG88] Benny Chor and Oded Goldreich. Unbiased bits from sources of weak randomness and probabilistic communication complexity. *SIAM J. Comput.*, 17(2):230–261, 1988.

[CV08] Kai-Min Chung and Salil Vadhan. Tight bounds for hashing block sources, 2008. http://www.citebase.org/abstract?id=oai:arXiv.org:0806.1948.

[GS02] Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70:419, 2002.

[ILL89] Russell Impagliazzo, Leonid A. Levin, and Michael Luby. Pseudo-random generation from one-way functions (extended abstracts). In *Proceedings of the Twenty First Annual ACM Symposium on Theory of Computing*, pages 12–24, Seattle, Washington, 15–17 May 1989.

[Knu98] Donald E. Knuth. *The art of computer programming, Volume 3: Sorting and Searching*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1998.

[MV08] Michael Mitzenmacher and Salil Vadhan. Why simple hash functions work: Exploiting the entropy in a data stream. In *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '08)*, pages 746–755, 20–22 January 2008.

[Mut03] S. Muthukrishnan. Data streams: algorithms and applications. In *SODA*, pages 413–413, 2003.

[NT99] Noam Nisan and Amnon Ta-Shma. Extracting randomness: A survey and new constructions. *J. Comput. Syst. Sci.*, 58(1):148–173, 1999.

[NZ96] Noam Nisan and David Zuckerman. Randomness is linear in space. *Journal of Computer and System Sciences*, 52(1):43–52, February 1996.

[RT00] Jaikumar Radhakrishnan and Amnon Ta-Shma. Bounds for dispersers, extractors, and depth-two superconcentrators. *SIAM Journal on Discrete Mathematics*, 13(1):2–24 (electronic), 2000.

[Rey04] Leonid Reyzin. A note on the statistical difference of small direct products. Technical Report BUCS-TR-2004-032, Boston University Computer Science Department, 2004.

[Sha04] Ronen Shaltiel. Recent developments in extractors. In G. Paun, G. Rozenberg, and A. Salomaa, editors, *Current Trends in Theoretical Computer Science*, volume 1: Algorithms and Complexity. World Scientific, 2004.

[Vad07] Salil Vadhan. The unified theory of pseudorandomness. *SIGACT News*, 38(3), September 2007.

[Zuc96] David Zuckerman. Simulating BPP using a general weak random source. *Algorithmica*, 16(4/5):367–391, October/November 1996.